ABSTRACT

Title of Thesis: **IMPACTS OF WEIGHTING SCHEMES AND TRANSFORMED ENVIRONMENTAL VARIABLES ON BIODIVERSITY MODELING WITH PRESENCE-ONLY DATA**

Kavya Pradhan, Master of Science, 2017

Thesis Directed By: Associate Professor Matthew C. Fitzpatrick, UMCES Appalachian Laboratory

Biodiversity modeling techniques at the community- and species-level can be used to address questions in ecology, management, and conservation. I addressed aspects of community-level and specie-level models using virtual and inventoried species in North and South America. Firstly, I assessed the effectiveness of two weighting schemes in reducing impacts (if any) of five sampling routines (simulating unrepresentative sampling in presence-only data) on the model performance of Generalized dissimilarity model (GDM). Unrepresentative sampling lowers model performance, but weighting species can reduce this negative impact to a certain extent. However, PO data severely impacts GDM's ability to detect the relative contribution of environmental gradients. Secondly, I examined the potential of (GDM) transformed environmental variables in improving the performance of Maxent models (presence-only) along with the influence of range size, sample size, and species dependence type. Transformed environmental variables improved model

performance, especially when used with small-ranged species and/or low sample sizes.

IMPACTS OF WEIGHTING SCHEMES AND TRANSFORMED ENVIRONMENTAL VARIABLES ON BIODIVERSITY MODELING WITH PRESENCE-ONLY DATA

by

Kavya Pradhan

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2017

Advisory Committee:
Associate Professor Matthew C. Fitzpatrick, Chair
Associate Research Professor Helen Bailey
Associate Research Professor Katia Engelhardt

Acknowledgements

First and foremost, I want to thank my advisor, Dr. Matt Fitzpatrick, whose guidance, insight, and endless support made this thesis possible. I also wish to thank Drs. Helen Bailey and Katia Engelhardt for being a part of my brilliant committee. My experience in the Appalachian lab would not have been the same without members of my lab group whose kind critique and generous advice always pushed me in the right direction. I would like to thank Matt Lisk for sharing his R wisdom, especially when I had been stubbornly staring at the same line of code for hours. And lastly, I want to acknowledge the tremendous amount of support that I got from the wonderful graduate student community at the Appalachian lab and Frostburg State University. They kept me sane through an insane experience.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under the receiver operator Curve
CLM	Community-level Model
ENA	Eastern North America
FIA	Forest Inventory Analysis
GDM	Generalized Dissimilarity Modeling
PA	Presence-absence
PCA	Principle Components Analysis
PO	Presence-only
SACA	northern South America and Central America
SDM	Species Distribution Model
USFS	United States Forest Service

Overview

Biodiversity modeling techniques are used to quantify and map various aspects of biodiversity – from species distributions to assemblage-level patterns such as community composition or species richness – across a geographic area of interest using empirical relationships between distributions of species (expressed as a set of point locations at which the species is known to occur) and coincident environmental variables (typically derived from digital maps of interpolated climatic data, Elith *et al.*, 2006; Elith & Leathwick, 2009; McMahon *et al.*, 2011; D'Amen *et al.*, 2015). Predictions and projections from these models not only provide insight into the state of biodiversity in the present and the future, but have also been used as tools for resource management and conservation planning (Franklin 2013, Guisan et al. 2013). However, the success of these applications is contingent upon the legitimacy of the relationships and patterns derived from the models, which are in turn dependent upon proper use of the model themselves. This thesis is divided into two chapters, each of which addresses aspects of biodiversity models for predicting patterns of biodiversity and species distributions.

Chapter 1 focuses on Generalized dissimilarity modeling (GDM; Ferrier et al. 2007) a community-level technique that relates community composition turnover between sites to environmental differences between the same sites. Although GDM has been used with PO data in the past (Fitzpatrick et al. 2013, Mokany et al. 2015), the affect of unrepresentative sampling present in PO datasets on model performance has not been previously assessed. Methods to mitigate the influence of

unrepresentative sampling in the form of weights for site-pairs (Ferrier et al. 2007) are present, but have not been evaluated either. As such, *this chapter assesses the suitability of using GDM with PO data and tests methods to mitigate any issues that may arise.*

Chapter 2 analyzes the impacts of using preprocessed environmental variables as predictors on the model performance of a species-level model (Maxent; Phillips et al. 2004, 2006). Although species distribution models (SDMs) are generally fit using abiotic variables (such as climate, soils, topography, etc), there are limitations to their explanatory ability. Using environmental variables that have been preprocessed using information from community-level patterns may help address several primary challenges related to fitted SDMs, such as the lack of species co-occurrence information (Elith et al. 2006, Maguire et al. 2016), the use of only abiotic variables (Wisz et al. 2013), and sample size limitations (Hernandez et al. 2006, Wisz et al. 2008, Feeley and Silman 2011, Bean et al. 2012).

Chapter 1: Impacts of presence-only data on community-level modeling and the effectiveness of weighting schemes in mitigating these impacts

Abstract

Biodiversity modeling techniques at the community level can be used to address questions in ecology, resource management, and conservation. Generalized dissimilarity modeling (GDM) is one such technique that models compositional dissimilarity between sites as a function of their geographic separation and environmental conditions. Though GDM has been used with presence-only (PO) data, the impacts of unrepresentative sampling on model performance are unknown. Additionally, weighting site-pairs has been used to mitigate impacts of PO data, but the effectiveness of weighting schemes remains untested. In this study, I assessed the impacts of five sampling routines (sampling biases) and the effectiveness of two weighting schemes (richness and expected-vs-observed species richness ratio) in reducing the impacts of biases using virtual communities and inventory data in North and South America. Unrepresentative sampling lowered model explanatory power and resulted in misidentification of the contribution of environmental gradients to compositional turnover. However, weighting by richness improved model explanatory power while using observed-vs-expected species richness ratio resulted in spatial patterns of turnover similar to unbiased models, especially in cases when the sampling bias was inversely related to species richness. As such, PO data can be used with GDM to understand explanatory power of variables used to model

community characteristics or to examine spatial patterns of community turnover. However, using PO data with GDM to understand the relationships between compositional turnover and environmental characteristics can lead to misleading results.

Introduction

Ongoing and anticipated impacts of global change on natural systems have led to increased concerns regarding the future of biodiversity (Dillon et al. 2010, McMahon et al. 2011, Bellard et al. 2012). Advances in computation, biodiversity databases, and the availability of comprehensive gridded data layers characterizing abiotic conditions have supported the development and application of spatial modeling to understand and predict biodiversity patterns – from species distributions to assemblage-level patterns such as community composition or species richness – and how they might be altered by human activities. The dominant paradigm is to model species individually and in isolation of co-occurring taxa using species distribution models (SDMs, also called Environmental Niche Models or ENMs; Guisan and Thuiller 2005). However, recent work has demonstrated that community-level models (CLMs; Ferrier and Guisan 2006, D'Amen et al. 2015), which simultaneously consider all species in an assemblage, may offer some benefits over SDMs (Elith et al. 2006, Ferrier and Guisan 2006, D'Amen et al. 2015, Maguire et al. 2016).

CLMs use biological records from multiple species to model community-level characteristics either instead of, or in addition to, species-level information (Ferrier and Guisan 2006, D'Amen et al. 2015). CLMs can be divided into three strategies– (i) "assemble [the community] first, predict later", which treats communities as static sets of co-occurring species that are classified into community types and then predicted, (ii) "predict [each species] first, assemble [the community] later" that treats communities as coincidental assemblages of individual species such that species are

first modeled and predicted individually, and then classified or aggregated to get community level information, and (iii) "assemble [all species] and predict together" that is an intermediate between the former two (Ferrier and Guisan 2006, D'Amen et al. 2015). The "assemble and predict together" strategy is of particular interest for modeling biodiversity as it incorporates all available species occurrence data into a single modeling process, while offering an effective balance between assuming communities are fixed entities (strategy and that species exist in isolation (strategy 2) (Ferrier and Guisan 2006, D'Amen et al. 2015). Incorporating flexibility in what constitutes a community into CLMs is important as changes across the global landscape can lead to alterations in biotic interactions, interacting organisms, and composition of the community (Montoya and Raffaelli 2010, Walther 2010, Mokany et al. 2015, Maguire et al. 2016). This is supported by evidence from analysis of fossil records showing that communities do not remain fixed through time (Williams et al. 2013).

Generalized dissimilarity modeling (GDM) is one example of the "assemble and predict together" strategy that is used to model compositional dissimilarity between locations (i.e., beta diversity) as a function of geographic separation and differences in environmental characteristics between locations. It is a non-linear, generalized extension of matrix regression that incorporates two types of non-linearities commonly observed in biological data (Ferrier et al. 2007). First, GDM incorporates the non-linear relationship between ecological separation and compositional dissimilarity by following a generalized linear model approach on a matrix regression (Ferrier et al. 2007). Second, variation in compositional turnover

along environmental gradients is represented by fitting non-linear monotonic combinations of I-spline basis functions (Ramsay 1988) to each predictor variable (Ferrier et al. 2007). GDM thus allows flexibility in the formulation of the models based on available data that are especially important for analysis conducted over large spatial extents and extrapolation across unsampled regions or times (Ferrier et al. 2007, Fitzpatrick et al. 2011).

GDM is increasingly being used in assessing biodiversity patterns at the community level (Jones et al. 2013, Valdujo et al. 2013, Bell et al. 2013, Fitzpatrick et al. 2013, Loiseau et al. 2017), analyzing genetic dissimilarity (Fitzpatrick and Keller 2015, Hermoso et al. 2016), comparing compositional turnover across time (Blois et al. 2013), incorporating phylogenetic information into biodiversity analyses (Rosauer et al. 2014), and informing conservation and management decisions (Leathwick et al. 2011, Thomassen et al. 2011, Willis et al. 2012, Prober et al. 2012). However, few studies have examined how the quality of species occurrence data impacts GDM. Although GDM facilitates the use of all data and can incorporate record-poor species (Ferrier and Guisan 2006, D'Amen et al. 2015), it may be particularly sensitive to incomplete sampling of modeled communities as unobserved species will artificially inflate biologically distances between sites. For instance, when calculating dissimilarity between two sites, each containing three species, the difference between the actual and the estimated compositional dissimilarity could be substantial if one species is not represented in the data. For this reason, GDM ideally should be fit using high-quality abundance or presence-absence (PA) data, which document both presence and absence of all study species at a site.

Although PA data are considered more robust from a statistical modeling perspective, presence-only (PO) data, which contain only species presence information and are often collected by ad hoc surveys rather than systematic sampling, are considerably more common (Suarez and Tsutsui 2004, Graham et al. 2004) and, therefore, more often used in biodiversity modeling, including GDM. When PO data are used to fit GDM, taxa are considered to be absent at a site if there are no presence records for that species at that location. In essence, the lack of an observation is equated with absence (Ferrier et al. 2007). Although this assumption may be valid for comprehensive inventory data, where failure to observe implies absence, it is unlikely to be valid for PO data. PO data may be unrepresentative of actual communities due to biases associated with sampling design (or lack thereof) and/or preferential sampling of certain locations and/or taxa over others (Meyer et al. 2016). Although PO data have been used in some studies (Fitzpatrick et al. 2011, Mokany et al. 2015), how such data influence GDM remains largely unknown.

To reduce the influence of biases associated with the use of PO data, individual site-pairs can be weighted to alter their relative contribution in model fitting (Ferrier et al. 2007). The effect of PO data on GDM and the handling of PO data with or without weighting schemes has not been formally studied. Weighting sites according to species richness is the standard method for mitigating incomplete/biased sampling in PO data. Weighting by species richness reduces the influence of sites with fewer species, which are most sensitive to PO sampling bias from a distance metric perspective (Ferrier et al. 2007). However, while species richness observed at a site may be related to sampling intensity, it also varies as a

function of the environment. In cases where low richness sites are well (or even completely) sampled and contain information valuable to the modeling process, naïve richness weighting would lead to loss of information. Thus, sampled richness at a site might be a result of the environment or unrepresentative sampling. The lack of discrimination here can be especially problematic when the richness patterns arise due to the environment rather than sampling bias. As an alternative, I propose to compare richness weighting to an approach that represents the disparity between real and observed community composition using the ratio of observed (species richness in sampled data) to expected richness (total richness) as a proxy of sampling completeness at each site. This method would therefore give greater weight to sites that are more fully sampled rather than sites with the most species. A downside is that the index of sampling completeness requires an estimate of species richness at each site.

In this study, I assessed how unrepresentative PO data influence GDM and the relative merits of different weighting schemes for dealing with these biases. I fit GDMs using both virtual communities constructed from a large set of simulated species and inventory data across North and South America. By using both real and simulated community data in two regions that differ in environmental gradients and patterns of biodiversity, I was able to more fully assess the performance of GDM fit with PO data. I aimed to answer two broad questions:

1) How does the degree and type of unrepresentative sampling influence model performance?

2) Which weighting scheme (no weights, weighting by species richness, weighting by observed/actual species richness) best corrects for unrepresentative sampling in terms of model explanatory power, ability to identify the contribution of the environmental gradients to compositional turnover, and accurately map spatial patterns?

I predict that unrepresentative sampling will reduce model performance and that weighted models will perform significantly better than unweighted models with regard to both explanatory power and ability to map spatial patterns. I also predict that weighting by sampling completeness (ratio of observed to actual species richness) will outperform naïve species richness weighting because it will serve as a proxy for sampling completeness, thereby ensuring the inclusion of information from well-sampled low richness sites.

Materials and Methods

Generalized Dissimilarity Modeling

GDM quantifies the relationship between species and environmental turnover and can predict spatial patterns of compositional dissimilarity. The compositional dissimilarity between all pairs of sites (site-pairs, d_{ij}) is calculated using any distance metric scaled between 0 and 1, mostly commonly Sorensen's distance or Sorenson similarity index is used for species composition data. The calculated dissimilarity can be weighted equally for all site-pairs, by richness, or using a custom weight. Here I used the presence-absence version of Bray-Curtis dissimilarity (eq1), which is 1 – Sorenson similarity index:

$$d_{ij} = 1 - \frac{2A}{2A+B+C} \quad \text{(eq1)}$$

where d_{ij} = compositional dissimilarity between sites *i* and *j*, A = number of species in common between the two sites, B = number of species at site *i*, C = number of species at site *j*.

To relate compositional dissimilarity to environmental gradients, I-spline basis functions are first derived for each environmental variable (x_1 to x_n). I-splines allow for the incorporation of non-linearity while maintaining monotonicity ($a_{pk} \geq 0$) and allowing for greater or lesser complexity depending on the number of knots used (three knots is the default). The maximum height of I-splines for each predictor indicates the relative contribution of that predictor to explaining species turnover, and the shape demonstrates the variation in the rate of turnover along the predictor's gradient. The value of the environmental variables at each site is derived using the fitted I-splines and the pairwise differences at all site-pairs are calculated (eq2). In addition to the environmental variables, geographical distance can also be included as a predictor. These I-splines can be used to transform environmental variables to a biological scale and include their biological importance. Finally, a non-negative iteratively re-weighted least squares regression is fitted using the compositional dissimilarity as the response variable and the pairwise differences of the environmental predictors and geographical distance previously derived using I-splines (eq2):

$$-\ln(1 - d_{ij}) = a_o + \sum_{p=1}^{n} \sum_{k=1}^{m_p} a_{pk} \, | \, I_{pk}(x_{pi}) - I_{pk}(x_{pj}) | \quad \text{(eq2)}$$

where a_o is the intercept, n = number of environmental variables, m = number of splines used, $I_{pk} = k^{th}$ spline for variable x_p, and a_{pk} = fitted coefficient for I_{pk} such that $a_{pk} \geq 0$.

Study region

I fit GDMs using two study regions that differ in climate and therefore the length and structure of environmental gradients. I used two study regions to enable the comparison of the impact of unrepresentative sampling and weighting methods in regions with different environmental characteristics. I selected Eastern North America (ENA) to represent a region with comparatively low environmental turnover across space and northern South America and Central America (SACA) to represent a region with higher environmental turnover (Buckley & Jetz 2008). It should be noted that the two regions also differ in extent, with SACA being approximately three times larger than ENA.

Environmental data

I used four climate variables from the WorldClim database (Hijmans et al. 2005) at 10 arc minute resolution: annual mean temperature (bio1), temperature seasonality (bio4), annual precipitation (bio12), and precipitation seasonality (bio15) to characterize environmental gradients. These variables were selected for simulating species habitat preferences (probability of occurrence) because of their known relationship with species richness patterns, distributions, and community composition (McCain 2007, Buckley and Jetz 2008, Wang et al. 2009, Ulrich et al. 2014). These

four variables were used both to simulate individual species distributions (from which virtual communities were created) and to fit the GMDs.

Community composition data

Simulated species distributions

The "virtual ecologist" approach (Zurell et al. 2010) is an effective means of generating "virtual species" (and communities; Hirzel et al. 2001) based on ecological processes and rules, which can be used to test ecological theory or methodological approaches (Meynard and Kaplan 2013, Leroy et al. 2016). Virtual species and community data, while not necessary reflective of real biodiversity patterns, allows control and complete knowledge of the factors determining species distributions. These "perfect" data can also be subsampled in different ways to mimic field sampling and observational biases present in PO data (Zurell et al. 2010). Comparisons of models fit with biased and unbiased data are therefore akin to traditional experimentation.

I used the 'virtualSpecies' package (Leroy et al. 2016) in R (RStudio Team 2016, R Core Team 2015) to simulate two sets of 500 species in each study region. The virtualSpecies package allows users to create individual species with either user defined or random responses to environmental (either real or simulated) patterns. To create a community using the virtualSpecies package, it is necessary to generate multiple species using environmental data and some information regarding the response of each virtual species to these variables (Fig.1.1). Once species

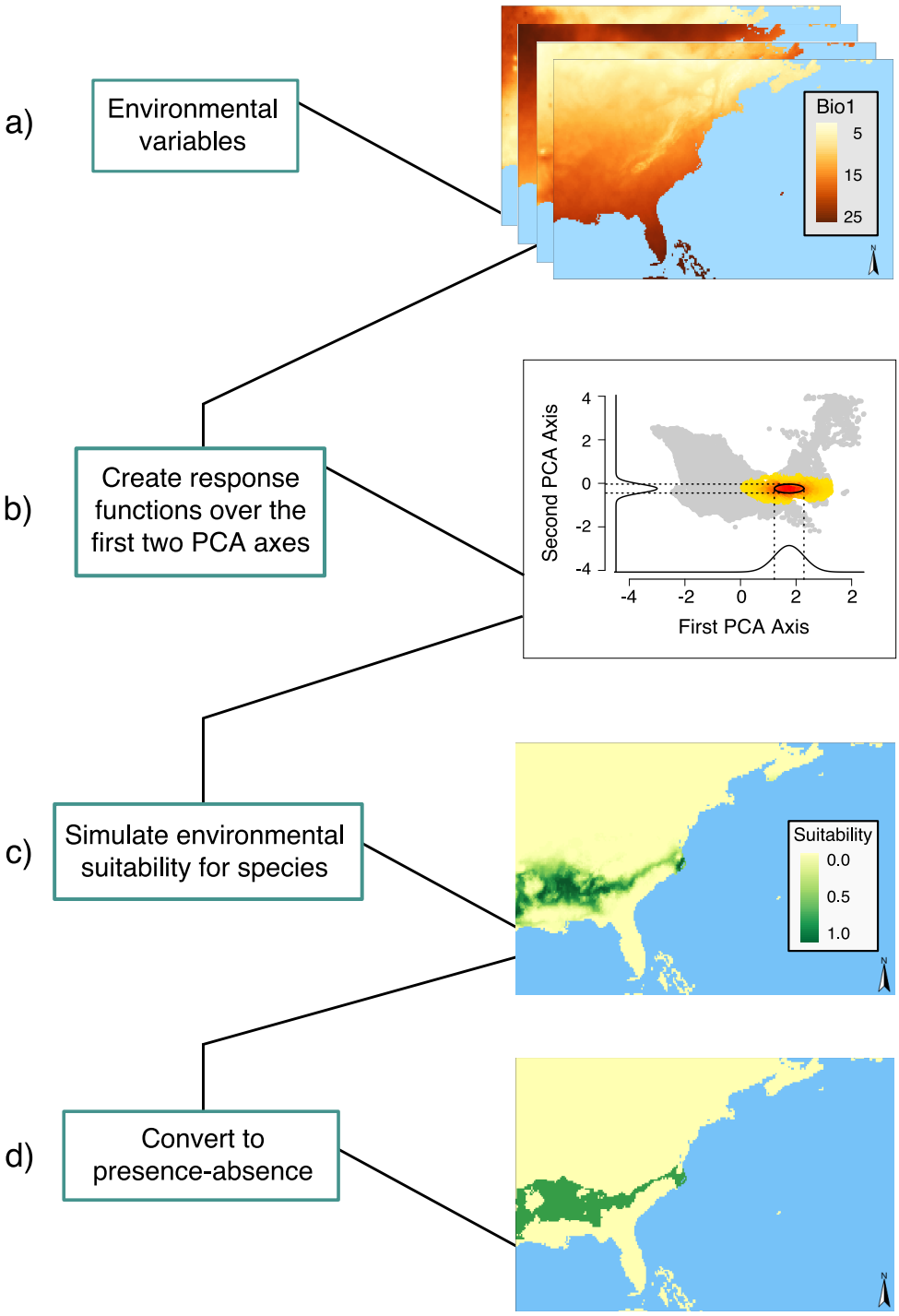


Figure 1. 1 Procedure used to simulate the presence-absence of one virtual species. (a) Four environmental variables are used to constrain the niche of the species, with niche width being determined by (b) the standard deviation of a Gaussian function over the first two axes of a PCA on the four environmental predictors. This function is then used to generate (c) the probability of occurrence of the species across the study region. Last, the probability of occurrence is converted to (d) presence-absence (green indicates presence).

distributions have been simulated, species occurring in the same location are combined to form a ommunity. When communities are generated using this method, interactions between species are not incorporated into the creation of a community unless some proxy of interaction is included as an environmental variable (for inclusion of interactions in community simulation, see below).

