GEOGRAPHIC RANGE SIZE: MEASURING THE FUNDAMENTAL UNIT OF BIOGEOGRAPHY AND EVALUATING CLIMATIC FACTORS THAT MAY INFLUENCE LONGITUDINAL RANGE SIZE GRADIENTS IN NORTH AMERICAN TREES

by

John Donoghue

A Thesis Submitted to the Faculty of the

SCHOOL OF NATURAL RESOURCES AND THE ENVIRONMENT

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE
WITH A MAJOR IN NATURAL RESOURCES

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

STATEMENT BY AUTHOR

The thesis titled Measuring the Fundamental Unit of Biogeography and Evaluating Climatic Factors That May Explain Longitudinal Range Size Gradients in North American Trees prepared by John C. Donoghue II has been submitted in partial fulfillment of requirements for a master's degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

SIGNED: John Donoghue

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Steven Archer
Professor of Natural Resources

15 Dec 2016
Date

ACKNOWLEDGEMENTS

Portions of this research were made possible through funding provided by the iPlant Collaborative, National Center for Ecology Analysis and Synthesis (NCEAS), the Botanical Information and Ecology Network (BIEN), the School of Natural Resources and the Environment (SNRE), the University of Arizona Graduate College, and my committee members.

I was fortunate to be involved in some remarkable collaborative research efforts that shaped my individual research program. This research could not have been completed without the assistance and support of many people, far more than can be mentioned individually; forgive me. Nonetheless, I wish to thank my committee members, Drs. Steve Archer, Brian McGill and Stuart Marsh, who selflessly provided mentorship, guidance, funding, and occasional moral support throughout my research. I especially want to thank Dr. Brian Enquist for his mentorship, guidance and generosity in providing me a home within his lab and inviting me to join his brilliant research team.

Finally, I wish to acknowledge all my wonderful colleagues in the Enquist and Archer labs at the University of Arizona, the McGill lab at the University of Maine, and the Svenning lab at Arhus University, Denmark for their friendship and support. They are by far the best scientists, and best people, I have had the pleasure of working with.

DEDICATION

To my wife, Christine Jacobs-Donoghue, who has always shared my interest in nature and has been my friend and foundation throughout our shared life together. This is for you.

and

To my family, who instilled in me the value of hard work, the desire for higher education and an appreciation and respect for the natural world.

and

To James de Lauriers, who first introduced me to the study of ecology and fueled me with his infectious fascination for the unanswered questions of nature.

TABLE OF CONTENTS

ABSTRACT 6

INTRODUCTION 7

PAPER 1: MEASURING THE FUNDAMENTAL UNIT OF BIOGEOGRAPHY 8
- Abstract 8
- Introduction 9
- Estimating Geographic Range Size 11
- Materials and Methods 12
- Results 16
- Discussion 19
- Acknowledgements 23
- Boxes 24
- Tables 29
- Figure Legends 31
- Figures 33
- Supplemental Tables 38
- Supplemental Figure Legends 42
- References 48

PAPER 2: DOES THE CLIMATIC VARIABILITY HYPOTHESIS EXPLAIN THE LONGITUDINAL RANGE SIZE GRADIENT IN NORTH AMERICAN TREES? 52
- Abstract 52
- Introduction 53
- Materials and Methods 54
- Results 56
- Discussion 58
- Conclusion 59
- Tables 61
- Figures 64
- References 69

CONCLUSION 71

ABSTRACT

This research seeks to advance our understanding of how to make better informed species conservation decisions on a global scale and advance our understanding of how species' spatial distributions (their geographic ranges) may be respond to climate change, so we can know which areas should be set aside to ensure their present and future conservation

To understand how species' geographic ranges may change, it's important to first assess how geographic ranges are defined and measured. The quantifiable measurement of a species' geographic range, (its geographic range size), is a key criterion the International Union for the Conservation of Nature uses to determine the conservation status and prioritization of species worldwide. Thus, part one of this thesis evaluates different measures for how geographic range size is commonly quantified in the conservation community, to determine whether some range size measures are more reliable than others.

Further, to evaluate how species' geographic ranges may respond to climate change, I examine the climatic factors influencing observable longitudinal range size gradients in the North American tree species range maps from E.L. Little's Atlas of North American Trees.

INTRODUCTION

One of the main goals of my research was to advance knowledge of topics concerning global species conservation. My interests in biogeography, and knowledge of geographic information systems, databases and programming, found large-scale ecology a good foundation from which to pursue studies that advanced our understanding of how to make better informed species conservation decisions on a global scale. I also wanted to advance our understanding of how species' spatial distributions (their geographic ranges) may be respond to climate change, so we can know which areas should be set aside to ensure their present and future conservation

To understand how species' geographic ranges may change, it's important to first assess how geographic ranges are defined and measured. The quantifiable measurement of a species' geographic range, (its geographic range size), is a key criterion the International Union for the Conservation of Nature (IUCN) uses to determine the conservation status and prioritization of species worldwide (IUCN, 2001; Baillie, Hilton-Taylor & Stuart 2004; Gaston and Fuller, 2008). Thus, part of my research program sought to evaluate different measures for how geographic range size is commonly quantified in the conservation community, to determine whether some range size measures are more reliable than others. This research is presented herein in the chapter entitled "Paper 1".

Second, one way to evaluate how species' geographic ranges may respond to climate change is by examining the factors that influence the current geographic distributions of species. Thus, a second part of my research program sought to understand the factors influencing the geographic range sizes of North American tree species by examining the climatic factors influencing observable longitudinal range size gradients in the North American tree species range maps from E.L. Little's Atlas of North American Trees (Critchfield and Little, 1966; Little, 1971; Little 1976-1978). This research is presented herein in the chapter entitled "Paper 2".

This thesis details the methods, materials, and results of these separate investigations.

PAPER 1: MEASURING THE FUNDAMENTAL UNIT OF BIOGEOGRAPHY

Authors: John C. Donoghue II[1], N. Morueta-Holme[2], B. Boyle[3], L. L. Sloat[1], B. J. Enquist[1], B. J. McGill[4], J.- C. Svenning[2], & The BIEN Working Group[5]

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.

Abstract

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.

Introduction

A species' geographic range defines the spatial distribution of the species on the earth. The quantifiable 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. Thus, 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, that combine 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 also reviewed for accuracy and suitability for this analysis (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 was 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 to facilitate comparison of the geographic range size distribution estimates with those from expert maps. 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 ($r^2 = 0.547$, $r^2 = 0.600$ for the combined temperate and tropical datasets, respectively, $p < 2.22e^{-16}$). 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 temperate or tropical 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 ($r^2 = 0.6052$, $p < 2.22e^{-16}$). 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 ($r^2 = 0.570$, $p < 2.22e^{-16}$). 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 ($r^2 = 0.4275$, $p < 2.22e^{-16}$), 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 ($r^2 = 0.3208$, $p < 2.22e^{-16}$, Table S1), but was the worst fitting model for tropical palms ($r^2 = 0.1803$, $p < 2.22e^{-16}$, 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 ($r^2 = 0.1455$, $p < 2.22e^{-16}$). Similarly, Maxent models parameterized with only the 19 climate layers, and using a fixed threshold, were not much better ($r^2 = 0.2009$, $p < 2.22e^{-16}$). 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 ($r^2 = 0.4459$, $p < 2.22e^{-16}$), and were consistently the third highest ranking model after range areas derived from convex 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 for all predictor sets, 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 r^2 values (Table 1).

Accuracy of Predicted Distribution

In quantifying the spatial agreement between the geographic ranges obtained from each method with those of the expert maps, the convex hull 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).

We found strong differences between models with respect to the dependence of accuracy on sample size, as measured by the Jaccard Similarity Index. 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 samples. Beyond 75 samples 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 samples, they continued to exhibit a subsequent decline beyond 75 samples.

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 profound implications for distribution modeling, as it implies that at least for macroecological studies, complex models such as Maxent may not be necessary to obtain a good estimate of a species geographic range size and, potentially, its 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. Moreover, the use of convex hulls to estimate geographic range size can mitigate potential circularity problems with calculating range sizes based on models using environmental variables, and subsequently using the same environmental variables to explain range size distribution or richness patterns.

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. This is not surprising, as occupied cells reflect a species "occupied range" rather than its potential range shown in the expert maps. 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 (r^2 = 0.3208; Table S1), but fit tropical palms poorly (r^2 = 0.1803; Table S1), likely because 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 a 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 results that constrain the over-prediction bias resulting from the use of solely environmental data We further recommend clipping any non-contiguous regions more than 1,000 km away from the main continuous range area of presence/absence maps generated from Maxent models to account for over-prediction of each species' geographic distribution. 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).

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 area 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. defined using the 1% training presence threshold above.

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 could 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 many 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.

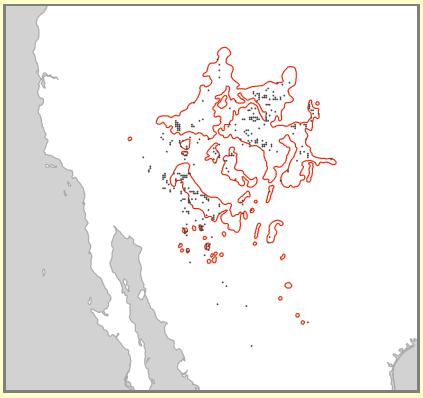
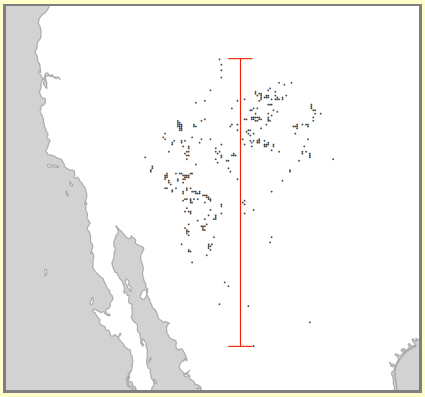
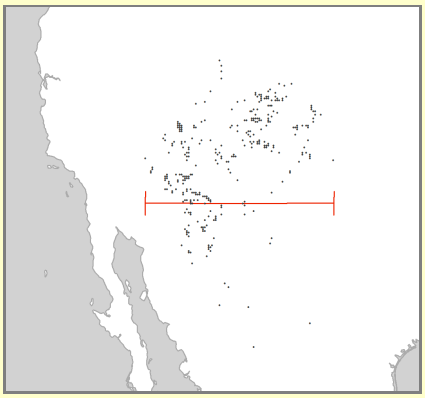
Lastly, combining occurrence data from many sources is not without problems, and we expended considerable effort validating both the geographic and taxonomic status of the occurrence data used for this analysis. Geographic latitude and longitude values for each occurrence record were plotted and compared 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. It is also possible that some species' geographic ranges have changed within the past 40 years. 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.

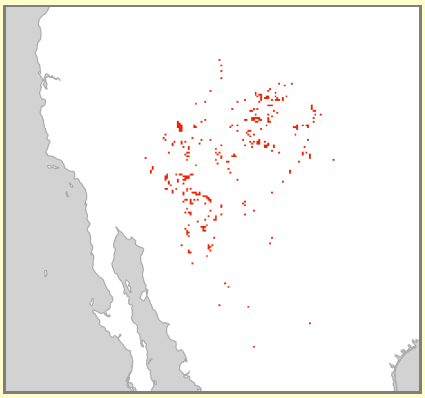
Acknowledgements

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.

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.

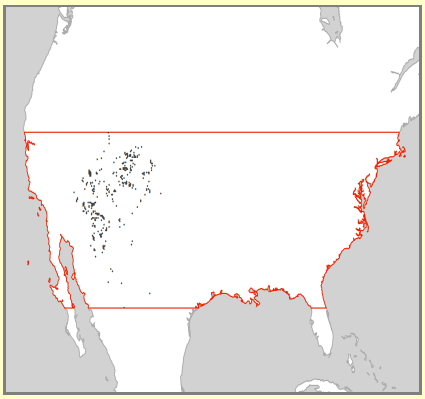
	**Expert Map** **Inputs:** Species occurrence data points **Output:** Total geographic area of one or more polygons manually drawn based on species occurrences and expert opinion.
	**Latitudinal Extent:** **Inputs:** Species occurrence data points **Output:** Linear distance between the latitude of the northernmost and southernmost geographic locations in your dataset.
	**Longitudinal Extent** **Inputs:** Species occurrence data points **Output:** Linear distance between the longitude of the westernmost and easternmost geographic locations in your dataset.



