Received: 29 July 2016 | Revised: 17 August 2017 | Accepted: 29 August 2017

DOI: 10.1111/geb.12678

MACROECOLOGICAL METHODS

WILEY Global Ecology and Biogeography A Journal of Macroecology

Creating individual accessible area hypotheses improves stacked species distribution model performance

Jacob C. Cooper | **Jorge Soberón**

University of Kansas Biodiversity Institute
and Department of Ecology and
Evolutionary Biology, 1345 Jayhawk
Boulevard, Lawrence, KS 66045, USA

Correspondence
Jacob C. Cooper, University of Kansas Biodiversity Institute and Department of Ecology and Evolutionary Biology, 1345 Jayhawk Boulevard, Lawrence, KS 66045, USA.
Email: jccooper@uchicago.edu

Present address
Jacob C. Cooper, Committee on Evolutionary Biology, The University of Chicago, Culver Hall 402, Chicago, IL 60637, USA
and
Life Sciences Section, Integrative Research Center, The Field Museum, 1400 South Lake Shore Drive, Chicago, IL 60605, USA.

Funding information
National Science Foundation, Grant/Award Number: 1208472; University of Kansas Biodiversity Institute; University of Kansas Ecology and Evolutionary Biology Department

Editor: Naia Morueta-Holme

Abstract

Aim: Stacked species distribution models (SDMs) are an important step towards estimating species richness, but frequently overpredict this metric and therefore erroneously predict which species comprise a given community. We test the idea that developing hypotheses about accessible area a priori can greatly improve model performance. By integrating dispersal ability via accessible area into SDM creation, we address an often-overlooked facet of ecological niche modelling.

Innovation: By limiting the training and transference areas to theoretically accessible areas, we are creating more accurate SDMs on the basis of a taxon's explorable environments. This limitation of space and environment is a more accurate reflection of a taxon's true dispersal properties and more accurately reflects the geographical and environmental space to which a taxon is exposed. Here, we compare the predictive performance of stacked SDMs derived from spatially constrained and unconstrained training areas.

Main conclusions: Restricting a species' training and transference areas to a theoretically accessible area greatly improves model performance. Stacked SDMs drawn from spatially restricted training areas predicted species richness and community composition more accurately than non-restricted stacked SDMs. These accessible area-based restrictions mimic true dispersal barriers to species and limit training areas to the suite of environments to those which a species is exposed to in nature. Furthermore, these restrictions serve to 'clip' predictions in geographical space, thus removing overpredictions in adjacent geographical regions where the species is known to be absent.

KEYWORDS
accessible area, dispersal, ecological niche models, hummingbirds, macroecological models, presence–absence matrix, stacked niche models

1 | INTRODUCTION

Conservation and biodiversity research worldwide has generally been limited and biased towards biodiversity hotspots and regions that are relatively easy to access and well known (Ponder, Carter, Flemons, & Chapman, 2001), but gaps in our knowledge of the area of distribution of species (the Wallacean shortfall) are substantial (Whittaker et al., 2005). Most of what is known about species distributions is affected by this incomplete sampling and is therefore spatially limited and biased from the outset (Reddy & Dávalos, 2003). Species distribution models (SDMs) are seen as an important method towards estimating a species' distributional extent (Feria & Peterson, 2002; Peterson et al., 2011; Soberón & Peterson, 2005). Moreover, SDMs can be aggregated and stacked for multiple taxa to predict richness and community composition for a given locality (Aranda & Lobo, 2011; Graham & Hijmans, 2006; Peterson et al., 2011), although this method is still being developed (Calabrese, Certain, Kraan, & Dormann, 2013).

One of the most common ways of estimating a species' distribution is by creating an ecological niche model (ENM). ENMs use climatic (or other environmental) associations of observed localities to estimate the environmental space (i.e., ecological niche) occupied by the species; this 'niche model' can then be projected into geographical space to find areas with environments similar to those where the species has been observed (i.e., 'homocline matching'; Peterson et al., 2011). This represents a potential distribution based on an index of environmental similarity, and not a 'probability of presence' (Peterson et al., 2011;

Soberón & Peterson, 2005). Many algorithms can be used to provide such indices of environmental similarity (Chefaoui & Lobo, 2008; Elith et al., 2006, 2011; Peterson et al., 2011; Phillips et al., 2009), as reviewed by Franklin (2010), which in turn can then be thresholded to create binary correlative models that predict a species' potential distribution in geographical space (Liu, White, & Newell, 2013).

When true absence data are available, it is possible to estimate statistically the probability of presence given an environmental combination (Pearce & Boyce, 2006), and therefore an estimation of the actual occupied area can be derived. In such cases, the projection of the niche model to geographical space represents a hypothesis about the actual distribution of the species (a realized species distribution model). It is also possible to estimate an occupied area by creating a 'hybrid' or 'process-oriented' model that combines a mechanistic model of the fundamental niche with a dispersal kernel and with inferred interactions with syntopic species (Kissling et al., 2012), although this method requires substantial ancillary data to parameterize models. Alternatively, even without explicit absence data, it is also possible to post-process the predictions of presence-only correlative niche models that estimate potential areas by resorting to hypotheses about dispersal capacities to 'clip' estimates (Soberón, 2010). This method is based on the ideas of the **BAM** diagram (Soberón & Peterson, 2005), which relate a species' biotic interactions (**B**) to its abiotic environment (**A**) and to its motility and accessible area (**M**). In this framework, hypothesized dispersal limitations (an **M** hypothesis) can be established based on biogeographical considerations (Soberón, 2010), thus clipping potential distribution to the **M** hypothesis as a proxy for the species' historical dispersal area. Addressing a species' biogeographical history in this way removes assumptions of local extinction or potential colonization, and thus constrains a species' model to a known accessible area.