To ensure that the simulated species have realistic responses to environmental gradients, the environmental suitability for each species was based on Gaussian response functions with varying means and standard deviations over the first two axes of a Principal Component Analysis. Leroy et al. (2016) recommend that this method be used in cases where multiple species are being simulated based on multiple predictor variables since defining response functions for each of the predictor variables individually can result in unrealistic environmental combinations. Instead of specifying the means and standard deviations of the response functions for each species, I varied the species niche-breadths (i.e., the standard deviation of the Gaussian response functions) such that the two regions had different proportions of narrow-niche-width (low standard deviation) and wide-niche-width (high standard deviation) species. Approximately 60% of the species had narrow niches in SACA and 40 % had wide niches (vice versa for ENA). Niche-breadth variation was implemented to reflect the patterns of niche breadth across various latitudes. The resultant probabilities of occurrence were subsequently converted to presence-absence using a probabilistic approach where a logistic curve with random values for the slope and the inflection point was used to determine the relationship between the

environmental suitability of a species and its probability of occurrence (Leroy et al. 2016).

Of the 500 species in each region, the distribution of 400 depended solely on environmental characteristic (termed "climate-dependent virtual species"). The remaining 100 species were simulated such that their distributions depended on both environmental characteristics and a proportion of the climate-dependent species present in that location (termed "community-dependent virtual species"). To generate the 100 community-dependent virtual species, I first generated the 400 climate-dependent species and then selected a random proportion of the 400 climate-dependent virtual species in each location. I summed the presences of this random subset to create communities in each location (grid cell), resulting in a raster that represented the number of selected species in each cell. Next, I used a range of thresholds from 0.2 to 0.8 in increments of 0.2 to create variation in the degree of dependence on other species. The resulting raster was then used as an additional niche axis for simulating the remaining 100 species. This process allowed me to create a subset of species that depend both on the abiotic environment as well as the set of species present at each location, thus introducing some level of interaction among species in the simulated communities (Fig.1.2).

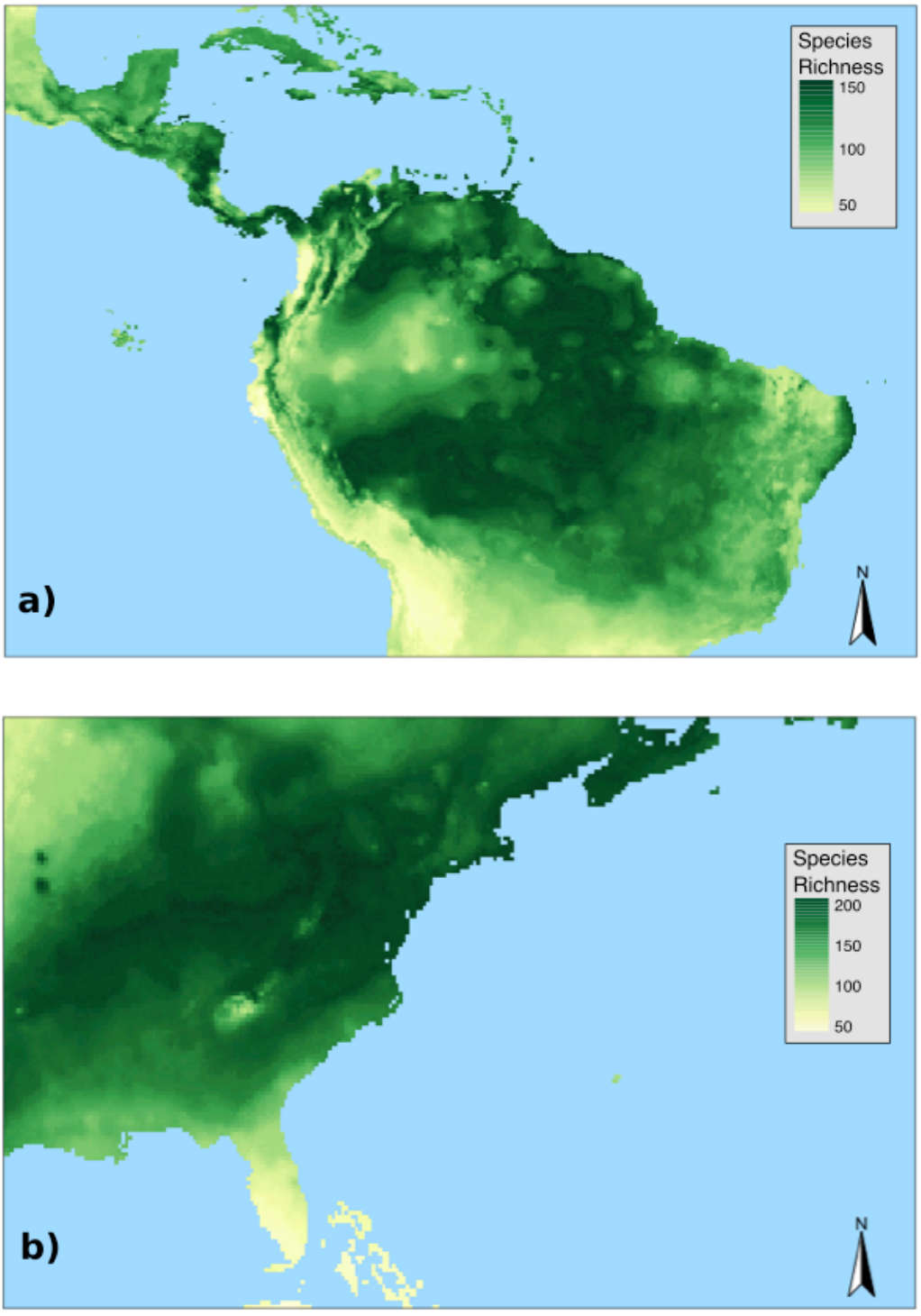


Figure 1. 2 Simulated pattern of species richness produced by assembling communities from the distributions of 500 virtual species in (a) northern South America and Central America (SACA) and (b) eastern North America (ENA)

Species inventory data

Though simulated communities offer great flexibility in conducting model experiments, they will not necessarily mimic natural patterns. Therefore, the simulations were complemented with analyses of actual biological survey data from the United States Forest Service Forest (USFS) Inventory Analysis (FIA) in the eastern United States (Fig.1.3). The FIA data contain species-level inventory information of permanent forested plots (>= 0.4 ha and >= 10% canopy cover) coordinated by the USFS, where all tree species with greater than 12.7 cm diameter at breast height are inventoried (Woudenberg et al. 2010, Gray et al. 2012). These plots are distributed across a majority of the United States in forested lands with varying ownership types (Woudenberg et al. 2010, Gray et al. 2012).

I used the 2003-2008 plot inventory data (Fig. 3), which consists of 77,734 sites and 143 tree species after the full dataset was cropped to a region comparable to that of the simulated ENA community. I first aggregated the information from all the sites to 10 arcmin grid cells (based on the climate data) by taking the mean, thus reducing the number of sites to 10477. After this, I converted counts (abundance) to PA such that a species was considered present if at least one individual of that species had been identified at a site. Although some sites in the FIA database have degraded spatial accuracy (an offset of 1.6 km or switched site survey within forest-class, owner-class and county) to maintain privacy of private landowners and integrity of the plot (Woudenberg et al. 2010, Gray et al. 2012), my aggregation of plot data to 10 arcmin grid cells should reduce the influence of this lack of spatial accuracy.

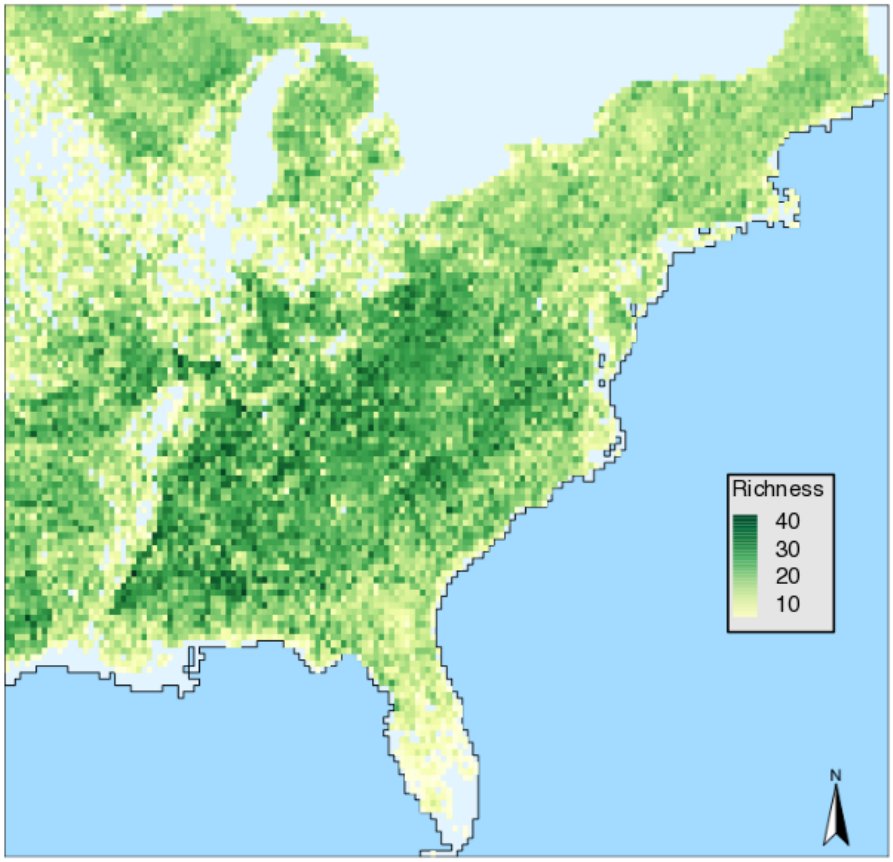


Figure 1. 3 Observed tree species richness based on USFS Forest Inventory Analysis (FIA) data for eastern North America. The value in each pixel is the sum of all observed tree species obtained by aggregating FIA plots falling within the same 10 arcmin grid cell.

Biasing species occurrence data

I compared models fit with "unbiased" and "biased" species occurrence data. The unbiased datasets contained observations of all species at a site, whereas the biased datasets contained observations for a subset of species at each site (see below; Fig.1.4). To create the unbiased datasets for both the virtual communities and the inventory data, I selected 1000 sites at random from each geographic region (hereafter siteSets). I repeated this procedure 100 times to create 100 different

random siteSets. Each of these 100 unbiased datasets were used to fit GDM (hereafter unbiased models) against which the models fitted using the biased data were compared.

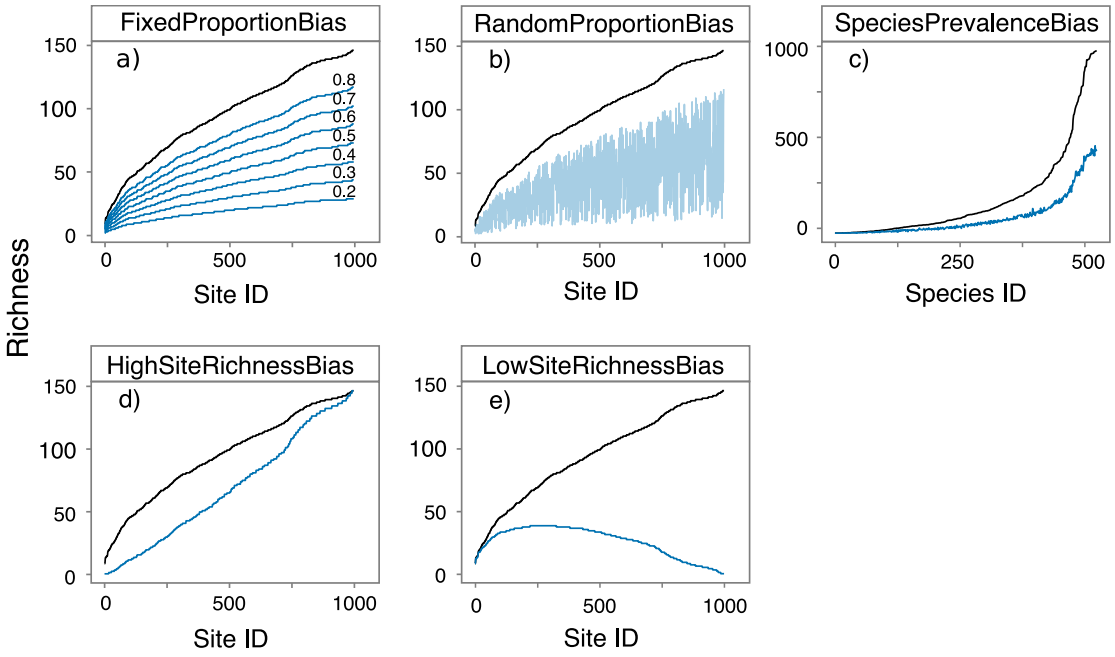


Figure 1. 4 Subsampling routines applied to the completely sampled communities to simulate unrepresentative sampling. This plot shows one siteSet of 1000 sites in northern South America and Central America. The unbiased data (the true richness) is represented by the black line while the blue lines indicate observed richness after the sampling routine has been applied. For a) FixedProportionBias, each blue line represents a fixed proportion of the species retained per site with the proportion written above the lines.

To replicate unrepresentative sampling present in PO data, I degraded the unbiased siteSets by randomly subsampling species occurrences with or without a specific bias. Each of the following routines were performed 100 times for every

siteSet, thereby creating 100 random biased datasets for each type of sampling bias described below.

1. *FixedProportionBias* – I removed a fixed proportion of species at random from each site using proportions from 0.2 to 0.8 at increments of 0.1. This created datasets where all sites had only 20% of species observed, 40% of species observed, etc in each grid cell. This bias was implemented to assess the impacts of degrading the unbiased dataset and the ability of different weighting schemes to correct for incomplete sampling of community composition.
2. *RandomProportionBias* – I removed a different random proportion (between 0.2 and 0.9) of species from each of the 1000 sites. This created datasets where sites had anywhere from between 10% to 80% of species observed in each grid cell. This sampling routine was implemented to simulate a case where sampling completeness is random across sites and probably best replicates sampling bias in PO data.
3. *HighSiteRichnessBias* – Communities were subsampled such that the number of species observed at each site was proportional to the species richness of that site. HighSiteRichnessBias simulates a situation where higher richness sites are of greater interest or more easily accessibly and therefore are more completely sampled.
4. *LowSiteRichnessBias* – The sampling completeness of each site is inversely proportional to the species richness at that site. The implementation of LowSiteRichnessBias is the opposite that of HighSiteRichnessBias and

simulates a case where low richness sites are more completely sampled than high richness sites.

5. *SpeciesPrevalenceBias* – I calculated the prevalence of each species in a siteSet as the sum of all cells occupied by the species. I then converted presences to absences for a random set of occupied cells for each species based upon the prevalence of that particular species. The probability that a species is observed at a site is proportional to the prevalence of the species in the dataset, such that common species are more likely to be observed. For this routine, sampling completeness is based on the species prevalence rather than properties of the community at a site. This simulates sampling were common species are more likely to be observed than rare species.

Model fitting

GDMs were fit to both biased and unbiased data using the 'gdm' package (Manion et al. 2016) in R. Because I was interested in the impact of unrepresentative sampling and weighting scheme combinations on the fitted relationship between the environmental and compositional turnover, I did not include geographic distance as a predictor. The degree to which the sample data accurately reflect community composition will affect model fitting. When complete community composition data are available, model performance should be highest because Bray-Curtis distance will not be artificially inflated due to missing species observations. Because community composition is perfectly known in the unbiased datasets, there is no need to implement weighting to account for bias. Biases associated with incomplete sampling

can be partially corrected for using weights. For example, if available, prior information regarding the sampling completeness can be used to weight the influence of the site on the model. Here, I compared three weighting schemes to attempt to correct for biased data:

1. *No weights (W_{none})* – Sites are weighted equally and so contribute equally to the model. This is a demonstration of a case where the incompleteness of the community composition data is ignored.
2. *Weighting by richness ($W_{richness}$)* – Sites with higher richness have a greater influence on the model and reflects the assumption that higher richness sites are more completely sampled than sites with lower richness. Additionally, unrepresentative sampling will affect high richness sites less than low richness sites.
3. *Weighting by the ratio of observed to expected species richness ($W_{obs/exp}$)* – Here the ratio of the observed species richness to the expected (or actual) species richness acts as a proxy for sampling completeness. Actual species richness was calculated as the count of the number of species present in each cell, whereas observed richness was the number of species present after the application of sampling biases described above.

Analysis of weighting schemes

I evaluated model performance using three methods: (i) model fit using percent deviance explained, (ii) predictor contribution using sums of coefficients of I-splines, and (iii) ability to accurately predict spatial patterns of compositional turnover using

Procrustes analysis. I assessed model explanatory power by comparing the percent deviance explained amongst biased models using ANOVA or Kruskal-Wallis test based on the heteroscedastcity of the data followed by multiple comparisons (Games-Howell and Mann-Whitney-Wilcoxon tests respectively) and to the percent deviance explained of the unbiased models.

In addition to model explanatory power, I wanted to assess how well biased models fit using different weights can correctly identify the primary environmental gradients associated with compositional turnover. To do this, I summed the coefficients of the I-splines to quantify the relative contribution of different environmental variables. I then examined how the relative contributions of variables changed in the presence of unrepresentative sampling and with the three weighting schemes.

Finally, to examine the congruence between the biased and unbiased models in terms of spatial predictions of compositional turnover, I implemented a Procrustes analysis using the 'vegan' package (Oksanen et al. 2017) in R. Procrustes analysis involves the superimposition of two datasets by rotating, scaling and translating the data to minimize the sum of squared deviations between them. The measure of fit of a Procrustes analysis is the m^2 statistic, with lower values of m2 indicating higher concordance between two datasets (Jackson 1995, Peres-Neto and Jackson 2001). Additionally, examining the vector residuals can quantify similarity of the two datasets. In order to assess the concordance between models fitted to biased vs. unbiased data, I followed the procedure outlined in Pitcher et al. (2012) for the comparison of model predicted dissimilarity. I first obtained the transformed

environmental variables to biological space using the fitted models and subjected the transformed environmental variables to a scaled Principle Component Analysis (PCA). The results of the PCA were then subjected to a Procrustes superimposition.

Results

Unbiased (completely sampled and fully representative) data

The explanatory power of GDM varied by data type (virtual vs. survey) and region. Explanatory power (percent deviance explained) was greatest for models fit using virtual species in ENA (94.22 ± 0.04 %) followed by SACA (65.92 ± 0.14 %) and lowest for survey data (FIA; 36.27 ± 0.14 %; Fig.1.5). Models fitted with fully sampled datasets (unbiased models) explained a higher percentage of the deviance than biased models regardless of the communities used or the weighting scheme employed, with the exception of models fit in SACA to data with HighSiteRichnessBias.

The relative contribution of each environmental gradient in explaining turnover also varied by region and dataset. The sums of coefficients of I-splines fitted to the unbiased data in ENA indicate that compositional turnover was greatest along gradients of mean annual precipitation (bio12; 1.09 ± 0.00 %), followed by precipitation seasonality (bio15; 1.08 ± 0.01 %), temperature seasonality (bio4; 1.04 ± 0.01 %), and mean annual temperature (bio1; 0.35 ± 0.01 %; Fig.1.6). In contrast, for SACA, compositional turnover was greatest along the precipitation seasonality gradient (bio15; 0.78 ± 0.01 %) and least along temperature seasonality (0.38 ± 0.00

%; Fig.1.6), whereas for the FIA data, this variable (2.29 ± 0.02 %) was associated with the most turnover and precipitation seasonality the least (1.55 ± 0.03 %; Fig.1.6).