Occupied Cell Area

Inputs: Species occurrence data points

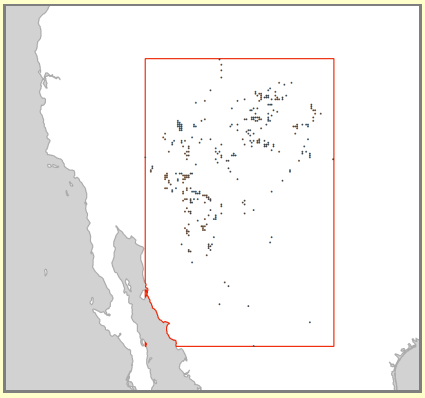
Output: Total geographic area of all cells containing at least one occurrence of your focal species. Cells are created by converting species occurrence points to a raster surface at a particular scale and spatial resolution.



Latitudinal Band

Inputs: Species occurrence data points

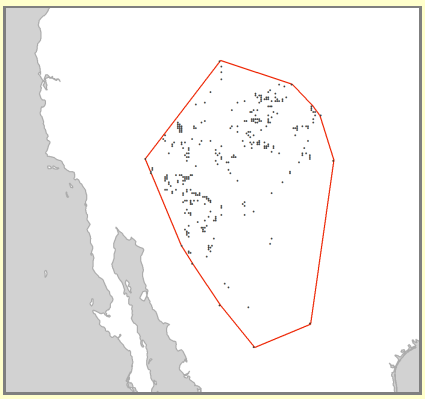
Output: Total geographic area of a single polygon defined by the latitude of the northernmost and southernmost geographic locations in your dataset and the western and eastern limits of the continent.



Bounding Box

Inputs: Species occurrence data points

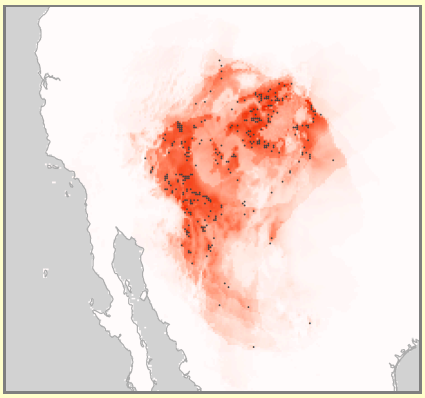
Output: Total geographic area of a single polygon defined by the latitude of the northernmost and southernmost geographic locations and the longitude of the westernmost and easternmost geographic locations in your dataset.



Convex Hull

Inputs: Species occurrence data points

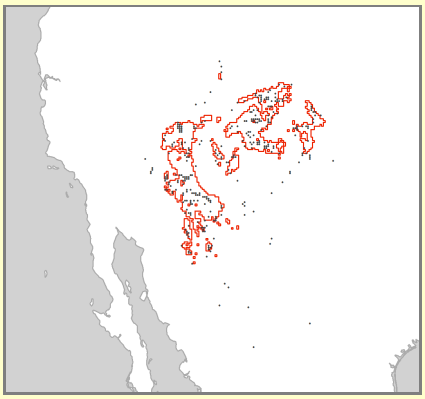
Output: Total geographic area of the smallest single polygon that encompasses all the species occurrence locations in your dataset.



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.



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	Threshold that 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 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 two combined datasets Henderson's tropical palms and Little's temperate trees. 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)	R^2	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
	0.972	0.001	1.000	0.002	0.001	0.001
Occupied Cells	+/- 0.096	+/- 0.008	+/- 0.00	+/- 0.016	+/- 0.008	+/- 0.008
	0.894	0.834	0.893	0.192	0.726	0.138
Latitudinal Band	+/- 0.168	+/- 0.447	+/- 0.178	+/- 0.365	+/- 0.407	+/- 0.304
	0.955	0.739	0.958	0.390	0.696	0.284
Bounding Box	+/- 0.126	+/- 0.529	+/- 0.132	+/- 0.459	+/- 0.499	+/- 0.391
	0.971	0.568	0.979	0.413	0.547	0.302
Convex Hull	+/- 0.086	+/- 0.575	+/- 0.079	+/- 0.472	+/- 0.550	+/- 0.403
	0.968	0.367	0.990	0.278	0.356	0.179
Maxent Bioclim	+/- 0.092	+/- 0.48	+/- 0.045	+/- 0.331	+/- 0.466	+/- 0.235
	0.969	0.423	0.991	0.308	0.415	0.204
Maxent Spatial	+/- 0.089	+/- 0.579	+/- 0.038	+/- 0.386	+/- 0.569	+/- 0.287
Maxent	0.973	0.301	0.997	0.291	0.298	0.189
Bioclim+Spatial	+/- 0.087	+/- 0.475	+/- 0.010	+/- 0.354	+/- 0.473	+/- 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 evaluating the effect of sampling intensity, by 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.

Figures

Figure 1:

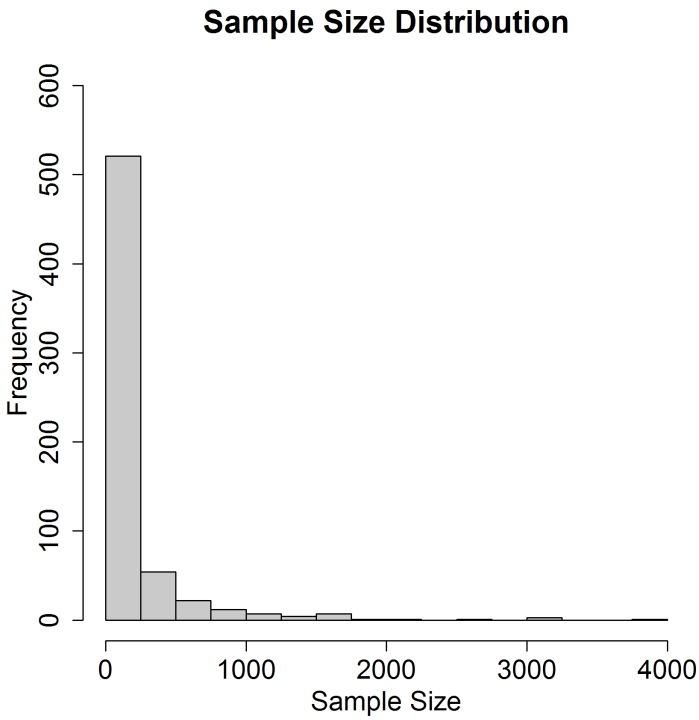


Figure 2:

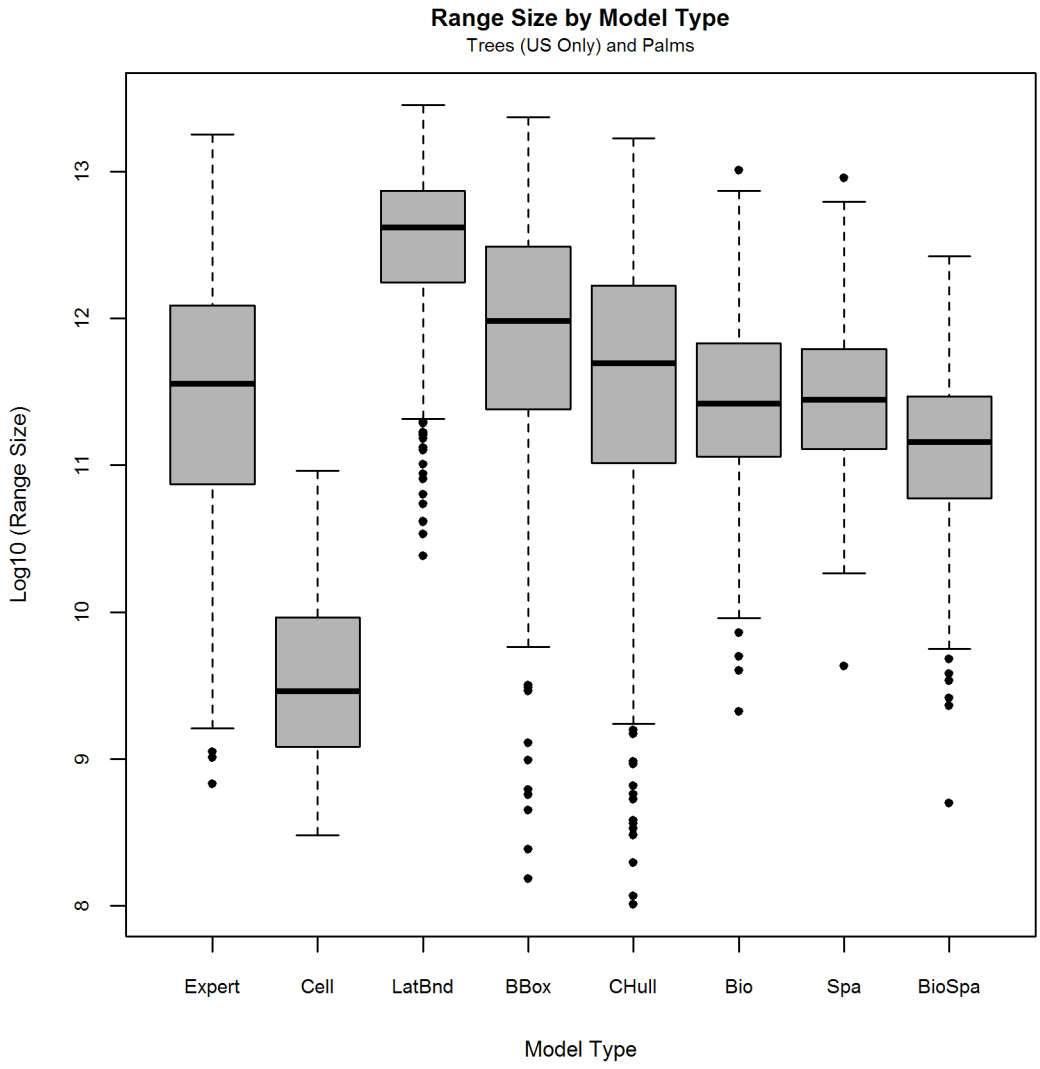

Range Size by Model Type
Trees (US Only) and Palms
Log10 (Range Size)
8
9
10
11
12
13
Expert
Cell
LatBnd
BBox
CHull
Bio
Spa
BioSpa
Model Type

Figure 3:

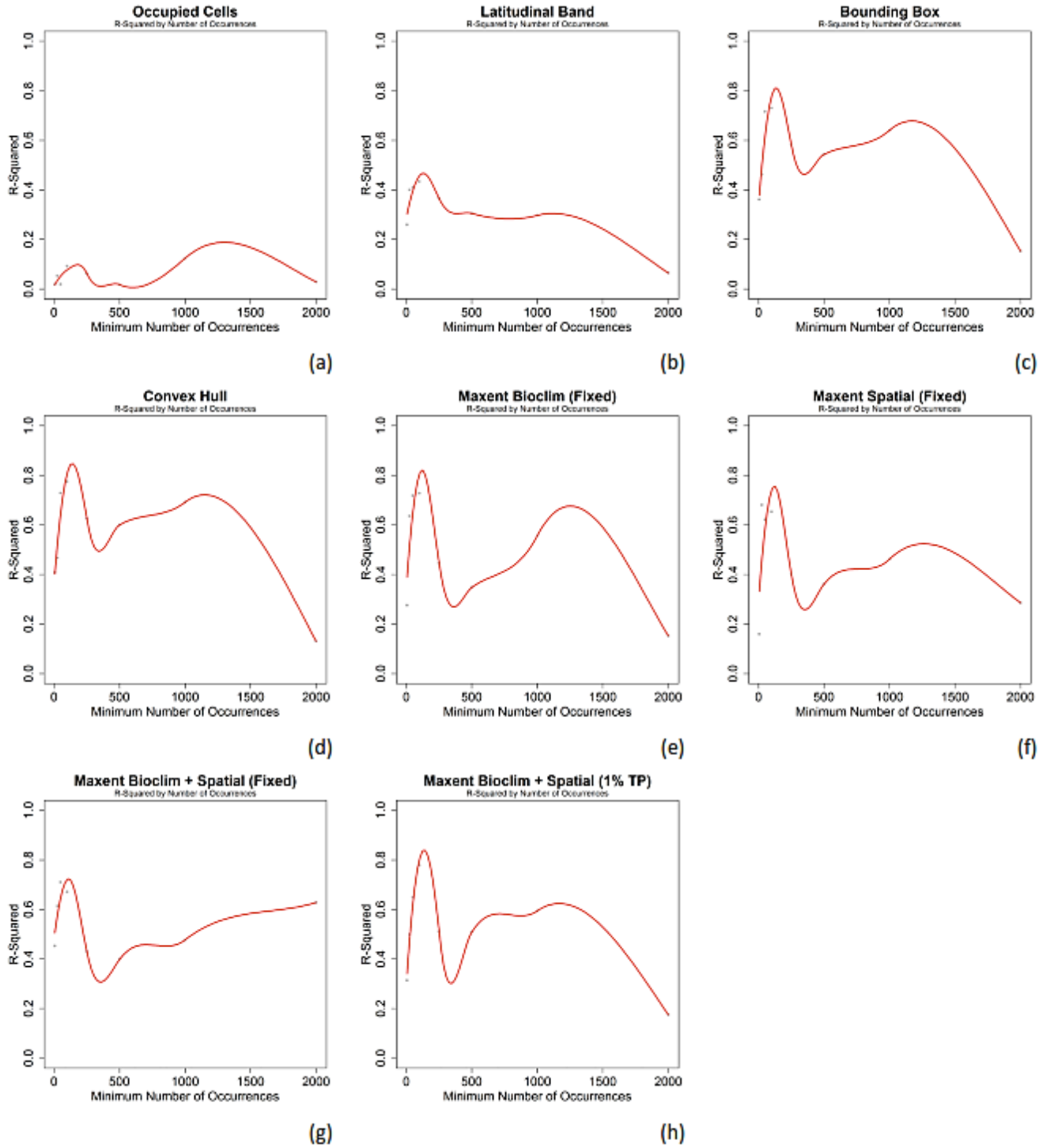

Occupied Cells
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(a)
Latitudinal Band
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(b)
Bounding Box
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(c)
Convex Hull
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(d)
Maxent Bioclim (Fixed)
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(e)
Maxent Spatial (Fixed)
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(f)
Maxent Bioclim + Spatial (Fixed)
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(g)
Maxent Bioclim + Spatial (1% TP)
R-Squared by Number of Occurrences
R-Squared
Minimum Number of Occurrences
(h)

Figure 4:

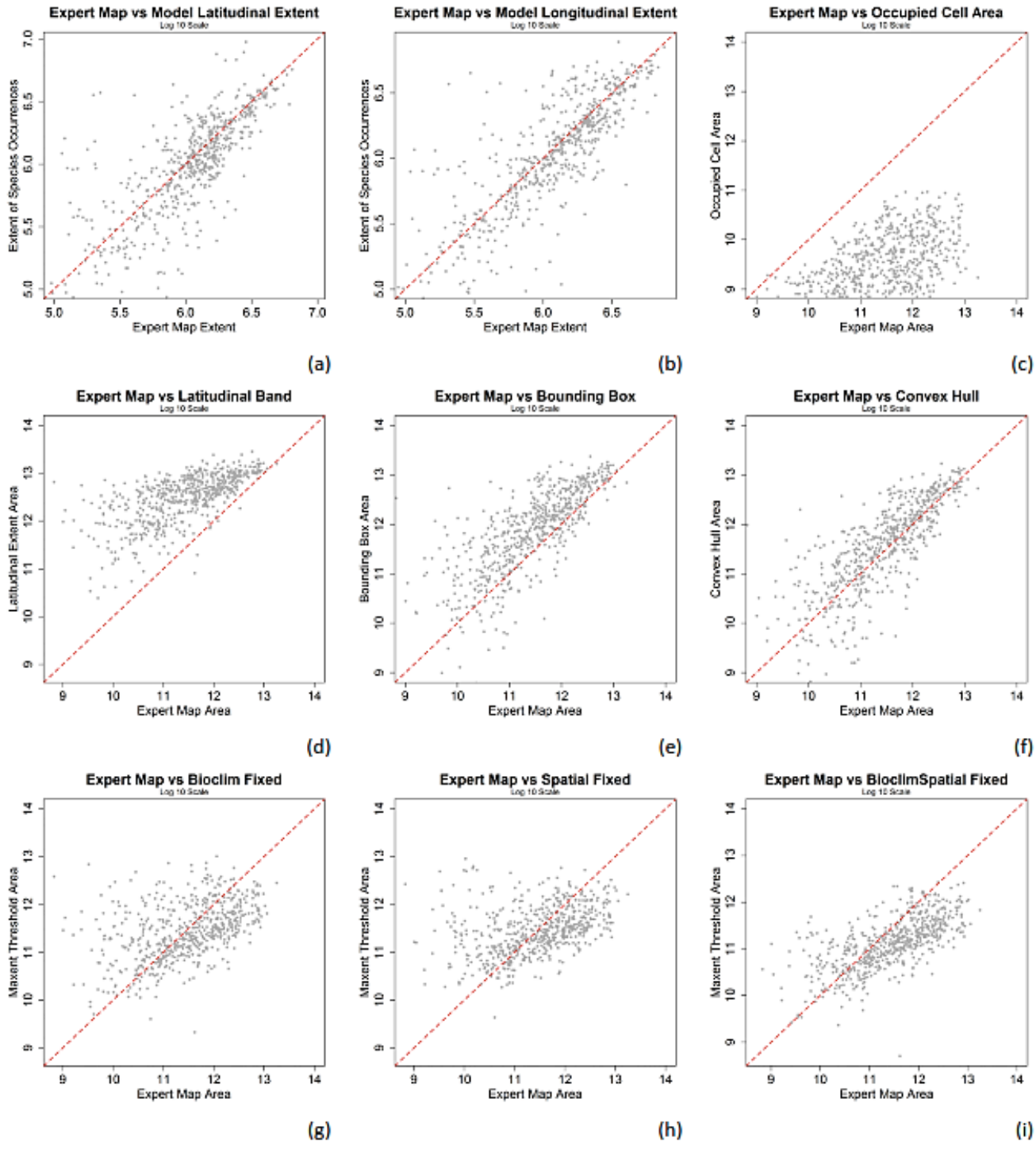
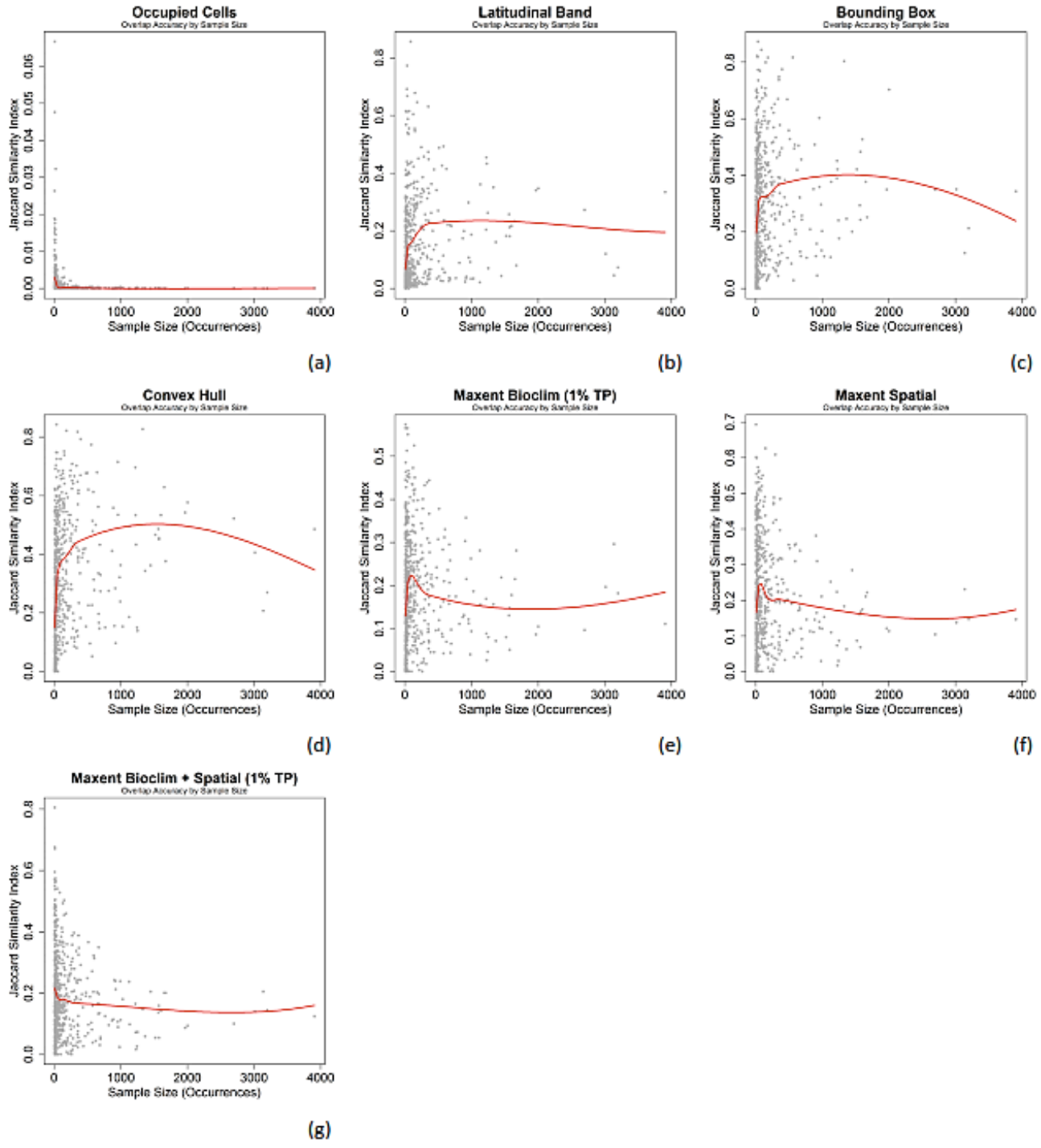

Expert Map vs Model Latitudinal Extent
Log 10 Scale
Extent of Species Occurrences
Expert Map Extent
(a)
Expert Map vs Model Longitudinal Extent
Log 10 Scale
Extent of Species Occurrences
Expert Map Extent
(b)
Expert Map vs Occupied Cell Area
Log 10 Scale
Occupied Cell Area
Expert Map Area
(c)
Expert Map vs Latitudinal Band
Log 10 Scale
Latitudinal Extent Area
Expert Map Area
(d)
Expert Map vs Bounding Box
Log 10 Scale
Bounding Box Area
Expert Map Area
(e)
Expert Map vs Convex Hull
Log 10 Scale
Convex Hull Area
Expert Map Area
(f)
Expert Map vs Bioclim Fixed
Log 10 Scale
Maxent Threshold Area
Expert Map Area
(g)
Expert Map vs Spatial Fixed
Log 10 Scale
Maxent Threshold Area
Expert Map Area
(h)
Expert Map vs BioclimSpatial Fixed
Log 10 Scale
Maxent Threshold Area
Expert Map Area
(i)

Figure 5:


Occupied Cells
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(a)
Latitudinal Band
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(b)
Bounding Box
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(c)
Convex Hull
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(d)
Maxent Bioclim (1% TP)
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(e)
Maxent Spatial
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(f)
Maxent Bioclim + Spatial (1% TP)
Overlap Accuracy by Sample Size
Jaccard Similarity Index
Sample Size (Occurrences)
(g)

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).