If one is interested in the composition of an entire fauna or flora, 'stacks' of thresholded individual SDMs can be used to generate a list. It should be noted that a floristic or faunistic list is in itself a 'thresholded' object (i.e., a list of presences and implicit absences determined by a likelihood of presence); therefore, some form of thresholding is inherent to list estimation. Thresholded potential distribution models represent a set of coordinate-referenced presence and absence predictions that can be spatially aligned with cells to create presence–absence matrices that predict the community composition and species richness at any given locality (D'Amen, Dubuis et al., 2015; D'Amen, Rahbek, Zimmermann, & Guisan, 2015; Dubuis et al., 2011; Peterson et al., 2011; Pineda & Lobo, 2009). It is important to note, however, that by the above arguments, unless the modelling process is specifically aimed at estimating the occupied area, presence-only correlational models will approximate the larger potential area (Soberón, 2010). Previous studies that use presence-only data and unclipped SDMs to estimate community composition should be expected to overpredict richness, despite being correlated with observed species numbers (Calabrese et al., 2013; Pineda & Lobo, 2009). Alternative species richness estimates have been derived from macroecological models, but these estimates lack the ability to predict community composition and are therefore limited in their usefulness when a list is needed (e.g., when the presence of a particular assemblage of rare or threatened species is sought; Calabrese et al., 2013; Dubuis et al., 2011).

Whether stacks of binary SDMs can create useful, realistic presence–absence matrices is still an open question. Some attempts have been made to improve SDM accuracy by means of adding information (e.g., absences) in high-quality presence–absence data (D'Amen, Rahbek, et al., 2015; Guisan & Rahbek, 2011). Even using presence–absence data, the overall feasibility of these procedures has been questioned (Calabrese et al., 2013; Gastón & García-Viñas, 2013), and such methods are data demanding if aimed for large taxonomic groups or geographical scales. It has become apparent that when the purpose is to estimate the realized distribution of a species, using methods that fail to incorporate the dispersal constraints is likely to bias predicted distributions by including suitable environments that are inaccessible to study taxa (Soberón & Peterson, 2005).

In this study, we resort to the method of hypothesizing **M** regions to clip ENM outputs (estimated from presence-only data) to 'reduce' the larger estimated potential area to a hypothesis about the accessible area containing the occupied area. We compare two sets of SDMs (one clipped, the other unclipped) to address how **M** affects the predictive ability for both species richness and community composition. An initial set of ENMs was estimated using classic presence-only approaches that use a calibration area that is either a continent or the area accessible to the entire clade (i.e., an unconstrained model). The second set of models aims to estimate occupied niches by calibrating ENMs within species-specific **M**s defined by expert range maps, biogeographical areas and reliable digitally accessible knowledge (Barve et al., 2011; Owens et al., 2013; Soberón & Peterson, 2005). This procedure also 'clips' the predictions in geographical space, explicitly incorporating abiotic barriers (when definable) and implicitly incorporating biotic barriers to range expansion. This **M**-based approach, which has never been applied to a large group of organisms, should yield estimates closer to the occupied area rather than the potential area for individual species, and thus should produce better estimates of realized species richness and community composition from the stacked models.

2 | METHODS

2.1 | Data and test locality description

Primary occurrence data were downloaded for all hummingbird species (Aves: Trochilidae) for which data were available from the Global Biodiversity Information Facility (GBIF; Global Biodiversity Informatics Facility, 2014) using the R 3.0.0 (R Core Team, 2015) package rgbif (Chamberlain, Ram, Barve, & Mcglinn, 2015) and directly from eBird in December 2013 (eBird, 2012; Sullivan et al., 2009). We followed the taxonomy of the 2013 International Ornithological Congress checklist (Gill & Donsker, 2013) with the exception of the west Ecuadorian *Anthracothorax prevostii iridescens*, which was treated separately from other *A. prevostii* populations (see discussion in Ridgely & Greenfield, 2001). eBird data were parsed to remove all records that listed distances > 10 km or > 900 min of effort as a precaution against data drawn from checklists that covered multi-day effort or checklists covering areas larger than the locality of interest (e.g., observers often cluster data near cities, well-known sites or country/state centroids). Merged eBird and GBIF data had duplicate localities removed and remaining localities rarefied to a minimal inter-point distance of 10 km for each species in R 3.0.0 (J. D. Manthey, unpublished code 2015; R Core Team, 2015; Vavrek, 2011). We performed the following procedures to remove erroneous localities manually. Localities were compared with existing range maps from online databases (BirdLife International, 2014; Ridgely et al., 2005), regional field guides (Hilty, 2003; Howell & Webb, 1995; Raffaele, Wiley, Garrido, Keith, & Raffaele, 2003; Ridgely & Greenfield, 2001; Schulenberg, Stotz, Lane, O'Neill, & Parker, 2007) and species accounts in the *Handbook of the Birds of the World* (Schuchmann, 1999) to identify and remove erroneous localities and genuine vagrants (i.e., singleton records far from established populations). Migratory species with discrete breeding and wintering ranges or with breeding ranges extending broadly out of the tropics were reduced to their winter distributions (eBird, 2012; Howell & Webb, 1995; Schuchmann, 1999), which we considered more representative of their distributions at our tropical testing localities (Guisan et al., 2006; Martínez-Meyer, Peterson, & Navarro-Sigüenza, 2004; Nakazawa, Peterson, Martínez-Meyer, & Navarro-Sigüenza, 2004). Non-migratory tropical species with a history of long-distance dispersal had areas of known dispersal included inside their training regions, but observations representing these motile individuals were excluded from the model dataset. Questionable locality records were researched to determine whether they were truly erroneous locality records or valid (i.e., substantiated, or experienced and trustworthy observer) reports (eBird, 2012; M. J. Anderson, personal communication 2014; M. Brady, personal communication 2014; M. G. Harvey, personal communication 2014; P. Hosner, personal communication 2014; J. D. Wolfe, personal communication 2014; see aforementioned literature).