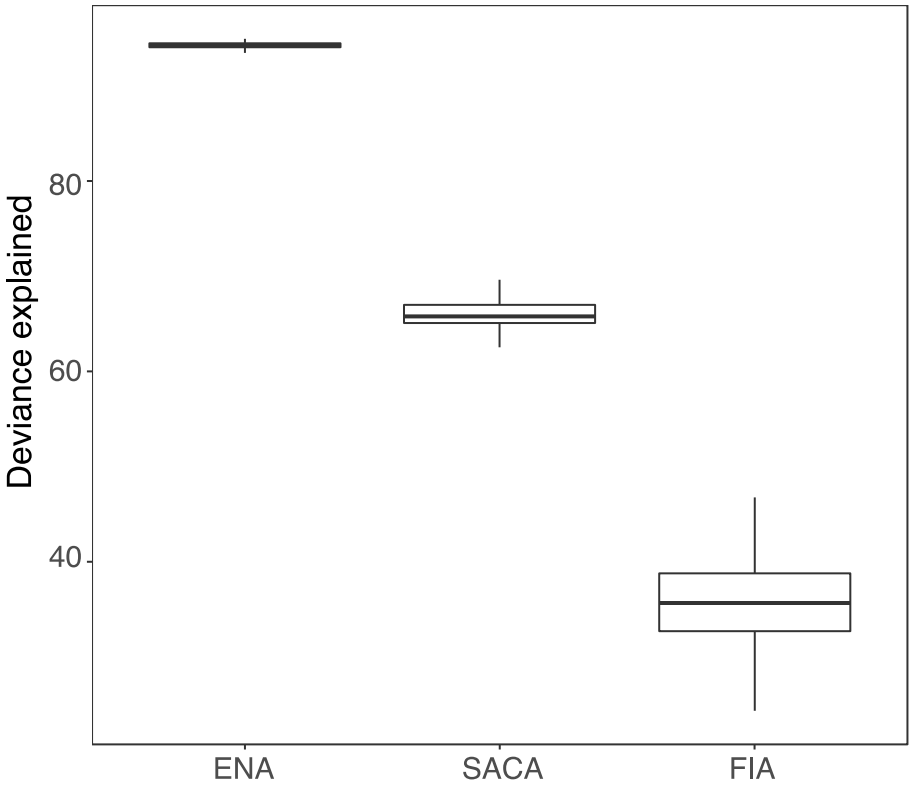


Figure 1. 5 Deviance explained. The explanatory power of the unbiased models for all siteSets in ENA, SACA and FIA. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

FixedProportionBias

As expected, there was a positive relationships between the proportion of species observed at a site and the explanatory power of GDM, with the least amount of degradation (80% of species retained at each site) producing percent deviance explained values closest to those of the unbiased models in all regions, data types, and weighting schemes (Figures 1.7, 1.8, 1.9). Weighting by species richness produced the best models for all fixed proportions except for 20% species retention in FIA, for

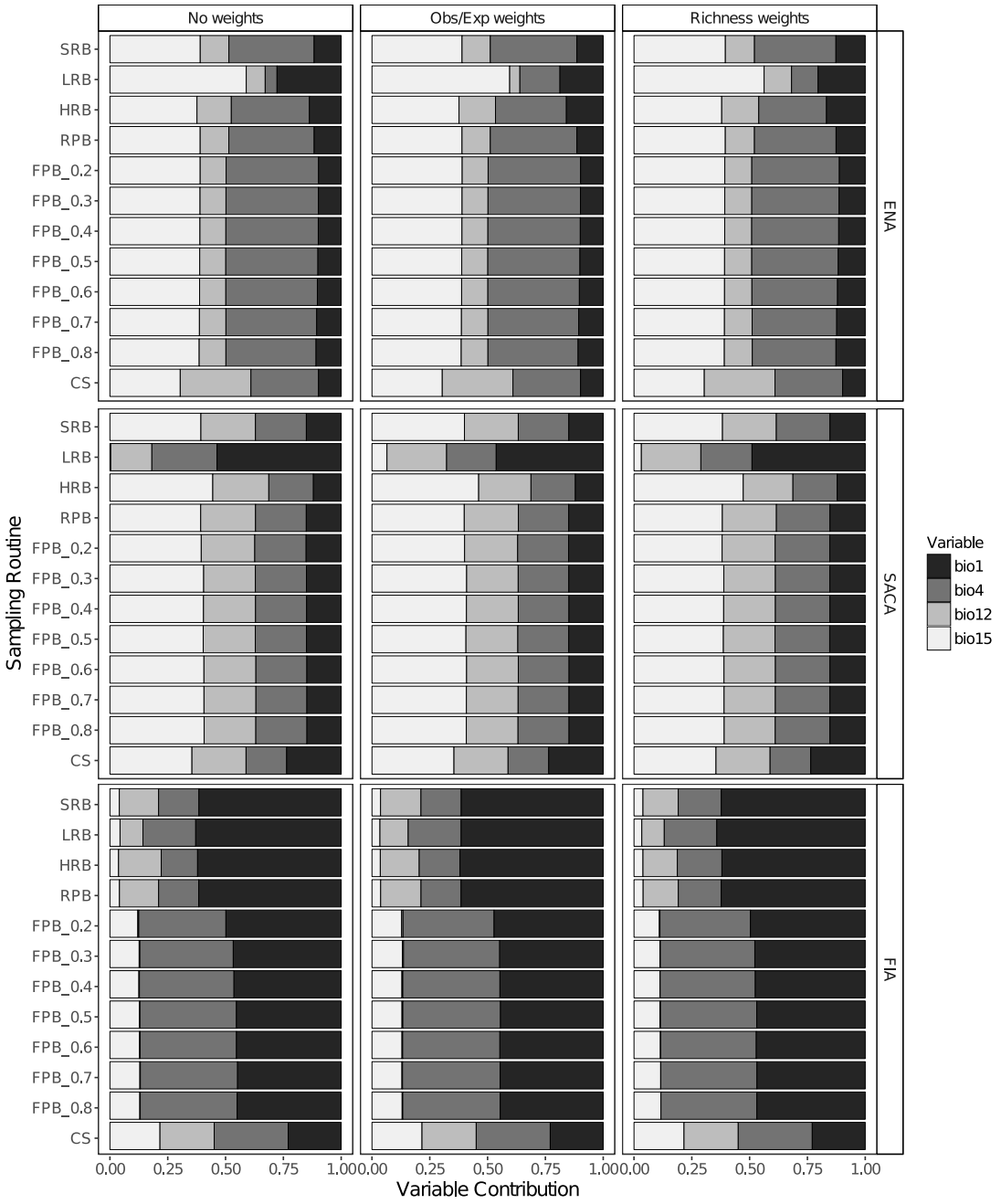


Figure 1. 6 Relative variable importance. Variable importance was inferred from the sums of coefficients for each variable for models fitted in ENA, SACA, and FIA. The sampling routines are along the y-axis and represent the unbiased or completely sampled data (CS), FisedProportionBias (FPB), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). FixedProportionBias is further divided into the various fixed proportions used. Each column represents a weighting scheme.

which $W_{obs/exp}$ produced the best model (6.28 ± 0.07 %; Fig.1.10). While weighting by sampling completeness ($W_{exp/obs}$) was statistically better than no weighting (W_{none}) for the FIA data, for most cases in the simulated communities there was no statistically significant difference between W_{none} and $W_{richness}$ (Fig.1.10).

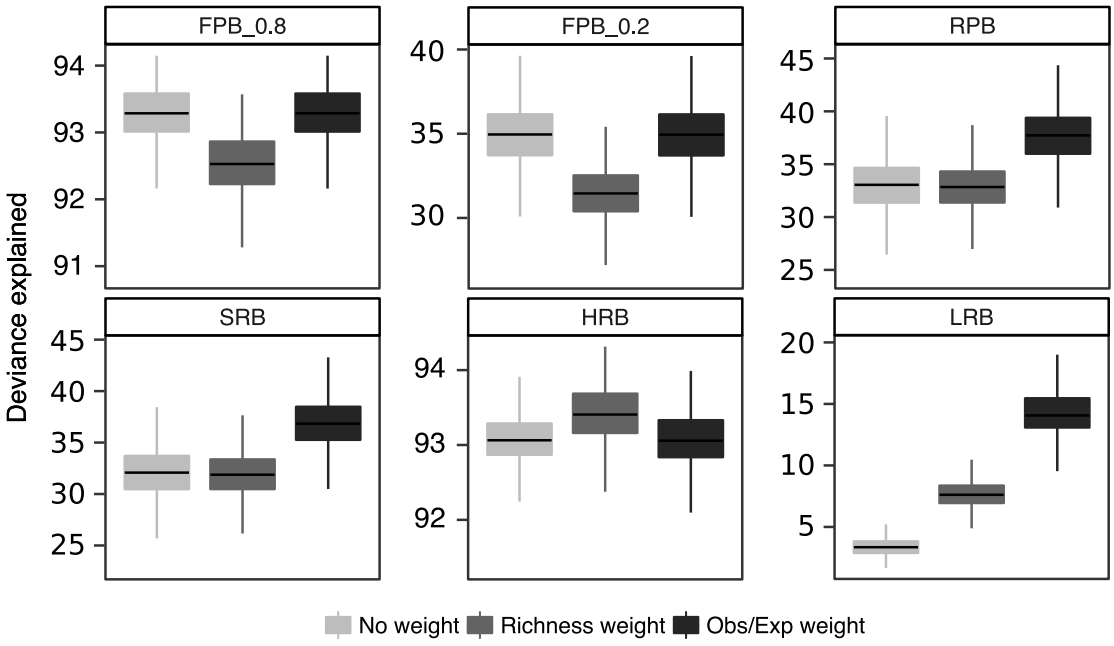


Figure 1. 7 Deviance explained by biased models fitted in ENA. The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). The y-axes are variable in order to display the results with clarity. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

Fixed proportion sampling also influenced the ability of GDM to correctly quantify the relative contribution of each environmental gradient to species turnover, as compared to models fitted with the unbiased data (Fig.1.6). The relative contribution

of predictors changed the least for models fit in SACA; the contribution of bio15 increased while that of bio1 decreased. For biased models in both ENA and FIA, the greatest difference was a decrease in relative contribution of bio12. Concordance between mapped spatial patterns of compositional dissimilarity using the unbiased and the biased data was greatest for GDMs fit using richness weighting produced the most similar patterns for models fitted in ENA (Fig.1.11; Appendix 1.1). For models fitted in SACA (Fig.1.12) and using the FIA data (Fig.1.13), on the other hand, using Wnone tended to result in the highest concordance, with some exceptions (Appendix 1.1).

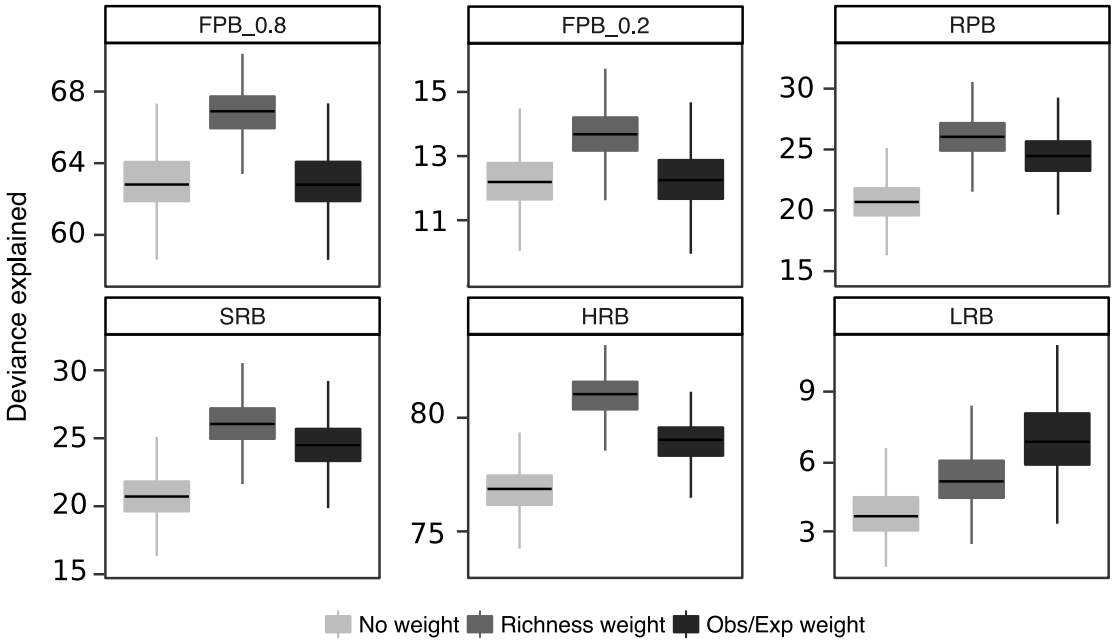


Figure 1. 8 Deviance explained by biased models fitted in SACA. The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). Note that the plot contains variable y-axes. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges

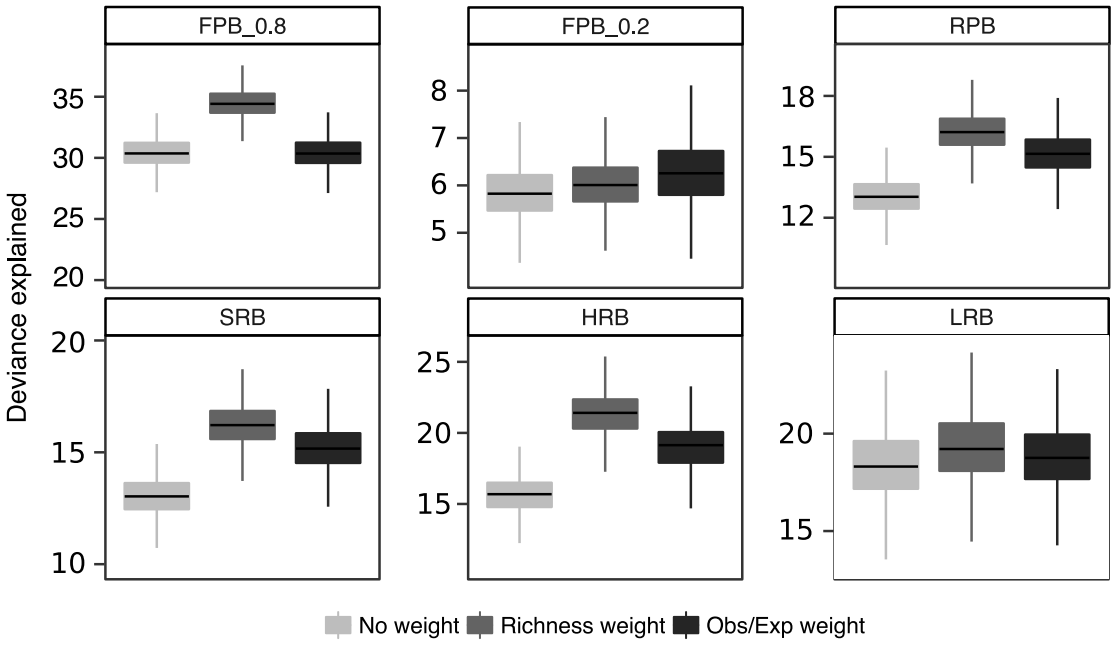


Figure 1. 9 Deviance explained by biased models fitted in FIA. The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). Note that the plot contains variable y-axes. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

RandomProportionBias

RandomProportionBias reduced percent deviance explained when compared to unbiased models by a larger degree for the simulated communities than for the inventory data. Biased models fitted in ENA had significantly higher percent deviance explained when fit with $W_{obs/exp}$ (36.59 ± 0.03 %) than W_{none} (32.10 ± 0.02 %; Fig.1.7), but $W_{richness}$ produced models with significantly higher percent deviance explained both in SACA (26.13 ± 0.02 %; Fig.1.8) and for the FIA dataset (16.24 ± 0.01 %; Fig.1.9). The relative contribution of predictors was consistent across

weighting schemes for all datasets and similar to that of models produced using FixedProportionBias in ENA and SACA (Fig.1.6).

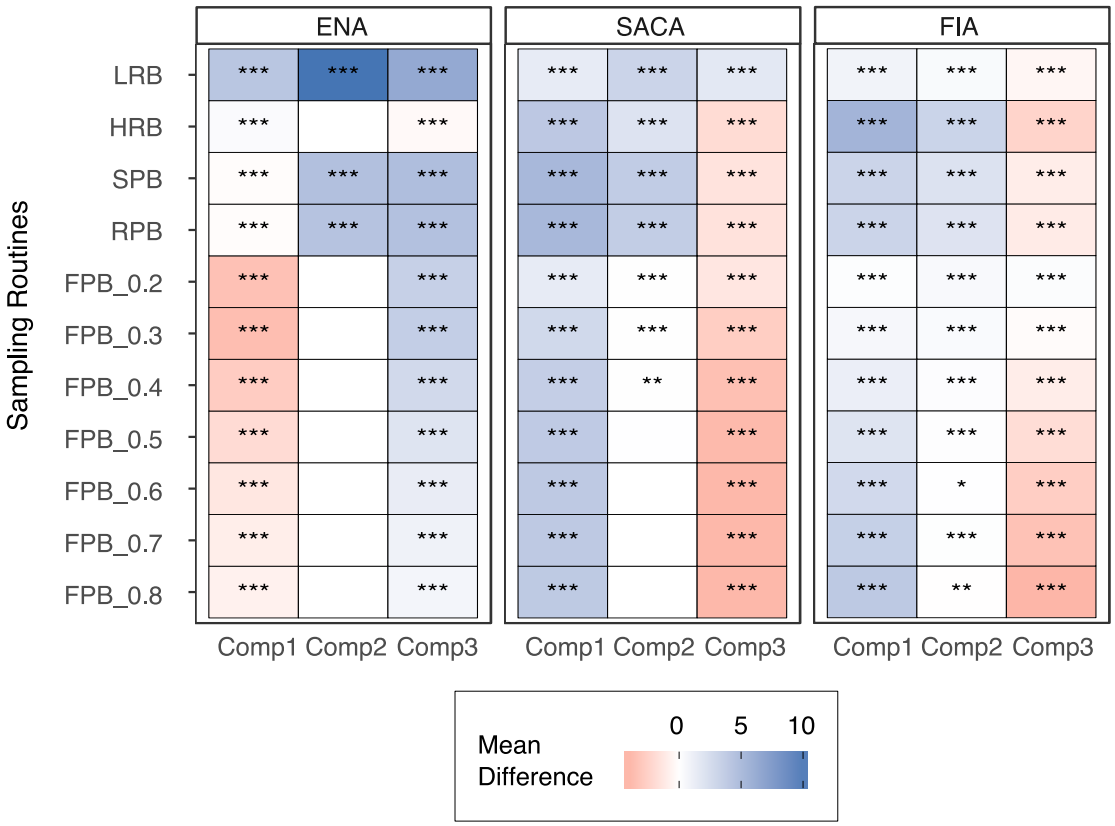


Figure 1. 10 Heatmap of results from Games-Howell test for deviance explained by models fitted in ENA, SACA, and FIA. The colors indicate the mean differences between comparisons of models (Comp1: W_{none} / $W_{richnes}$; Comp2: W_{none} / $W_{obs/exp}$; Comp3: $W_{richness}$ / $W_{obs/exp}$) and the asterisks represent the significance level of the differences(* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB).

RandomProportionBias greatly increased the importance of bio1 and reduced that of bio15 compared to the unbiased models when used on the FIA dataset (Fig.1.6); these variables were fairly similar in contribution for the unbiased models.

Procrustes analysis of the mapped spatial patterns of compositional dissimilarity from unbiased models to biased models (Appendix 1.1) showed that $W_{obs/exp}$ had the greatest degree of concordance in both ENA and SACA, and W_{none} the least (Figures 1.11 and 1.12). For models fitted using the FIA data, W_{none} resulted in the highest concordance between the unbiased and the biased models while $W_{richness}$ produced the lowest (Fig.1.13).

