# How to best measure the fundamental unit of biogeography

**Authors**: John C. Donoghue II1\*, N. Morueta-Holme2, B. Boyle3, L. L. Sloat1, B. J. Enquist1, B. J. McGill4, J.­C. Svenning2, & The BIEN Working Group5

**Affiliations:**

1 Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA.

2 Ecoinformatics and Biodiversity Group, Department of Bioscience, Aarhus University, DK-8000 Aarhus C, Denmark.

3 The iPlant Collaborative, Thomas W. Keating Bioresearch Building, 1657 East Helen Street, Tucson, AZ 85721, USA.

4 School of Biology and Ecology / Sustainability Solutions Initiative, University of Maine, Orono, ME 04469, USA.

5 National Center for Ecological Analysis and Synthesis, University of California, 735 State Street, Suite 300, Santa Barbara, USA.

\* Correspondence to: mail@johndonoghue.net

**Article Type:** Research Paper

**Short Title:** Measuring the fundamental unit of biogeography

**Keywords:** geographic range, range size, species distribution modeling, biodiversity, biogeography, species richness, maxent

***GEB Guidelines***

*The body of a manuscript (excluding references, tables and figures) should not normally exceed 5000 words, and there should not be more than 50 cited references. The manuscript must include (i) a short running title, (ii) a list of 6-10 key words (or phrases), and (iii) an abstract of no more than 300 words structured under the headings: Aim, Location, Methods, Results, Main conclusions. The order of material should be as follows: title page (including article title, author name(s), author research address(es), correspondence author address and e-mail, article type, short running title, abstract, key words), main text, references, biosketch, tables with their captions, figure legends, figures.*

**Abstract (300 words):**

Geographic range size is a fundamental property of a species and a key criterion in determining conservation status and prioritization. Yet, many methods of measuring range size exist in the literature, with some methods focusing on capturing the outer limits of the species’ occurrences, while other methods focusing on the smaller area within those outer limits that is actually occupied by the species. As each of these methods can generate widely different range size estimates, how we measure geographic range size has profound implications for our assessment of a species’ extinction risk and our prioritization of its conservation needs. To demonstrate the potential implications from using different methods to estimate the geographic range size and distribution for a species, we use a very large dataset of New World plant species to compare several common approaches of estimating the geographic range sizes and distributions of species. For each approach, we compute geographic range size and spatial distribution and compare them with those of expert-drawn maps. Our methods range from those estimating the size of a species’ geographic extent, through methods approximating the size of a species’ occupied range, to methods estimating both a species’ occupied range and distribution. We also explore how combinations of environmental layers and different thresholds influence the range size estimates derived from presence/absence maps of species distribution models. We find that for range size measures derived solely from occurrence data, the area defined by plotting a convex hull around the species’ occurrences is the best predictor of geographic range size, independent of sample size or geographic position (temperate or tropical). Our results highlight the value of a simple metric such as convex hull, over more computationally and comprehensibly difficult species distribution models, for obtaining a reasonably accurate estimate of geographic range size.

## Main Text (5000 words):

I suggest the following structure for the intro:

* What is the geographic range, why is it important [finish with one sentence stating the problem and linking to next paragraph – that there are many ways of measuring it and it’s unclear which is best]
* Measuring the range size – there are many methods, they are different in their ecological meaning, and this has implications for understanding and for conservation [i.e. lump current paragraph 2 + 3]
* Broad overview of the current methods – from simple occurrence-based ones to SDMs – pros and cons related to range size estimation and geographic position
* Short paragraph on expert-maps and how they are our best option for comparison
* What we do in this paper [mention BIEN, temperate trees and palms, the specific questions we test]

A species’ geographic range defines the spatial distribution of the species on the earth. The measurement of this range, termed the species’ geographic range size, is a fundamental property of a species and may vary as much as twelve orders of magnitude among species (Brown *et al.*, 1996). Ecologically, different geographic range sizes of species are believed to relate to specialist/generalist strategies (Wilson and Yoshimura, 1994) and the interaction of dispersal and competition factors (Pulliam, 2000). Thus, different geographic range sizes may provide insight into the processes that affect the distributions of species and community diversity (Stevens, 1989; Gaston, 2003). Geographic range size is also a key criterion in determining the conservation status and prioritization of species (IUCN, 2001; Baillie, Hilton-Taylor & Stuart 2004; Gaston and Fuller, 2008). For example, 47% of the assessed species on the IUCN Red List of Threatened Species are listed solely due to their geographic range features (Gaston and Fuller, 2008).

As Brown *et al.* (1996) stated, “If the geographic range of a species is a basic unit of geography, then biogeographic research will depend on how species and their ranges are characterized”. Yet, there is no established method for measuring species’ geographic range size (Gaston, 1994). Instead, a suite of different methods exists in the literature that captures different aspects of a species’ geographic distribution, and occasionally confuses measures that capture the outer limits of the species’ occurrences with measures that focus on the smaller area within those outer limits in which the species actually occurs (cf. Gaston and Fuller, 2009). These range size measurement methods can be generally divided into measures that focus on a) quantifying range size in terms of the linear distance between widely separated occurrences, b) measures quantifying range size as a two-dimensional measurement of area within the limits of the species’ occurrences, and c) measures that quantify range size in terms of the number of areas (i.e. grid cells or sites) occupied by the species (Gaston 1996).

The use of different methods to quantify geographic range size has led to heated discussions, since each method generates a different estimate for the geographic range size of the same species [REFS]. As a result, the various methods have the potential to drastically influence the extinction risk estimates and conservation prioritization status for a species, depending on the metrics and the assumptions the methods are based upon (Hubbell *et al.*, 2008a; Feeley and Silman, 2008; Hubbell *et al.*, 2008b; Feeley and Silman, 2009). Further, since species richness maps are commonly generated by overlapping the geographic range maps for various species in a study area (REF?), the different methods for deriving those geographic range maps can have profound consequences for our estimates of biodiversity and our understanding of its correlation with climatic variables (Swenson *et al.* 2004).

Despite the fundamental nature of geographic range size, a quantitative assessment of the best method for estimating range size has not yet been conducted. With a variety of methods used in the literature, we might expect that some methods are substantially better than others, with the optimal method providing a high degree of accuracy in its estimate of geographic range size, independent of sample size or the species’ geographic position (e.g. temperate or tropical location). Further, some methods may not be suitable for estimating the physical location of the species’ geographic range, but could be very useful for accurately estimating the species’ geographic range size – especially if those methods are easier to compute than generating a species distribution model.

In this methodological study, we use a very large dataset of New World plants, containing a temperate group of North American trees and a generally tropical group of New World palms, to investigate several popular approaches for estimating the geographic range sizes and distributions of species to ascertain whether a “best” estimation method exists and which methods are generally more accurate than others. Our approaches include a continuum of methods for estimating the geographic range for a species (Box 1), extending from very simple ways of estimating the overall spread of a species, (e.g. latitudinal and longitudinal extent and bounding box), through simple methods that gradually approximate the size of a species’ occupied range (e.g. convex hull), to methods designed to estimate both the size of a species’ occupied range and its spatial distribution (e.g. occupied grid cells and species distribution models).

Of course, there are few species for which we know their exact geographic range boundaries. Thus, for each approach, we compute the geographic range size and spatial distribution, and compare the results with those of expert-drawn maps. In doing so we must assume that the expert-drawn maps are a true representation of each species’ geographic range. While this assumption is obviously questionable, without knowing the exact geographic range of each species in our study, the use of scientifically vetted expert-drawn range maps is the best comparison method that we, or anyone for that matter, have at present.

With respect to species distribution modeling (SDM) algorithms, we focus on Maxent (Phillips *et al.*, 2004) because it is one of the most commonly employed ecological niche modeling algorithms (Warren and Siefert, 2011). Similarly, we parameterize our Maxent models with full suites of commonly used environmental data layers and presence-only occurrences. While this is a rather crude way to estimate the potential geographic distribution for a species, this modeling method is a fairly common practice (Warren and Siefert, 2011). Thus, we also aim to ascertain the extent to which combining presence-only data with generic environmental datasets and a commonly used SDM algorithm can accurately estimate the geographic range size and distribution for any species, independent of geographic position (temperate or tropical) or sample size.

**Estimating Geographic Range Size**

Since there is no established method for measuring species’ geographic range size, studies have utilized a variety of methods (Gaston, 1994). For some species, their range boundaries were previously interpolated onto hand-drawn maps using expert knowledge, to create “expert” range maps. Many of these maps have since been digitized and are widely available and still in use today. For these species, their geographic range size can easily be obtained using a GIS to capture the area attribute of their digitized polygons.

For the remaining species, their geographic range size must be estimated by applying one or more of the common estimation methods (Box 1). For methods that rely solely on species occurrence data, range size can be estimated either as a linear distance between two widely dispersed occurrences (i.e. latitudinal or longitudinal extent), as the area of a one or more polygons that contain the species occurrences (i.e. bounding box, convex hull), or the number of areas (i.e. grid cells or sites) occupied by the species (Gaston 1996).

Finally, the geographic range size of a species can be estimated by developing a species distribution model, which combines species’ occurrence data and environmental data to develop correlative models of the environmental conditions associated with species occurrences to predict the relative probability of observing those environmental conditions in each portion of the study landscape. The map generated by a species distribution model is a raster layer in which each cell contains a value representing the probability of that cell having suitable environmental conditions for that species. Thus, to obtain a defined geographic range from this map, the probability surface must be converted to a presence/absence map by transforming the decimal value probability estimates of each cell to a binary 0 or 1 value, in which cells having a value of 0 represent areas of potentially unsuitable environmental conditions to support the presence of the species, while cells having a value of 1 represent areas containing potentially suitable environmental conditions for the species. This is performed by setting a threshold value at which cells having probability values less than the threshold are assigned a value of 0, while cells having a value equal to or greater than the threshold value are assigned a value of 1. The species’ geographic range size can then be calculated as the total area of all cells having a value of 1.

## Materials and Methods:

Digitized expert-drawn maps were available for both New World palms (n=645) (Henderson *et al.*, 1995) and temperate North American trees (n=679) (Little, 1971; 1976; 1977; 1978). We reviewed expert maps for temperate North American trees and excluded maps in which a species’ geographic range was unnaturally truncated at a political boundary and maps having geographic range boundaries that appeared to be overly generalized. Expert maps for New World palms were reviewed for accuracy (Balslev, 2011). For all species with expert maps, we extracted all species occurrence records from the BIEN database (BIEN, 2012), which contains a combination of data from herbarium records and vegetation plots. All observations were assigned standardized taxon names using the Taxonomic Name Resolution Service (Boyle 2013; TNRS, 2012) and the World Checklist of Palms (Govaerts *et al.*, 2005). The geographic location of each observation was validated using the Global Administrative Areas dataset version 2.0 (GADM, 2012). We excluded any observations that could not be resolved and species for which there were fewer than five occurrence records, which is the minimum number required for computing a species distribution model (REF).