To provide independent testing data, 13 localities regarded as well sampled (therefore providing information about true presences and absences) were selected based on data availability and the amount of effort (i.e., complete eBird checklists as of April 2015) associated with each (eBird, 2012). Localities possessing high species richness and extensive effort were preferred, with effort deemed a more important metric for validating estimations. These localities were selected non-randomly between 25° N and 25° S, with one locality per 5° band per continent and two additional equatorial (i.e., within $^{1}/_{2}$° of the Equator) localities (Figure 1). These localities were restricted to the tropics to focus on areas where the most species of hummingbirds occur and to minimize geographical distortion from latitude on our model projections. Species lists for each locality were drawn from the eBird data and existing checklists when available (see Arvin, 2001; Copalinga Lodge, 2013; Costa Rica Gateway, 2015; Lees et al., 2013; Serra dos Tucanos Birding Tours, 2015; Tropical Birding, 2005; Wilderness Explorers, 2008). eBird occurrence data from an expanded *c*. 20 km cell around localities were aggregated to represent a locality's observational data, with more recent (i.e., post-2013 download) data gleaned directly from the eBird website (eBird, 2012; Vavrek, 2011). Two localities possess published checklists that may exceed the 20 km definition: Municipio Gómez Farías, Tamaulipas, México and Rio Cristalino (a.k.a. Alta Floresta), Mato Grosso, Brazil. Municipio Gómez Farías was considered acceptable for comparison because it occurs at a rather northern latitude with few microendemics and rather homogeneous bird communities, and Rio Cristalino was considered acceptable because it is in a climatically homogeneous lowland region.

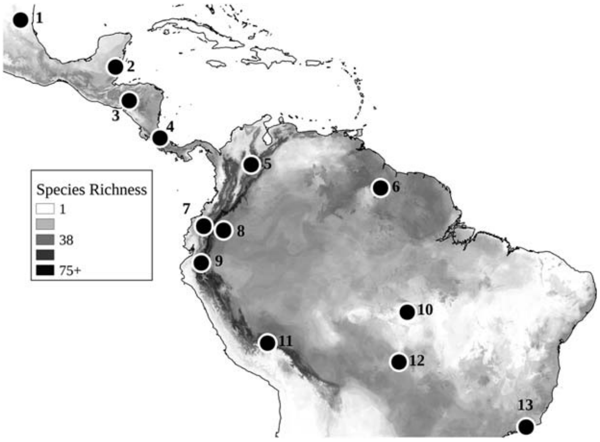

FIGURE 1 A map of estimated species richness within the study area, as derived from stacking **M**-restricted ecological niche models. Test localities' names and species lists can be found in Supporting Information Appendix S2. The map was created with QGIS 2.8.6 (Quantum GIS Development Team, 2017) using data from NaturalEarthData.com

All locality data within 20 km of test localities were removed prior to modelling, in order to create absences for testing the models. This action was performed to ensure that we had 'true' absences of each species at our study localities for model training and testing. Species for which there were fewer than five occurrences remaining after data cleaning and preparation were excluded from the modelling process (Pearson, Raxworthy, Nakamura, & Peterson, 2007). A full list of species modeled ($n = 293$) or excluded ($n = 49$) is in Supporting Information Appendix S1, with codes and data available via www.github.org (see Data Accessibility statement).

2.2 | Species modelling

Climate data were drawn from the BIOCLIM v. 1.4 dataset (covering 1960–1990) available through WorldClim at a spatial resolution of 2.5′ (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005). We used the layers 'as is' and refrained from transforming them into an equal area projection because all of our testing localities are in the tropics and the latitudinal distortion is $<$ 10%. Four 'bioclimatic' layers were used to calibrate models: mean temperature of the warmest quarter, mean temperature of the coldest quarter, precipitation of the wettest quarter, and precipitation of the driest quarter. These four layers were chosen because they represent the climatic extremes that often constrain species distributions and because most other bioclimatic layers are derived from different combinations of or are tightly correlated with these variables (Root, 1988). Initial tests of ENMs using only these four variables created distribution estimates that matched known distributions (Howell & Webb, 1995; Ridgely et al., 2005; Schuchmann, 1999; Schulenberg et al., 2007).

Two training areas were used for each species to obtain alternative predictions. In the first approach, the entire area inhabited by the Trochilidae was used to calibrate and project models. This area was defined as the area over which hummingbirds are known to have occurred within recent geological history, including all of continental North and South America, the Caribbean, the Juan Fernández Islands and the Aleutian and Bering Sea Islands. Island archipelagos that do not possess hummingbirds (i.e., the Canadian Arctic Archipelago, the Falkland Islands, etc.) were excluded.

In the second approach, locality data were overlaid on terrain base maps in ArcGIS 10.2 (Environmental Systems Research Institute, 2011) together with a world ecoregions layer (World Wildlife Fund, 2011). These were used to identify breaks in habitat and ecological regions in topographically homogeneous areas. For each species, we hypothesized an accessible area (**M**) by using topography, ecoregions and known occurrences (Soberón & Peterson, 2005). We performed initial tests of different calibration areas on *Phaethornis striigularis* to explore the effects of training area extent. These tests indicated that overly broad training areas produce errors that vary in effect from overfitting presence suitability to drastically increasing commission error in inaccessible regions. Conversely, extremely narrow training areas can overpredict or wrongly predict distributions as a result of having identical environments for drawing presence and absence localities. Restricting calibration areas to regions bounded by significant abiotic barriers (e.g., large rivers, mountain ranges) and known or hypothesized dispersal distances yielded more accurate models and reduced these errors (Barve et al., 2011; Owens et al., 2013; Royle, Chandler, Yackulic, & Nichols, 2012; Saupe et al., 2012). Thus, in our study, **M**s were constrained by deep valleys (e.g., the Marañon Valley), the crests of mountains (e.g., the Andes) and other distinct features likely to act as barriers to species distributions (e.g., the llanos of northern South America). We did not create multiple **M**s for each species to test the delimitation of **M** that performed best; rather, with the idea of using hard biogeographical barriers whenever possible, we created an **M** for each species with criteria as consistent as possible (e.g., multiple species bounded by a river/mountain range were assumed to have identical or near-identical limits at said barrier). Although this sometimes resulted in less accurate range maps, it did allow us to 'batch process' all species and look only at the effects of restricting models to accessible areas. In uninterrupted lowland regions, a *c.* 200–300 km buffer was used to approximate dispersal from regions where a given species is known. The unique **M** regions for each species were then used as the calibration and projection regions for modelling.