SpeciesPrevalenceBias

Results for data biased with SpeciesPrevalenceBias were similar to that of RandomProportionBias for all aspects of model performance assessed in this study for the two simulated communities. Percent deviance explained by biased models fitted in ENA was intermediate across all sampling routines and $W_{obs/exp}$ had the highest values (36.82 ± 0.02%; Fig.1.7). For models in SACA, the highest percent deviance explained was for the models fit with $W_{richness}$ (26.15 ± 0.02%), followed by $W_{exp/obs}$ (24.60 ± 0.02%; Fig.1.8). The explanatory power for models for the FIA data was also the highest for $W_{richness}$ (16.23 ± 0.01%; Fig.1.9). The relative contribution of predictors showed greatly increased the importance of bio1 and reduced that of bio15 compared to the unbiased models for the FIA data (Fig.1.6) compared to unbiased models. Spatial patterns of disagreement of compositional turnover between

unbiased models and models using data with SpeciesPrevalenceBias were also similar to those of RandomProportionBais (Figures 1.11, 1.12, and 1.13, Appendix 1.1).

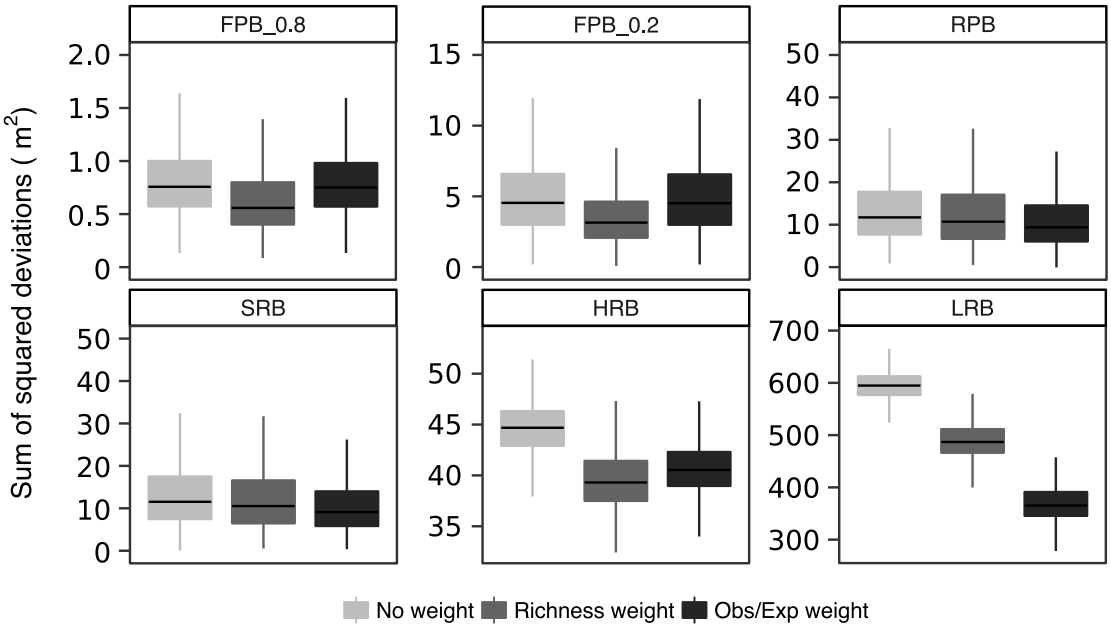


Figure 1. 11 Sum of squared deviations (m^2) from Procurstes analysis of models fitted in ENA. The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). Note that the plot contains variable y-axes. The lower and upper hinges represent the 25^{th} and 75^{th} percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

HighSiteRichnessBias

Of all the sampling bias types, LowSiteRichnessBias had the least influence on explanatory power and produced models with percent deviance explained values comparable to that of the unbiased models for the two simulated communities. Weighting by richness (93.36 ± 0.00 %) produced significantly better models than both W_{none} (93.03 ± 0.00 %, $p << 0.001$) and $W_{obs/exp}$ (93.03 ± 0.00 %, $p << 0.001$) in

ENA (Fig.1.7). In SACA, LowSiteRichnessBias produced models with greater explanatory power than unbiased models; $W_{richness}$ had the highest percent deviance explained (80.88 ± 0.01 %) followed by $W_{obs/exp}$ (78.89 ± 0.01 %) and W_{none} (76.80 ± 0.01 %; Fig.1.8). In contrast, FIA data subjected to high richness bias produced models with much lower percent deviance explained compared to unbiased models. $W_{richness}$ performed the best (21.32 ± 0.01 %) followed by $W_{obs/exp}$ (18.92 ± 0.02 %) and W_{none} (15.69 ± 0.01 %; Fig.1.9). Variables contributions for all communities were altered relative to the unbiased model (Fig.1.6). The variable contributions in ENA and SACA were similar to FixedProportionBias, RandomProportionBias and SpeciesPrevalenceBias across all weighing schemes (Fig.1.6). For FIA data, variable contributions were comparable to RandomProportionBias and SpeciesPrevalenceBias but not FixedProportionBias (Fig.1.6). Comparing the sums of squared deviations from the Procrustes analysis among weighting schemes indicates that unweighted biased models performed the best both in SACA and for the FIA data (Figures 1.12 and 1.1; Appendix 1.1). On the other hand, Procrustes analysis showed that $W_{richness}$ resulted in the highest concordance between biased and unbiased models, followed by $W_{obs/exp}$ in ENA (Fig.1.11; Appendix 1.1).

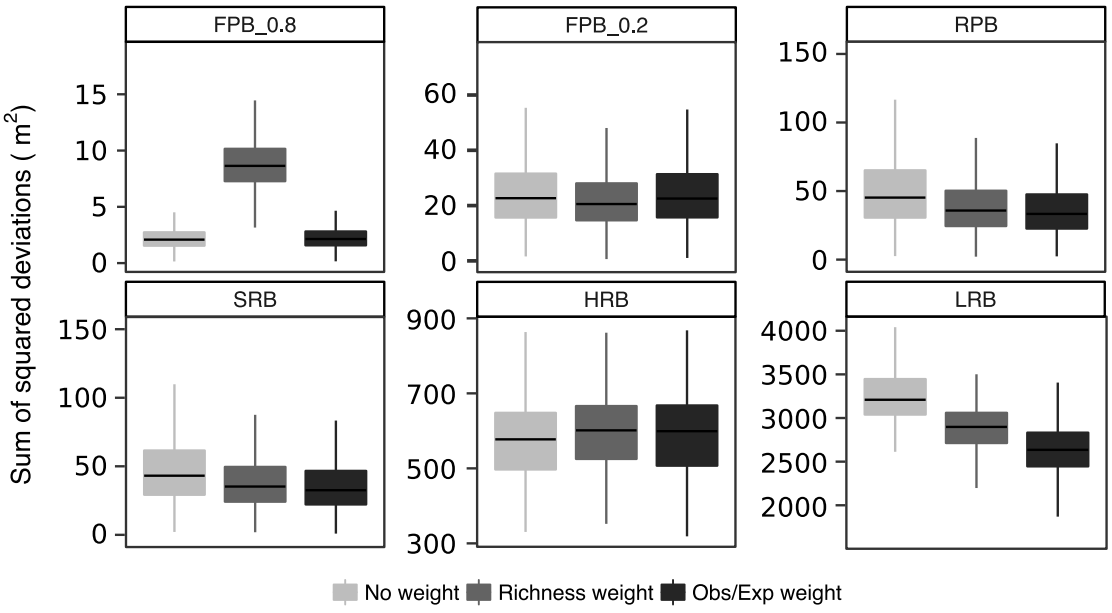


Figure 1. 12 Sum of squared deviations (m^2) from Procurstes analysis of models fitted in SACA. The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). Note that the plot contains variable y-axes. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

LowSiteRichnessBias

LowSiteRichnessBias resulted in the lowest percent deviance explained values for all weighting schemes and regions/datasets except FIA. $W_{obs/exp}$ (14.26 ± 0.02 % in ENA; 6.94 ± 0.01 % in SACA) performed better than $W_{richness}$ (7.73 ± 0.01 % in ENA; 5.24 ± 0.01 % in SACA), both of which were significantly better than W_{none} (3.54 ± 0.01 % in ENA; 3.77 ± 0.01 % in SACA; Figures 1.7 and 1.8). Biased models for FIA data were not as comparatively poor as those for the other two regions and $W_{richness}$ (19.31 ± 0.02 %) outperformed the other weighting schemes (Fig.1.9). The predictor

contributions were the most different from those of the unbiased models for this bias in both ENA and SACA; in ENA, the contribution of both bio1 and bio15 increased and that of bio4 and bio12 decreased while in SACA the contribution of bio15 dramatically decreased accompanied by increases in the contribution of bio1 and bio4 (Fig.1.6). The impact on coefficients of the I-splines was similar to RandomProportionBias, SpeciesPrevalenceBias and LowSiteRichnessBias (Fig.1.6). Sum of squared deviations for biased models weighted with $W_{obs/exp}$ had the lowest values across all communities studied (Figures 1.11, 1.12, and 1.13).

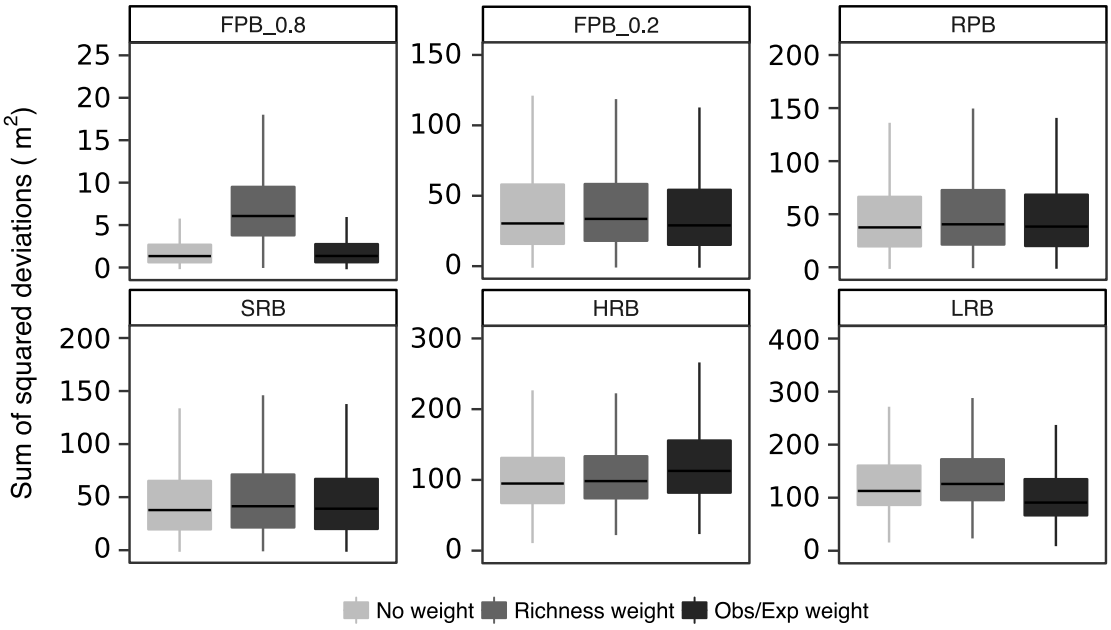


Figure 1. 13 Sum of squared deviations (m^2) from Procurstes analysis of models fitted in FIA. The sampling routines represented here are FixedProportionBias with 80% species retained at each site (FPB_0.8), FixedProportionBias with 20% species retained at each site (FPB_0.2), RandomProportionBias (RPB), HighSiteRichnessBias (HRB), LowSiteRichnessBias (LRB), and SpeciesPrevalenceBias (SPB). Note that the plot contains variable y-axes. The lower and upper hinges represent the 25^{th} and 75^{th} percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

Discussion

Given the volume and accessibility of PO species occurrence records, it is important to assess how well such data can be used to understand and predict patterns of biodiversity. PO data are plagued by a number of issues, and in particular a better understanding of how sampling biases impact community-level models will help improve the application of biodiversity modeling for management purposes. GDM uses community-level data and distances matrices to model compositional patterns. Unrepresentative sampling in PO datasets can artificially inflate community dissimilarity, the response variable of GDM, and while weighting schemes (Ferrier et al. 2007) have been used to remediate biases, their effectiveness remains unknown. In this study, I examined how biases in PO data influence GDM and the ability of different weighting schemes to correct for these biases. Overall, I found that all types of PO biases I simulated using different sampling routines reduced the performance of GDM, with sampling biased inversely to richness causing the greatest decline in model performance. I also found that the use of weights can partially mitigate the impacts of unrepresentative sampling, but no single weighting scheme proved appropriate for all situations. Richness weighting tended to perform better than weighing by the ratio of observed versus expected species richness in terms of explanatory power for a majority of the sampling routines. Weighting site-pairs by the observed versus expected species richness ratio also improved model explanatory power relative to no weighting and in some cases provided predictions that had greater similarity to those of the unbiased models compared to richness weighting. However, none of the biased models – either without any weights or with the two

weighting schemes – were able to correctly assess the relative contribution of different environmental gradients to compositional turnover.

Impacts of unrepresentative sampling on model performance

As expected, unrepresentative sampling lowered model explanatory power (with a few exceptions) and resulted in misidentification of the contribution of environmental gradients to compositional turnover. FixedSamplingBias, in which the same proportion of species are removed at random from all sites, demonstrates that incomplete sampling reduces model robustness and that, regardless of the identity of the species (not) observed, a more complete dataset results in higher explanatory power. Subsampling data using either RandomProportionBias (random proportion of species retained at each site) or SpeciesPrevalenceBias (common species are more completely detected) has a similar effect on GDM and reduced model performance to intermediate levels when compared to the other sampling routines. The comparable performance of GDM may be attributed to the similarity in how RandomProportionBias or SpeciesPrevalenceBias altered richness and species prevalence. Models fit with data biased using HighSiteRichnessBias (sampling bias is proportional to site richness) had some of the highest percent deviance explained values, while those fit with data biased using LowSiteRichnessBias (sampling bias is inversely proportional to site richness) had the lowest. Though models fitted with biased data had lower deviance explained than models fitted with unbiased data, GDM models fitted with the biased data produce spatially similar results to the unbiased models. The regions of higher disagreement (higher residuals) were

concentrated to regions with lower species richness. However, whether sampling bias was directly or inversely proportional to site richness, the concordance of spatial patterns of compositional turnover was the lowest for all biased models, suggesting that when sampling is biased by species richness (either directly or inversely), the fitted relationship between the environmental and compositional turnover is altered. Of all the sampling routines, biasing data by LowSiteRichnessBias had the greatest affect on the contribution of environmental variables for the two virtual communities. However, when data is biased in accordance with site richness or even randomly, the impact of unrepresentative sampling, while still present, is reduced. The lack of agreement between unbiased and biased models indicates that failure to fully document community composition can lead to a misidentification of the gradients of turnover. The most striking effect of the sampling routines, in terms of ranking the relative contribution of environmental variables, was observed on communities in the FIA dataset where the contribution of precipitation related variables decreased dramatically and the contribution of mean annual temperature increased. The predictor variables used in modeling communities in the FIA dataset were not selected specifically for this dataset. As such, the variable set used may not include gradients most important to turnover in tree compositions or does not contain all variables that explain community turnover patterns in FIA, as demonstrated by the low percent deviance explained even by the unbiased models.

Impacts of weighting schemes on model performance

Although no weighting scheme was superior for all cases, both richness and the observed versus expected richness ratio had unique advantages and disadvantages. Taken together, my findings support the use of richness weights to improve the explanatory power of GDM as richness weights produced models with highest percent deviance explained in the majority number of community-sampling routine combinations. However, improvements in explanatory power using richness weighting did not translate into an improved ability to map spatial patterns of species turnover. For mapping of spatial patterns of compositional turnover, observed versus expected richness ratio performed intermediately compared to the other weighting schemes. However, a major downside of observed-vs-expected-richness weighting is that is requires an estimate of site richness. Though there have been attempts to quantify species richness (Kreft and Jetz 2007, Kier et al. 2009, Jenkins et al. 2013, Pimm et al. 2014, Jenkins et al. 2015, Jenkins and Van Houtan 2016), these analyses are limited to a handful of taxa and regions, which limits the application of this weighting scheme for broader assessments of biodiversity. However, methods developed for estimating species richness (eg, using occupancy modeling (e.g., using occupancy modeling; Guillera-Arroita 2017) may aid in the use of observed versus expected species richness ratio as weights.

Recommendations for use of GDM with PO data

Modeling biodiversity patterns enables the understanding of ecological relationship between organisms and environmental characteristics and the application

of this knowledge for management purposes (Ferrier and Guisan 2006, D'Amen et al. 2015). When the goal of a study is to understand explanatory potential of variables used to model community characteristics or examine spatial patterns of community turnover PO data can be used with GDM with caution. Weighting by richness improved model explanatory power while, $W_{obs/exp}$ emerges as a much better weighting scheme with respect to the concordance of spatial patterns of turnover between biased and unbiased models, especially in cases when the sampling bias was inversely related to species richness.

On the other hand, when GDM is used to assess the relationship between compositional turnover and environmental characteristics, PO data can lead to misleading results. None of the weighting schemes were able to correctly identify the relative contribution of environmental gradients regardless of community studied or the types of bias present in the data. Although the richness at each site is an important aspect of communities, the identity of species present and absent is also important. Weights implemented in this study were solely associated with the richness present at each site and are unable to fully rectify sampling issues with respect to which species are sampled. As such, weighting schemes that address both site richness and species identity may be better suited to applications where identification of environmental contributions is the main aim.

Conclusion

Overall, this study suggests that PO data can be used with GDMs with caution as PO data always impacts model performance even when used with weighting schemes.

Biased data will have poorer model performance when explanatory power, identification of contribution of explanatory variable and spatial mapping are all taken into account. However, weighting the influence of site-pairs can overcome issues associated with unrepresentative sampling in PO data to a certain extent. The use of both weighting schemes assessed in this study lead to improved explanatory power in a majority of the biased models. However, if the types of biases or inaccuracies present in the sampling data are known, especially with respect to the identity of the species present or absent at each site, then using a weighting scheme that reflects this information will lead to the most robust models. Additionally, I recommend using caution when assessing the relative contributions of environmental variables as the type of bias in the data and the environmental variables used for model fitting will impact GDM.

Chapter 2: Transforming raw environmental variables for improved species distribution modeling

Abstract

Species distribution models (SDMs) can enable the understanding of relationships between species distributions and environmental variables and support assessment of impacts of global changes at the species-level. Environmental variables that have been preprocessed or transformed using community-level information can be used as predictors in SDMs so that the predictor variables reflect ecological patterns. However, the effectiveness of transformed variables in improving model performance has not been assessed. Generalized dissimilarity models (a community level model) can be used to transform raw environmental variables into ecological space based on community compositional turnover patterns. In this study, assessed whether the transformed environmental variables obtained from GDM can improve the performance of Maxent models (an SDM) using virtual and inventoried species in regions of North and South America. I also assessed the influence of species range size, sample size, and species dependence type on the ability of transformed environmental variables to improve model performance. Overall, using transformed environmental variables as predictors in Maxent models improved model discrimination and ability to map habitat suitability, especially for species with small ranges and/or fewer occurrences. The differences between model performance of the two predictor types, though significant, were relatively small.

Introduction

Impacts of global changes can be assessed at the species-level with statistical modeling techniques referred to collectively as species distribution models (SDMs). SDMs relate information on species occurrence to concurrent environmental conditions to model their probability of occurrence, probability of presence, or relative habitat suitability (Guisan and Zimmermann 2000, Guisan and Thuiller 2005, Elith and Leathwick 2009, Franklin 2010). These models are based on the niche concept (Hutchinson 1957, Guisan and Thuiller 2005, Elith and Leathwick 2009) and can be used to assess species distributions and their relationships with environmental variables for the purposes of ecological and biogeographical research, and as tools for decision making in conservation and resource management (Franklin, 2013; Guisan et al., 2013).

SDMs like any model are approximations of reality and are sensitive to the information used to fit the models and the algorithm used to establish relationships between species and environmental characteristics and are subject to a number of working assumptions and caveats. Numerous studies have examined how data and statistical assumption influence model performance. For instance, the number of records required to produce reliable models (Hernandez et al. 2006, Wisz et al. 2008, Feeley and Silman 2011, Bean et al. 2012), selection of pseudoabsences (Phillips et al. 2009, Barbet-Massin et al. 2012), and spatial autocorrelation (F. Dormann et al. 2007, Václavík et al. 2012, Crase et al. 2012, Record et al. 2013), among other issues, have all been previously studied and led to recommendations for effective implementation of SDMs.