For each species, we used all geographically distinct occurrence records to calculate the species’ geographic range size based on metrics derived solely from the occurrence data (Box 1), such as (i) latitudinal extent, (ii) longitudinal extent, (iii) bounding box, (iv) latitudinal band, (v) convex hull, and (vi) area of grid cells occupied. Latitudinal extent is defined as the linear distance between the minimum and maximum latitude of all species occurrences, while longitudinal extent is the linear distance between the minimum and maximum longitude of all species occurrences. Bounding box is the area bounded by the minimum and maximum latitude and minimum and maximum longitude of all species occurrences, while the latitudinal band is the area bounded by the minimum and maximum latitude and the western and eastern continental boundaries of all species occurrences. The convex hull area is based on the minimum-fitting polygon that can be drawn to encompass all species occurrences. Finally, the area of occupied cells is defined as the total area of all cells containing one or more occurrences of a species. We note that latitudinal and longitudinal extents do not give area estimates, but they can still be compared to the remaining metrics by correlations.

In addition to metrics derived solely from occurrence data, we calculated the species’ geographic range size using (vii) the species distribution model Maxent (Phillips *et al.*, 2004). Maxent works by combining species’ occurrence and environmental data to develop correlative models of the environmental conditions associated with species occurrences to predict the relative probability of observing those environmental conditions in each portion of the study area. We included a set of 19 global layers of climate variables derived from monthly temperature and rainfall values (climate layers) from WorldClim 1.4 at 30-arc seconds (1 km) resolution (Hijmans *et al.*, 2005). Projecting distribution models across two continents solely based on climatic factors will likely result in overpredictions of species’ actual ranges, as many species may not be present in all of their potential climatic range due to dispersal limitations, historical factors, biotic interactions and other environmental factors (Gaston, 2009). Therefore, we also included 19 spatial filter layers as eigenvectors computed from a geographical distance matrix as additional predictors in the model. These spatial filter layers were computed using the same approach as in Blach-Overgaard *et al.* (2010) and represent relatively broad- to medium scale spatial patterns across the geometry of the continent’s area and have been shown to effectively capture non-environmental spatial constraints caused by dispersal-limited non-equilibrium range dynamics (De Marco *et al*. 2008).

Using Maxent, we explored the influence of different model parameters on the resulting range size estimates. We computed (a) a “standard” model with a set of all 19 climate layers, (b) a Maxent model using only 19 spatial filter layers, and (c) a Maxent model that combined a balanced set of the 19 climate layers with the 19 spatial filter layers (Blach-Overgaard *et al.*, 2010). All data were projected and resampled to a lambert-equal area grid at 10-km resolution, and work was performed using ArcGIS 9.3 (ESRI, 2011) and R 2.15.1 (R Development Core Team, 2007).

To obtain a defined geographic range from the output of a Maxent model, the probability surface generated from the model must be converted to a presence/absence map by converting the decimal value probability estimates of each cell to a binary 0 or 1 value, representing areas of potentially unsuitable or suitable environmental conditions to support the species, respectively. This is performed by setting a threshold value at which cells having probability values less than the threshold are assigned a value of 0, while cells having a value equal to or greater than the threshold value are assigned a value of 1. The simple method of creating a presence/absence map by using fixed threshold (typically 0.5) to the probably surface has been widely used in ecology (see references in Liu *et al.*, 2005). However, the results from a fixed threshold are highly influenced by the prevalence of the occurrence data, and the method is not recommended in comparative studies of threshold selection (Freeman and Moisen, 2008; Liu *et al.*, 2005). Therefore, we also applied an individual threshold for each species. Since our purpose was to estimate geographic range sizes accurately, we examined the effect of threshold choice on range size estimates applying the following thresholds: (Box 2), (i) a fixed threshold of 0.5, (ii) the maximum training sensitivity plus specificity threshold, (iii) the minimum training presence threshold, (iv) the maximum Kappa threshold, (v) the 1 percent training presence threshold, (vi) the 5 percent training presence threshold, and (vii) the 10 percent training presence threshold.

All computed geographic range areas were clipped to the continental land areas. The performance of each method was assessed through pairwise comparisons of the output for a specific species with its corresponding expert-drawn map, which was assumed to represent the species’ true presences and absences. The agreement between the geographic range maps was quantified by computing several accuracy measures widely used for model assessment (Box 3; Liu *et al.*, 2005): the Jaccard similarity index, overall accuracy, sensitivity, specificity, Kohen’s Kappa statistic and the True Skill Statistic) (Allouche *et al.*, 2006).

Finally, we also assessed whether the performance of each geographic range size metric differed between temperate and tropical species (represented by North American trees and palms, respectively). To examine the effect of sample size on range size model performance, we plotted how each performance measure varied with sample size. We separated the effects of poor sampling from the effects of rarity (low species occurrence) by analyzing subsamples of well-sampled species to test how sample size affected the geographic range size estimates. This allowed us to assess (i) which range size models perform best at low and high sample sizes, (ii) which range size models are resilient to sample size and (iii) what minimum sample size produces acceptable range size estimates.

## Results:

The sample sizes of the occurrence data for each species used to develop each model varied considerably (min = 5, max = 3907, mean = 182). While most species in the analysis had 150 occurrence points or less, some species had over 3,000 occurrence points (Figure 1). Estimated geographic range areas were highly variable with respect to model type (Figure 2). For all models, range areas appeared to show a slight curvilinear relationship with sample size, with most models approaching a range area asymptote at a sample size of roughly 1,000 occurrences (Figure S1). However, range areas based on the area of occupied cells had a stronger curvilinear relationship and required a higher sample size to approach a range area asymptote, at a sample size of roughly 2,000 occurrences (Figure 2).

***Pairwise Range Extent and Area***

For linear models of pairwise comparisons of geographic range extent, both the latitudinal and longitudinal extents measured from the species occurrence points were fairly well correlated with the extents measured from the expert map ranges (r2 = 0.547, r2 = 0.600 for the combined temperate and tropical datasets, respectively, p < 10-17). However, the longitudinal extent consistently provided a better fit to the expert map range extent than the latitudinal extent, and was among the top three best performing models, independent of geographic position (Table 1).

For pairwise comparisons of geographic range area (Figure 3), we found that areas defined by convex hulls drawn around the species occurrence points consistently provided the best fit to expert map areas independent of geographic position (r2 = 0.6052, p < 10-17). Moreover, range areas defined by the bounding box of the species occurrence points consistently provided the second-best fit to expert map range areas independent of geographic position (r2 = 0.570, p < 2.22e-16). Range areas computed from the species’ latitudinal band area (the full continental area between the minimum and maximum latitudes of the species occurrence points), had surprisingly high fit to expert map range areas (r2 = 0.4275, p < 2.22e-16), while the pairwise fitness of range areas derived from the area of occupied cells was among the worst performing models, and highly dependent on geographic position (Table 1). For example, occupied cell area fit expert map range area fairly well for temperate tree (r2 = 0.3208, p < 2.22e-16, Table S1), but was the worst fitting model for tropical palms (r2 = 0.1803, p < 2.22e-16, Table S1).

Maxent models exhibited highly variable performance in pairwise comparisons of range area. Range areas derived from presence/absence maps of Maxent models parameterized with only spatial filter data, and using a fixed threshold, were among the worst two models (r2 = 0.1455, p < 2.22e-16). Similarly, Maxent models parameterized with only the 19 Bioclim layers, and using a fixed threshold, were not much better (r2 = 0.2009, p < 2.22e-16). However, range areas derived from presence/absence maps of Maxent models parameterized with both Bioclim and spatial filters, and using the 1% training presence threshold, exhibited very good fit to expert map range areas independent of geographic position (r2 = 0.4459, p < 2.22e-16), and were consistently the third highest ranking model after range areas derived fromconvex hulls and bounding boxes around the species occurrence points (Table 1).

Range areas derived from presence/absence maps of Maxent models combining both Bioclim and spatial filter layers consistently outperformed Maxent models using either set of layers separately, independent of geographic position (Table S2). While, occasionally presence/absence maps derived using the Maximum Training Presence threshold, parameterized with either Bioclim or spatial filter data independently, provided a better fit with temperate tree datasets (Table S2), in general range areas derived using the 1% training presence threshold frequently exhibited a better fit to expert map range areas than range areas derived from presence/absence maps using other thresholds, independent of geographic position or which layer data were used to parameterize the Maxent model (Table S2). However, clipping any non-contiguous regions more than 1,000 km away from the main continuous region (to account for over-prediction of each species’ geographic distribution) generated a consistently best-fitting Maxent model, independent of geographic position or environmental data used (Table S2). However, this model did not outperform range areas derived by drawing convex hulls and bounding boxes around the species occurrence points (Table 1).

R-square values of the accuracy of pairwise comparisons of geographic range area were highly variable with respect to sample size (Figure 3). In general all models exhibited a bimodal curve, in which the first r-squared maxima for each model was obtained at a sample size between 100-250 occurrences, and a second maxima was obtained at a sample size of approximately 1250 occurrences. The r-squared value of each maxima differed among models, with the convex hull and bounding box models having highest r2 values (Table 1).

***Overlap Accuracy***

In quantifying the spatial agreement between the geographic ranges obtained from each method with those of the expert maps, the bounding box model had the highest Kappa (0.413 ± 0.472 CI) and Jaccard Similarity Index (JSI) scores (0.302 ± 0.403 CI), followed by the bounding box model (Kappa = 0.390 ± 0.459 CI, JSI = 0.284 ± 0.391 CI). However, True Skill Statistic (TSS) results were mixed, with the latitudinal band model having the highest TSS score (0.726 ± 0.407), followed by bounding box (0.696 ± 0.499), and convex hull (0.547 ± 0.550). Among the Maxent models, those parameterized with only spatial filters had higher Kappa, TSS and JSI scores (Kappa = 0.308 ± 0.386, TSS = 0.415 ± 0.569, JSI = 0.204 ± 0.287) than models parameterized with only Bioclim layers (Kappa = 0.278 ± 0.331, TSS = 0.356 ± 0.466, JSI = 0.179 ± 0.235) or both Bioclim and spatial filters combined (Kappa = 0.291 ± 0.354, TSS = 0.298 ± 0.473, JSI = 0.189 ± 0.262). In this case, range areas were obtained from Maxent presence/absence maps using a 1% training presence threshold (Table 2).