For both approaches, data layers were clipped to the same extent as calibration regions in R 3.0.0 (R Core Team, 2015) using the packages ENMGadgets (Barve & Barve, 2014), maptools (Bivand & Lewin-Koh, 2015), raster (Hijmans, 2015), rgdal (Bivand, Keitt, & Rowlingson, 2015), shapefiles (Stabler, 2013) and sp (Pebesma & Bivand, 2005). Models were created using MAXENT 3.3 (Phillips, Dudík, & Shapire, 2004) within R 3.0.0 (R Core Team, 2015) using the packages dismo (Hijmans, Phillips, Leathwich, & Elith, 2015) and rJava (Urbanek, 2013). Five replicates of each model were conducted with no clamping or extrapolation and with all the default 'features' used; the average raw output of these models was saved. All models used the default number of background points ($n = 10{,}000$), and the five replicates were completely random. We limited the amount to which we manipulated MAXENT settings to ensure that we were truly looking at the effects of restricting training areas to accessible areas. We used all points to train the models because we created artificial absences (i.e., removed all observations from our test localities) to test model performance. Outputs were thresholded using fixed sensitivities of 99, 95 and 90% to create binary outputs. There are arguments for maximizing the sum of sensitivity and specificity in thresholding such predictions, but this assumes equal weight for false positives (i.e., false presences) and false negatives (i.e., false absences), which is debatable (Peterson et al., 2011). Instead, fixed presence thresholds were chosen to reflect confidence in data quality and comparatively higher certainty of presence over absence localities (Liu et al., 2013; Peterson et al., 2011). Then we performed extractions for 99, 95 and 90% thresholds to compare their performance in predicting richness and validate our choice of threshold based on data quality. Thresholding is a contentious procedure, but unavoidable given our objective of obtaining presence–absence matrices and lists. Moreover, as the range of values of the outputs of MAXENT changes with the size of the training region, thresholding provides a way of standardizing such outputs (Barve et al., 2011; Owens et al., 2013; VanDerWal, Shoo, Graham, & Williams, 2009).

SDMs were successfully created using both approaches for 293 species of hummingbird, representing 85.6% of all described species (Gill & Donsker, 2013; Supporting Information Appendix S1) and resulting in two comparable presence–absence matrices. Within these matrices, all species predicted as present in $\geq$ 25% of grid cells in the 20 km expanded test locality were predicted as present at that locality in the final species list. These presence–absence matrices, their derived richness maps and their derived species lists were directly contrasted to determine how altering only the geographical training area affects the accuracy of list estimation. The full list of species predicted and observed for each locality can be found in Supporting Information Appendix S2.

2.3 | Contrasting predictions methods

Predictions of the entire list of species (i.e., community composition) were compared using two statistics derived from a confusion matrix as defined by D'Amen, Dubuis et al. (2015). Both these metrics draw from the number of true positives (*tp*; species predicted as present and has been observed), the number of true negatives (*tn*; species predicted as absent and has not been observed), the number of false negatives (*fn*; species predicted as absent but has been observed), the number of false positives (*fp*; species predicted as present but has not been observed), and the total species pool (*sp*). These metrics are the prediction success (p) and the Sørensen index (s_i), defined as follows:

$$p = \frac{tp + tn}{sp}$$

$$s_i = \frac{2tp}{2tp + fn + fp}$$

The overall success of generated lists to estimate gross richness was evaluated using a completeness index (C), derived from the number of observed species (o) divided by the number of expected species

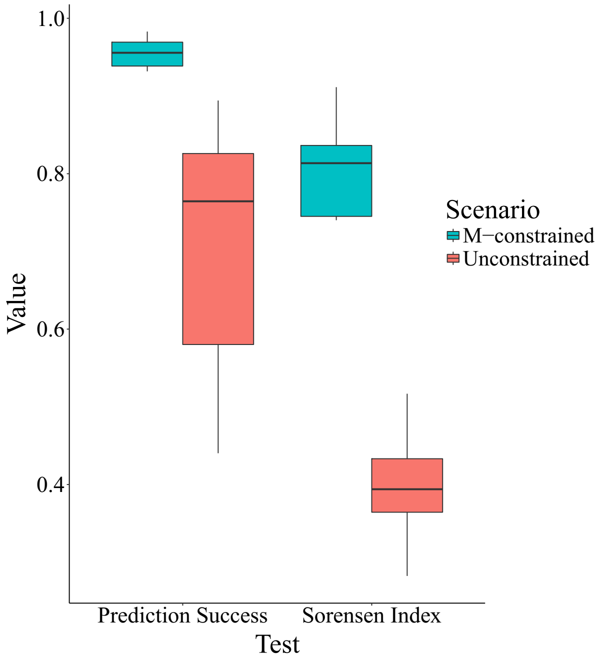

FIGURE 2 Comparisons of richness (prediction success) and community composition (Sørensen index) estimations between dispersal (**M**)-constrained and unconstrained training regions. The dark bar within the box represents the median of the data, while the edges of the box are the upper and lower quartile limits, and the 'whiskers' denote points within 1.5 times the interquartile range of the upper and lower quartile limits. Outliers from the main data distribution are denoted with points. The plot was created with R 3.4.0 (R Core Team, 2017) and 'ggplot2' (Wickham, 2009)

(*e*) from the models (Hawkins, Norris, Hogue, & Feminella, 2000). An ideal prediction would have a completeness index of one.

3 | RESULTS

Comparing species presence lists derived from 2013 and 2015 data revealed that twenty-two 'false predictions' have been validated as 'true predictions' (Supporting Information Appendix S2). Given the amount of data added, only the 2015 data were used to assess model effectiveness. Values presented here include their 95% confidence intervals, appended with '±'.

Our initial comparisons of threshold performance via the completeness index of **M**-derived models found that 90% thresholds underpredict species richness ($C = 1.38$), whereas 99% thresholds overpredict species richness ($C = .73$). The best average completeness index was recovered for the 95% threshold ($C = .90$); thus, all results discussed hereafter refer to the 95% threshold.

3.1 | Comparing model performance

The prediction success of models that were and were not constrained by accessible area hypotheses was significantly different (Wilcoxon rank sum test, $W = 1$, $p < .001$), with the prediction success of constrained models ($p = .95 \pm .01$) being greater than that of unconstrained (i.e., with no dispersal limitations) models ($p = .71 \pm .08$; Figure 2). Sørensen indices indicated that unconstrained models differed significantly from their constrained counterparts (Wilcoxon rank sum test, $W = 2$, $p < .001$), with **M**-constrained models providing better estimates of community composition ($s_i = .77 \pm .06$) than unconstrained models ($s_i = .40 \pm .05$; Figure 2).