One aspect of the SDM framework that has received relatively less attention than others is the selection of predictor variables and incorporation of additional information in the form of altered predictor variables. Studies that have assessed environmental variable selection have mostly focused on the impacts of scale and resolution of the environmental data (e.g., Guisan et al. 2007, Franklin et al. 2013) and few have focused on the type of predictors (e.g., climate, soil, topography, etc.) to be selected (Williams et al. 2012). SDMs primarily employ abiotic variables (e.g., temperature, precipitation, soils, etc) to model the distribution of species, and have often been criticized for the lack of realism therein (Araújo and Luoto 2007, Wisz et al. 2013) as they may poorly represent actual conditions that the organisms are responding to and incompletely explain the variance in the dataset. There have been advances in incorporating population level demographic processes like dispersal (Engler and Guisan 2009, Midgley et al. 2010, Bocedi et al. 2014, Dytham et al. 2014) and biotic predictors (Heikkinen et al. 2007, Meier et al. 2010, Araújo et al., 2014) in predictions of species responses to changes in the environment. However, incorporating individual species level demographic and biotic information still poses a major challenge due to the lack of complete knowledge and availability of data for most species.

An alternative, but largely untested, approach to incorporating biotic information into SDMs involves 'preprocessing' abiotic variables using community-level information such that raw abiotic variables better reflect ecological patterns (Ferrier et al 2007). Most SDM frameworks do not incorporate species co-occurrence information (Elith et al. 2006, Maguire et al. 2016), but previous studies have found

congruence between the community composition turnover of different taxa (Buckley and Jetz 2008, Jones et al. 2013, Duan et al. 2016) and demonstrated that the inclusion of community dissimilarity of one group (e.g., ferns) as a predictor for another (e.g., trees) can increase the explanatory power (Jones et al. 2013) of community-level models (CLM). Although the underlying cause behind the congruence in community turnover of different taxa is not well understood, it can be attributed broadly to a similarity in response to either abiotic or biotic conditions (Duan et al. 2016). In either case, turnover in community composition can represent some pertinent features of the environment that may not be captured by abiotic variables typically used in biodiversity modeling (Elith et al. 2006, Maguire et al. 2016).

Preprocessed variables can include aspects of the environment not characterized by abiotic variables or have an implicit biotic component. In either case, using preprocessed variables as predictors in SDMs have the potential to increase the variance explained by the models. Although there is a lack of consensus regarding the aspect of the niche that is modeled by SDMs (Guisan and Thuiller 2005, Elith and Leathwick 2009), most studies suggest that SDMs quantify the realized niche rather than the fundamental niche (Hutchinson 1957) or even the potential niche (Ackerly 2003) because of the impacts of biotic interactions and resource limitations already present in species observations (Guisan and Thuiller 2005, Elith and Leathwick 2009). Within this context, the "preprocessed" environmental variables that contain some aspect of community responses to the environmental characteristics would be part of the realized niche of a species.

One method to incorporate additional information into species level modeling approaches is through the use of Generalized Dissimilarity Modeling (GDM; Ferrier et al. 2007), a CLM that relates compositional turnover to environmental turnover. In addition to predicting compositional dissimilarity between sites, GDM identifies the primary environmental variables that contribute to the variation of turnover across space and transforms these variables to reflect their role in driving community turnover (Ferrier and Guisan 2006, Ferrier et al. 2007). As such, GDM uses community information to transform the environmental predictors such that they better reflect ecological patterns. Ferrier et al. (2007) suggested that these gradients could be used as predictor variables for individual species distributions, and Elith et al. (2006) demonstrated that "preprocessing" of environmental variables using GDM resulted in robust predictive performance for individual species. This improvement in predictive performance can be attributed to the additional information contained in the underlying response of the communities as a whole to abiotic or biotic drivers (Elith et al. 2006, Maguire et al. 2016), which is absent from the "raw" environmental information typically used in fitting SDMs. Using transformed variables to fit SDMs is expected to include biological responses of communities to alterations in climatic conditions. Additionally, combining GDM and SDMs could be beneficial for the modeling of low-sampled species because the response of the community as a whole can potentially supplement the lack of occurrence records.

In this study, I assess whether the use of transformed variables leads to improved predictions of habitat suitability at the species level using Maxent (Phillips et al. 2006). To accomplish this, I fit SDMs using complementary simulated and

inventory data in Eastern North America (ENA) and northern South America and Central America (SACA) with untransformed and transformed variables. Specifically, I aimed to answer the following questions with respect to models fitted with untransformed and transformed environmental variables:

1) Does the use of transformed environmental variables improve SDMs in terms of discriminatory ability, model quality, and ability to map spatial patterns of environmental suitability?
2) How do sample size, species range sizes, and species dependence influence model performance when they are used with untransformed or transformed variables?

I expect transformed environmental variables will improve fit of all models with the degree of improvement varying with characteristics of the species and the input data. Overall, models for species with low prevalence will exhibit greater improvement than those with higher prevalence and models for species with high dependence on other species will exhibit greater improvement than models for species with low or no dependence on other species. Furthermore, models fitted with fewer observations will show a larger difference in performance with regard to simulated species.

Materials and Methods

Study area

I fit both GDMs and Maxent in Eastern North America (ENA) and northern South America and Central America (SACA). These two regions were selected because of their difference in climatic characteristic. ENA has relatively low environmental turnover across space while SACA has higher turnover (Buckley & Jetz 2008). In addition to this, the varied climatic gradients also allowed for differences in the characteristics of the species that were simulated.

Environmental data

I used a subset of the 19 bioclimatic variables from the WorldClim database (Hijmans et al. 2005) at 10 arc minute resolution for both the simulation of virtual species and model fitting. Annual mean temperature (bio1), temperature seasonality (bio4), annual precipitation (bio12), and precipitation seasonality (bio15) were selected for simulating species habitat preferences (probability of occurrence) because of their known relationship with species richness patterns, distributions, and community composition (McCain 2007, Buckley and Jetz 2008, Wang et al. 2009, Ulrich et al. 2014). These variables were not used in the model-fitting step. Instead, variable selection was carried out such that the final variable set would have minimal collinearlity. As such, GDMs were fit in each community using a unique set of explanatory variables (Table 1). For fitting the Maxent models, two sets of predictor variables were used – either climatic variables in their original state or after transformation using GDM.

Species and community data

I used both observational (real) and simulated (virtual) data to balance their strengths and weaknesses.

Virtual data allows the user to control the factors driving species distributions and community patterns (Zurell et al. 2010), whereas observational data are subject to errors and uncertainty. I used simulated communities in ENA and SACA, each consisting of 500 species, to assess model performance with the use of two types of predictor variables. Detailed explanation of species simulation was presented in the previous chapter (see the methods section of chapter 1, pg 16-17) but briefly, I simulated species based on a PCA of four environmental variables with differing niche-breadths using the "virtualSpecies" R package (Leroy et al. 2016). Of the 500 species in each community, 400 were based solely on the PCA of the environmental conditions (hereafter referred to as "climate-dependent virtual species"). The remaining 100 species were dependent on the presence of a variable proportion of the climate-dependent species (hereafter referred to as "community-dependent virtual species"). I introduced this variation to assess whether the interactions of species leads to any differences in model performance.

Though virtual species offer great flexibility and control, the lack of realism can be problematic as results obtained using virtual data may not be applicable to studies of real patterns, an issue that is addressed in this study with the use of "real" inventory data. Biological survey data from the United States Forest Service Forest (USFS) Inventory Analysis (FIA) in the eastern United States was used to complement the

analyses conducted on the simulated species. I used the 2003-2008 plot inventory data, averaged to 10 arcmin grid cells and converted to presence-absence. For further detail in the inventory data, see Methods section of chapter 1 (pg 18-19).

I created sampled data for each community by selecting a random set of 2000 sites in order to ensure a large enough sample size for modeling the species distributions. These "sampled communities" were then treated as inventoried sites such that presences were absolute presences and absences were true absences. I used these sampled sites to fit the GDMs and the Maxent models.

Table 2. 1 Predictor variables obtained from WorldClim. The analysis step that the variables were used in is indicated by an "X" in the table.

Bioclimatic variable	Community simulation	ENA Models	SACA Models	FIA Models
BIO1 = Annual Mean Temperature	X			
BIO2 = Mean Diurnal Range		X		X
BIO3 = Isothermality			X	
BIO4 = Temperature Seasonality	X			
BIO7 = Temperature Annual Range		X	X	
BIO8 = Mean Temperature of Wettest Quarter		X		X
BIO10 = Mean Temperature of Warmest Quarter		X	X	X
BIO12 = Annual Precipitation	X			
BIO15 = Precipitation Seasonality	X			
BIO16 = Precipitation of Wettest Quarter		X		X
BIO17 = Precipitation of Driest Quarter		X		X
BIO18 = Precipitation of Warmest Quarter			X	
BIO19 = Precipitation of Coldest Quarter			X	

Statistical modeling

GDM quantifies the relationship between species and environmental turnover across pairs of sites (site-pairs) and can predict spatial patterns of compositional dissimilarity. The compositional dissimilarity between all site-pairs is calculated using any distance metric (here Bray-Curtis distance is used) scaled between 0 and 1 and is related to environmental gradients using a non-negative iteratively re-weighted least squares regression fitted using the compositional dissimilarity as the response variable and the pairwise differences of the environmental predictors and geographical distance as the covariates. The pairwise differences of the predictors are obtained using I-spline basis functions that allow for the incorporation of non-linearity and flexibility in complexity while maintaining monotonicity. As such, in addition to quantifying the relationship between compositional and environmental turnover, GDM creates turnover functions that can be used to transform environmental variables into a biologically relevant scale (Ferrier et al. 2007). These functions describe the relationship between environmental turnover and the turnover of community composition with respect to that predictor variable (Fig.2.1). To accomplish this, raw environmental distances between each pair of sites are converted into I-splines and the information regarding how these splines relate to the compositional turnover is used to transform the predictor variables. The transformed predictor variables represents the spatial variation in species composition as a function predictor gradients and their relative importance in driving these biological patterns. These transformed variables can then be used as predictor variables in SDMs like Maxent.

Maxent is an implementation of the principle of maximum entropy on a set of presences (PO data) and background locations (where presence is unknown) from a defined landscape along with predictor variables across this landscape to get the species' potential geographic distribution (Phillips et al. 2006). The Maxent function is fitted over multiple features (transformations of predictors) such that the coefficients match the constraints put on their means without overfitting. To do this, Maxent maximizes the gain function (a penalized maximum likelihood function) and finds a model that can differentiate between presences and absences. Constraints are set by the characteristics of the environmental predictors in the background locations; in the environmental space Maxent constrains the mean of the environmental predictors at the presence and the background locations to be close to one another while minimizing the distance between the conditional density of the predictors at the occurrence sites and their marginal density at the background locations (Elith et al. 2011). In terms of the geographical space, Maxent derives the probability that a species in found in each cell or pixel in the landscape such that there is maximum entropy in the geographical space, i.e., a distribution that is most spread out and closest to uniform (Phillips et al. 2006, Merow et al. 2013).

Model fitting

I first fit GDMs using the selected variable set using the "gdm" package (Manion et al. 2016) in R (RStudio Team 2016, R Core Team 2015) . The GDM model was fit with the default settings with respect to number of knots and geographic distance was not used as a predictor. I then used the turnover functions obtained from the fitted

models to transform the environmental variables into a biologically relevant scale with respect to compositional turnover.

As the second step, I fit Maxent models (using default setting) using the "biomod2" R package (Thuiller et al. 2016). I first selected all species present in the 2000 sampled sites that had greater than ten occurrences. I also selected 50 presences and 50 absences from the total community for each species to create an evaluation dataset that was used to assess the model; sites present in the sampled community were excluded from being selected to ensure independence of the evaluation dataset. Species for which this dataset could not be created were also excluded from the analysis. As such, the distribution of 470 (374 climate-dependent and 96 community-dependent virtual species), 367 (273 climate-dependent and 94 community-dependent virtual species), and 122 species were modeled for ENA, SACA, and FIA respectively. To assess the influence of sample size on model performance of models fitted with either type of predictor variable, for these species, the number of occurrences was divided into three sample sizes – 10 to 20, 21 to 50, and 51 to 100. Species with a total number of occurrences between 10 and 20 were modeled once using all records. Those with a total number of occurrences between 21 and 50 were modeled twice, once with 20 observations and the second time with all observations. Those with greater than 50 total occurrences were modeled three times – with 20, 50 and 100 (or all) observations. Selection of the occurrences at each sample size was repeated 5 times for each species and Maxent models were fitted with 15 training testing splits (75% training and 25% testing) of the occurrence data once with each predictor type resulting in two types of models – model fit using the untransformed

variables (untransformed models or M_U and models fit using the transformed variables (transformed models or M_T).

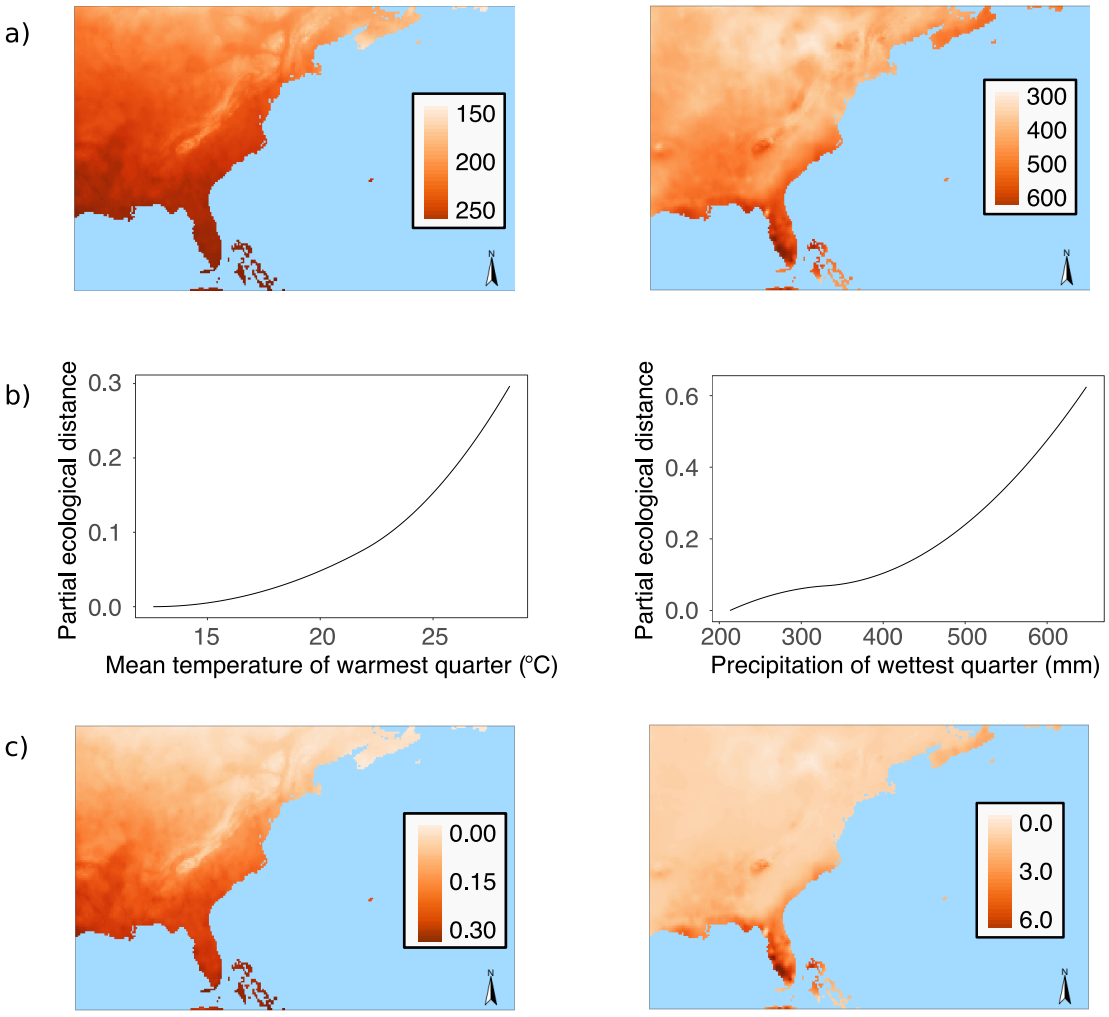


Figure 2. 1 Examples of transformation of environmental variable. Spatial pattern of a) mean temperature of warmest quarter (left) and precipitation of wettest quarter (right), is transformed using the using the relationship presented in b), a plot of the I-spline basis function for each variable. In b), the x-axis contains the raw values for the environmental variable that will be transformed to values along the y-axis, which is in units of Bray-Curtis distance. Thus, the c) transformed variables for Eastern North America contains information regarding community turnover (y-axis of plot b) based on the environmental gradient (x-axis of plot b).

Analysis of model performance

I assessed the ability of the model to discriminate between known presences and background data and to map patterns of environmental suitability for the species as measures of model performance. To assess model discrimination, I used the continuous Boyce index, a presence-only method that assesses how much model predictions deviate from randomness (Boyce et al. 2002, Hirzel et al. 2006). The habitat suitability range obtained from Maxent is classified into multiple bins using a moving window and the ratio of predicted frequency and expected frequency (expected from a random distribution) of the evaluation points is calculated. The Boyce index ranges from -1 to 1 with negative values indicating an incorrect model, values close to zero indicating a model that is no different than random, and values close to 1 indicating a model whose predictions are compatible with the evaluation data. The Boyce index also enables the assessment of model quality in terms of model robustness across cross-validation sets and habitat suitability resolution. To assess the model's ability to map spatial patterns of environmental suitability, I used the I similarity statistic that ranges from 0 to 1, where higher values indicate greater similarity (Warren et al. 2008). Although the I similarity statistic was conceived as a method to measure niche equivalency, it is an effective way of comparing the pairwise differences between the true habitat suitability to the predicted habitat suitability of the species. I calculated the Boyce index using the "ecospat" package (Di Cola et al. 2017) and the I similarity statistic using the "SDMTools" package (VanDerWal et al. 2014) in R.

I compared the overall value of the evaluation metrics for M_U and M_T using the Mann-Whitney-Wilcoxon test. I also assessed the relationship between predictor type, sample size, species range size, and dependence (only for simulated species) of the species on model performance using linear mixed models (LMMs) fit to a normal distribution in R using the "nlme" package (Pinhiero et al. 2017). For the LMMs, I used Boyce index, and I similarity statistic (for simulated species only) as the response variables. Species were considered random effects, while predictor type, sample size, range size, and dependence or independence were the fixed effects.

Results

The explanatory power of GDM varied amongst the three communities with the model in ENA having the highest percent deviance explained (87.38 %), followed by GDM for the FIA dataset (42.19 %) and finally in SACA (34.33 %). The accuracy of SDMs over all regions with respect to Boyce index varied widely from -1 to 1 (Figures 2.2, 2.3, and 2.4). This was also the case for I similarity statistic (0.033 to 1.00, Figures 2.2, 2.3). Although some models performed worse than random, the majority of models (> 90%) performed well, with a mean Boyce index of 0.60 $\pm$ 0.00 for species in ENA (Fig.2.2), 0.74 $\pm$ 0.00 SACA (Fig.2.3), and 0.78 $\pm$ 0.00 FIA (Fig.2.4). The fitted models also produced mapped patterns that were similar to the true habitat suitability of the simulated communities. The mean vales of the I similarity statistic for species in ENA and SACA were 0.88 $\pm$ 0.00 (Fig.2.2) and 0.75$\pm$ 0.00 (Fig.2.3) respectively.

Comparisons of models fitted with untransformed and transformed predictors

Assessment of the accuracy of the models across predictor type using the Mann Whitney Wilcoxon test showed that M_T preformed better for all metrics in ENA, but the differences between the models were relatively small (Table.2.2). In SACA M_T performed better with respect to I similarity statistic ($p < 0.001$; Table.2.2) but the differences between the two predictor types were not statistically significant for Boyce index ($p = 0.887$, Table.2.2). For SDMs fit to the inventory data, using the transformed predictors resulted in lower values of Boyce Index ($p < 0.001$; Table2.2). For the simulated communities, when the species were separated into climate-dependent and community-dependent virtual species, in all cases using the transformed predictors lead to better model performance (Table.2.3).

Effects of range size, sample size, and community dependence

When model performance was assessed as a function of species range sizes, number of observations, and species dependence (for simulated species only), predictor type had a significant (positive) impact on model performance though the impact varied by data type, study region, and evaluation metric. Species range size had a significant and negative influence on model performance with respect to Boyce index in all communities; in other words, models for species with larger ranges had lower discrimination (Fig.2.5). The affect of species range was positive for the I similarity statistic for both simulated communities (Fig.2.6). For species in both FIA and SACA, models fitted with fewer observations had reduced model performance (Figures 2.5 and 2.6). In ENA, SDMs with sample sizes between 21and 50 had higher

Boyce indices than models fit using between 51 and 100 occurrence records; in all other cases, lower sample sizes reduced model performance (Fig.2.5). Species dependence in the simulated communities did not affect the model when considered in isolation.