Table S1: COMBINED PALMS AND TREES (N=634)

	Threshold	R^2	Intercept	P
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				
Using on Climate Layers	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
	1TP	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
Using Only Spatial Layers	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
	1TP	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
Using Both Climate &	Fixed	0.4459	0.969	< 2.22e-16

Spatial Layers	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
	1TP	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

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 (Log 10 km^2) 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:

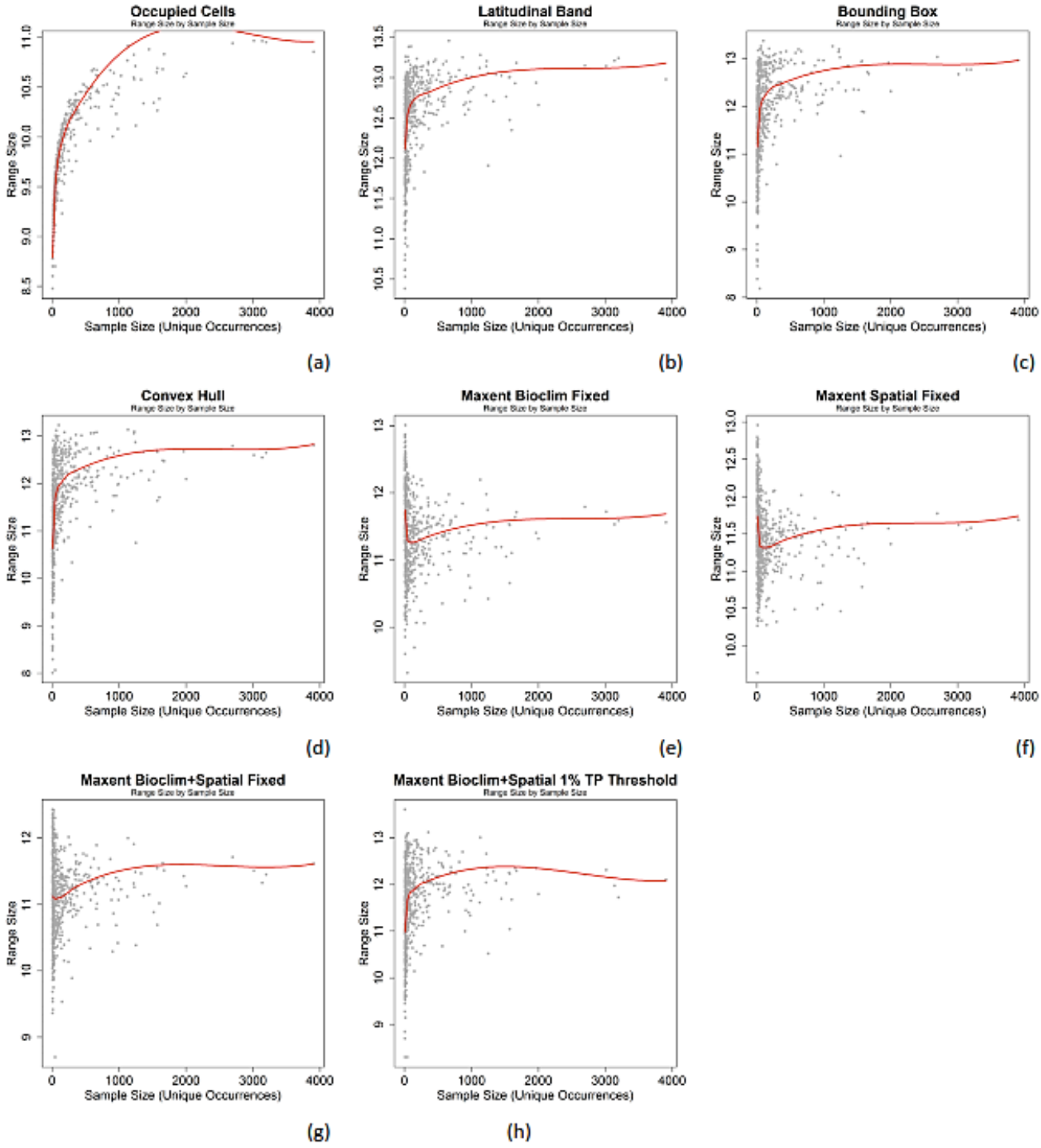


Figure S2:

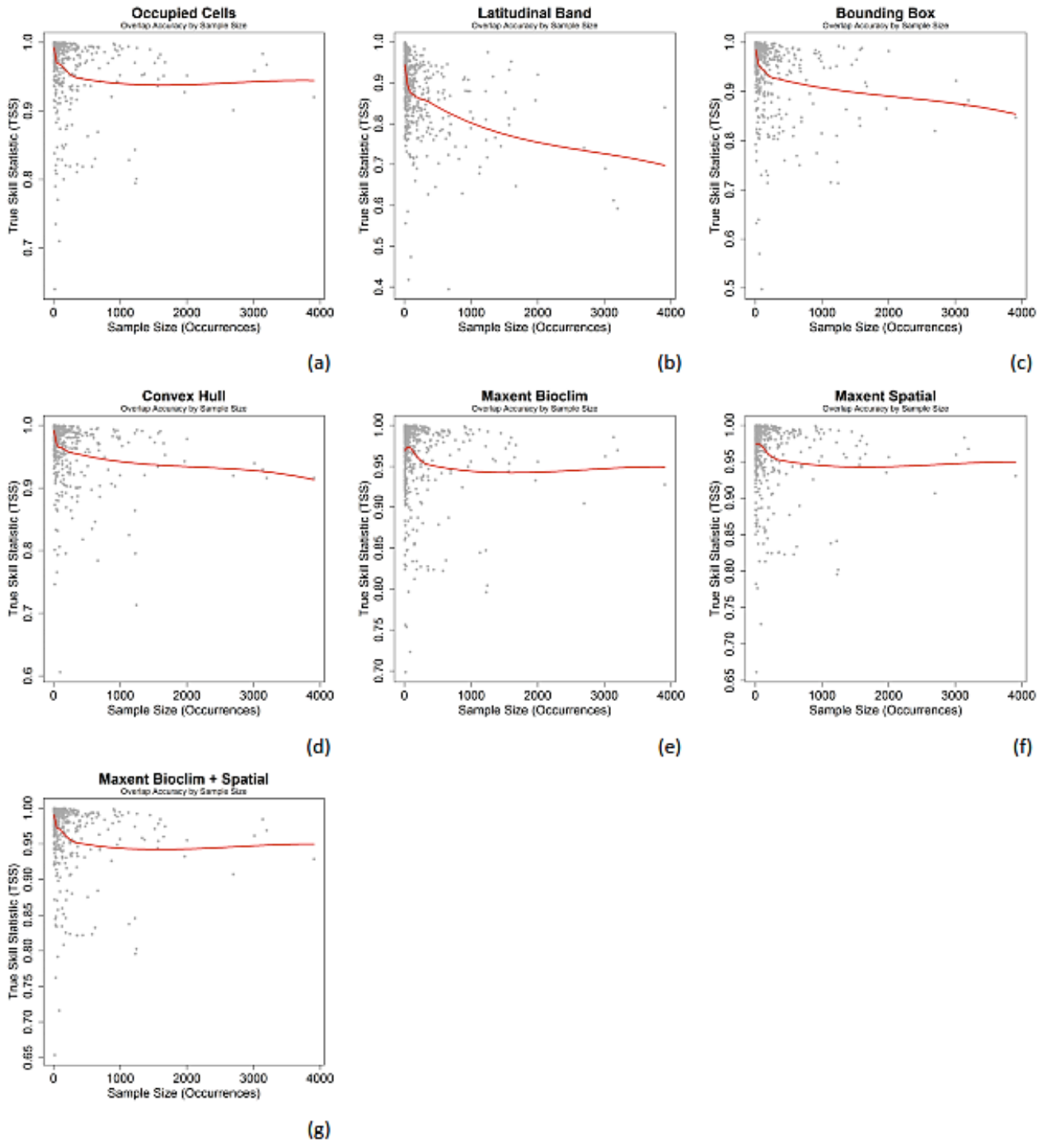

Occupied Cells
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(a)
Latitudinal Band
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(b)
Bounding Box
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(c)
Convex Hull
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(d)
Maxent Bioclim
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(e)
Maxent Spatial
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(f)
Maxent Bioclim + Spatial
Overlap Accuracy by Sample Size
True Skill Statistic (TSS)
Sample Size (Occurrences)
(g)

Figure S3:

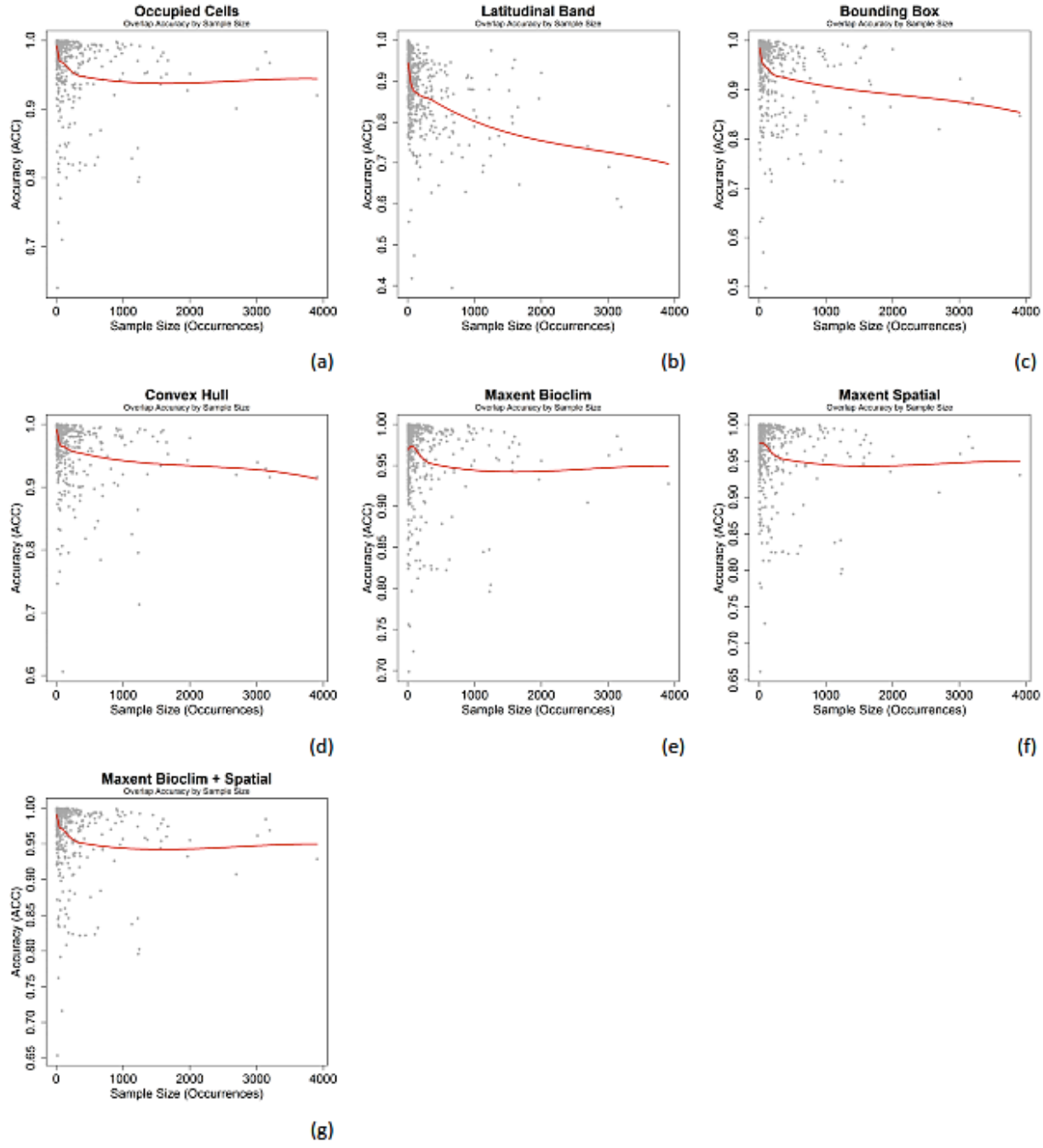
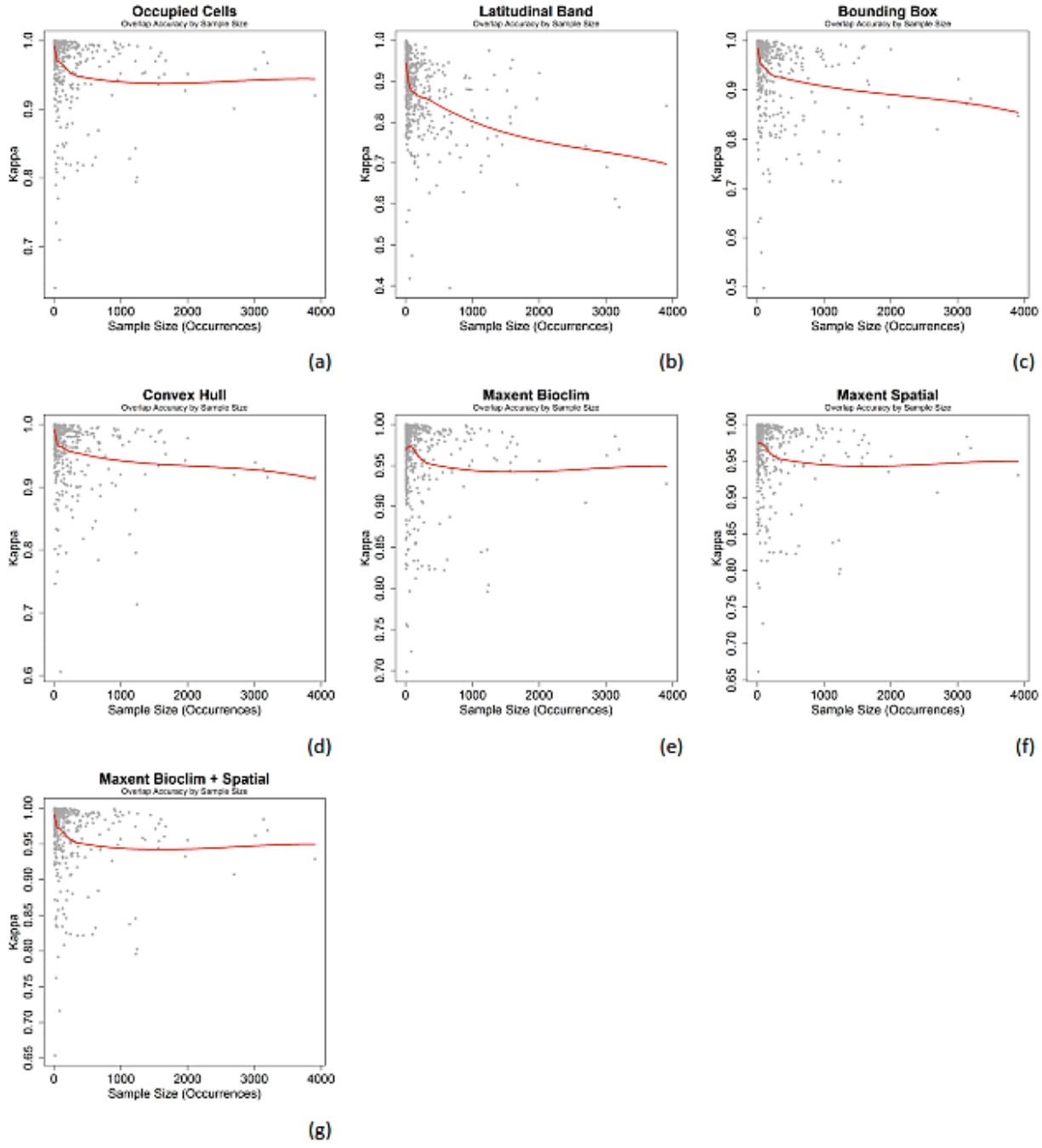

Occupied Cells
Latitudinal Band
Bounding Box
Convex Hull
Maxent Bioclim
Maxent Spatial
Maxent Bioclim + Spatial
Accuracy (ACC)
Sample Size (Occurrences)
(a)
(b)
(c)
(d)
(e)
(f)
(g)

Figure S4:


Occupied Cells
Latitudinal Band
Bounding Box
Convex Hull
Maxent Bioclim
Maxent Spatial
Maxent Bioclim + Spatial
Kappa
Sample Size (Occurrences)
(a)
(b)
(c)
(d)
(e)
(f)
(g)

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.

PAPER 2: DOES THE CLIMATIC VARIABILITY HYPOTHESIS EXPLAIN THE LONGITUDINAL RANGE SIZE GRADIENT IN NORTH AMERICAN TREES?

Authors: John C. Donoghue II[1], B. J. McGill[2]

Affiliations:
[1] School of Natural Resources and the Environment, University of Arizona, Tucson, AZ 85721, USA.
[4] School of Biology and Ecology / Sustainability Solutions Initiative, University of Maine, Orono, ME 04469, USA.

Abstract

Latitudinal range size gradients have long been documented for North American tree species, and recent work has shown that latitudinal range size gradients are consistent with the climatic variability hypothesis. While longitudinal range size gradients for North American tree species have also been documented, it is uncertain whether this pattern can also be explained by the climatic variability hypothesis. In this study, I evaluated whether the longitudinal gradient in geographic range sizes can be explained by climatic variability.

Using digital representations of range maps from E.L. Little's Atlas of North American Trees, information from the USDA Plant Database, and climatic data from WorldClim, I performed a principal components analysis on 19 bioclimatic variables extracted from the geographic limits of each species' range, to ascertain how geographic range size fluctuated with multidimensional climatic variability. I also used the PCA results to identify which bioclimatic factors most influenced climatic variability over geographic ranges, and tested how those combined factors varied with range size.

Results show that variation in PCA 1 was driven by factors related to temperature, while variation in PCA 2 was driven by factors related to precipitation. PCA 1 exhibited a moderately statistically significant negative relationship with range size for all species ($r^2 = 0.3675$, $p = 0.0$). The pattern was somewhat weaker for gymnosperms ($r^2 = 0.2483$, $p < 0.0001$) but stronger for angiosperms ($r^2 = 0.4222$, $p < 0.0001$). PCA 2 showed only a moderately statistically significant positive relationship with range size for gymnosperms ($r^2 = 0.2127$, $p < 0.0001$). A regression

model using five bioclimatic factors most influential in determining climatic variation across species' ranges was statistically significant, explaining 66% of the variation in range size ($r^2 = 0.66$, $p < 0.0001$).

This study shows that the longitudinal range size gradient for North American trees is consistent with the climatic variability hypothesis. Species with small geographic ranges sample a climatic space with more variable temperature ranges and precipitation regimes than large range species. The results also seem to contradict the notion that species' geographic range sizes are driven by whether they are climatic specialists or generalists. Large range species may not be climatic generalists; they may be simply exploiting a more homogenous climate.

Introduction

Geographic range size is a basic and fundamental unit of biogeography (Brown et al., 1996) and the geographic patterns and distribution of species range size can inform our understanding of the processes that limit species distributions (Gaston, 2003) and our projections of how species distributions may be influenced by climate change.

One commonly cited geographic pattern of species' range size is Rapoport's rule (Rapoport, 1975; Rapoport 1982; Stevens, 1989), in which range sizes of plants and animals have been observed to be smaller at lower latitudes and increase with higher latitudes. Morin and Lechowicz (2011) tested whether the Rapoport effect of a latitudinal gradient of range size could result from selectively favoring species with greater climatic tolerances at higher latitudes. Their results showed significant positive correlations between range size and latitudinal midpoint, thus demonstrating that the climatic variability hypothesis is consistent with the latitudinal range size gradient observed for many species (Morin and Lechowicz, 2011).

Similarly, longitudinal gradients of range size (using longitudinal midpoint) have been observed for North American tree species, with geographic range sizes east of the Rocky Mountains being, generally, larger than range sizes in the western portion of the continent (Morin and Lechowicz, 2011). Possible explanations for this pattern include differing degrees of isolation and

paleofloristic history between the western and eastern portions of the continent (Qian, 2001) and greater heterogeneity of topography and edaphic climatic gradients in the western portion of the continent that could restrict range expansion and increase rates of speciation (Morin and Lechowicz, 2011). It is also possible that the overall longitudinal gradient of range size across the continent is consistent with the climatic stability hypothesis, which has yet to be investigated, though Morin and Lechowicz (2011) proposed that the contrasting patterns of range size in the west or eastern portions of the continent might be observed through phylogenetic differences in climatic tolerance.

For a climatic stability hypothesis to be valid, range sizes would have to be inversely proportional to overall climatic variability. Further, the individual factors influencing the climatic variability should also exhibit variation that contributes to either climatic variability or stability and the direction of this variation should be evident along a west to east gradient. Accordingly, this study aims to test these hypotheses by 1) evaluating the degree to which climatic variability is related to geographic range size, 2) determining which climatic factors, if any, best explain the variability of range sizes, and 3) ascertaining how the factors individually influence climatic variability such that they help explain the differences in range sizes between the western and eastern portions of the United States.

Materials and Methods

For this study, I used tree species from E.L. Little's Atlas of North American Trees (Critchfield and Little, 1966; Little, 1971; Little 1976-1978). Using information from the USDA Plant Database (USDA, NRCS, 2011), I reviewed the current taxonomic status of each tree species and excluded those with sequent taxonomic changes that make it difficult to compare the species' currently recognized taxonomic status with the corresponding species represented in E.L. Little's Range Maps. I also grouped each species by family, taxonomic group (angiosperm or gymnosperm), primary growth habit (tree, shrub, etc.) and growth duration (annual or perennial) with data obtained from the USDA Plant Database.

I obtained digital range maps for each species represented in Little's atlases from the USGS in shapefile format (U.S. Geological Survey, 1999) and excluded species with range maps that were unnaturally truncated at political boundaries. Species with range maps that seemed overly generalized (as being represented by simple circles, ovals or triangular shapes) were marked for exclusion in subsequent analyses. To do this, each species' geographic range was mapped in a Lambert Equal Area projection and each species' mean perimeter-area fractal dimension (PAFRAC) was calculated. The PAFRAC provides an index of how much a feature's shape deviates from simple Euclidean geometry, with 0 indicating a feature is Euclidean and higher values representing more complex shapes. Species with geographic range areas having PAFRAC values below 1 were marked for exclusion, as their shapes did not deviate from simple Euclidean geometry types. Thus, they were not likely realistic representations of species' ranges, which are expected to be more complex shapes.

For the remaining 500 species, the total area of each species' range map (in km^2) was calculated using ArcGIS (ESRI, 2011 and the center point of each range (in northing and easting metric units) was recorded. I obtained climatic data from WorldClim (Hijmans et al., 2005) in 2.5 km raster resolution and transformed it to the Lambert Equal Area projection system for analysis with the range maps and used R (R Development Core Team, 2007) to clip each bioclimatic layer obtained from WorldClim by each species range boundary. I then computed the minimum, maximum, mean and variance of each bioclimatic layer's measurements across each species' range. Each measure was then log transformed for analysis.

A series of linear models were developed to investigate the relationship between range size and the longitudinal center point of each range, and evaluate the relationship between range size and each bioclimatic factor's mean and variance. In addition, a measure of multidimensional climatic variability was obtained by performing a principal components analysis on the mean and variance for all 19 bioclimatic factor measurements. The individual loadings of the first two PCA components were employed in subsequent regression models to examine how each PCA component varied with geographic range size. This provided a means of determining whether there was any relationship between multidimensional climatic variability and range size. Finally, I developed a series of regression models relating the individual loadings of the first two PCA

components with the mean and variance of each bioclimatic factor, as calculated across each species' range. This provided a means of determining which bioclimatic factors were most responsible for the variation exhibited in range size, from which I also constructed a multiple linear regression model using the five bioclimatic factors that were observed to most influence range size.