Regressions of the observed versus predicted number of species (i.e., the completeness index, *C*) further illustrated the overpredictions present in unconstrained models. Unconstrained models yield a slope of $.27 \pm .03$ (adjusted $R^2 = .96$), indicating consistent overprediction of species richness when these ENMs are used. Conversely, **M**-constrained models predicted species richness with a slope of $.77 \pm .13$ (adjusted $R^2 = .93$; the inverse of *C* is shown in Figure 3), much closer to the value that would characterize a perfect prediction. Both regressions had high adjusted R^2 values, and both reflect a clear relationship between the number of species known at a locality and the number predicted. Despite this, **M**-constrained models were far more accurate at predicting actual species richness than unconstrained models.

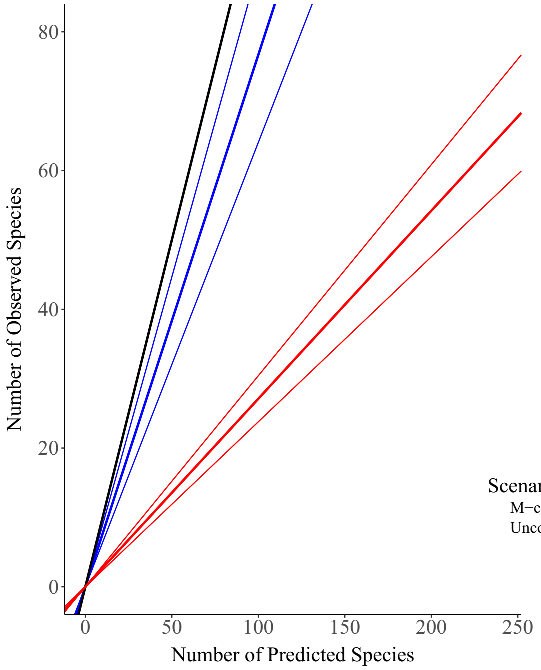

FIGURE 3 Regressions of the inverse completeness index (i.e., the number of expected species versus the number of observed species) with and without **M** constraints. Slopes are presented with their 95% confidence intervals. The black line indicates a 'perfect' prediction (i.e., a line with a slope of one); points above this line are underpredictions, and points below this line are overpredictions. Non-inverse equations for these lines are as follows: **M**-constrained, $y = (1.22 \pm .20)x$, adjusted $R^2 = .93$; unconstrained, $y = (3.55 \pm .44)x$, adjusted $R^2 = .96$. The plot was created with R 3.4.0 (R Core Team, 2017) and 'ggplot2' (Wickham, 2009)

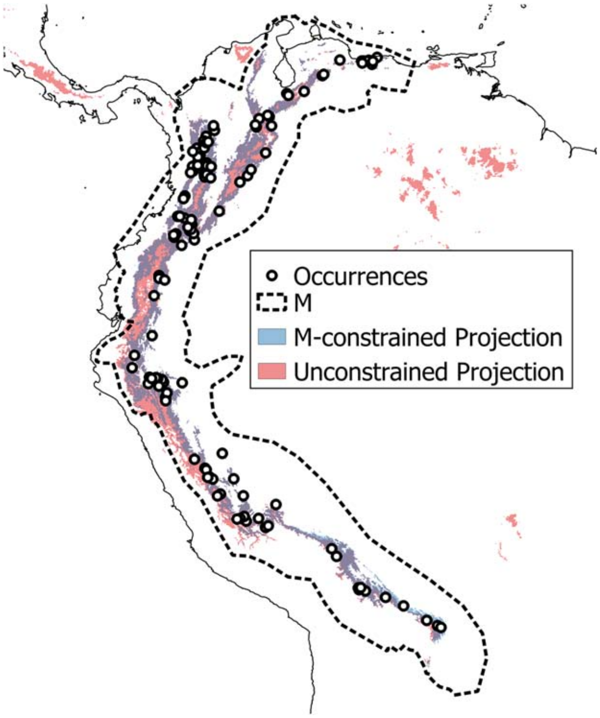

FIGURE 4 A comparison between unconstrained (i.e., modelled throughout North and South America) and **M** (i.e., accessible area, indicated by dotted line)-constrained species distribution models for *Coeligena coeligena*. Red areas indicate areas in which the unconstrained model predicted species presence, blue areas those in which the **M**-constrained model predicted species presence, and purple areas the area in which the models overlap. Points indicate presences used to create both models. The map was created with QGIS 2.8.6 (Quantum GIS Development Team, 2017) using data from NaturalEarthData.com

4 | DISCUSSION

This study is a case of a 'predict first, assemble later' strategy to estimate the composition of communities (D'Amen, Rahbek, et al., 2015), but it is the first large-scale case where stacking was constrained by hypotheses about species dispersal and accessible regions in geographical space. As such, it constitutes a novel and practical approach to stacking as a method to model the geographical pattern of species composition (Peterson et al., 2011; Soberón, 2010; Soberón & Peterson, 2005). Lack of detailed data on dispersal has been cited as a major factor hindering a more widespread use of process-oriented SDMs (Hortal, de Marco, Santos, & Diniz-Filho, 2012), but our method is less demanding because there is no need to calculate dispersal kernels. Nevertheless, stacked SDMs constrained a priori in their geographical extent perform better in anticipating geographic patterns of observed species richness than models that lack dispersal-based constraints (Barve et al., 2011; Guisan et al., 2006; Soberón, 2010; Soberón & Peterson, 2005). Restricting the calibration area to a species' **M** creates more accurate predictions of the species' true occupied geographical distribution (i.e., the occupied area), increasing the predictive power of SDMs in environmental space and eliminating overlap of allopatric species complexes (Figure 4; Barve et al., 2011; Owens et al., 2013; Soberón & Peterson, 2005). Furthermore, these models reduce bias in ENMs and their derived SDMs by removing regions in which estimations would be created via extrapolations from known environmental tolerances (Owens et al., 2013; Saupe et al., 2017) and offer improved confidence in individual model performance.