Jaccard Similarity Index values with respect to sample size were highly dependent on the model used to compute the geographic range. Most models (excluding the area of occupied cells) exhibited an initial sharp increase in accuracy for the first 50-75 occurrences. However, beyond 75 occurrences, geographic ranges computed using Maxent models showed a slight concave curve at higher sample sizes, while ranges computed from bounding box and convex hull models exhibited a convex curve at higher sample sizes. Finally, geographic ranges computed from occupied cells exhibited a very low JSI accuracy that was essentially invariant to sample size (Figure 5).

In contrast, accuracy estimates measured by TSS (Figure S2), ACC (Figure S3), and Kappa (Figure S4) for occupied cell area and Maxent models initially declined sharply within approximately 50-75 occurrences. Beyond 75 occurrences they appeared invariant with increasing sample size. While TSS (Figure S2), ACC (Figure S3), and Kappa (Figure S4) accuracy estimates for latitudinal band, convex hull and bounding box models also initially declined sharply within approximately 50-75 occurrences, they continued to exhibit a subsequent decline beyond 75 occurrences.

## Discussion:

Overall, our results suggest that the best way to estimate both the geographic range size and distribution for a species, given broadly dispersed presence-only data with generic environmental datasets such as Bioclim, is by applying a convex hull around the species’ occurrence points. In our analysis, the convex hull consistently provided the best pairwise fit of geographic range area to expert map range area, and had the highest Jaccard Similarity Index value in assessments of how well the convex hull range maps spatially matched expert range maps. This finding has some fairly profound implications for distribution modeling, as it implies that complex models such as Maxent may not be necessary to obtain a good estimate of a species geographic range size and, potentially, distribution. In most cases, convex hulls are likely to provide the best estimate, and they are also conceptually intuitive, computationally easy to generate, and not dependent on correlation with environmental data.

In addition, our results illustrate some general patterns with some of the modeling methods that should be kept in mind when estimating geographic ranges. First, range areas estimated from the area of occupied cells are consistently much lower than those of both expert range maps and the maps obtained from other estimation methods, and highly dependent on very high sample sizes for better results. Pairwise fitness r-squared values were consistently low and varied with the geographic position of the species occurrence dataset. For example, the area from occupied cells fit the area of expert maps moderately well for temperate trees (r2 = 0.3208; Table S1), but fit tropical palms poorly (r2 = 0.1803; Table S1), likely due to the fact that our temperate tree sample sizes were much greater than those of our tropical palms. Thus, the accuracy of range area estimates from the area of occupied cells is highly reliant on having a very large sample size.

In contrast, range areas estimated from latitudinal band models appear to consistently over-predict range area, due to the coarseness with which the latitudinal band model represents species’ geographic range. The model appears to be only moderately useful for predicting range sizes and distributions for large-ranged species that cover a large continental area. This is illustrated in figure 4 which shows the accuracy of the latitudinal band model increasing at greater expert map range sizes.

With respect to geographic range models made using Maxent models, we found that including spatial filters (eigenvectors) as proxies for environmental spatial constraints (Blach-Overgaard et al., 2010), resulted in almost consistently much higher pairwise fit results than the use of either Bioclim or spatial filters alone, independent of geographic position (temperate or tropical) or sample size. Thus, we recommend that large-scale geographic range models using Maxent should likely include an equal number of spatial filters into their models to obtain better results that constrain the over-prediction bias resulting from the use of solely environmental data.

Further, we found that range areas from presence/absence maps defined using a fixed threshold appeared to over-predict small range areas and under-predict large range areas (Figure 4). Range areas from presence/absence maps defined using the maximum Kappa threshold were consistently the worst performing Maxent model, independent of geographic position or environmental layer used. In contrast, range areas from presence/absence maps defined from Maxent models using Bioclim and spatial filter layers and a 1% training presence threshold were nearly always the best performing Maxent model independent of geographic position or layer data used. Our best Maxent model was obtained by removing any non-contiguous regions more than 1,000 km away from the main continuous region generated by the presence/absence map defined using the 1% training presence threshold above. Removing the non-contiguous range areas helped account for over-prediction of each species’ geographic distribution into regions beyond its dispersal capacity and resulted in the consistently best-fitting Maxent model, independent of geographic position or environmental data used (Table S2).

We were surprised to find that, while the individual pairwise accuracy measures of range area estimates varied significantly by model type, the accuracies for all models peaked at sample sizes of approximately 100-250 occurrences (Figure 3; Table 1). Additional samples beyond 250 occurrences did not appear to increase the accuracy of any model higher than the r-squared value achieved between 100-250 occurrences. These findings suggest that, for modeling with presence-only data, no more than 250 occurrences of each species are needed to establish a useful geographic range area. Thus, biodiversity data collection efforts should strive to obtain additional occurrence data for species with fewer than 250 occurrences to ensure their geographic range areas can be adequately modeled.

**Caveats**

We recognize that our convex hull model results may be highly influenced by the fact that we are comparing the range maps obtained from each modeling method to those of expert range maps - that were likely created by drawing irregular polygons around known occurrences. Thus, it may be that expert range maps themselves are essentially modified convex hull models. Moreover, by using species occurrences from an amalgamated dataset BIEN, we were able to obtain very broadly sampled species occurrence data. Had our occurrence data come from only a small portion of a species’ geographic range, the convex hull measure would have performed much more poorly. Therefore, the accuracy of methods like convex hull is highly dependent on having broadly distributed species occurrence data.

As with any effort that analyzes large amounts of data, we encountered a number of issues during this study that are worth mentioning for their value in aiding further biogeographic research on large, combined datasets. For example, we found that the maps from E.L. Little Atlas of North American Trees (Little, 1971; 1976; 1977; 1978) provided good coverage of species occurring within the United States and Canada, but poor coverage of species occurring largely in the United States and Mexico. However, the BIEN database (BIEN, 2012) we used for our analysis contained millions of records for the United States and thousands for Mexico, but relatively few records for Canada. To account for the uneven coverage, we restricted our analysis to only temperate North American tree species for which the geographic range area drawn in the expert map was within the continental United States boundary, ensuring we could compare range area estimates for a geographic area that our models and the expert maps had in common. Moreover, we found that some of the digitized E.L. Little maps depicted species’ geographic ranges as relatively simple Euclidean shapes that appeared to be rather coarse estimates of the species’ geographic range. As we surmised that these very coarse approximations may confound our ability to the compare geographic range size and distribution estimates from our models with those of expert map, we excluded any species having its expert map depicted this way.

Lastly, combining occurrence data from many sources is not without problems, and we expended considerable effort validating both the geographic and taxonomic status of the occurrence data used for this analysis. Geographic latitude and longitude values for each occurrence record were plotted and compared with the record’s associated locality description and any records with geographic locations that did not match their locality descriptions were excluded. Further, some species required taxonomic name validation/correction to ensure their expert map could be paired with their corresponding occurrence data. This was particularly prevalent in the temperate tree species data, for which the expert maps were drawn more than 40 years ago and in some cases reflected old taxonomy. Any species having had extensive splits or merges which could influence its overall geographic range were excluded to ensure parity between its expert map and associated occurrence data. Finally, we found that our combined data sources contain numerous occurrence records containing cultivated specimens from herbaria and botanical gardens. Since the accidental inclusion of a cultivated specimen existing far outside of a species’ natural range would likely inflate our geographic range models by directly enlarging the total spatial extent of the occurrence data and/or including additional climates in which the species does not normally occur, we made every effort to identify and exclude such records from our analysis.

FINAL THOUGHTS?

## Acknowledgements:

This work was conducted as a part of the Botanical Information and Ecology Network (BIEN) Working Group (PIs BJE, RC, BB, SD, RKP) supported by the National Center for Ecological Analysis and Synthesis, a center funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California. The BIEN Working Group was also supported by iPlant (National Science Foundation #DBI-0735191; URL: www.iplantcollaborative.org). We thank all the contributors (see full list in supporting material) for the invaluable data provided to BIEN. N M-H was supported by an EliteForsk Award and the Aarhus University Research Foundation. JCS acknowledges support from the Center for Informatics Research on Complexity in Ecology (CIRCE), funded by the Aarhus University Research Foundation under the AU Ideas program, and the European Research Council (ERC Starting Grant: HISTFUNC). JCD was supported by the NSF-funded iPlant Collaborative.

## 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.

Baillie, J.E.M., Hilton-Taylor, C. and Stuart, S.N. (Editors) 2004. 2004 IUCN Red List of Threatened Species. A Global Species Assessment. IUCN, Gland, Switzerland and Cambridge, UK. xxiv + 191 pp.

Balslev, Henrik. 2011. Personal communication.

Blach-Overgaard, A., Svenning, J. C., Dransfield, J., Greve, M., & Balslev, H. (2010) Determinants of palm species distributions across Africa: the relative roles of climate, non-climatic environmental factors, and spatial constraints. *Ecography*, **33**, 380-391.

Botanical Information and Ecology Network version 2.0. http://bien.nceas.ucsb.edu/bien. accessed on 15 May 2012.

Boyle, B. et.al. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics. **14**:16. 1-14.

Brown, J. H., Stevens, G. C., & Kaufman, D. M. (1996) The geographic range size: size, shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics*, **27**, 597-623.

Cohen, J. (1960) A coefﬁcient of agreement of nominal scales. Educational and Psychological Measurement, 20, 37 –46.

De Marco, P. *et al.* 2008. Spatial analysis improves species distribution modelling during range expansion. *Biology Letters*. **4,** 577-580.

ESRI (Environmental Systems Resource Institute). 2011. ArcGIS Desktop: Release 10. Redlands, California.

Feeley, K. J. & Silman, M. R. (2008) Unrealistic assumptions invalidate extinction estimates. *Proceedings of the National Academy of Sciences*, **105**, E121.

Freeman, E. A. & Moisen, G. G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modeling*, **217**, 48-58.

Gaston, K. J. (1994) *Rarity*. Chapman & Hall

Gaston, K. J. (1996) Species-range-size distributions: patterns, mechanisms and implications. *Trends in Ecology and Evolution*, 11:5, 197–201.

Gaston, K. J. (2003) The structure and dynamics of geographic ranges. Oxford Univ. Press.

Gaston, K. J. (2009) Geographic range limits of species. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 1391-1393.

Gaston, K. J. & Fuller, R. A. (2008) The sizes of species' geographic ranges. *Journal of Applied Ecology*, **46**, 1-9.

Global administrative areas dataset version 1.0. http://www.gadm.org. accessed 15 May 2012.

Govaerts, R., Dransfield, J., Zona, S. F., Hodel, D. R., & Henderson, A. (2005) World Checklist of palms. Royal Botanic Gardens Kew, Richmond.

Graham, C.H. and Hijmans, R.J. (2006), A comparison of methods for mapping species ranges and species richness, *Global Ecology and Biogeography*, **15**, 578–587

Henderson, A., Galeano, G., & Bernal, R. (1995) Field guide to the palms of the Americas. Princeton University Press, Princeton, New Jersey.