Results

A simple linear regression model of the variation in range size with the longitude of the center point of each species' range provides results that are consistent with the general observation of larger ranges in the eastern half of the United States (Figure 1). However, the relationship is weak ($r^2 = 0.1917$, P = 1.19 e-24, n=496).

Simple linear models relating geographic range size with individual mean bioclimatic factors show moderate correlations, though the strength of each relationship and importance of individual bioclimatic factors differed between aggregate species, gymnosperms and angiosperms (Table 1). In general, when analyzing mean bioclimatic measurements, isothermality stood out as the single most important factor, while mean annual temperature and mean temperature of the driest quarter were secondary important factors to gymnosperms and angiosperms respectively. In contrast, analysis of the variance of bioclimatic factor measurements provided mixed results (Table 2). Temperature seasonality and mean temperature of the coldest month were the most important factors for gymnosperms and angiosperms respectively. Mean temperature of the coldest quarter was consistently the second most important factor among aggregate species, gymnosperms and angiosperms.

Results of the principal components analysis show that many species are oriented along axes determined by two or more bioclimatic factors, with most species clustering around a few factors. In addition, this orientation appears to be strong for most species, as they are oriented some distance away from the plot center. Further, the orientation of the variation in bioclimatic factors is also strong, as the arrows originating from the plot center are long (Figure 2).

Results of the regression models built to evaluate the strength of the relationship between PCA component 1 and range size show a moderately statistically significant negative correlation of PCA component 1 with range size for all species (Figure 3a), (r^2= 0.3675, p = 0.0, n=500). The pattern is somewhat weaker for only gymnosperms (Figure 3b), (r^2= 0.2483, p = 5.8 e-7, n=90) but stronger for angiosperms (figure 3c), (r^2= 0.4222, p = 2.0 e-50, n=409). In contrast, regression models evaluating the strength of the relationship between PCA component 2 and range size were not statistically significant (Figures 3d, 3f), except for gymnosperms (figure 3e), which showed a moderately significant positive correlation with PCA 2 (r^2= 0.2127, p = 4.76 e-6, n=90).

When evaluating the results of the PCA on mean BIOCLIM variables (Table 3a), PCA 1 appears to be most strongly correlated with BIOCLIM layer 7 (Temperature Annual Range), layer 4 (Temperature Seasonality) and layer 3 (Isothermality). In contrast PCA 2 is consistently strongly correlated with BIOCLIM layer 17 (Precipitation of Driest Quarter), followed by layer 12 (Annual Precipitation) and layer 14 (Precipitation of the Driest Month).

In contrast, examination of the results of the PCA on the variance of BIOCLIM variables (Table 3b) shows that PCA 1 is consistently moderately correlated with BIOCLIM layer 9 (Mean Temperature of Driest Quarter), while PCA 2 is consistently correlated with BIOCLIM layer 14 (Precipitation of Driest Month) and layer 17 (Precipitation of Driest Quarter), though the correlations are less strong.

Based on these results I developed a parsimonious multiple linear regression using five bioclimatic variables that appeared to be most consistently influential in determining range size as derived from the PCA results. These variables were BIOCLIM layer 7 (Temperature Annual Range), Layer 9 (Mean Temperature of the Driest Quarter), layer 12 (Annual Precipitation), layer 14 (Precipitation of Driest Month) and layer 17 (Precipitation of Driest Quarter). The model result was statistically significant and the five bioclimatic variables combine to explain 66% of the variation in range size (r^2= 0.66, p = 2.2 e-16, DF=491, n=497).

Discussion

The moderately significant negative correlation between PCA component 1 and range size is consistent with the hypothesis that range size should be inversely proportional to climatic variability. Mean range size is observed to decrease with increasing climatic variability, with slightly different rates of decrease between gymnosperms and angiosperms.

The PCA of bioclimatic factor means results show that species range sizes appear to be primarily influenced by variation in temperature factors that drives PCA 1, and secondarily by factors relating to the variation in precipitation that drives PCA 2. This general pattern is maintained for both gymnosperms and angiosperms, though each taxonomic group appears to be influenced by slightly different individual bioclimatic factors.

Gymnosperm range-size fluctuations with PCA 1 are related to temperature and are most influenced by isothermality, temperature seasonality, and annual temperature range, respectively. Gymnosperm range-size fluctuations with PCA 2 are related to precipitation, and are most influenced by precipitation of the driest quarter, mean diurnal range of monthly minimum and maximum temperatures, and precipitation of the driest month, respectively.

Angiosperm range-size fluctuations with PCA 1 also appear to be related to temperature, and are most influenced by annual temperature range, temperature seasonality, and minimum temperature of the coldest month. Angiosperm range-size fluctuations with PCA 2 are also related to precipitation, and are most influenced by precipitation of the driest quarter, annual precipitation and precipitation of the driest month.

The PCA on the variance of bioclimatic measures results also demonstrate that both gymnosperm and angiosperm range-size fluctuations with PCA 1 are generally influenced by temperature, with mean temperature of the driest quarter consistently the most important factor and mean temperature of the wettest quarter a secondary factor of importance to angiosperms only. Both gymnosperm and angiosperm range-size fluctuations with PCA 2 are influenced by precipitation, with precipitation of the driest month and precipitation of the driest quarter of importance to both taxonomic groups.

Since the results of the parsimonious multiple linear regression analysis show that these bioclimatic factors can explain 66% of the variation in range size, these factors are also expected to individually vary along a west to east transect, with an increase or decrease in the variation along the transect consistent with each factor's expected contributions to either climatic variability or stability, thus influencing range size. For example, for the five bioclimatic factors in the parsimonious multiple linear regression analysis, Table 4 shows the direction that each factor is expected to vary to positively influence range size by increasing climatic stability (or decreasing variability), thus supporting the bioclimatic variability hypothesis.

Temperature annual range (BIO 7) is expected to decrease along a west-east transect because a declining range would be indicative of more stable temperatures and reduced variability in temperature extremes. Similarly, mean temperature of the driest quarter (BIO 9) is also expected to decrease, as cooler temperatures during the driest quarter of the year are expected for eastern continental areas with more moderate temperatures, and hence lower variability, than the western continental areas. Similarly, precipitation of both the driest month (BIO 14) and driest quarter (Bio 17) are expected to increase along the transect, because more stable (i.e. less variable) environments should have more stable precipitation gradients which would have higher precipitation in the driest portions of the year relative to areas of more variable climate. Finally, the more consistent precipitation regimes of the east are expected to result in higher overall annual precipitation (BIO 12) than the west.

When each bioclimatic factor is plotted separately, the direction of its variation is consistent with the expectations listed in Table 4. This is illustrated below (Figures 4a-j), in which each bioclimatic factor is plotted on a map and paired with a graph showing a profile of each factor's measured value along a west-east transect that includes a trend line.

Conclusion

By extracting the bioclimatic factors within each species' geographic range, we sample the specific climatic environment that each species experiences. Accordingly, we see that the

longitudinal range size gradient for North American trees is consistent with the climatic variability hypothesis. Species with small geographic ranges sample a climatic space with more variable temperature and precipitation regimes than species with large geographic ranges. This finding is somewhat surprising as it contradicts the general notion that species' geographic range sizes are determined by whether they are climatic specialists or generalists, with large range species assumed to be climatic generalists that can tolerate variable climates. According to the results presented in this study, large range species are not necessarily climatic generalists, as they sample a less variable climatic space than small range species, which have small ranges but sample a wider range of the environment. Instead, large range species may simply be exploiting a more homogenous climate.

Tables

Table 1: R-Squared values for linear regressions of range size and mean bioclimatic factors. Cells highlighted in green indicate strong correlations (above 0.4). Cells highlighted in red indicate weak correlations (between 0.2 and 0.4)

	All Species		Gymnosperms		Angiosperms	
Bioclimatic Layer	**n**	**R-Squared**	**n**	**R-Squared**	**n**	**R-Squared**
BIO1 = Annual Mean Temperature	489	0.2221	85	0.2993	404	0.3015
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))	500	0.0115	90	0.1207	409	0.0011
BIO3 = Isothermality (BIO2/BIO7) (* 100)	500	0.3906	90	0.4265	409	0.3713
BIO4 = Temperature Seasonality (standard deviation *100)	500	0.2563	90	0.239	409	0.2638
BIO5 = Max Temperature of Warmest Month	500	0.1335	90	0.1587	409	0.1579
BIO6 = Min Temperature of Coldest Month	251	0.0602	36	0.0092	215	0.0869
BIO7 = Temperature Annual Range (BIO5-BIO6)	500	0.2464	90	0.2914	409	0.2348
BIO8 = Mean Temperature of Wettest Quarter	495	0.0024	85	0.005	409	0.0196
BIO9 = Mean Temperature of Driest Quarter	442	0.2725	77	0.2328	365	0.3111
BIO10 = Mean Temperature of Warmest Quarter	500	0.1114	90	0.1085	409	0.15
BIO11 = Mean Temperature of Coldest Quarter	380	0.1319	52	0.017	328	0.1638
BIO12 = Annual Precipitation	500	0.0009	90	0.0392	409	0.0001
BIO13 = Precipitation of Wettest Month	500	0.049	90	0.1281	409	0.032
BIO14 = Precipitation of Driest Month	498	0.0486	90	0.0647	407	0.0473
BIO15 = Precipitation Seasonality (Coefficient of Variation)	500	0.0719	90	0.0611	409	0.0796
BIO16 = Precipitation of Wettest Quarter	500	0.0319	90	0.114	409	0.0171
BIO17 = Precipitation of Driest Quarter	500	0.0409	90	0.052	409	0.041
BIO18 = Precipitation of Warmest Quarter	500	0.0091	90	0.022	409	0.0062
BIO19 = Precipitation of Coldest Quarter	500	0.0002	90	0.0146	409	0.001

Table 2: R-Squared values for linear regressions of range size and variance of bioclimatic factors. Cells highlighted in red indicate weak correlations (between 0.2 and 0.4).

	All Species		Gymnosperms		Angiosperms	
Bioclimatic Layer	**n**	**R-Squared**	**n**	**R-Squared**	**n**	**R-Squared**
BIO1 = Annual Mean Temperature	499	0.1863	90	0.1182	408	0.2319
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))	500	0.0028	90	0	409	0.0053
BIO3 = Isothermality (BIO2/BIO7) (* 100)	500	0.0442	90	0.0019	409	0.0694
BIO4 = Temperature Seasonality (standard deviation *100)	500	0.1439	90	0.2662	409	0.1448
BIO5 = Max Temperature of Warmest Month	500	0.0711	90	0.0059	409	0.1066
BIO6 = Min Temperature of Coldest Month	500	0.3	90	0.2046	409	0.3486
BIO7 = Temperature Annual Range (BIO5-BIO6)	500	0.193	90	0.1776	409	0.21
BIO8 = Mean Temperature of Wettest Quarter	499	0.1078	90	0.0674	408	0.1324
BIO9 = Mean Temperature of Driest Quarter	499	0.2578	90	0.1922	408	0.2964
BIO10 = Mean Temperature of Warmest Quarter	499	0.0832	90	0.0078	408	0.1219
BIO11 = Mean Temperature of Coldest Quarter	499	0.2701	90	0.2445	408	0.3088
BIO12 = Annual Precipitation	500	0.0599	90	0.0002	409	0.1113
BIO13 = Precipitation of Wettest Month	500	0.0023	90	0.0734	409	0.0007
BIO14 = Precipitation of Driest Month	498	0.1963	90	0.1458	407	0.2163
BIO15 = Precipitation Seasonality (Coefficient of Variation)	500	0.0809	90	0.061	409	0.0886
BIO16 = Precipitation of Wettest Quarter	500	0.0001	90	0.0538	409	0.0043
BIO17 = Precipitation of Driest Quarter	500	0.1957	90	0.15	409	0.2153
BIO18 = Precipitation of Warmest Quarter	500	0.0178	90	0.0067	409	0.024
BIO19 = Precipitation of Coldest Quarter	500	0.0768	90	0.0024	409	0.1244