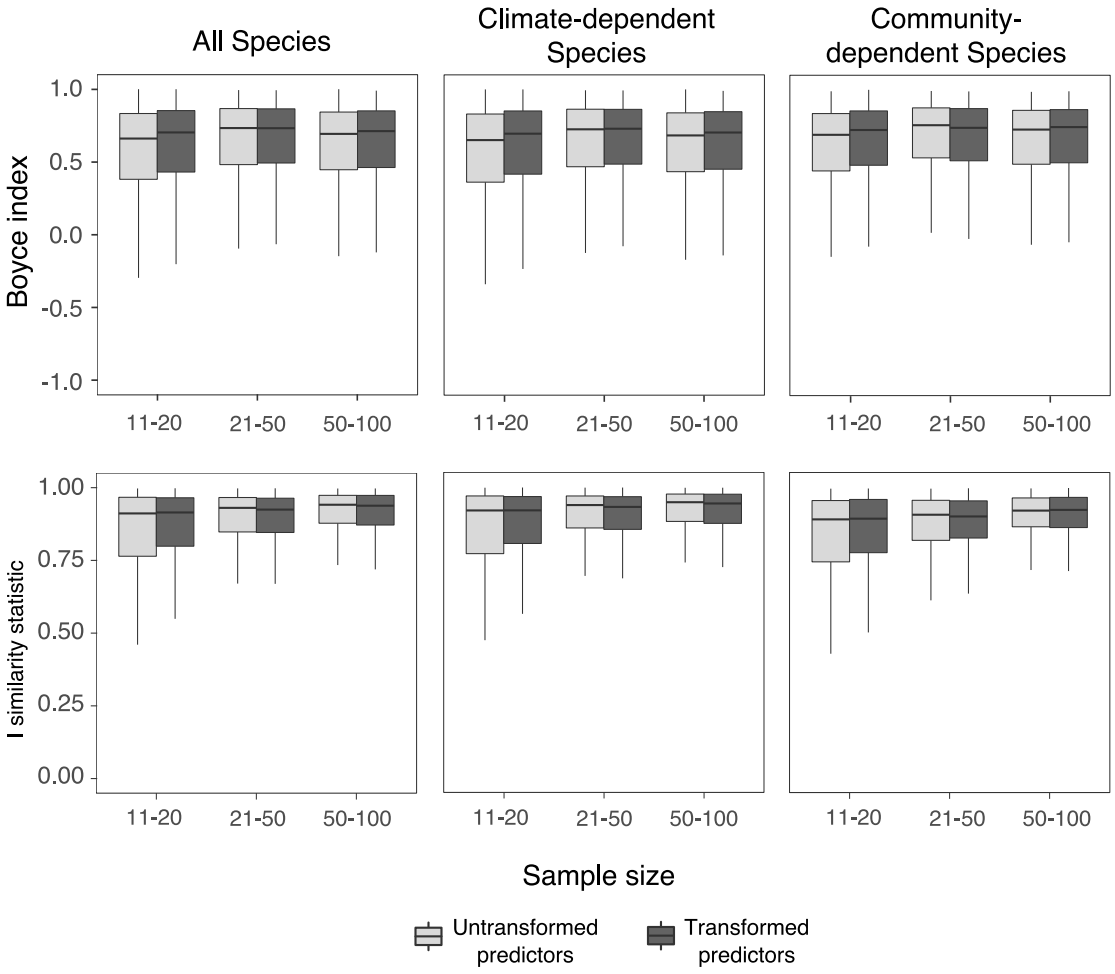


Figure 2. 2 Comparisons of model performance for models fitted with untransformed and transformed predictor variables in ENA. The values of the evaluation metrics (Boyce index and I similarity statistic) are given on the y-axis and the x-axis shows the sample sizes. Values of both metrics closer to 1 indicate a good model. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

Interactive effects of predictor type, and species and data characteristics

The LMMs were also used to assess the interactive affects of predictor type and other explanatory variables. The lowest sample sizes performed better when used with transformed variables for all communities when I similarity statistic were considered.

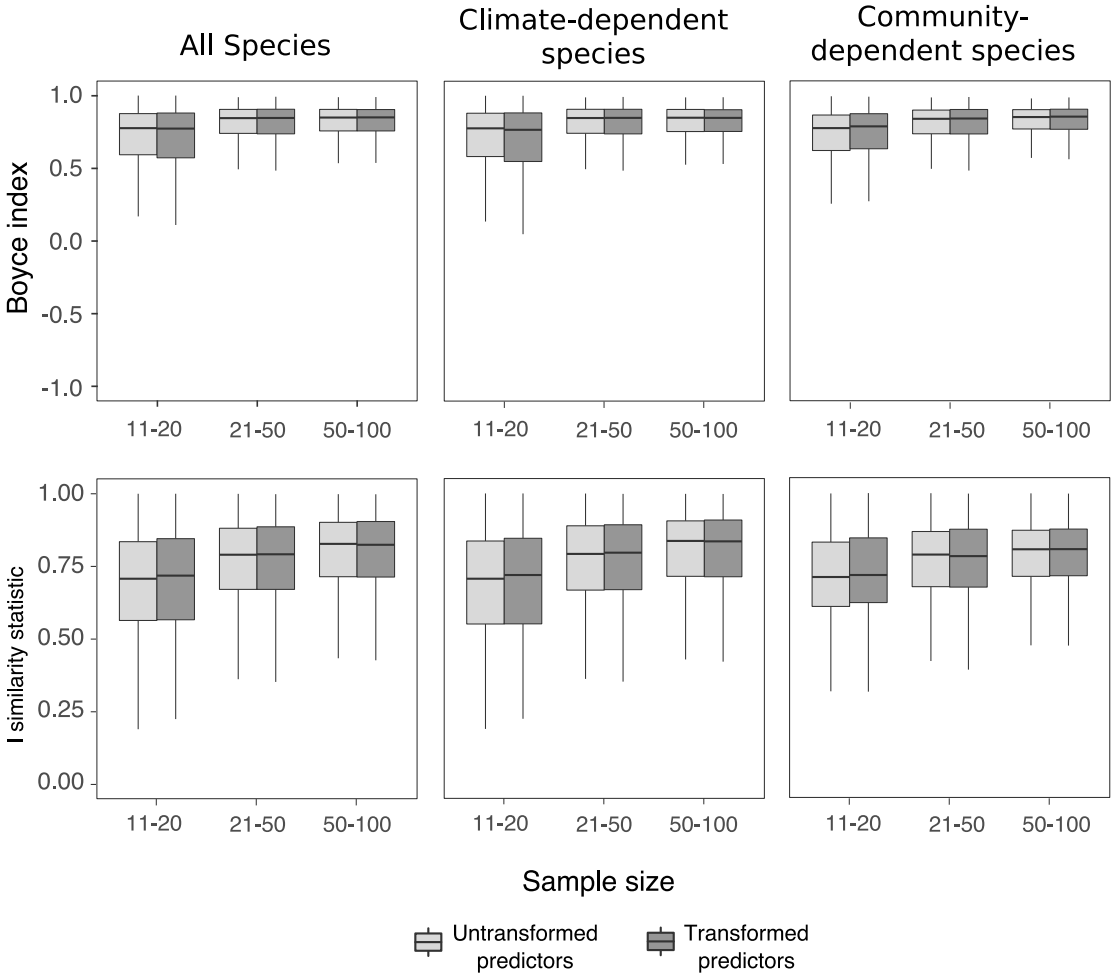


Figure 2. 3 Comparisons of model performance for models fitted with untransformed and transformed predictor variables in SACA. The values of the evaluation metrics (Boyce index and I similarity statistic) are given on the y-axis and the x-axis shows the sample sizes. Values of both metrics closer to 1 indicate a good model. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

However, the Boyce index showed mixed results for the influence of both the smallest and lowered model performance of medium sample sizes with transformed predictor variables (Fig.2.5). Larger ranges were associated with lowered Boyce index values when used with transformed variables in FIA and ENA (Fig.2.5). Boyce indices in SACA were not affected by range size, but were negatively related to I similarity statistic when used with transformed predictors (Figures 2.5 and 2.6). Species dependence characteristics had different results for the two virtual communities – in SACA, using transformed predictors for modeling climate-dependent species consistently lowered model performance, while in ENA, it increased Boyce indices and I similarity statistics (Figures 2.5 and 2.6).

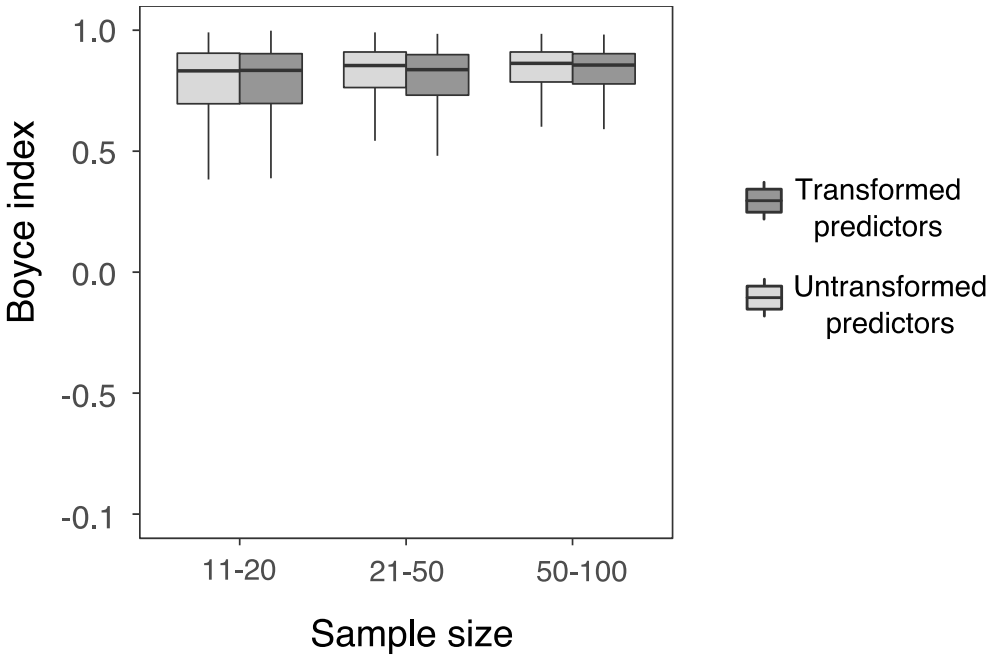


Figure 2. 4 Comparison of model performance for models fitted with untransformed and transformed predictor variables in FIA. The Boyce index is given on the y-axis and the sample sizes on x-axis. Values closer to 1 indicate a good model. The lower and upper hinges represent the 25th and 75th percentiles respectively, and the whiskers extend no more then the furthest values 1.5* inter-quartile range from the hinges.

Table 2. 2 Summary statistics from Mann-Whitney-Wilcoxon test on model evaluation metrics. Comparison was made between models fitted with untransformed climate variables (M_U) and models fitted with transformed variables (M_T) for all species regardless of range, dependence, or number of observations.

Community	Evaluation metrics	W	Mean M_U	Mean M_T
ENA	Boyce Index	4.43E+09 ***	0.921	0.933
ENA	I similarity statistix	5.18E+09	0.879	0.882
SACA	Boyce Index	2.47E+09 ***	0.842	0.848
SACA	I similarity statistix	2.84E+09 ***	0.749	0.753
FIA	Boyce Index	3.11E+08	0.838	0.841

*** $p << 0.001$

Table 2. 3 Summary statistics from Mann-Whitney-Wilcoxon test on model evaluation metrics based on species dependence. Comparison was made between models fitted with untransformed climate variables (M_U) and models fitted with transformed variables (M_T) for all species regardless of range or number of observations.

Community	Species dependence	Evaluation metric	W	Mean M_U	Mean M_T
ENA	Climate-dependent	Boyce Index	2.08E+09 ***	0.594	0.606
ENA	Community-dependent	Boyce Index	6.37E+08 ***	0.605	0.607
ENA	Climate-dependent	I similarity statistic	3.19E+09 *	0.883	0.887
ENA	Community-dependent	I similarity statistic	2.41E+08 *	0.862	0.865
SA	Climate-dependent	Boyce Index	1.04E+09 **	0.734	0.722
SA	Community-dependent	Boyce Index	4.61E+08 ***	0.746	0.751
SA	Climate-dependent	I similarity statistic	1.61E+09 ***	0.750	0.753
SA	Community-dependent	I similarity statistic	1.73E+08 ***	0.747	0.752

* $p < 0.05$, ** $p < 0.01$, *** $p << 0.001$

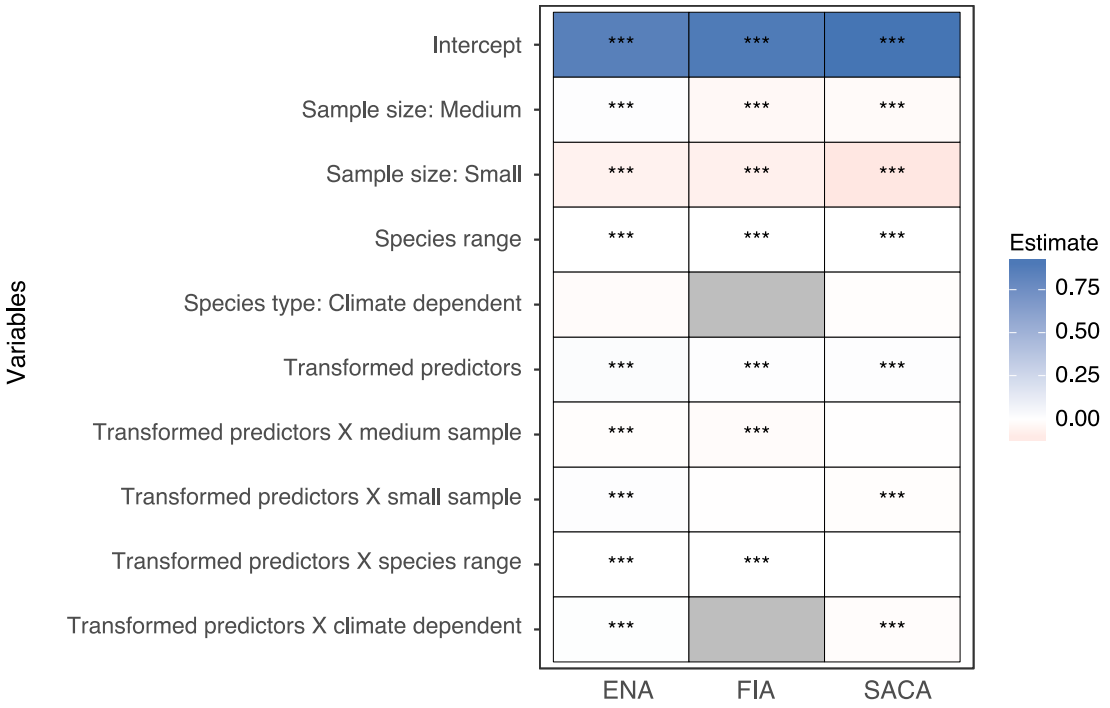


Figure 2. 5 Summary of results from Linear Mixed Models fitted with Boyce index as the response variable. The estimate for each predictor variable (listed on the y-axis) is represented by the colors and the significance of the estimate is given by the *s such that *** = $p < 0.001$.

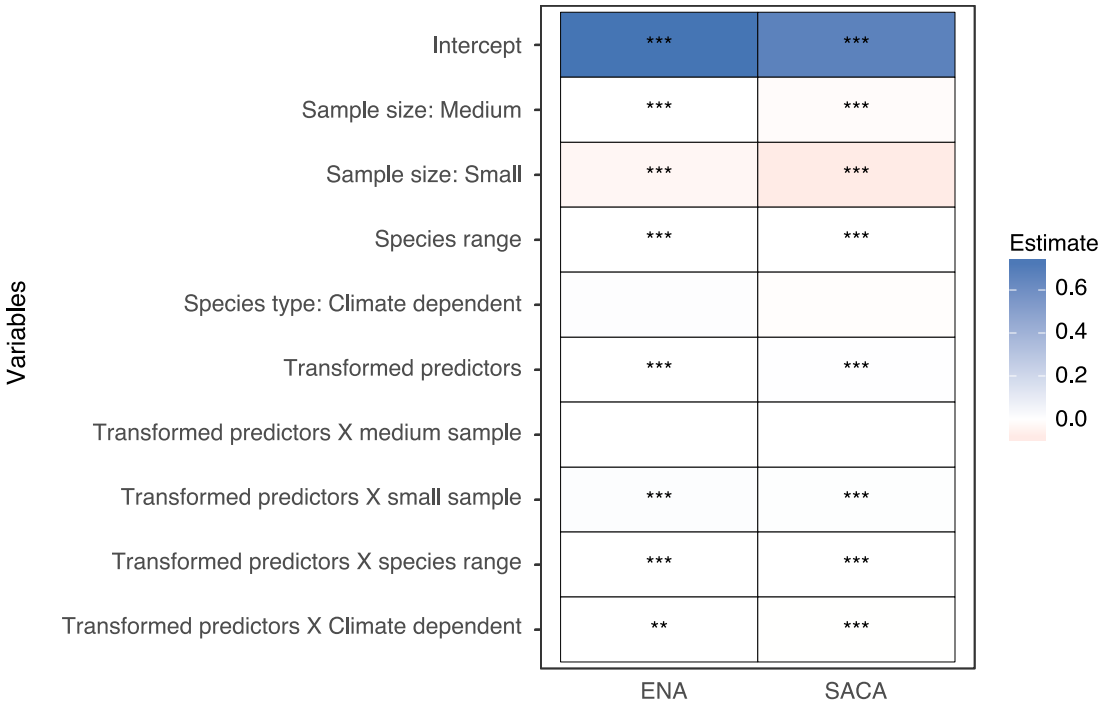


Figure 2. 6 Summary of results from Linear Mixed Models fitted with I similarity statistic as the response variable. The estimate for each predictor variable (listed on the y-axis) is represented by the colors and the significance of the estimate is given by the *s such that ** = $p < 0.01$ and *** = $p < 0.001$. Note that this analysis was only conducted for the simulated communities.

Discussion

Although most SDM frameworks do not incorporate species co-occurrences and community characteristics, this additional information can be useful in describing the relationship between environmental variables and species responses (Elith et al. 2006, Maguire et al. 2016). Here, I compared SDMs fit with and without GDM transformed environmental variables to test whether community-level information improves predictions. In general, the results showed support for the use of transformed predictor variables in SDMs, especially when sample sizes are smaller; however, the

differences between model performance of models using the untransformed or transformed predictor variables, while significant, were generally small.

Influence of transformed variables on model performance

In all communities, models fit using the transformed environmental variables performed better. Instances of poorer performance were limited to the discrimination ability (Boyce index) of models fit to the FIA. This shows that using transformed environmental variables can improve model quality and lead to spatial predictions of habitat suitability that more closely match actual species distributions. However, the differences in the mean values of model performance were small. Maxent is a robust model that has been shown to outperform other SDMs (Elith et al. 2006, Hernandez et al. 2006). However, the robustness of Maxent can sometimes come at the cost of overfitting due to high model complexity (Warren and Seifert 2011, Radosavljevic and Anderson 2014), especially when the default settings are used. Further study is required to determine whether other SDMs algorithms also realize improvements in model performance. As such, future analysis that includes a variety of modeling algorithms and with varying complexity may provide further insight into the degree of improvement that using transformed variables can provide.

Influence of species characteristics and sample size

Maxent models for species in all communities were influenced by species range size and the number of occurrences used to fit the models. The influence of number of occurrences on the model performances of various SDMs, including Maxent, has

been previously documented (Hernandez et al. 2006, Pearson et al. 2007, Wisz et al. 2008). Though Maxent has been shown to perform well with sample sizes as low as 5, lowered sample size can impact the individual sites that are included in the sample and affect model performance as a reduction (Hernandez et al. 2006, Pearson et al. 2007). My results show that higher sample sizes lead to better model performance and agree with results from previous studies (Stockwell and Peterson 2002, Hernandez et al. 2006, Pearson et al. 2007, Wisz et al. 2008).

SDMs fitted with the smallest sample size performed better when used in concert with transformed environmental variables. This improvement in performance likely can be attributed to the additional ecological information given by transformed environmental variables (compared to untransformed variables), which may supplement information that is missing due to the lower sample size. For instance, species in a community might follow similar distribution patterns as a result of which the transformed environmental variables can supplement the information provided by presences.

In addition to sample size, range size of a species was also important, though its effect was relatively small. Previous studies have found that models perform better with smaller-ranged species (Hernandez et al. 2006, Tessarolo et al. 2014), and suggest that species with larger ranges also need a larger number of observations (van Proosdij et al. 2016). I found that SDMs of species with smaller ranges performed better, especially when used with transformed environmental predictors. As such, models of small-ranged species can be improved with the use of transformed

environmental variables. Although transformed variables improve model performance, this is not uniformly true for all regions and range sizes. For example, in this study, models for species in SACA did not benefit from the use of transformed environmental variables, while those for ENA performed better with respect to Boyce index and I similarity statistic. The differences in regional characteristics may also be exaggerated by the niche-breadths of the species simulated in ENA and SACA. As such, the affect of range sizes may be small, but still needs to be taken into consideration when modeling species distributions.