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.

Hubbell, S. P., He, F. L., Condit, R., Borda-de-Água, L., Kellner, J., & ter Steege, H. (2008a) How many tree species and how many of them are there in the Amazon will go extinct? *Proceedings of the National Academy of Sciences*, **105**, 11498-11504.

Hubbell, S. P., He, F., Condit, R., Borda-de-Água, L., Kellner, J., & ter Steege, H. (2008b) Reply to Feeley and Silman: Extinction risk estimates are approximations but are not invalid. *Proceedings of the National Academy of Sciences*, **105**, E122.

IUCN (2001) IUCN Red List Categories and Criteria: Version 3.1. IUCN, Gland, Switzerland.

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

Little, E. L. (1971) Atlas of United States trees, volume 1, conifers and important hardwoods. U.S. Department of Agriculture Miscellaneous Publication 1146, 9 p., 200 maps.

Little, E. L. (1976) Atlas of United States trees, volume 3, minor Western hardwoods. U.S. Department of Agriculture Miscellaneous Publication 1314, 13 p., 290 maps.

Little, E. L. (1977) Atlas of United States trees, volume 4, minor Eastern hardwoods. U.S. Department of Agriculture Miscellaneous Publication 1342, 17 p., 230 maps.

Little, E. L. (1978) Atlas of United States trees, volume 5, Florida. U.S. Department of Agriculture Miscellaneous Publication 1361, 262 maps.

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.

Phillips, S. J., Dudik, M., & Schapire, R. E. (2004) A maximum entropy approach to species distribution modeling. Proceedings of the Twenty-First international Conference on Machine Learning ACM, New York.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

R Development Core Team (2007). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Stevens, G. C. (1989) The latitudinal gradient in geographical range: how so many species coexist in the tropics. *The American Naturalist*, **133**, 240-256.

Swenson, N.G., Boyle, B., Pither, J., Weisner, M.D. & Enquist, B.J. (2004) Can Ecological Niche Models Accurately Predict Species Richness and Range Size?. Unpublished.

The Taxonomic Name Resolution Service version 3.0. http://tnrs.iplantcollaborative.org. accessed on 10 December 2012.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., Ferreira de Siqueira, M., 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.

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.

Wilson, D.S. & Yoshimura, J. (1994) On the coexistence of specialists and generalists. *The American Naturalist*, **144**, 692– 707.

## Biosketch:

TBD

## Boxes

**Box 1:** Illustrations of geographic range sizes and areas resulting from our modeling approaches. Areas in white represent continental land mass, while black dots represent locations of species occurrences. Red lines represent boundaries of geographic ranges areas or lengths of geographic range extents.

|  |  |  |
| --- | --- | --- |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Expert_Map.png |  | **Expert Map**  **Inputs:** Species occurrence data points  **Output:** Total geographic area of one or more polygons manually drawn based on species occurrences and expert opinion. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Lat_Extent.png |  | **Latitudinal Extent:**  **Inputs:** Species occurrence data points  **Output:** Linear distance between the latitude of the northernmost and southernmost geographic locations in your dataset. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Long_Extent.png |  | **Longitudinal Extent**  **Inputs:** Species occurrence data points  **Output:** Linear distance between the longitude of the westernmost and easternmost geographic locations in your dataset. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Occupied_Cells.png |  | **Occupied Cell Area**  **Inputs:** Species occurrence data points  **Output:** Total geographic area of all cells containing at least one occurrence of your focal species. Cells are created by converting species occurrence points to a raster surface at a particular scale and spatial resolution. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Long_Band.png |  | **Latitudinal Band**  **Inputs:** Species occurrence data points  **Output:** Total geographic area of a single polygon defined by the latitude of the northernmost and southernmost geographic locations in your dataset and the western and eastern limits of the continent. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Bounding_Box.png |  | **Bounding Box**  **Inputs:** Species occurrence data points  **Output:** Total geographic area of a single polygon defined by the latitude of the northernmost and southernmost geographic locations and the longitude of the westernmost and easternmost geographic locations in your dataset. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Convex_Hull.png |  | **Convex Hull**  **Inputs:** Species occurrence data points  **Output:** Total geographic area of the smallest single polygon that encompasses all the species occurrence locations in your dataset. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Maxent_BioSpa_Raw.png |  | **Maxent Logistic**  **Inputs:** Species occurrence data and environmental layers  **Output:** Raster layer in which the value of each cell represents the proportional probability of that cell containing suitable environmental conditions to support the presence of a species. |
| C:\Users\jdonoghue2\Dropbox\School\Projects\BIEN\Manuscripts\Range Size\Boxes\Quercus_gambelii_Maxent_BioSpa.png |  | **Maxent Presence/Absence**  **Inputs:** Species occurrence data and environmental layers  **Output:** Total geographic area of one or more polygons defined by setting a value to the logistic surface at which any cells having probability values less than the threshold are given a value of 0 (areas of potentially unsuitable environmental conditions for the species), while cells having a value equal to or greater than the threshold are given a value of 1 (areas containing potentially suitable environmental conditions for the species). |

**Box 2:** Common threshold values used to reclassify Maxent logistic output into binary presence/absence maps.

|  |  |  |
| --- | --- | --- |
| **Threshold** |  | **Description** |
| Fixed |  | An arbitrary fixed threshold of 0.5 is very commonly used in modeling studies and represents the minimum probability at which suitable habitat is predicted to be present. |
| Maximum training sensitivity plus specificity |  | The maximum training sensitivity plus specificity threshold balances errors of omission (false absences) and commission (false presences). |
| Minimum training presence |  | The minimum training presence threshold minimizes errors of omission (false absences). |
| Maximum Kappa |  | The maximum Kappa threshold is the probability threshold at which Cohen’s Kappa Statistic is the highest - where Kappa is the greatest difference between the observed classified locations and those locations classified by chance. |
| 1 percent training presence |  | The 1 percent training presence threshold represents the threshold values that would result in excluding approximately 1 percent of the presence records. |
| 5 percent training presence |  | The 5 percent training presence threshold represents the threshold values that would result in excluding approximately 5 percent of the presence records. |
| 10 percent training presence |  | The 10 percent training presence threshold represents the threshold values that would result in excluding approximately 10 percent of the presence records. |

**Box 3:** Description of the methods used to assess the accuracies of each modeled geographic range map with its corresponding expert range map.

|  |  |  |
| --- | --- | --- |
| **Accuracy Metric** |  | **Description** |
| Accuracy |  | Accuracy or ‘correct classification rate’ is the percentage of sites that are correctly predicted. |
| Sensitivity |  | Sensitivity or ‘true positive fraction’ is the percentage of observed presences that are correctly predicted. It quantiﬁes errors of omission. |
| Specificity |  | Speciﬁcity or ‘true negative fraction’ is the percentage of observed absences that are correctly predicted. It quantiﬁes errors of commission. |
| Kohen’s Kappa Statistic |  | Cohen’s kappa is a very popular statistic that combines both commission and omission errors in one parameter. It is similar to ‘Accuracy’ but takes into account the proportion of correct predictions expected by chance. Values range from 1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random (Cohen 1960). |
| True Skill Statistic (TSS) |  | Kappa has been shown to be sensitive to species’ prevalence (or the proportion of sampled sites in which the species is recorded present). Thus TSS is a measure of accuracy that has all the advantages of the Kappa statistic, but is insensitive to species’ prevalence. |
| Jaccard Similarity Index |  | The Jaccard Similarity Index quantifies the amount of overlap between maps. It is measured as the size of the intersection of both maps divided by the size of their union. Values range from 0 for maps that have no areas in common, to 1 for maps having all areas in common. |

## Tables

**Table 1**: Summary of linear model results from pairwise comparisons of expert

range area and range areas obtained from each model for three datasets: Henderson’s tropical palms and Little’s temperate trees combined. Temperate trees were restricted to species with geographic ranges that are confined to the contiguous United States. Models included (1) area of occupied cells, (2) bounding box area, (3) convex hull area, (4) latitudinal band area, and the area obtained from Maxent modeled presence maps using (5) Bioclim layers only, (6) spatial layers only, and (7) Bioclim+spatial layers. All presence/absence maps were derived using the fixed threshold setting. For each dataset, the model with the best fit is in indicated in bold type.

|  |  |  |  |
| --- | --- | --- | --- |
| **COMBINED PALMS AND TREES (N=630)** | **R2** | **Intercept** | **Coefficient** |
| Expert Map vs Model Latitudinal Extent | 0.5466 | 0.994 | < 2.22e-16 |
| Expert Map vs Model Longitudinal Extent | 0.6000 | 0.994 | < 2.22e-16 |
| Expert Map vs Occupied Cell Area | 0.2421 | 0.831 | < 2.22e-16 |
| Expert Map vs Bounding Box | 0.5701 | 1.040 | < 2.22e-16 |
| **Expert Map vs Convex Hull** | **0.6052** | **1.010** | **< 2.22e-16** |
| Expert Map vs Latitudinal Band | 0.4375 | 1.090 | < 2.22e-16 |
| Expert Map vs Bioclim Fixed | 0.2009 | 0.995 | < 2.22e-16 |
| Expert Map vs Spatial Fixed | 0.1455 | 0.997 | < 2.22e-16 |
| Expert Map vs Bioclim + Spatial Fixed | 0.4459 | 0.969 | < 2.22e-16 |

**Table 2:** Mean values and 95% confidence intervals for total accuracy (ACC), Sensitivity (Sens), Specificity (Spec), Kappa, True Skill Statistic (TSS), and the Jaccard Similarity Index (JSI) obtained by calculating contingency matrices from the overlap of expert range maps for tropical palms and temperate trees (n=630; DF=629) with range maps obtained from the following models: Area of occupied cells, Latitudinal band, Bounding box, Convex hull, and from Maxent modeled presence maps using Bioclim layers only, Spatial layers only, and Bioclim+spatial layers. The Maxent modeled binary presence/absence maps were derived using the 1% training presence threshold setting.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ACC** | **Sens** | **Spec** | **Kappa** | **TSS** | **JSI** |
| Occupied Cells | 0.972  +/- 0.096 | 0.001  +/- 0.008 | 1.000  +/- 0.00 | 0.002  +/- 0.016 | 0.001  +/- 0.008 | 0.001 +/- 0.008 |
| Latitudinal Band | 0.894  +/- 0.168 | 0.834  +/- 0.447 | 0.893  +/- 0.178 | 0.192  +/- 0.365 | 0.726  +/- 0.407 | 0.138 +/- 0.304 |
| Bounding Box | 0.955  +/- 0.126 | 0.739  +/- 0.529 | 0.958  +/- 0.132 | 0.390  +/- 0.459 | 0.696  +/- 0.499 | 0.284 +/- 0.391 |
| Convex Hull | 0.971  +/- 0.086 | 0.568  +/- 0.575 | 0.979  +/- 0.079 | 0.413  +/- 0.472 | 0.547  +/- 0.550 | 0.302 +/- 0.403 |
| Maxent Bioclim | 0.968  +/- 0.092 | 0.367  +/- 0.48 | 0.990  +/- 0.045 | 0.278  +/- 0.331 | 0.356  +/- 0.466 | 0.179 +/- 0.235 |
| Maxent Spatial | 0.969  +/- 0.089 | 0.423  +/- 0.579 | 0.991  +/- 0.038 | 0.308  +/- 0.386 | 0.415  +/- 0.569 | 0.204 +/- 0.287 |
| Maxent  Bioclim+Spatial | 0.973  +/- 0.087 | 0.301  +/- 0.475 | 0.997  +/- 0.010 | 0.291  +/- 0.354 | 0.298  +/- 0.473 | 0.189 +/- 0.262 |

## Figure Legends

**Figure 1:** Frequency distribution in the sample sizes for both tropical palms and temperate trees (n=630) used in this analysis.