Table3a: Results of regression models evaluating how the individual loadings of principal components 1 and 2 vary with the mean values of each bioclimatic layer. Values with grey backgrounds indicate $r^2 > 0.4$

	All Species		Gymnosperms		Angiosperms	
Bioclimatic Layer (Mean Value)	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**
BIO1 = Annual Mean Temperature	0.6519	0.0018	0.7225	0.0673	0.6390	0.0002
BIO2 = Mean Diurnal Range	0.0327	0.4267	0.0919	0.7004	0.0894	0.3760
BIO3 = Isothermality	0.6749	0.1879	0.8193	0.2824	0.6773	0.1710
BIO4 = Temperature Seasonality	0.7101	0.0264	0.7952	0.0779	0.7689	0.0147
BIO5 = Max Temperature of Warmest Month	0.3794	0.0289	0.3876	0.1740	0.3673	0.0311
BIO6 = Min Temperature of Coldest Month	0.6699	0.0341	0.6313	0.0666	0.6921	0.0250
BIO7 = Temperature Annual Range	0.7459	0.0048	0.7453	0.0019	0.7740	0.0087
BIO8 = Mean Temperature of Wettest Quarter	0.1743	0.0060	0.1011	0.1257	0.2469	0.0003
BIO9 = Mean Temperature of Driest Quarter	0.4452	0.0207	0.4793	0.0222	0.4435	0.0230
BIO10 = Mean Temperature of Warmest Quarter	0.4923	0.0009	0.4700	0.0506	0.5238	0.0021
BIO11 = Mean Temperature of Coldest Quarter	0.6465	0.0181	0.6880	0.0335	0.6412	0.0193
BIO12 = Annual Precipitation	0.0665	0.7854	0.1432	0.5589	0.0580	0.8384
BIO13 = Precipitation of Wettest Month	0.2911	0.3392	0.4884	0.0912	0.2779	0.4325
BIO14 = Precipitation of Driest Month	0.0089	0.7849	0.1422	0.6400	0.0033	0.8139
BIO15 = Precipitation Seasonality	0.1704	0.4861	0.3449	0.3657	0.1717	0.5115
BIO16 = Precipitation of Wettest Quarter	0.2128	0.3979	0.4323	0.1191	0.1968	0.5019
BIO17 = Precipitation of Driest Quarter	0.0077	0.8182	0.1392	0.7031	0.0028	0.8413
BIO18 = Precipitation of Warmest Quarter	0.0668	0.3828	0.0240	0.1487	0.0713	0.4549
BIO19 = Precipitation of Coldest Quarter	0.0151	0.4489	0.0527	0.3214	0.0079	0.5039

Table3b: Results of regression models evaluating how the individual loadings of principal components 1 and 2 vary with the variance of each bioclimatic layer. Values with grey backgrounds indicate $r^2 > 0.4$

	All Species		Gymnosperms		Angiosperms	
Bioclimatic Layer (Variance)	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**
BIO1 = Annual Mean Temperature	0.3474	0.0520	0.0368	0.0000	0.4262	0.0601
BIO2 = Mean Diurnal Range	0.0038	0.1530	0.0137	0.0005	0.0002	0.2118
BIO3 = Isothermality	0.0019	0.0679	0.0145	0.0119	0.0078	0.1115
BIO4 = Temperature Seasonality	0.2248	0.0325	0.1941	0.0148	0.2421	0.0451
BIO5 = Max Temperature of Warmest Month	0.2894	0.2134	0.0191	0.0753	0.3549	0.2362
BIO6 = Min Temperature of Coldest Month	0.2990	0.0005	0.1131	0.0651	0.3413	0.0032
BIO7 = Temperature Annual Range	0.2024	0.0450	0.2034	0.0173	0.1949	0.0693
BIO8 = Mean Temperature of Wettest Quarter	0.3921	0.0022	0.2129	0.0950	0.4497	0.0061
BIO9 = Mean Temperature of Driest Quarter	0.4925	0.0135	0.3693	0.0807	0.5422	0.0101
BIO10 = Mean Temperature of Warmest Quarter	0.2840	0.2276	0.0018	0.1352	0.3648	0.2422
BIO11 = Mean Temperature of Coldest Quarter	0.3713	0.0024	0.1119	0.0573	0.4351	0.0055
BIO12 = Annual Precipitation	0.0338	0.0012	0.0168	0.0921	0.0471	0.0001
BIO13 = Precipitation of Wettest Month	0.0187	0.0442	0.1760	0.0004	0.0172	0.0601
BIO14 = Precipitation of Driest Month	0.1451	0.3659	0.1357	0.5221	0.1565	0.3374
BIO15 = Precipitation Seasonality	0.0545	0.1491	0.0422	0.0059	0.0578	0.2043
BIO16 = Precipitation of Wettest Quarter	0.0102	0.0316	0.1435	0.0011	0.0104	0.0419
BIO17 = Precipitation of Driest Quarter	0.1568	0.3105	0.1338	0.5701	0.1697	0.2721
BIO18 = Precipitation of Warmest Quarter	0.0442	0.0269	0.0709	0.0621	0.0397	0.0205
BIO19 = Precipitation of Coldest Quarter	0.2312	0.0367	0.1077	0.2112	0.2525	0.0273

Table4: Expected West-to-East direction of variation for the five important bioclimatic factors.

Bioclimatic Factor	**Direction**
BIO 7 Temperature Annual Range	Decrease
BIO 9 Mean Temperature of the Driest Quarter	Decrease
BIO 12 Annual Precipitation	Increase
BIO 14 Precipitation of Driest Month	Increase
BIO 17 Precipitation of Driest Quarter	Increase

Figures

Figure 1: Simple linear regression of range size and longitude of range center point

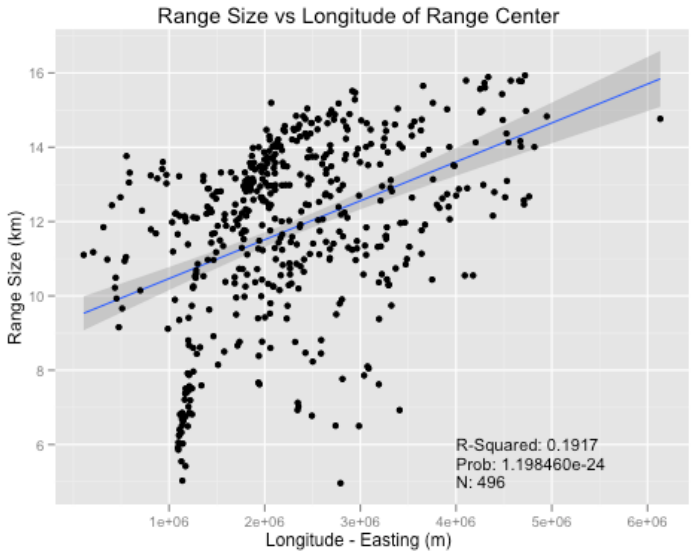


Figure 2a-f: Linear regression models of PCA 1 and 2 vs range size for combined species, gymnosperms and angiosperms.

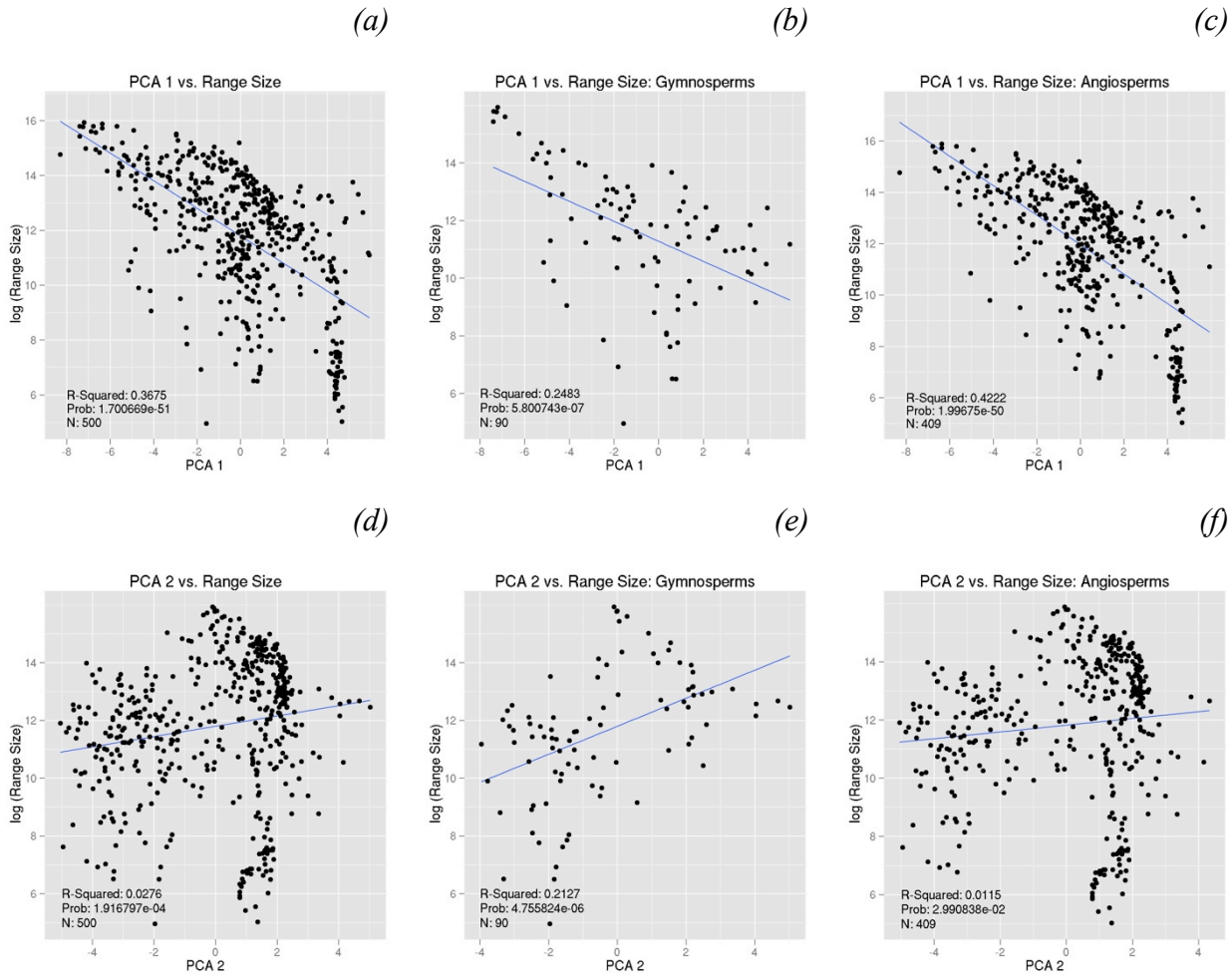


Figure 3: Principal components analysis of mean bioclimatic factors across all species ranges.