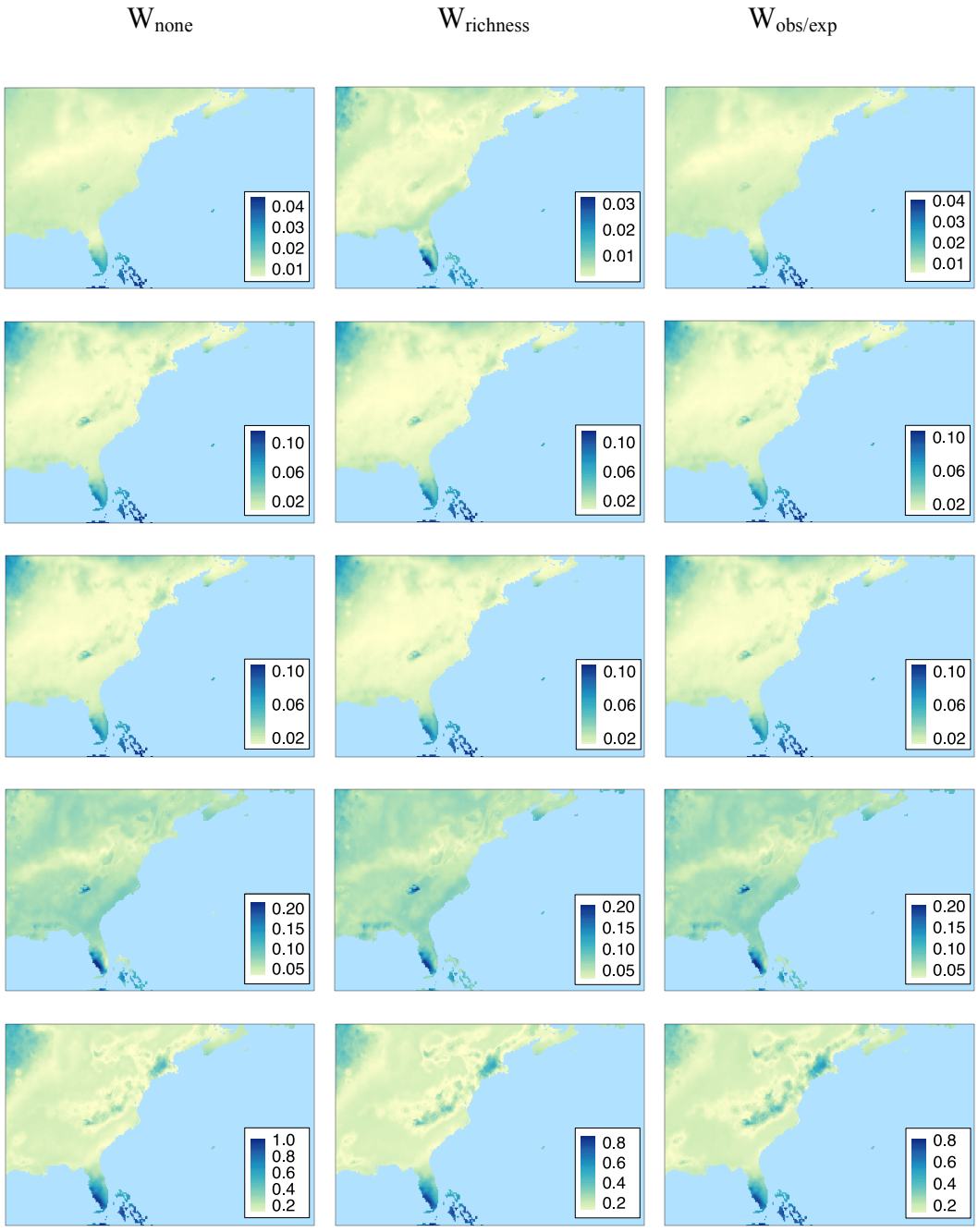
Species dependence also impacted model performance, but not uniformly for the two virtual communities. In SACA, SDMs for community-dependent species performed better when used with transformed variables, while those for climate-dependent species saw no improvement. On the other hand, SDMs for climate-dependent species in ENA performed better with transformed predictor variables.

Conclusion

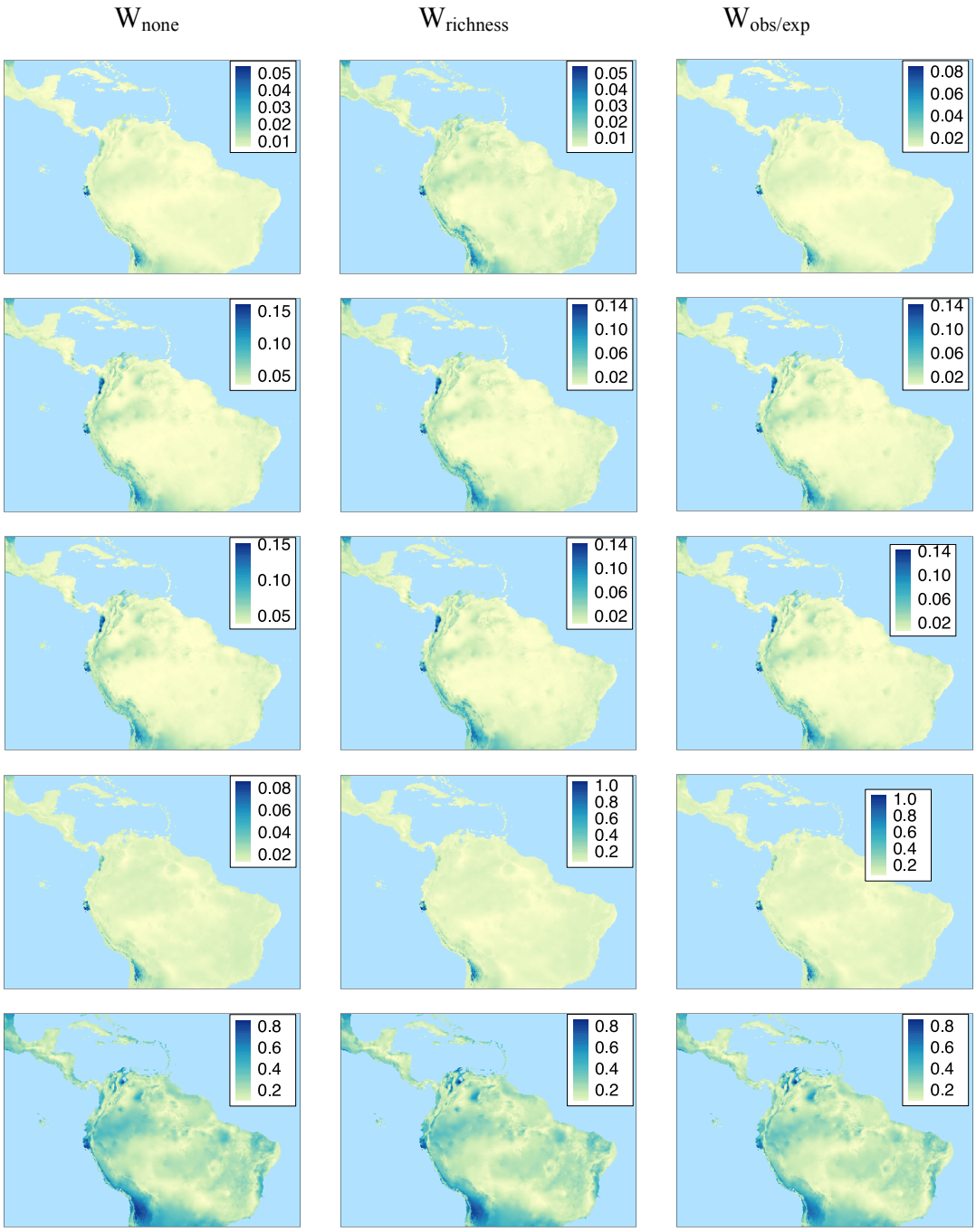
Overall, this study suggests that transformed environmental variables can improve Maxent models, especially when modeling the distributions of species with relatively small ranges and/or relatively few occurrence records. Future studies that consider other SDM algorithms would help provide insight into which algorithms would most benefit from the use of transformed predictor variables. More broadly, my results suggest a potential means of harnessing community-level information for the prediction of individual species distributions.

APPENDICES

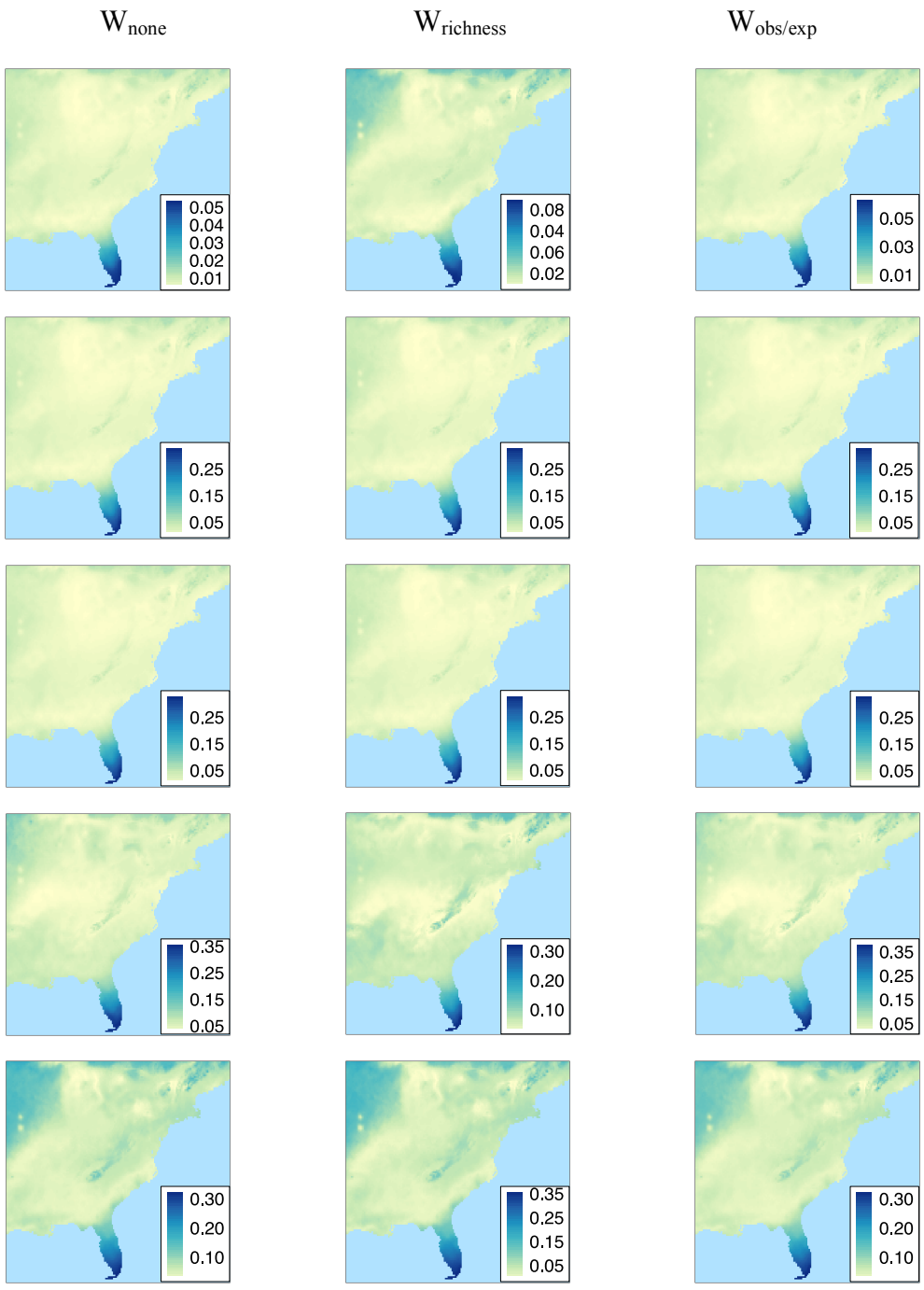
Appendix 1.1. Results from Procrustes analysis



Appendix 1.1.1. Mapped patterns of residuals of Procrustes analysis for biased and unbiased models in ENA. Higher values show the areas of lower agreement between models fitted to the unbiased and the biased data.

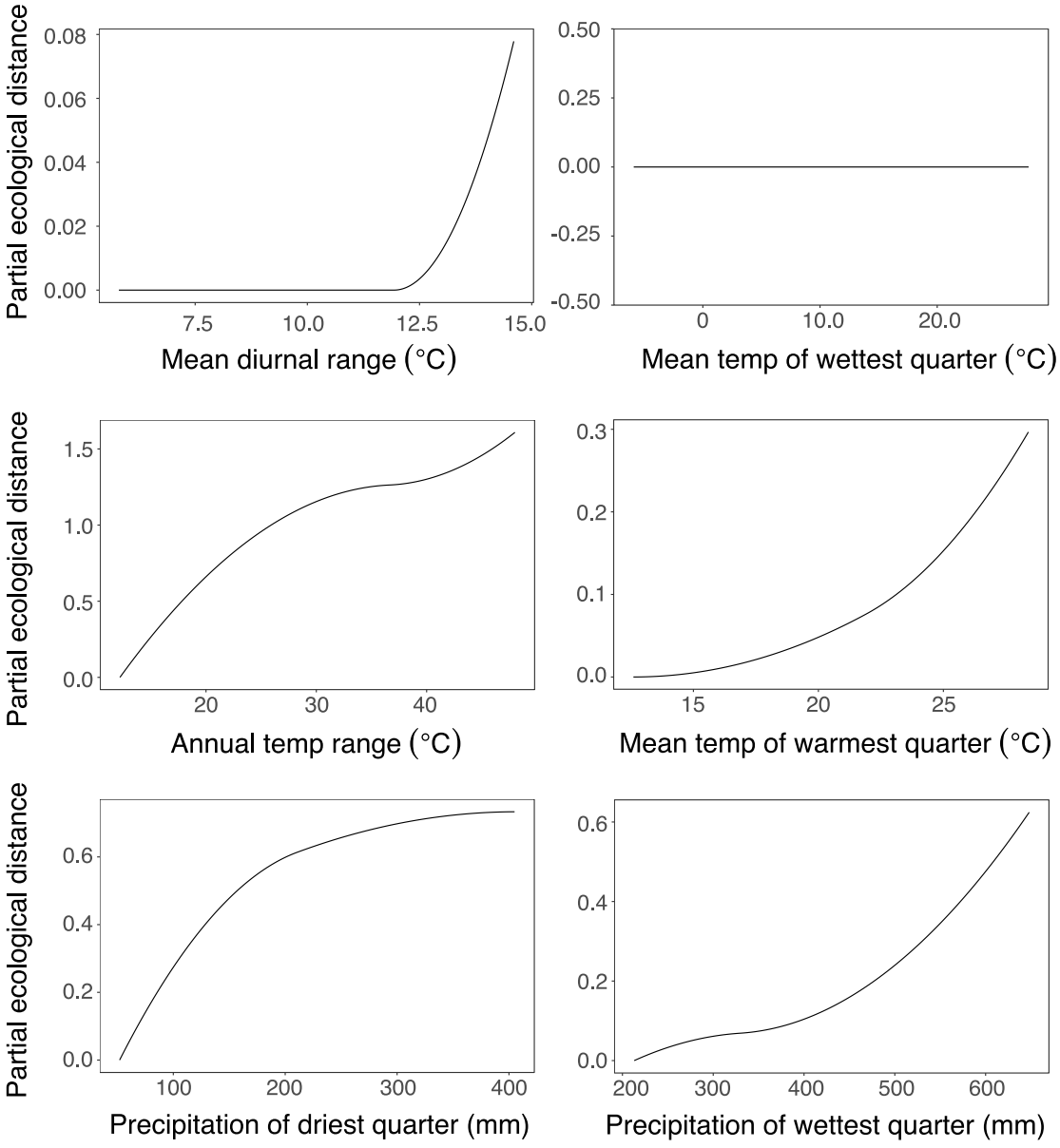


Appendix 1.1.2. Mapped patterns of residuals of Procrustes analysis for biased and unbiased models in SACA. Higher values show the areas of lower agreement between models fitted to the unbiased and the biased data.

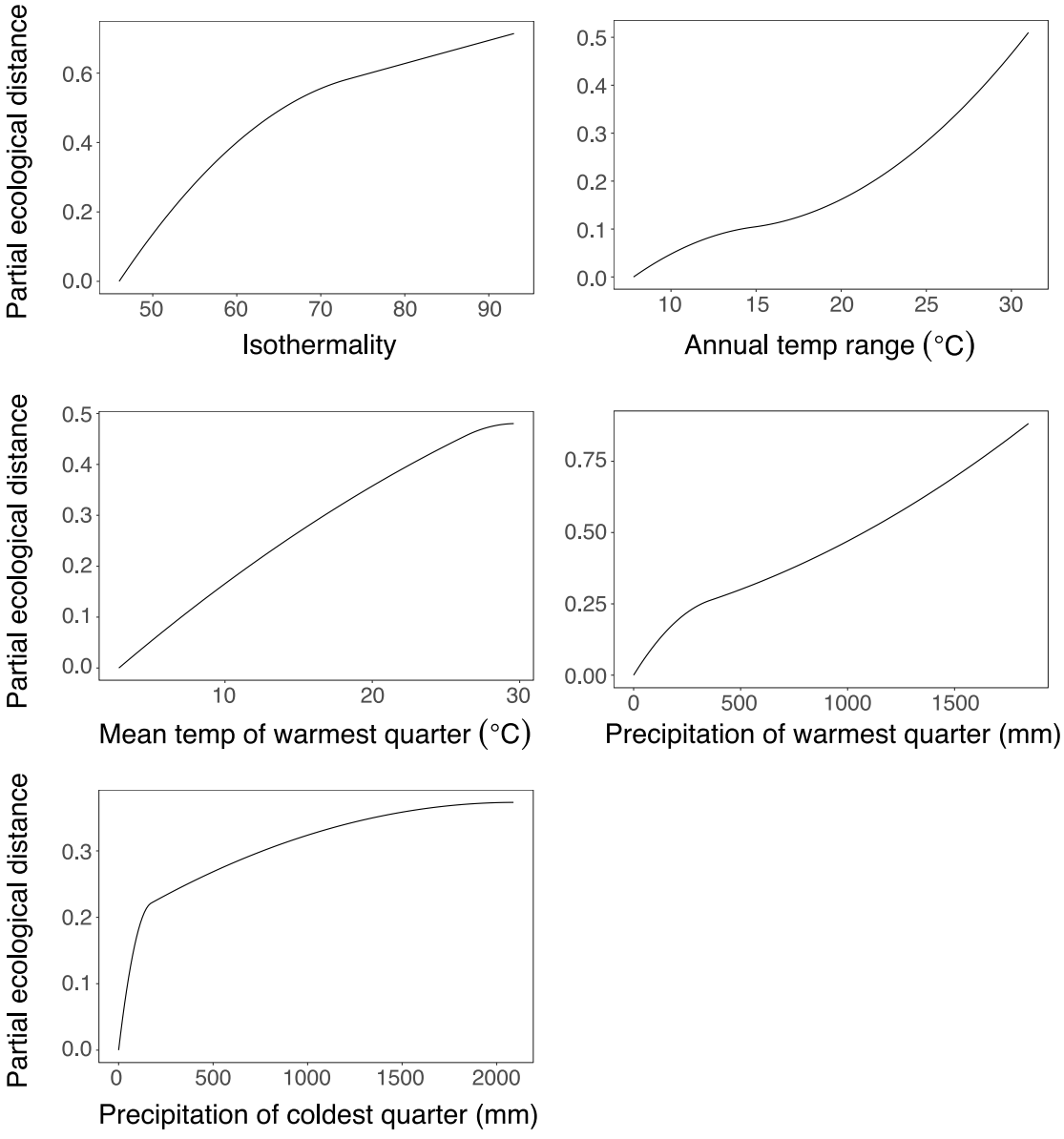


Appendix 1.1.3. Mapped patterns of residuals of Procrustes analysis for biased and unbiased models in FIA. Higher values show the areas of lower agreement between models fitted to the unbiased and the biased data.