These factors are further reinforced by the consistent overpredictions present in unconstrained models in many montane systems, where similar regions with similar species assemblages are often separated by large canyons or lowlands. Enforcing dispersal limitations results in a better reflection of the occupied niche not only by restricting abiotic variables, but also by implicitly accounting for some biotic interactions by limiting the environments in which species can co-occur. In instances where there is interest in projecting models, our method provides a way to assess a species' occupied niche accurately in the present, in order to estimate distributions in other geographical and temporal scenarios better. The performance of these models rejects the notion that an unconstrained biogeographical model of the invasible niche is appropriate for creating a single training area for multiple evolutionary units and highlights the importance of this often-neglected part of the modelling process.

Likewise, these restrictions have resulted in more accurate predictions of community composition at test localities. Predicted distributions are more constrained and narrower within the species invasible area when **M**s are used as training areas (Figure 4). Unconstrained models that incorporated inaccessible regions within the training region possessed a large amount of error and had a lower predictions success than **M**-constrained models. Much of this error was commission error derived from species that are ecologically similar to test localities but from different biogeographical realms.

M-derived presence–absence matrices still possess more commission than omission error (i.e., generally result in more false presences than false absences when predicting community compositions), but have much lower rates of commission error than unconstrained models. Commission error is usually considered preferable for community composition estimations, as it is less likely that species of conservation concern will be excluded from the derived species list. Despite this, both commission and omission can be harmful to conservation efforts; further reducing commission error in the future is a necessity to ensure the proper allocation of conservation funds and for the creation of effective reserves (Loiselle et al., 2003). In all regions for which predictions of community composition are made, a certain amount of survey work is also required to confirm the model's accuracy and identify possible errors (Pineda & Lobo, 2009).

Although our **M** method appears to be promising, much work remains to be done in refining and identifying objective definitions of **M**, perhaps on the basis of geographical features. This work is in progress. Our evaluation methods did not explicitly test for differences between restricting the training area a priori or post-processing by clipping the area of the model projection. The effects of restricting the training area probably depend on the SDM algorithm used, something that we will explore in detail in the future. However, we suspect that **M**-trained models result in more spatially restrictive SDMs than models trained in an unrestricted area. Unfortunately, for many species, the density of testing localities necessary for explicitly analysing the effects of training areas within a species' individual SDM may be insufficient; this should be tested in the future with well-sampled species for which high-resolution data are available.

4.1 | Effects of effort

In this study, three localities were underpredicted in terms of species richness when **M** training areas were used: Municipio Gómez Farías, Tamaulipas, México; Rancho Naturalista, Cartago, Costa Rica; and Río Cristalino, Mato Grosso, Brazil, which falls within the well-surveyed Alta Floresta region (Lees et al., 2013; Zimmer, Parker, Isler, & Isler, 1997). All of these localities have seen an enormous amount of effort ($>$ 140 complete eBird checklists per centroid) and possess official published checklists (Arvin, 2001; Costa Rica Gateway, 2015; Lees et al., 2013). False negatives at these locations are likely to be related to three factors: occasional migrants and vagrants that occur in adjacent regions but that are not resident at the test locality (e.g., *Amazilia rutila*, Rancho Naturalista), highly localized species that were predicted at a locality but fell below the 25% threshold (e.g., *Atthis heliosa*, Municipio Gómez Farías), and genuine prediction errors by the models (e.g., *Topaza pella*, Rio Cristalino).

Conversely, many overpredictions appear to be attributable to insufficient survey effort. Many validated predictions were in regions of topographic complexity, suggesting that survey effort is concentrated in accessible regions, rather than covering the entire site (Ponder et al., 2001). At Reserva Natural de las Aves (RNA) Reinita Cielo Azul, Santander, Colombia, six predictions were confirmed (Supporting Information Appendix S2); many of the remaining false-positive predictions refer to montane species perhaps yet to be found in rugged areas away from the reserve's access points (e.g., *Ensifera ensifera*). Future research on these metrics is needed to examine how much overprediction is related to sampling effort.

4.2 | Improving upon M SDMs

Many factors that limit species distributions in the Neotropics (e.g., elevational effects on metabolism, biotic interactions and specificity for different habitat types) were not accounted for in these basic ecological models (Altshuler & Dudley, 2002; Buermann et al., 2011). Despite acceptable overall model performance using only four BIOCLIM variables, several species that are limited by other factors had poorly performing ENMs. One glaring example is *Leucippus chlorocercus*, restricted to riparian scrub along large rivers in western Amazonia (Ridgely & Greenfield, 2001; Schuchmann, 1999; Schulenberg et al., 2007): this species' SDM does not conform to rivers within the region and omits the test locality from which the species is known to occur (Sacha Lodge, Sucumbíos, Ecuador).

Unique biotic relationships can be used to refine SDMs further. Hummingbirds are well known to be associated with certain species of flowers (e.g., *Heliconia*), and many possess unique bill morphologies reflecting these close relationships (Schuchmann, 1999; Stiles, 1975). Furthermore, interactions between species may also be restricting distributions, as many erroneous predictions overlapped with related allo- or parapatric taxa (e.g., *Coeligena coeligena* versus *Coeligena wilsoni* in the western Andes). More research into how to mitigate these contingencies is necessary to create better models of species distributions, especially in areas of parapatry.

Initial studies have shown that using remote-sensing data, such as the normalized difference vegetation index (NDVI), can improve stacked SDM models originally based solely on climatic data layers (Cord, Klein, Gernandt, Pérez de la Rosa, & Dech, 2014; Jiang et al., 2014). These data show great potential for species with poor dispersal abilities (e.g., *Pinus*), but are extremely difficult to correlate spatio-temporally with motile animal occurrences. Even when this is accomplished, restricting these additional layers to the regions that the species can explore is still imperative, and using such data without caution may result in egregious errors. Others have mathematically taken into account spatial heterogeneity in species occupancy and dispersal (De Marco, Diniz-Filho, & Bini, 2008); using these approaches in conjunction with an **M** adhering to hard biogeographical barriers for a species could greatly improve SDM predictions.