**Figure 2:** Box-plot showing variance in the geographic range sizes for both tropical palms and temperate trees (n=630) by model type from the following models: expert maps (Expert), area of occupied cells (Point), latitudinal band area (LatBnd), bounding box area (BBox), convex hull area (CHull), and the area obtained from Maxent fixed threshold modeled presence maps using Bioclim layers only (Bio), spatial layers only (Spa), and Bioclim + spatial layers (BioSpa).

**Figure 3:** Plots comparing r-squared values for each model at various sample sizes (for tropical palms and temperate trees; n=630). RMSEs were calculated from linear models comparing expert map range area with range areas obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers using the fixed threshold setting, as well as (h) Bioclim + spatial layers using 1% training presence threshold setting. The red line represents a smoothed loess function.

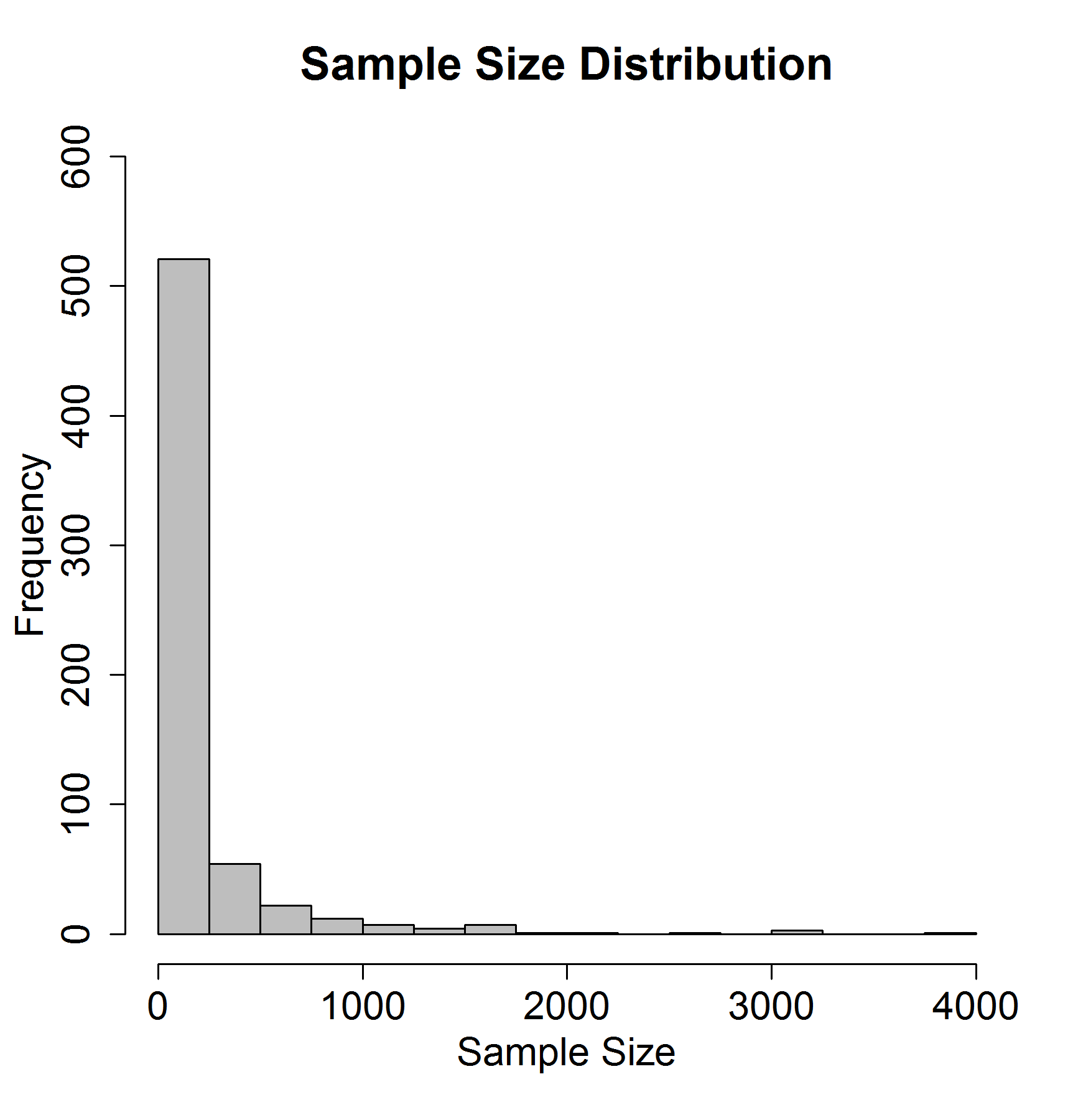
**Figure 4:** Pairwise comparisons of expert map range area with tropical palm and temperate trees range areas (n=607) obtained from the following models: occupied cells, latitudinal band, bounding box, convex hull, and areas obtained from Maxent model presence maps using the fixed threshold setting, and 1% training presence threshold setting with both Bioclim and spatial layer inputs.

**Figure 5:** Jaccard Similarity Index (JSI) by sample size. JSI is computed from the overlap of expert range maps for both tropical palms and temperate trees (n=630) and range maps obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers. The Maxent modeled binary presence/absence maps were derived using the 1% training presence threshold setting.

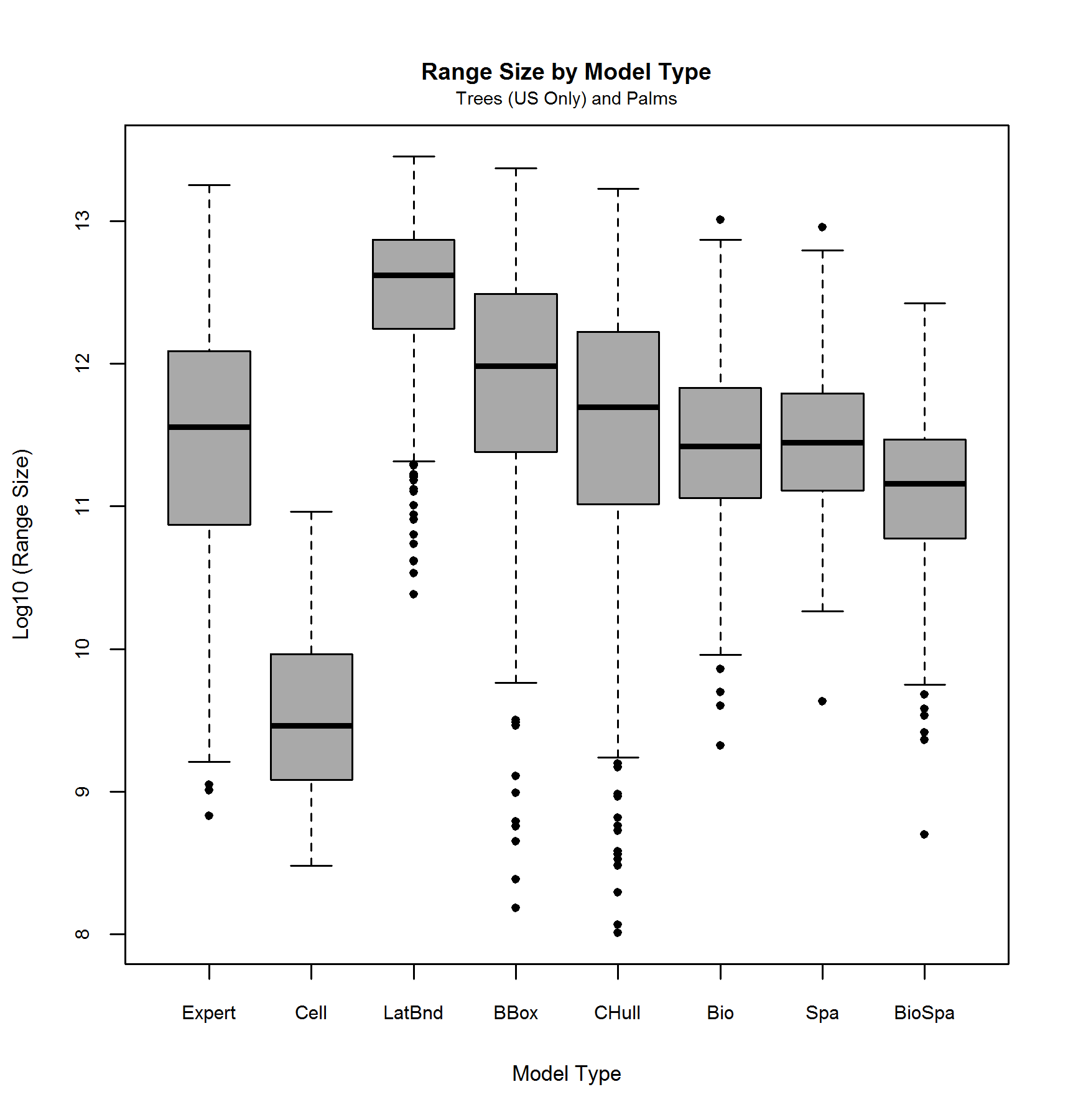
**Figure 6:** Species richness maps calculated from the range maps generated with geographic ranges obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers using the fixed threshold setting, as well as (h) Bioclim + spatial layers using 1% training presence threshold setting.