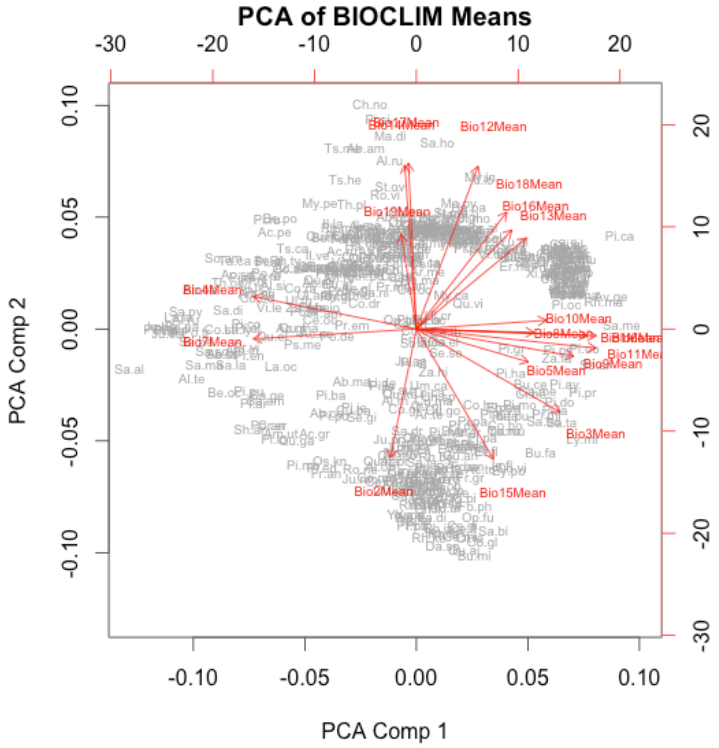
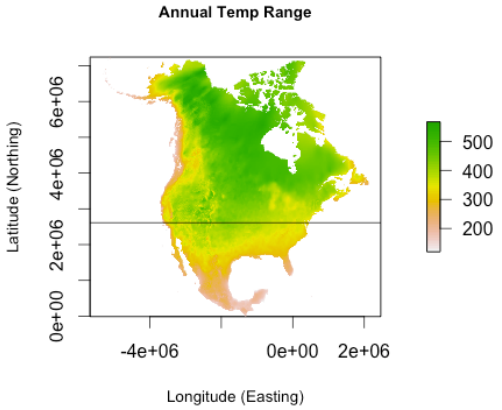
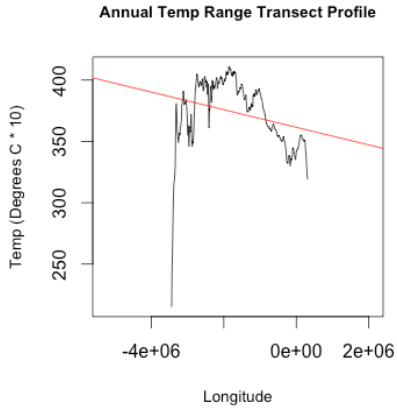


Figure 4a-j: Plots of five important bioclimatic factors paired with plots showing the profile of each factor's measured value along a west-east transect (black), along with a trend line (red).

(a)



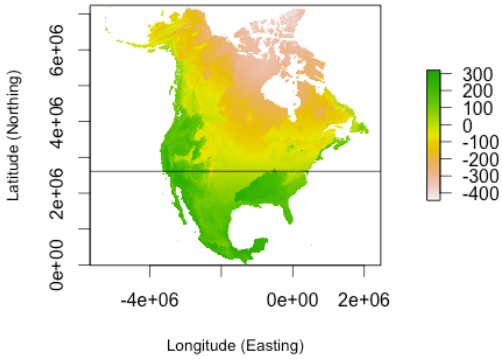
(b)



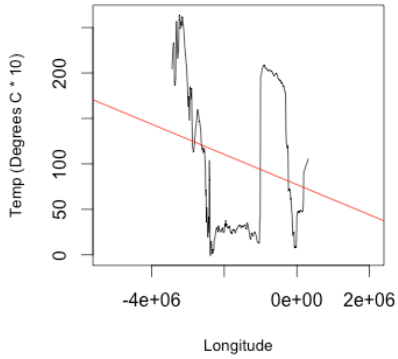
(c)

(d)

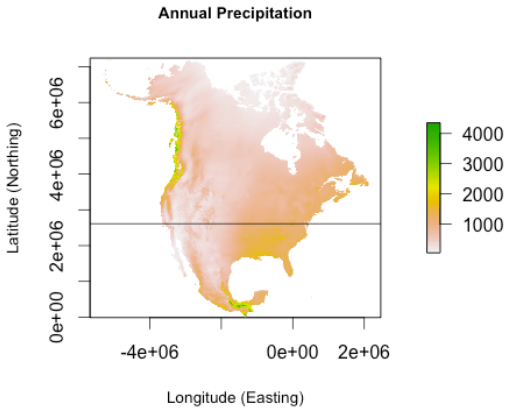
(e)



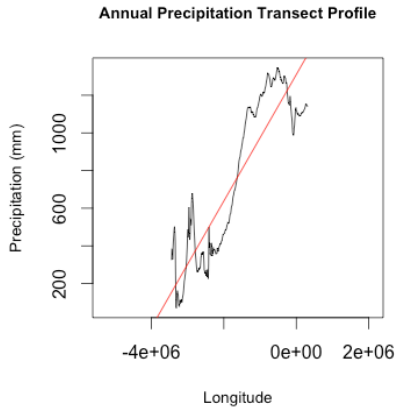
(f)

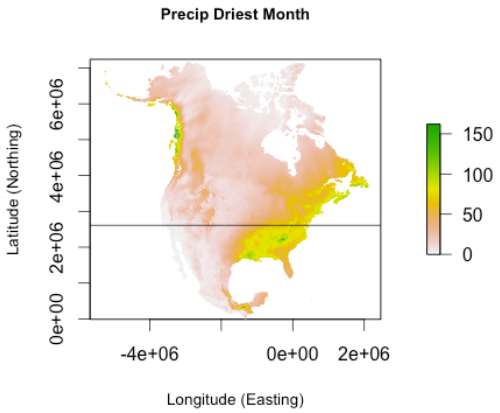


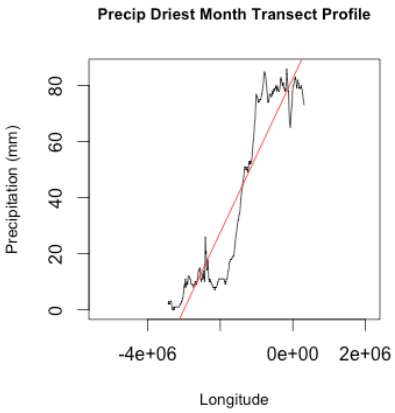
(g)



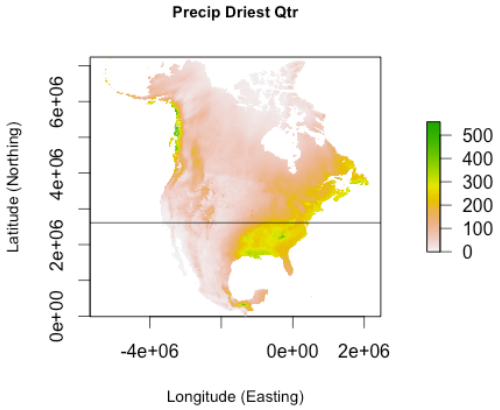
(h)

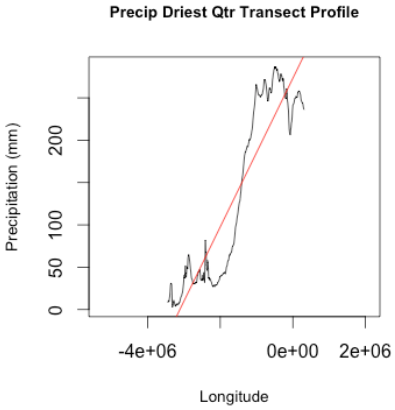



Precip Driest Month
Latitude (Northing)
0e+00
2e+06
4e+06
6e+06
-4e+06
0e+00
2e+06
Longitude (Easting)
150
100
50
0


Precip Driest Month Transect Profile
Precipitation (mm)
0
20
40
60
80
-4e+06
0e+00
2e+06
Longitude

(i) (j)


Precip Driest Qtr
Latitude (Northing)
0e+00
2e+06
4e+06
6e+06
-4e+06
0e+00
2e+06
Longitude (Easting)
500
400
300
200
100
0


Precip Driest Qtr Transect Profile
Precipitation (mm)
0
50
100
200
-4e+06
0e+00
2e+06
Longitude

References

Critchfield, W.B., and Little, E.L., Jr., (1966), Geographic distribution of the pines of the world: U.S. Department of Agriculture Miscellaneous Publication 991, p. 1-97.

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

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

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

Morin et al, 2007. Process-Based Modeling Of Species' Distributions: What Limits Temperate Tree Species' Range Boundaries?. Ecology, 88:9, 2280–2291

Morin X. and M.J. Lechowicz, 2011. Geographical and ecological patterns of range size in North American trees. Ecography. http://dx.doi.org/10.1111/j.1600-0587.2010.06854.x

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

Rapoport, E. H. (1975). Areografía. Estrategias Geográficas de las Especies. Fondo de Cultura Económica, México

Rapoport, E. H. (1982). Areography. Geographical Strategies of Species. Trad. B. Drausal, Pergamon Press, Oxford. ISBN 978-0080289144

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

U.S. Geological Survey, 1999, Digital representation of "Atlas of United States Trees" by Elbert L. Little, Jr.. Online at: http://esp.cr.usgs.gov/data/atlas/little/

USDA, NRCS. 2011. The PLANTS Database (http://plants.usda.gov, 6 May 2011). National Plant Data Center, Baton Rouge, LA 70874-4490 USA.

CONCLUSION

Geographic ranges are a fundamental concept of biogeography, yet they are very difficult to study. While often represented as physical polygons that discretely define the limits of a species extent on the landscape, that characterization oversimplifies a highly complex phenomenon. Geographic range limits are not a fixed line, but a fuzzy one. They are dynamic and always in flux, and represent the product of constantly varying environmental and biotic factors that act to create differential survival, reproduction and dispersal rates at the edge of the species' range.

Moreover, the geographic ranges we observe today or perhaps documented 40 years ago in E.L. Little's Atlas of North American Trees, likely represent the product of historical environmental stressors and population demographics rather than present day conditions. For temperate trees, seedlings and juveniles have different physiological tolerances than adults. Thus, the present-day geographic range boundaries of some species may reflect successful recruitment and survival from historical climatic regimes, and the geographic range boundaries we study may not be in equilibrium with environmental conditions with which we use to explain or interpret the mechanisms producing the geographic range patterns we observe.

Finally, our characterization of geographic ranges is limited by our technical abilities to sufficiently model and represent the complex interactions of multiple abiotic and biotic factors that vary across both space and time, to determine the geographic extents of a species' range. While raster and vector representations of geographic ranges are useful for modeling range boundaries, they are constructs for phenomena that are limited by our ability to measure them at sufficient spatial and temporal resolutions.

Moving forward, we need more data. Many studies of abundance structure and demographic processes, and modeled habitat suitability that ultimately define a species' geographic range boundary have relied on relatively small sample sizes. While new databases are offering the ability to analyze these patterns and trends with very large sample sizes, we need to address gaps in those databases. Absence data is very lacking, and temporal data needs to be added so that we

can distinguish between historical and current data, as most current analyses treat all data contemporaneously.

Further, we need to accumulate a great deal of data on species abundance across space. Many studies assessing how species will respond to climate change focus on changes to range boundaries. However, shifts in the center of gravity of a species' range may be detectable in the abundance data long before a larger-scale alteration of a geographic range boundary may be evident. We also need finer spatial and temporal resolution data. Advanced modeling techniques like Maxent, regression trees, maximum likelihood, and machine learning methods now seem capable of producing models of relatively similar accuracy. Yet, our ability to differentiate significant distribution patterns may be limited by the finest resolution of our spatial data.

Moreover, the fuzzy lines of species' geographic range edges are under sampled; and they may be where the real action is. Most studies have inadequately sampled ranges by comparing a single peripheral site with a central site. We need more holistic assessments that examine processes among multiple peripheral sites, distributed along different range margins.

We also need to synthesize demographic and other approaches. Whether range boundaries are defined by abiotic or biotic conditions, they're ultimately created by differential reproduction (or sometimes mortality) among sites. Thus, we need more studies that combine assessments of environmental stressors and biotic interactions with demographic observations. This requires long-term studies across multiple portions of geographic ranges that may be possible by leveraging the spatial distribution of some species' occurrences within LTER sites.

In addition to spatially distributed abundance and demographic studies, we should also examine phenology across species' geographic ranges. If the demographic processes of differential reproduction and mortality act to create geographic range margins, we might have a chance of seeing patterns in shorter time frames by studying changes in phenology across a species' geographic range. This may provide an early signal of a larger-scale change.

Most importantly, our present day understanding of many patterns with respect to geographic ranges is limited to few well-sampled taxa (birds, trees, and palms). We need to expand efforts to integrate natural history collection databases and leverage networks such as HerpNet, VertNet, AntWeb, and other systems so we can begin to obtain and integrate data on taxa for which we have little geographically referenced information for study.