We therefore conclude that the efficacy of ecological niche models can be improved by incorporating dispersal limitations into models, a priori. These limitations should be based on known or estimated dispersal abilities combined with biogeographical barriers to create realistic training areas. In our case study, these methods improved distribution prediction, richness estimation and community assemblage estimation. Future work using ENMs and subsequently derived SDMs should incorporate known limitations to species accessible area in geographical space to reflect patterns found in nature better and thereby improve models' predictive performance.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation grant #1208472, the University of Kansas Biodiversity Institute and the University of Kansas Ecology and Evolutionary Biology department. Primary analyses and writing were completed by J.C.C. as part of his master's thesis under the guidance of J.S. We thank V. B. Arroyo Peña, Dr M. Ashcroft, Dr N. Barve, Dr V. Barve, Dr J. Elith, Dr J. D. Manthey, Dr N. Morueta-Holme, Dr A. T. Peterson, M. B. Robbins, Dr C. M. Witt and multiple editors and anonymous referees for comments and feedback about this research.

DATA ACCESSIBILITY

Codes and data for this project have been made available via J.C.C.'s GitHub page (https://github.com/jacobccooper/trochilidae).

ORCID

Jacob C. Cooper http://orcid.org/0000-0003-2182-3236
Jorge Soberón http://orcid.org/0000-0003-2160-4148

REFERENCES

Altshuler, D. L., & Dudley, R. (2002). The ecological and evolutionary interface of hummingbird flight physiology. *Journal of Experimental Biology*, *205*, 2325–2336.

Aranda, D. L., & Lobo, J. M. (2011). How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography*, *34*, 31–38.

Arvin, J. C. (2001). *An annotated checklist of the birds of the Gómez Farías region, southwestern Tamaulipas, Mexico*. Austin, Texas, USA: Texas Parks and Wildlife.

Barve, N., & Barve, V. (2014). *ENMGadgets: Tools for pre and post processing in ENM workflow* (R package version 0.0.5).

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., . . . Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*, 1810–1819.

BirdLife International. (2014). *Species factsheets: Trochilidae*. Retrieved from www.birdlife.org

Bivand, R., Keitt, T., & Rowlingson, B. (2015). *rgdal: Bindings for the geospatial data abstraction library* (R package version 0.9-3).

Bivand, R., & Lewin-Koh, N. (2015). *maptools: Tools for reading and handling spatial objects* (R package version 0.8–36).

Buermann, W., Chaves, J. A., Dudley, R., McGuire, J. A., Smith, T. B., & Altshuler, D. L. (2011). Projected changes in elevational distribution and flight performance of montane Neotropical hummingbirds in response to climate change. *Global Change Biology*, *17*, 1671–1680.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2013). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*, 99–112.

Chamberlain, S., Ram, K., Barve, V., & Mcglinn, D. (2015). *rgbif: Interface to the global biodiversity information facility API* (R package version 0.8.0).

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, *210*, 478–486.

Copalinga Lodge. (2013). *Birds seen or heard at the Copalinga restaurant & cabins*. Retrieved from www.copalinga.com

Cord, A. F., Klein, D., Gernandt, D. S., Pérez de la Rosa, J. A., & Dech, S. (2014). Remote sensing data can improve predictions of species richness by stacked species distribution models: A case study for Mexican pines. *Journal of Biogeography*, *41*, 736–748.

Costa Rica Gateway. (2015). *Rancho Naturalista checklist and nearby sites in the Tuis-Turrialba region of the Talamanca mountains, 1900–4500 ft*. Retrieved from www.ranchonaturalista.net

D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellisier, L., & Guisan, A. (2015). Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, *42*, 1255–1266.

D'Amen, M., Rahbek, C., Zimmermann, N., & Guisan, A. (2015). Spatial predictions at the community level: From current approaches to future frameworks. *Biological Reviews*, *92*, 169–187.

De Marco, P. Jr., Diniz-Filho, J. A. F., & Bini, L. M. (2008). Spatial analysis improves species distribution modelling during range expansion. *Biology Letters*, *4*, 577–580.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: A comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions*, *17*, 1122–1131.

eBird. (2012). *eBird: An online database of bird distribution and abundance [web application]*. Ithaca, NY. Retrieved from www.ebird.org

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., . . . Zimmerman, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57.

Environmental Systems Research Institute. (2011). *ArcGIS desktop: release 10.2*. Redlands, CA.

Feria, T. P., & Peterson, A. T. (2002). Prediction of bird community composition based on point-occurrence data and inferential algorithms: A valuable tool in biodiversity assessments. *Diversity and Distributions*, *8*, 49–56.

Franklin, J. (2010). *Mapping species distributions: Spatial inference and prediction*. Cambridge, U.K.: Cambridge University Press.

Gastón, A., & García-Viñas, J. I. (2013). Evaluating the predictive performance of stacked species distribution models applied to plant species selection in ecological restoration. *Ecological Modelling*, *263*, 103–108.

Gill, F., & Donsker, D. (Eds.). (2013). *IOC world bird list (v 3.5)*. Retrieved from www.worldbirdnames.org

Global Biodiversity Informatics Facility. (2014). *Aves: Trochilidae*. Retrieved from www.gbif.org

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

Guisan, A., & Rahbek, C. (2011). SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, *38*, 1433–1444.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M., Aspinall, R., & Hastie, T. (2006). Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, *43*, 386–392.

Hawkins, C. P., Norris, R. H., Hogue, J. N., & Feminella, J. W. (2000). Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications*, *10*, 1456–1477.

Hijmans, R. J. (2015). *raster: Geographic data analysis and modeling* (R package version 2.3–40).

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

Hijmans, R. J., Phillips, S. J., Leathwich, J., & Elith, J. (2015). *dismo: Species distribution modeling* (R package version 1.0–12).

Hilty, S. L. (2003). *Birds of Venezuela* (2nd Ed.). Princeton, NJ: Princeton University Press.

Hortal, J., de Marco, P. Jr., Santos, A. M. C., & Diniz-Filho, J. A. F. (2012). Integrating biogeographic processes and local community assembly. *Journal of Biogeography*, *39*, 627–628.

Howell, S. N. G., & Webb, S. (1995). *A guide to the birds of Mexico and northern Central America*. Oxford, U.K.: Oxford University Press.