## Figures

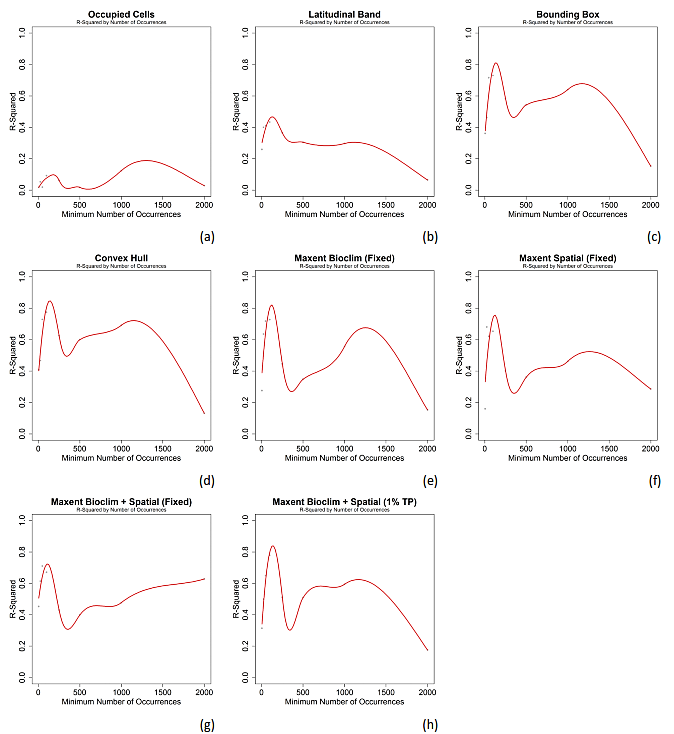
**Figure 1:**



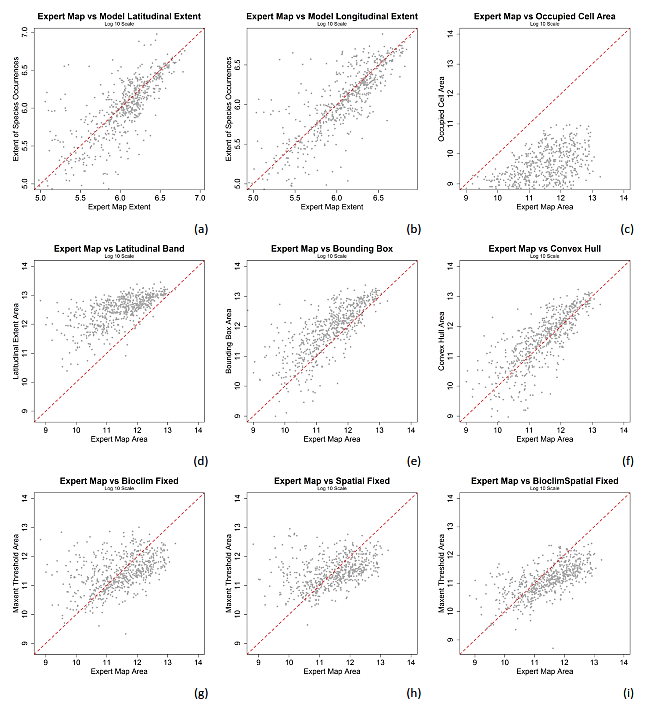
**Figure 2:**



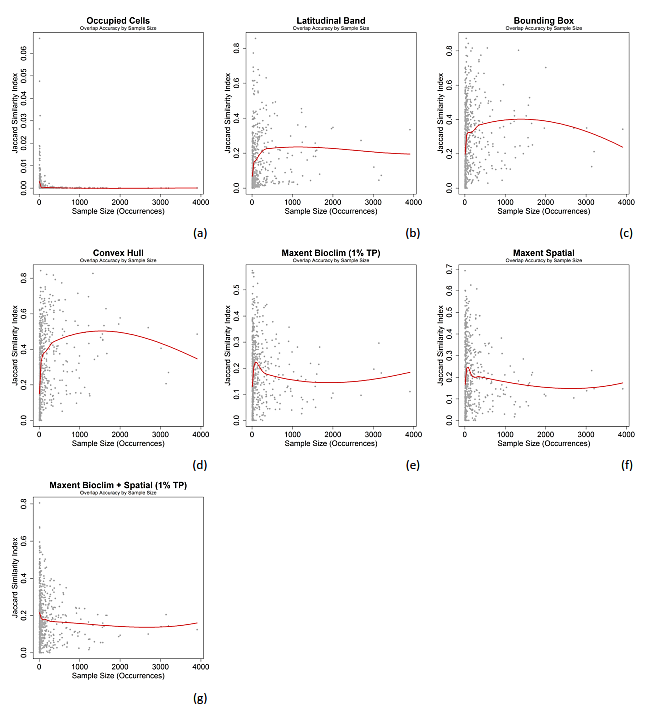
**Figure 3: LOW QUALITY PLACEHOLDER IMAGE BELOW**



**Figure 4: LOW QUALITY PLACEHOLDER IMAGE BELOW**



**Figure 5: LOW QUALITY PLACEHOLDER IMAGE BELOW**



## Supplemental Tables

**Table S1**: Summary of linear model results from pairwise comparisons of expert

range area and range areas obtained from all models for three datasets: Henderson’s tropical palms and Little’s temperate trees combined (n=634), tropical palms only (n=309), temperate trees only (n=329), and temperate restricted to species with geographic ranges that are confined to the contiguous United States (n=325).

|  |  |  |  |
| --- | --- | --- | --- |
| **COMBINED PALMS AND TREES (N=634)** | **R2** | **Intercept** | **Coefficient** |
| Expert Map vs Model Latitudinal Extent | 0.5466 | 0.994 | < 2.22e-16 |
| Expert Map vs Model Longitudinal Extent | 0.6000 | 0.994 | < 2.22e-16 |
| Expert Map vs Occupied Cell Area | 0.2421 | 0.831 | < 2.22e-16 |
| Expert Map vs Bounding Box | 0.5701 | 1.040 | < 2.22e-16 |
| Expert Map vs Convex Hull | 0.6052 | 1.010 | < 2.22e-16 |
| Expert Map vs Latitudinal Band | 0.4375 | 1.090 | < 2.22e-16 |
| Expert Map vs Bioclim Fixed | 0.2009 | 0.995 | < 2.22e-16 |
| Expert Map vs Bioclim Max SS | 0.2746 | 1.030 | < 2.22e-16 |
| Expert Map vs Bioclim Balanced SS | 0.2544 | 1.030 | < 2.22e-16 |
| Expert Map vs Bioclim Max Kappa | 0.0699 | 0.949 | < 2.22e-16 |
| Expert Map vs Bioclim Max TP | 0.3172 | 1.040 | < 2.22e-16 |
| Expert Map vs Bioclim 01TP | 0.3315 | 1.030 | < 2.22e-16 |
| Expert Map vs Bioclim 1 TP Clipped | 0.4053 | 1.020 | < 2.22e-16 |
| Expert Map vs Bioclim 5 TP | 0.3002 | 1.020 | < 2.22e-16 |
| Expert Map vs Bioclim 10 TP | 0.2606 | 1.010 | < 2.22e-16 |
| Expert Map vs Spatial Fixed | 0.1455 | 0.997 | < 2.22e-16 |
| Expert Map vs Spatial Max SS | 0.3548 | 1.020 | < 2.22e-16 |
| Expert Map vs Spatial Balanced SS | 0.3479 | 1.030 | < 2.22e-16 |
| Expert Map vs Spatial Max Kappa | 0.0172 | 0.950 | < 2.22e-16 |
| Expert Map vs Spatial Max TP | 0.3960 | 1.030 | < 2.22e-16 |
| Expert Map vs Spatial 01TP | 0.4027 | 1.030 | < 2.22e-16 |
| Expert Map vs Spatial 1 TP Clipped | 0.4513 | 1.020 | < 2.22e-16 |
| Expert Map vs Spatial 5 TP | 0.3400 | 1.020 | < 2.22e-16 |
| Expert Map vs Spatial 10 TP | 0.2812 | 1.010 | < 2.22e-16 |
| Expert Map vs BioclimSpatial Fixed | 0.4459 | 0.969 | < 2.22e-16 |
| Expert Map vs BioclimSpatial Max SS | 0.4777 | 1.010 | < 2.22e-16 |
| Expert Map vs BioclimSpatial Balanced SS | 0.4717 | 1.010 | < 2.22e-16 |
| Expert Map vs BioclimSpatial Max Kappa | 0.1607 | 0.931 | < 2.22e-16 |
| Expert Map vs BioclimSpatial Max TP | 0.4984 | 1.020 | < 2.22e-16 |
| Expert Map vs BioclimSpatial 01TP | 0.5198 | 1.010 | < 2.22e-16 |
| Expert Map vs BioclimSpatial 1 TP Clipped | 0.5229 | 1.010 | < 2.22e-16 |
| Expert Map vs BioclimSpatial 5 TP | 0.4853 | 0.997 | < 2.22e-16 |
| Expert Map vs BioclimSpatial 10 TP | 0.4346 | 0.988 | < 2.22e-16 |

Table S1 – [**COMBINED PALMS AND TREES (N=634)]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Threshold** | | **R2** | **Intercept** | **Coefficient** |
| **Basic measures** | | |  |  |  |
| Latitudinal Extent | | - | 0.5466 | 0.994 | < 2.22e-16 |
| Longitudinal Extent | | - | 0.6000 | 0.994 | < 2.22e-16 |
| Occupied Cell Area | | - | 0.2421 | 0.831 | < 2.22e-16 |
| Bounding Box | | - | 0.5701 | 1.040 | < 2.22e-16 |
| Convex Hull | | - | 0.6052 | 1.010 | < 2.22e-16 |
| Latitudinal Band | | - | 0.4375 | 1.090 | < 2.22e-16 |
| **Modeled measures** | |  |  |  |  |
| Climate only | | Fixed | 0.2009 | 0.995 | < 2.22e-16 |
| Max SS | 0.2746 | 1.030 | < 2.22e-16 |
| Balanced SS | 0.2544 | 1.030 | < 2.22e-16 |
| Max Kappa | 0.0699 | 0.949 | < 2.22e-16 |
| Max TP | 0.3172 | 1.040 | < 2.22e-16 |
| 01TP | 0.3315 | 1.030 | < 2.22e-16 |
| 1 TP Clipped | 0.4053 | 1.020 | < 2.22e-16 |
| 5 TP | 0.3002 | 1.020 | < 2.22e-16 |
| 10 TP | 0.2606 | 1.010 | < 2.22e-16 |
| Spatial only | | Fixed | 0.1455 | 0.997 | < 2.22e-16 |
| Max SS | 0.3548 | 1.020 | < 2.22e-16 |
| Balanced SS | 0.3479 | 1.030 | < 2.22e-16 |
| Max Kappa | 0.0172 | 0.950 | < 2.22e-16 |
| Max TP | 0.3960 | 1.030 | < 2.22e-16 |
| 01TP | 0.4027 | 1.030 | < 2.22e-16 |
| 1 TP Clipped | 0.4513 | 1.020 | < 2.22e-16 |
| 5 TP | 0.3400 | 1.020 | < 2.22e-16 |
| 10 TP | 0.2812 | 1.010 | < 2.22e-16 |
| Climate + spatial | | Fixed | 0.4459 | 0.969 | < 2.22e-16 |
| Max SS | 0.4777 | 1.010 | < 2.22e-16 |
| Balanced SS | 0.4717 | 1.010 | < 2.22e-16 |
| Max Kappa | 0.1607 | 0.931 | < 2.22e-16 |
| Max TP | 0.4984 | 1.020 | < 2.22e-16 |
| 01TP | 0.5198 | 1.010 | < 2.22e-16 |
| 1 TP Clipped | 0.5229 | 1.010 | < 2.22e-16 |
| 5 TP | 0.4853 | 0.997 | < 2.22e-16 |
| 10 TP | 0.4346 | 0.988 | < 2.22e-16 |