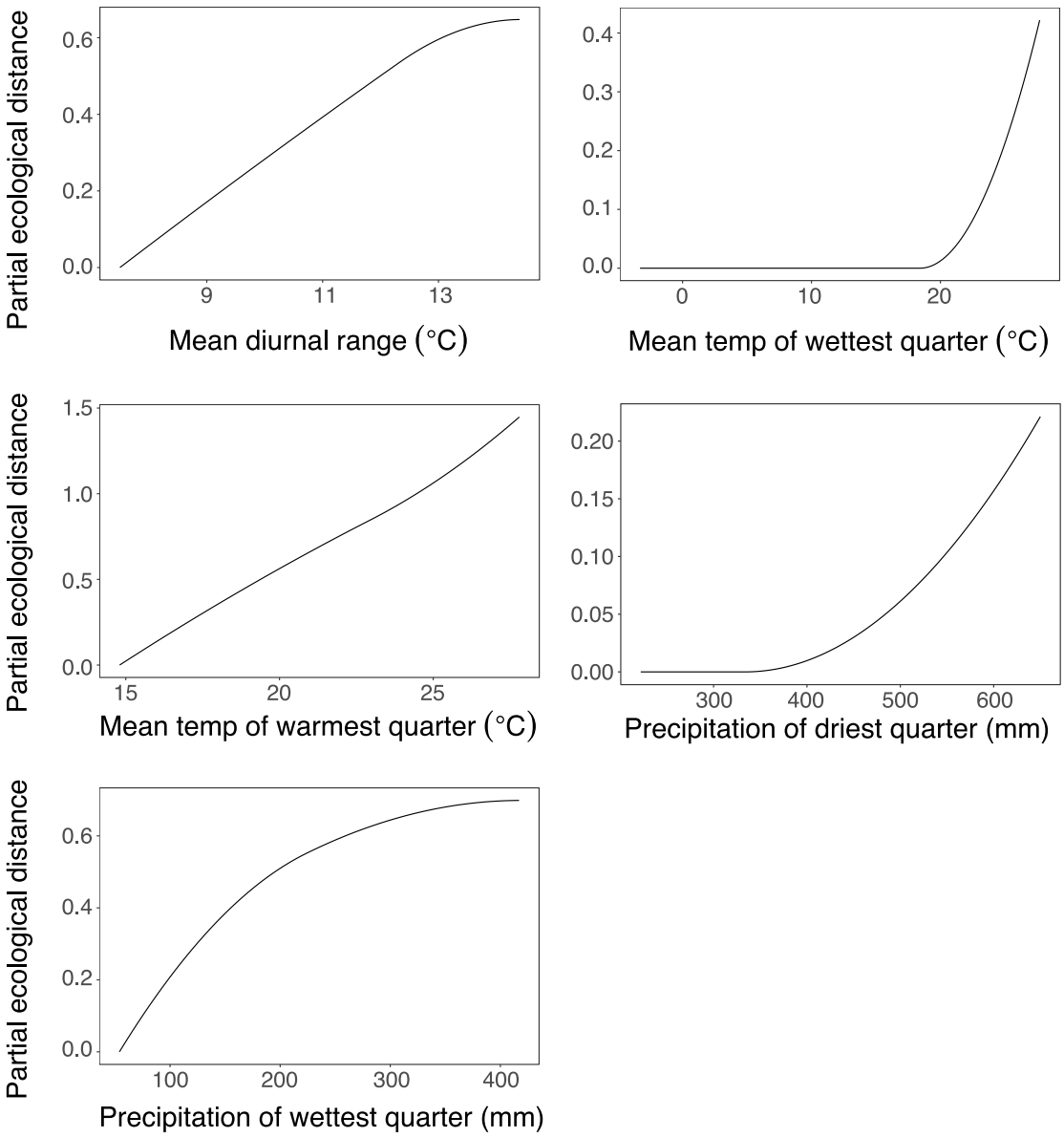
Appendix 2.1. Transformation functions obtained GDM models



Appendix 2.1.1 Splines obtained from GDM that were then used to transform raw environmental variables for ENA. The x-axis shows the raw values of the environmental variable and the y-axis shows the transformed value.



Appendix 2.1.2 Splines obtained from GDM that were then used to transform raw environmental variables for SACA. The x-axis shows the raw values of the environmental variable and the y-axis shows the transformed value.



Appendix 2.1.3 Splines obtained from GDM fitted data from FIA. These functions used to transform raw environmental variables. The x-axis shows the raw values of the environmental variable and the y-axis shows the transformed value.

Bibliography

Araújo, M. B., & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, *16*(6), 743–753. https://doi.org/10.1111/j.1466-8238.2007.00359.x

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Bean, W. T., Stafford, R., & Brashares, J. S. (2012). The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, *35*(3), 250–258. https://doi.org/10.1111/j.1600-0587.2011.06545.x

Bell, K. L., Heard, T. A., Manion, G., Ferrier, S., & van Klinken, R. D. (2013). The role of geography and environment in species turnover: phytophagous arthropods on a Neotropical legume. *Journal of Biogeography*, *40*(9), 1755–1766. https://doi.org/10.1111/jbi.12102

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, *15*(4), 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Jackson, S. T., & Ferrier, S. (2013). Space can substitute for time in predicting climate-change effects on biodiversity. *Proceedings of the National Academy of Sciences*, *110*(23), 9374–9379. https://doi.org/10.1073/pnas.1220228110

Bocedi, G., Palmer, S. C. F., Pe'er, G., Heikkinen, R. K., Matsinos, Y. G., Watts, K., & Travis, J. M. J. (2014). RangeShifter: a platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution*, *5*(4), 388–396. https://doi.org/10.1111/2041-210X.12162

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. (2002). Evaluating resource selection functions. *Ecological Modelling*, *157*(2), 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4

Buckley, L. B., & Jetz, W. (2008). Linking global turnover of species and environments. *Proceedings of the National Academy of Sciences*, *105*(46), 17836–17841. https://doi.org/10.1073/pnas.0803524105

Crase, B., Liedloff, A. C., & Wintle, B. A. (2012). A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, *35*(10), 879–888. https://doi.org/10.1111/j.1600-0587.2011.07138.x

D'Amen, M., Rahbek, C., Zimmermann, N. E., & Guisan, A. (2015). Spatial predictions at the community level: from current approaches to future frameworks. *Biological Reviews*, n/a-n/a. https://doi.org/10.1111/brv.12222

de Araújo, C. B., Marcondes-Machado, L. O., & Costa, G. C. (2014). The importance of biotic interactions in species distribution models: a test of the Eltonian noise hypothesis using parrots. *Journal of Biogeography*, *41*(3), 513–523. https://doi.org/10.1111/jbi.12234

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., … Guisan, A. (2017). ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography*, *40*(6), 774–787. https://doi.org/10.1111/ecog.02671

Dillon, M. E., Wang, G., & Huey, R. B. (2010). Global metabolic impacts of recent climate warming. *Nature*, *467*(7316), 704–706. https://doi.org/10.1038/nature09407

Duan, M., Liu, Y., Yu, Z., Baudry, J., Li, L., Wang, C., & Axmacher, J. C. (2016). Disentangling effects of abiotic factors and biotic interactions on cross-taxon congruence in species turnover patterns of plants, moths and beetles. *Scientific Reports*, *6*. https://doi.org/10.1038/srep23511

Dytham, C., Travis, J. M. J., Mustin, K., & Benton, T. G. (2014). Changes in species' distributions during and after environmental change: which eco-evolutionary processes matter more? *Ecography*, *37*(12), 1210–1217. https://doi.org/10.1111/ecog.01194

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Engler, R., & Guisan, A. (2009). MigClim: Predicting plant distribution and dispersal in a changing climate. *Diversity and Distributions*, *15*(4), 590–601. https://doi.org/10.1111/j.1472-4642.2009.00566.x

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Feeley, K. J., & Silman, M. R. (2011). Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, *17*(6), 1132–1140. https://doi.org/10.1111/j.1472-4642.2011.00813.x

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, *43*(3), 393–404. https://doi.org/10.1111/j.1365-2664.2006.01149.x

Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, *13*(3), 252–264. https://doi.org/10.1111/j.1472-4642.2007.00341.x

Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*(1), 1–16. https://doi.org/10.1111/ele.12376

Fitzpatrick, M. C., Sanders, N. J., Ferrier, S., Longino, J. T., Weiser, M. D., & Dunn, R. (2011). Forecasting the future of biodiversity: a test of single- and multi-species models for ants in North America. *Ecography*, *34*(5), 836–847. https://doi.org/10.1111/j.1600-0587.2011.06653.x

Fitzpatrick, M. C., Sanders, N. J., Normand, S., Svenning, J.-C., Ferrier, S., Gove, A. D., & Dunn, R. R. (2013). Environmental and historical imprints on beta diversity: insights from variation in rates of species turnover along gradients. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1768), 20131201. https://doi.org/10.1098/rspb.2013.1201

Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press. Retrieved from

http://books.google.com/books?hl=en&lr=&id=CkshAwAAQBAJ&oi=fnd&pg=PR15&dq=info:EQKVyx55qcQJ:scholar.google.com&ots=6pbXfvRD_q&sig=HzY3JnBZOMW9aDhh1SgEC8ZoFck

Franklin, J. (2013). Species distribution models in conservation biogeography: developments and challenges. *Diversity and Distributions*, *19*(10), 1217–1223. https://doi.org/10.1111/ddi.12125

Franklin, J., Davis, F. W., Ikegami, M., Syphard, A. D., Flint, L. E., Flint, A. L., & Hannah, L. (2013). Modeling plant species distributions under future climates: how fine scale do climate projections need to be? *Global Change Biology*, *19*(2), 473–483. https://doi.org/10.1111/gcb.12051

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, *19*(9), 497–503. https://doi.org/10.1016/j.tree.2004.07.006

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, *40*(2), 281–295. https://doi.org/10.1111/ecog.02445

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., & the NCEAS Species Distribution Modelling Group. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, *13*(3), 332–340. https://doi.org/10.1111/j.1472-4642.2007.00342.x

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435. https://doi.org/10.1111/ele.12189

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G., & Körber, J.-H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, *16*(6), 754–763. https://doi.org/10.1111/j.1466-8238.2007.00345.x

Hermoso, V., Kennard, M. J., Schmidt, D. J., Bond, N., Huey, J. A., Mondol, R. K., … Hughes, J. M. (2016). Species distributions represent intraspecific genetic diversity of freshwater fish in conservation assessments. *Freshwater Biology*, *61*(10), 1707–1719. https://doi.org/10.1111/fwb.12810

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

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*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, *145*(2–3), 111–121. https://doi.org/10.1016/S0304-3800(01)00396-9

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*(2), 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Jackson, D. A. (1995). PROTEST: A PROcrustean Randomization TEST of community environment concordance. *Écoscience*, *2*(3), 297–303. https://doi.org/10.1080/11956860.1995.11682297

Jenkins, C. N., Houtan, K. S. V., Pimm, S. L., & Sexton, J. O. (2015). US protected lands mismatch biodiversity priorities. *Proceedings of the National Academy of Sciences*, *112*(16), 5081–5086. https://doi.org/10.1073/pnas.1418034112

Jenkins, C. N., Pimm, S. L., & Joppa, L. N. (2013). Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences*, *110*(28), E2602–E2610. https://doi.org/10.1073/pnas.1302251110

Jenkins, C. N., & Van Houtan, K. S. (2016). Global and regional priorities for marine biodiversity protection. *Biological Conservation*, *204*, 333–339. https://doi.org/10.1016/j.biocon.2016.10.005

Jones, M. M., Ferrier, S., Condit, R., Manion, G., Aguilar, S., & Pérez, R. (2013). Strong congruence in tree and fern community turnover in response to soils and climate in central Panama. *Journal of Ecology*, *101*(2), 506–516. https://doi.org/10.1111/1365-2745.12053

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., … Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences*, *106*(23), 9322–9327. https://doi.org/10.1073/pnas.0810306106

Kreft, H., & Jetz, W. (2007). Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences*, *104*(14), 5925–5930. https://doi.org/10.1073/pnas.0608361104

Leathwick, J. R., Snelder, T., Chadderton, W. L., Elith, J., Julian, K., & Ferrier, S. (2011). Use of generalised dissimilarity modelling to improve the biological discrimination of river and stream classifications. *Freshwater Biology*, *56*(1), 21–38. https://doi.org/10.1111/j.1365-2427.2010.02414.x

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Loiseau, N., Legras, G., Kulbicki, M., Mérigot, B., Harmelin-Vivien, M., Mazouni, N., … Gaertner, J. c. (2017). Multi-component β-diversity approach reveals conservation dilemma between species and functions of coral reef fishes. *Journal of Biogeography*, *44*(3), 537–547. https://doi.org/10.1111/jbi.12844

Maguire, K. C., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., Williams, J. W., Ferrier, S., & Lorenz, D. J. (2016). Controlled comparison of species- and community-level models across novel climates and communities. *Proc. R. Soc. B*, *283*(1826), 20152817. https://doi.org/10.1098/rspb.2015.2817

Manion, G., Lisk, M., Ferrier, S., Nieto-Lugilde, D., & Fitzpatrick, M. C. (2016). gdm: Functions for Generalized Dissimilarity Modeling (Version 1.2.3). Retrieved from https://cran.r-project.org/web/packages/gdm/index.html

McCain, C. M. (2007). Could temperature and water availability drive elevational species richness patterns? A global case study for bats. *Global Ecology and Biogeography*, *16*(1), 1–13. https://doi.org/10.1111/j.1466-8238.2006.00263.x

McMahon, S. M., Harrison, S. P., Armbruster, W. S., Bartlein, P. J., Beale, C. M., Edwards, M. E., … Prentice, I. C. (2011). Improving assessment and modelling of climate change impacts on global terrestrial biodiversity. *Trends in Ecology & Evolution*, *26*(5), 249–259. https://doi.org/10.1016/j.tree.2011.02.012

Meier, E. S., Kienast, F., Pearman, P. B., Svenning, J.-C., Thuiller, W., Araújo, M. B., … Zimmermann, N. E. (2010). Biotic and abiotic variables show little

redundancy in explaining tree species distributions. *Ecography*, *33*(6), 1038–1048. https://doi.org/10.1111/j.1600-0587.2010.06229.x

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Meyer, C., Jetz, W., Guralnick, R. P., Fritz, S. A., & Kreft, H. (2016). Range geometry and socio-economics dominate species-level biases in occurrence information. *Global Ecology and Biogeography*, n/a-n/a. https://doi.org/10.1111/geb.12483

Meynard, C. N., & Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. *Journal of Biogeography*, *40*(1), 1–8. https://doi.org/10.1111/jbi.12006

Midgley, G. F., Davies, I. D., Albert, C. H., Altwegg, R., Hannah, L., Hughes, G. O., … Thuiller, W. (2010). BioMove – an integrated platform simulating the dynamic response of species to environmental change. *Ecography*, *33*(3), 612–616. https://doi.org/10.1111/j.1600-0587.2009.06000.x

Mokany, K., Thomson, J. J., Lynch, A. J. J., Jordan, G. J., & Ferrier, S. (2015). Linking changes in community composition and function under climate change. *Ecological Applications*, *25*(8), 2132–2141. https://doi.org/10.1890/14-2384.1

Montoya, J. M., & Raffaelli, D. (2010). Climate change, biotic interactions and ecosystem services. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1549), 2013–2018. https://doi.org/10.1098/rstb.2010.0114

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2017). vegan: Community Ecology Package (Version 2.4-3). Retrieved from https://cran.r-project.org/web/packages/vegan/index.html

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Townsend Peterson, A. (2007). ORIGINAL ARTICLE: Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, *34*(1), 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, *129*(2), 169–178. https://doi.org/10.1007/s004420100720

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

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*, 655–662.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., … Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, *344*(6187), 1246752. https://doi.org/10.1126/science.1246752

Prober, S. M., Hilbert, D. W., Ferrier, S., Dunlop, M., & Gobbett, D. (2012). Combining community-level spatial modelling and expert knowledge to inform climate adaptation in temperate grassy eucalypt woodlands and related grasslands. *Biodiversity and Conservation*, *21*(7), 1627–1650. https://doi.org/10.1007/s10531-012-0268-4

RStudio Team (2016). RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA. http://www.rstudio.com.

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. https://doi.org/10.1111/jbi.12227

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, *3*(4), 425–441. https://doi.org/10.1214/ss/1177012761

Record, S., Fitzpatrick, M. C., Finley, A. O., Veloz, S., & Ellison, A. M. (2013). Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. *Global Ecology and Biogeography*, *22*(6), 760–771. https://doi.org/10.1111/geb.12017

Roland Pitcher, C., Lawton, P., Ellis, N., Smith, S. J., Incze, L. S., Wei, C.-L., … Snelgrove, P. V. R. (2012). Exploring the role of environmental variables in

shaping patterns of seabed biodiversity composition in regional-scale ecosystems. *Journal of Applied Ecology*, *49*(3), 670–679. https://doi.org/10.1111/j.1365-2664.2012.02148.x

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Suarez, A. V., & Tsutsui, N. D. (2004). The Value of Museum Collections for Research and Society. *BioScience*, *54*(1), 66–74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Tessarolo, G., Rangel, T. F., Araújo, M. B., & Hortal, J. (2014). Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions*, *20*(11), 1258–1269. https://doi.org/10.1111/ddi.12236

Thomassen, H. A., Fuller, T., Buermann, W., Milá, B., Kieswetter, C. M., Jarrín-V., P., … Smith, T. B. (2011). Mapping evolutionary process: a multi-taxa approach to conservation prioritization. *Evolutionary Applications*, *4*(2), 397–413. https://doi.org/10.1111/j.1752-4571.2010.00172.x

Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). bimod2: Ensemble Platform for Species Distribution Modeling (Version 3.3-7). Retrieved from https://cran.r-project.org/web/packages/biomod2/index.html

Ulrich, W., Soliveres, S., Maestre, F. T., Gotelli, N. J., Quero, J. L., Delgado-Baquerizo, M., … Zaady, E. (2014). Climate and soil attributes determine plant species turnover in global drylands. *Journal of Biogeography*, *41*(12), 2307–2319. https://doi.org/10.1111/jbi.12377

Václavík, T., Kupfer, J. A., & Meentemeyer, R. K. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *Journal of Biogeography*, *39*(1), 42–55. https://doi.org/10.1111/j.1365-2699.2011.02589.x

Valdujo, P. H., Carnaval, A. C. O. Q., & Graham, C. H. (2013). Environmental correlates of anuran beta diversity in the Brazilian Cerrado. *Ecography*, *36*(6), 708–717. https://doi.org/10.1111/j.1600-0587.2012.07374.x

van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, *39*(6), 542–552. https://doi.org/10.1111/ecog.01509

VanDerWal, J., Falconi, L., Januchowski, S., & Storlie, L. S. and C. (2014). SDMTools: Species Distribution Modelling Tools: Tools for processing data associated

with species distribution modelling exercises (Version 1.1-221). Retrieved from https://cran.r-project.org/web/packages/SDMTools/index.html

version), J. P. (S, to 2007), D. B. (up, to 2002), S. D. (up, to 2005), D. S. (up, authors (src/rs.f), E., sigma), S. H. (Author fixed, … R-core. (2017). nlme: Linear and Nonlinear Mixed Effects Models (Version 3.1-131). Retrieved from https://cran.r-project.org/web/packages/nlme/index.html

Walther, G.-R. (2010). Community and ecosystem responses to recent climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1549), 2019–2024. https://doi.org/10.1098/rstb.2010.0021

Wang, Z., Brown, J. H., Tang, Z., & Fang, J. (2009). Temperature dependence, spatial scale, and tree species diversity in eastern Asia and North America. *Proceedings of the National Academy of Sciences*, *106*(32), 13388–13392. https://doi.org/10.1073/pnas.0905030106

Warren, D. L., Glor, R. E., & Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution; International Journal of Organic Evolution*, *62*(11), 2868–2883. https://doi.org/10.1111/j.1558-5646.2008.00482.x

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*(2), 335–342. https://doi.org/10.1890/10-1171.1

Williams, J. W., Blois, J. L., Gill, J. L., Gonzales, L. M., Grimm, E. C., Ordonez, A., … Veloz, S. D. (2013). Model systems for a no-analog future: species associations and climates during the last deglaciation. *Annals of the New York Academy of Sciences*, *1297*(1), 29–43. https://doi.org/10.1111/nyas.12226

Williams, K. J., Belbin, L., Austin, M. P., Stein, J. L., & Ferrier, S. (2012). Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science*, *26*(11), 2009–2047. https://doi.org/10.1080/13658816.2012.698015

Willis, K. J., Jeffers, E. S., Tovar, C., Long, P. R., Caithness, N., Smit, M. G. D., … Weissenberger, J. (2012). Determining the ecological value of landscapes beyond protected areas. *Biological Conservation*, *147*(1), 3–12. https://doi.org/10.1016/j.biocon.2011.11.001

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. *Diversity and*

Distributions, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., … Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, *88*(1), 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., … Grimm, V. (2010). The virtual ecologist approach: simulating data and observers. *Oikos*, *119*(4), 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x