Jiang, T., Wang, T., de Bie, C. A. J. M., Skidmore, A. K., Liu, X., Song, S., . . . Shao, X. (2014). Satellite-derived vegetation indices contribute significantly to the prediction of epiphyllous liverworts. *Ecological Indicators*, *38*, 72–80.

Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G. J., . . . O'Hara, R. B. (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, *39*, 2163–2178.

Lees, A. C., Zimmer, K. J., Marantz, C. A., Whittaker, A., Davis, B. J. W., & Whitney, B. M. (2013). Alta Floresta revisited: An updated review of the avifauna of the most intensively surveyed locality in south-central Amazonia. *Bulletin of the British Ornithologists' Club*, *133*, 178–239.

Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, *40*, 778–789.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G., & Williams, P. H. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, *17*, 1591–1600.

Martínez-Meyer, E., Peterson, A. T., & Navarro-Sigüenza, A. G. (2004). Evolution of seasonal ecological niches in the *Passerina* buntings (Aves: Cardinalidae). *Proceedings of the Royal Society B: Biological Sciences*, *271*, 1151–1157.

Nakazawa, Y., Peterson, A. T., Martínez-Meyer, E., & Navarro-Sigüenza, A. G. (2004). Seasonal niches of Nearctic-Neotropical migratory birds: Implications for the evolution of migration. *The Auk*, *121*, 610–618.

Owens, H. L., Campbell, L. P., Dornak, L. L., Saupe, E. E., Barve, N., Soberón, J., . . . Peterson, A. T. (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling*, *263*, 10–18.

Pearce, J., & Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, *43*, 405–412.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Peterson, A. T. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography*, *34*, 102–117.

Pebesma, E. J., & Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, *5*, 9–13.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*. Princeton, NJ: Princeton University Press.

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

Phillips, S. J., Dudík, M., & Shapire, R. E. (2004). A maximum entropy approach to species modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 655–662. Banff, Canada: Association for Computer Machinery Press.

Pineda, E., & Lobo, J. M. (2009). Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, *78*, 182–190.

Ponder, W. F., Carter, G. A., Flemons, P., & Chapman, R. R. (2001). Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, *15*, 648–657.

Quantum GIS Development Team. (2017). *Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project*. Retrieved from http://qgis.osgeo.org

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raffaele, H., Wiley, J., Garrido, O., Keith, A., & Raffaele, J. (2003). *Birds of the West Indies*. Princeton, NJ: Princeton University Press.

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*, 1719–1727.

Ridgely, R. S., & Greenfield, P. J. (2001). *The birds of Ecuador*. Ithaca, NY: Cornell University Press.

Ridgely, R. S., Allnutt, T. F., Brooks, T., McNicol, D. K., Young, B. E., & Zook, J. R. (2005). *Digital distribution maps of the birds of the Western Hemisphere* (version 2.1). Arlington, VA: NatureServe.

Root, T. (1988). Environmental factors associated with avian distributional boundaries. *Journal of Biogeography*, *15*, 489–505.

Royle, J. A., Chandler, R. B., Yackulic, C., & Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, *3*, 545–554.

Saupe, E. E., Barve, N., Owens, H. L., Cooper, J. C., Hosner, P. A., & Peterson, A. T. (2017). Reconstructing ecological niche evolution when niches are incompletely characterized. *Systematic Biology*, *accepted*. Retrieved from https://doi.org/10.1093/sysbio/syx084

Saupe, E. E., Barve, V., Myers, C. E., Soberón, J., Barve, N., Hensz, C. M., . . . Lira-Noriega, A. (2012). Variation in niche and distribution model performance: The need for *a priori* assessment of key causal factors. *Ecological Modelling*, *237–238*, 11–22.

Schuchmann, K. L. (1999). Family Trochilidae (Hummingbirds). In J. del Hoyo, A. Elliott, & J. Sargatal (Eds.), *Handbook of the birds of the world* (pp. 468–682). Barcelona, Spain: Lynx Ediciones.

Schulenberg, T. S., Stotz, D. F., Lane, D. F., O'Neill, J. P., & Parker, T. A. III. (2007). *Birds of Peru*. Princeton, NJ: Princeton University Press.

Serra dos Tucanos Birding Tours. (2015). *SerradosTucanos_lodgegrounds.pdf*. Retrieved from www.serradostucanos.com.br

Soberón, J. (2010). Niche and area of distribution modeling: A population ecology perspective. *Ecography*, *33*, 159–167.

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

Stabler, B. (2013). *shapefiles: Read and write ESRI shapefiles* (R package version 0.7).

Stiles, F. G. (1975). Ecology, flowering phenology, and hummingbird pollination of some Costa Rican *Heliconia* species. *Ecology*, *56*, 285–301.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*, 2282–2292.

Tropical Birding. (2005). *Tandayapa bird lodge annotated checklist*. Retrieved from www.tandayapabirdlodge.com

Urbanek, S. (2013). *rJava: Low-level R to Java interface* (R package version 0.9-6).

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, *220*, 589–594.

Vavrek, M. J. (2011). fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, *14*, 14.1.1T.

Whittaker, R. J., Araújo, M. B., Jepson, P., Ladle, R. J., Watson, J. E. M., & Willis, K. J. (2005). Conservation biogeography: Assessment and prospect. *Diversity and Distributions*, *11*, 3–23.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Wilderness Explorers. (2008). *Atta rainforest lodge & Iwokrama canopy walkway*. Retrieved from www.wilderness-explorers.com

World Wildlife Fund. (2011). *Terrestrial ecoregions*. Retrieved from www.worldwildlife.org

Zimmer, K. J., Parker, T. A. III, Isler, M. L., & Isler, P. R. (1997). Survey of a southern Amazonian avifauna: The Alta Floresta region, Mato Grosso, Brazil. *Ornithological Monographs*, *48*, 887–918.

BIOSKETCHES

Jacob C. Cooper is a graduate student at the University of Chicago and a research affiliate of The Field Museum, interested in biogeography and conservation.

Jorge Soberón is a professor at the University of Kansas, interested in how species interact with each other and their environments and how these interactions can predict species distribution and abundance.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Cooper JC, Soberón J. Creating individual accessible area hypotheses improves stacked species distribution model performance. *Global Ecol Biogeogr*. 2018;27:156–165. https://doi.org/10.1111/geb.12678