|  |  |  |  |
| --- | --- | --- | --- |
| **HENDERSON’S PALMS (N=309)** | **R2** | **Intercept** | **Coefficient** |
| Palm Map vs Latitudinal Extent | 0.5538 | 0.998 | < 2.22e-16 |
| Palm Map vs Longitudinal Extent | 0.5868 | 0.995 | < 2.22e-16 |
| Palm Map vs Occupied Cell Area | 0.1803 | 0.812 | < 2.22e-16 |
| Palm Map vs Bounding Box | 0.6414 | 1.020 | < 2.22e-16 |
| Palm Map vs Convex Hull | 0.6754 | 0.993 | < 2.22e-16 |
| Palm Map vs Latitudinal Band | 0.5289 | 1.090 | < 2.22e-16 |
| Palm Map vs Bioclim Fixed | 0.2960 | 1.010 | < 2.22e-16 |
| Palm Map vs Bioclim Max SS | 0.3127 | 1.040 | < 2.22e-16 |
| Palm Map vs Bioclim Balanced SS | 0.2856 | 1.040 | < 2.22e-16 |
| Palm Map vs Bioclim Max Kappa | 0.0827 | 0.957 | < 2.22e-16 |
| Palm Map vs Bioclim Max TP | 0.3126 | 1.050 | < 2.22e-16 |
| Palm Map vs Bioclim 01TP | 0.3479 | 1.040 | < 2.22e-16 |
| Palm Map vs Bioclim 1 TP Clipped | 0.4075 | 1.020 | < 2.22e-16 |
| Palm Map vs Bioclim 5 TP | 0.3694 | 1.030 | < 2.22e-16 |
| Palm Map vs Bioclim 10 TP | 0.3774 | 1.030 | < 2.22e-16 |
| Palm Map vs Spatial Fixed | 0.1916 | 1.020 | < 2.22e-16 |
| Palm Map vs Spatial Max SS | 0.3257 | 1.030 | < 2.22e-16 |
| Palm Map vs Spatial Balanced SS | 0.3224 | 1.030 | < 2.22e-16 |
| Palm Map vs Spatial Max Kappa | 0.0091 | 0.960 | < 2.22e-16 |
| Palm Map vs Spatial Max TP | 0.3378 | 1.040 | < 2.22e-16 |
| Palm Map vs Spatial 01TP | 0.3656 | 1.030 | < 2.22e-16 |
| Palm Map vs Spatial 1 TP Clipped | 0.4398 | 1.030 | < 2.22e-16 |
| Palm Map vs Spatial 5 TP | 0.3488 | 1.030 | < 2.22e-16 |
| Palm Map vs Spatial 10 TP | 0.3242 | 1.020 | < 2.22e-16 |
| Palm Map vs BioclimSpatial Fixed | 0.5497 | 0.979 | < 2.22e-16 |
| Palm Map vs BioclimSpatial Max SS | 0.5316 | 1.000 | < 2.22e-16 |
| Palm Map vs BioclimSpatial Balanced SS | 0.5448 | 1.010 | < 2.22e-16 |
| Palm Map vs BioclimSpatial Max Kappa | 0.1459 | 0.932 | < 2.22e-16 |
| Palm Map vs BioclimSpatial Max TP | 0.5421 | 1.010 | < 2.22e-16 |
| Palm Map vs BioclimSpatial 01TP | 0.5668 | 1.010 | < 2.22e-16 |
| Palm Map vs BioclimSpatial 1 TP Clipped | 0.5687 | 1.000 | < 2.22e-16 |
| Palm Map vs BioclimSpatial 5 TP | 0.5655 | 1.000 | < 2.22e-16 |
| Palm Map vs BioclimSpatial 10 TP | 0.5504 | 0.994 | < 2.22e-16 |

|  |  |  |  |
| --- | --- | --- | --- |
| **LITTLE’S TREES (N=329)** | **R2** | **Intercept** | **Coefficient** |
| Little Map vs Model Latitudinal Extent | 0.4668 | 0.990 | < 2.22e-16 |
| Little Map vs Model Longitudinal Extent | 0.5429 | 0.993 | < 2.22e-16 |
| Little Map vs Occupied Cell Area | 0.3317 | 0.847 | < 2.22e-16 |
| Little Map vs Bounding Box | 0.4656 | 1.050 | < 2.22e-16 |
| Little Map vs Convex Hull | 0.5265 | 1.020 | < 2.22e-16 |
| Little Map vs Latitudinal Band | 0.3047 | 1.090 | < 2.22e-16 |
| Little Map vs Bioclim Fixed | 0.1387 | 0.980 | < 2.22e-16 |
| Little Map vs Bioclim Max SS | 0.2339 | 1.020 | < 2.22e-16 |
| Little Map vs Bioclim Balanced SS | 0.2303 | 1.020 | < 2.22e-16 |
| Little Map vs Bioclim Max Kappa | 0.0488 | 0.942 | < 2.22e-16 |
| Little Map vs Bioclim Max TP | 0.3150 | 1.040 | < 2.22e-16 |
| Little Map vs Bioclim 01TP | 0.3146 | 1.030 | < 2.22e-16 |
| Little Map vs Bioclim 1 TP Clipped | 0.3952 | 1.020 | < 2.22e-16 |
| Little Map vs Bioclim 5 TP | 0.2368 | 1.010 | < 2.22e-16 |
| Little Map vs Bioclim 10 TP | 0.1625 | 0.997 | < 2.22e-16 |
| Little Map vs Spatial Fixed | 0.1497 | 0.981 | < 2.22e-16 |
| Little Map vs Spatial Max SS | 0.3358 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial Balanced SS | 0.3578 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial Max Kappa | 0.0178 | 0.942 | < 2.22e-16 |
| Little Map vs Spatial Max TP | 0.4182 | 1.030 | < 2.22e-16 |
| Little Map vs Spatial 01TP | 0.4026 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial 1 TP Clipped | 0.4084 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial 5 TP | 0.3062 | 1.010 | < 2.22e-16 |
| Little Map vs Spatial 10 TP | 0.2287 | 0.994 | < 2.22e-16 |
| Little Map vs BioclimSpatial Fixed | 0.3223 | 0.962 | < 2.22e-16 |
| Little Map vs BioclimSpatial Max SS | 0.3896 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial Balanced SS | 0.3624 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial Max Kappa | 0.1508 | 0.931 | < 2.22e-16 |
| Little Map vs BioclimSpatial Max TP | 0.4268 | 1.030 | < 2.22e-16 |
| Little Map vs BioclimSpatial 01TP | 0.4415 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial 1 TP Clipped | 0.4465 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial 5 TP | 0.3693 | 0.996 | < 2.22e-16 |
| Little Map vs BioclimSpatial 10 TP | 0.2887 | 0.983 | < 2.22e-16 |

|  |  |  |  |
| --- | --- | --- | --- |
| **LITTLE’S TREES - US ONLY (N=325)** | **R2** | **Intercept** | **Coefficient** |
| Little Map vs Model Latitudinal Extent | 0.4795 | 0.990 | < 2.22e-16 |
| Little Map vs Model Longitudinal Extent | 0.5560 | 0.993 | < 2.22e-16 |
| Little Map vs Occupied Cell Area | 0.3208 | 0.847 | < 2.22e-16 |
| Little Map vs Bounding Box | 0.4902 | 1.050 | < 2.22e-16 |
| Little Map vs Convex Hull | 0.5436 | 1.020 | < 2.22e-16 |
| Little Map vs Latitudinal Band | 0.3203 | 1.090 | < 2.22e-16 |
| Little Map vs Bioclim Fixed | 0.1506 | 0.979 | < 2.22e-16 |
| Little Map vs Bioclim Max SS | 0.2368 | 1.020 | < 2.22e-16 |
| Little Map vs Bioclim Balanced SS | 0.2327 | 1.020 | < 2.22e-16 |
| Little Map vs Bioclim Max Kappa | 0.0558 | 0.941 | < 2.22e-16 |
| Little Map vs Bioclim Max TP | 0.3164 | 1.040 | < 2.22e-16 |
| Little Map vs Bioclim 01TP | 0.3191 | 1.030 | < 2.22e-16 |
| Little Map vs Bioclim 1 TP Clipped | 0.3949 | 1.020 | < 2.22e-16 |
| Little Map vs Bioclim 5 TP | 0.2433 | 1.010 | < 2.22e-16 |
| Little Map vs Bioclim 10 TP | 0.1695 | 0.996 | < 2.22e-16 |
| Little Map vs Spatial Fixed | 0.1637 | 0.979 | < 2.22e-16 |
| Little Map vs Spatial Max SS | 0.4000 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial Balanced SS | 0.3984 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial Max Kappa | 0.0384 | 0.940 | < 2.22e-16 |
| Little Map vs Spatial Max TP | 0.4614 | 1.030 | < 2.22e-16 |
| Little Map vs Spatial 01TP | 0.4530 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial 1 TP Clipped | 0.4580 | 1.020 | < 2.22e-16 |
| Little Map vs Spatial 5 TP | 0.3570 | 1.000 | < 2.22e-16 |
| Little Map vs Spatial 10 TP | 0.2774 | 0.993 | < 2.22e-16 |
| Little Map vs BioclimSpatial Fixed | 0.3374 | 0.961 | < 2.22e-16 |
| Little Map vs BioclimSpatial Max SS | 0.3966 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial Balanced SS | 0.3677 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial Max Kappa | 0.1656 | 0.929 | < 2.22e-16 |
| Little Map vs BioclimSpatial Max TP | 0.4312 | 1.020 | < 2.22e-16 |
| Little Map vs BioclimSpatial 01TP | 0.4489 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial 1 TP Clipped | 0.4533 | 1.010 | < 2.22e-16 |
| Little Map vs BioclimSpatial 5 TP | 0.3797 | 0.995 | < 2.22e-16 |
| Little Map vs BioclimSpatial 10 TP | 0.2991 | 0.982 | < 2.22e-16 |

**Table S2:** R-squared values from pairwise comparisons of expert range area and range areas obtained from Maxent models parameterized with combinations of Bioclim layers and spatial filters and a variety of thresholds for three datasets: Henderson’s tropical palms and Little’s temperate trees combined (n=634), tropical palms only (n=309), temperate trees only (n=329), and temperate restricted to species with geographic ranges that are confined to the contiguous United States (n=325). The highest two r-squared values for each combination of threshold and environmental layer is shown in bold.

|  |  |  |  |
| --- | --- | --- | --- |
| **Combined Palms & Trees** | **Bioclim** | **Spatial** | **B+S** |
| Fixed Threshold | 0.2009 | 0.1455 | 0.4459 |
| Max Sensitivity & Specificity Threshold | 0.2746 | 0.3548 | 0.4777 |
| Balanced Sensitivity & Specificity Threshold | 0.2544 | 0.3479 | 0.4717 |
| Maximum Kappa Threshold | 0.0699 | 0.0172 | 0.1607 |
| Maximum Training Presence Threshold | 0.3172 | 0.3960 | 0.4984 |
| 1% Training Presence Threshold | **0.3315** | **0.4027** | **0.5198** |
| 1% Training Presence Threshold (Clipped) | **0.4053** | **0.4513** | **0.5229** |
| 5% Training Presence Threshold | 0.3002 | 0.3400 | 0.4853 |
| 10% Training Presence Threshold | 0.2606 | 0.2812 | 0.4346 |
|  |  |  |  |
| **Palms** | **Bioclim** | **Spatial** | **B+S** |
| Fixed Threshold | 0.2960 | 0.1916 | 0.5497 |
| Max Sensitivity & Specificity Threshold | 0.3127 | 0.3257 | 0.5316 |
| Balanced Sensitivity & Specificity Threshold | 0.2856 | 0.3224 | 0.5448 |
| Maximum Kappa Threshold | 0.0827 | 0.0091 | 0.1459 |
| Maximum Training Presence Threshold | 0.3126 | 0.3378 | 0.5421 |
| 1% Training Presence Threshold | 0.3479 | **0.3656** | **0.5668** |
| 1% Training Presence Threshold (Clipped) | **0.4075** | **0.4398** | **0.5687** |
| 5% Training Presence Threshold | 0.3694 | 0.3488 | 0.5655 |
| 10% Training Presence Threshold | **0.3774** | 0.3242 | 0.5504 |
|  |  |  |  |
| **Trees** | **Bioclim** | **Spatial** | **B+S** |
| Fixed Threshold | 0.1387 | 0.1497 | 0.3223 |
| Max Sensitivity & Specificity Threshold | 0.2339 | 0.3358 | 0.3896 |
| Balanced Sensitivity & Specificity Threshold | 0.2303 | 0.3578 | 0.3624 |
| Maximum Kappa Threshold | 0.0488 | 0.0178 | 0.1508 |
| Maximum Training Presence Threshold | **0.3150** | **0.4182** | 0.4268 |
| 1% Training Presence Threshold | 0.3146 | 0.4026 | **0.4415** |
| 1% Training Presence Threshold (Clipped) | **0.3952** | **0.4084** | **0.4465** |
| 5% Training Presence Threshold | 0.2368 | 0.3062 | 0.3693 |
| 10% Training Presence Threshold | 0.1625 | 0.2287 | 0.2887 |
|  |  |  |  |
| **Trees US Only** | **Bioclim** | **Spatial** | **B+S** |
| Fixed Threshold | 0.1506 | 0.1637 | 0.3374 |
| Max Sensitivity & Specificity Threshold | 0.2368 | 0.4000 | 0.3966 |
| Balanced Sensitivity & Specificity Threshold | 0.2327 | 0.3984 | 0.3677 |
| Maximum Kappa Threshold | 0.0558 | 0.0384 | 0.1656 |
| Maximum Training Presence Threshold | 0.3164 | **0.4614** | 0.4312 |
| 1% Training Presence Threshold | **0.3191** | 0.4530 | **0.4489** |
| 1% Training Presence Threshold (Clipped) | **0.3949** | **0.4580** | **0.4533** |
| 5% Training Presence Threshold | 0.2433 | 0.3570 | 0.3797 |
| 10% Training Presence Threshold | 0.1695 | 0.2774 | 0.2991 |

## Supplemental Figure Legends

**Figure S1:** Geographic range size for both tropical palms and temperate trees (n=630) plotted with sample size, for the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers. The Maxent modeled binary presence/absence maps were derived using the fixed and 1% training presence threshold settings.

**Figure S2:** True Skill Statistic (TSS) by sample size. TSS is calculated from contingency matrices calculated by the overlap of expert range maps for both tropical palms and temperate trees (n=630) and range maps obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers. The Maxent modeled binary presence/absence maps were derived using the 1% training presence threshold setting.

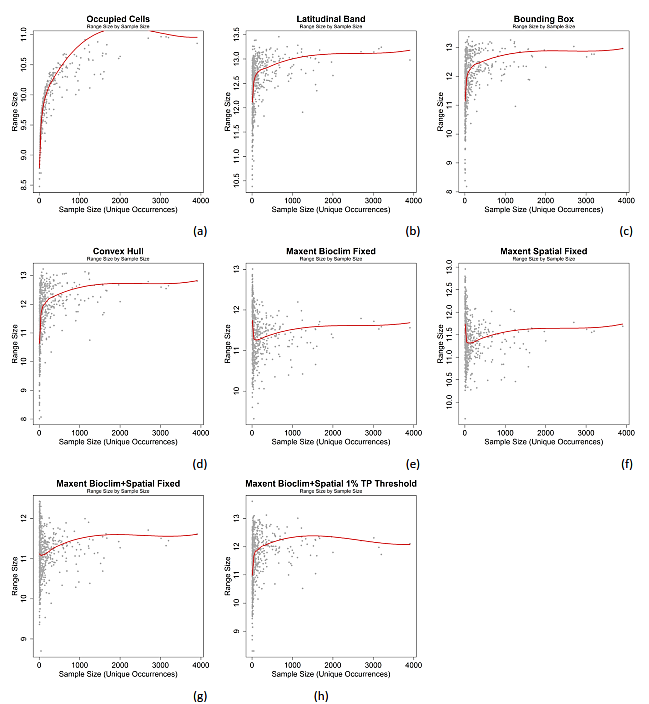
**Figure S3:** Accuracy (ACC) by sample size. Accuracy is calculated from contingency matrices calculated by the overlap of expert range maps for both tropical palms and temperate trees (n=630) and range maps obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers. The Maxent modeled binary presence/absence maps were derived using the 1% training presence threshold setting.

**Figure S4:** Kappa by sample size. Kappa is calculated from contingency matrices calculated by the overlap of expert range maps for both tropical palms and temperate trees (n=630) and range maps obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers. The Maxent modeled binary presence/absence maps were derived using the 1% training presence threshold setting.

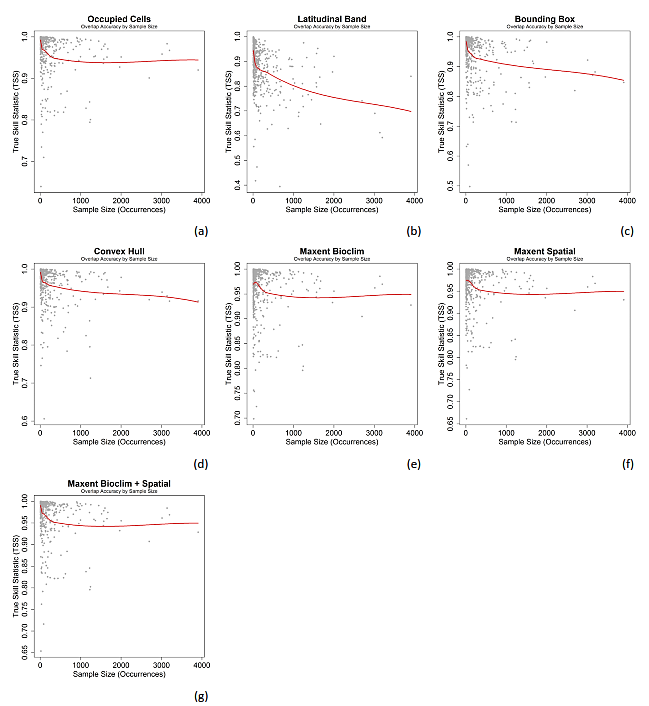
**Figure S5:** Delta maps comparing the absolute difference between species richness for both tropical palms and temperate trees (n=630) calculated from expert maps minus species richness from geographic range areas obtained from the following models: (a) area of occupied cells, (b) latitudinal band area, (c) bounding box area, (d) convex hull area, and the area obtained from Maxent modeled presence maps using (e) Bioclim layers only, (f) spatial layers only, and (g) Bioclim + spatial layers using the fixed threshold setting, as well as (h) Bioclim + spatial layers using 1% training presence threshold setting.

## Supplemental Figures

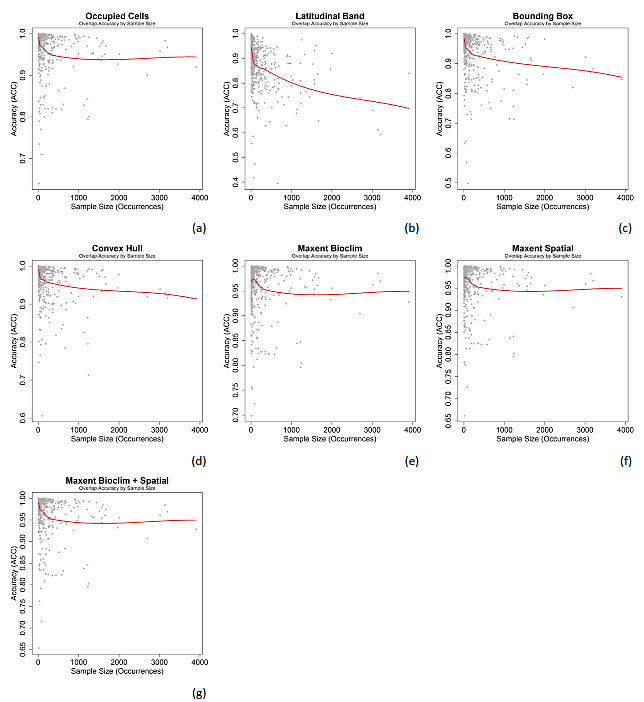
**Figure S1:**



**Figure S2: LOW QUALITY PLACEHOLDER IMAGE BELOW**



**Figure S3: LOW QUALITY PLACEHOLDER IMAGE BELOW**



**Figure S4: LOW QUALITY PLACEHOLDER IMAGE BELOW**

