ISSN 0305-0270

Journal of Biogeography

Contents

Volume 46, Number 6, June 2019

http://wileyonlinelibrary.com/journal/jbi

WILEY

This journal is available online at Wiley Online Library. Visit wileyonlinelibrary.com to search the articles and register for table of contents e-mail alerts.

JSTOR

Received: 10 November 2016 | Revised: 26 January 2019 | Accepted: 16 February 2019

DOI: 10.1111/jbi.13573

METHODS AND TOOLS

WILEY Journal of Biogeography

A new null model approach to quantify performance and significance for ecological niche models of species distributions

Corentin L. Bohl[1] | Jamie M. Kass[1,2] | Robert P. Anderson[1,2,3]

[1]Program in Biology, Graduate Center, City University of New York, New York, New York

[2]Department of Biology, City College of New York, City University of New York, New York, New York

[3]Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, New York

Correspondence
Corentin L. Bohl, Hadeland Videregående Skole, Gran, Norway.
Email: corentin.bohl@oppland.org

Funding information
National Science Foundation, Grant/Award Number: DEB-1119915

Editor: Daniel Chapman

Abstract

Aim: Ecological niche modelling requires robust estimation of model performance and significance, but common evaluation approaches often yield biased estimates. Null models provide a solution but are rarely used in this field. We implemented an important modification to existing null model tests, evaluating null models with the same withheld records that were used to evaluate the real model. We built and evaluated models across a range of modelling scenarios and for various performance measures using the algorithm Maxent and the monk parakeet (*Myiopsitta monachus*).

Location: Native range in Southern America and global invasions predominantly in North/Central America and Europe.

Methods: We tested the ability of models built under 15 scenarios (five sets of calibration records and three settings that varied the level of model complexity) to predict spatially independent evaluation data in the invaded range (in effect, testing the models under spatial transfer). We quantified performance with measures of discriminatory ability and overfitting based on area under the receiver operating characteristic curve (AUC) and the omission error rate. We estimated null distributions of these measures and calculated effect size and significance. We determined how these estimates varied across modelling scenarios, comparing with two tests existing in the literature.

Results: Performance varied starkly across modelling scenarios. As expected, the measures of overfitting agreed with each other and provided different information than that of discriminatory ability. However, high performance per se did not show strong association with high effect size and significance.

Main Conclusions: Ecological niche models should be assessed with measures of effect size and significance based on appropriate null distributions, in contrast to several approaches existing in the literature. The proposed approach using independent evaluation data, implemented with our accompanying code and R package, allows such estimates for either the same or a different region/time period, and it merits use and continued development.

KEYWORDS

AUC, discriminatory ability, effect size, Maxent, *Myiopsitta monachus*, null models, omission error rate, overfitting, significance, species distribution modelling

1 | INTRODUCTION

Models of species niches and distributions (niche models/modelling henceforth) are used commonly in biogeography, creating the need for unbiased and easily interpretable ways to assess their quality (Peterson et al., 2011, pp 150–181). Insufficient attention has focused on determining whether measures of performance for such models are statistically better than expected by chance (Beale, Lennon, & Gimona, 2008). Moreover, some common evaluation measures frequently yield inflated estimates of both performance and significance (Beale et al., 2008; Veloz, 2009). The fundamental challenge in arriving at unbiased estimates involves the formulation of appropriate null hypotheses. This requires realistically accounting for physical, historical, and biological constraints of the study system, so that the alternate hypothesis effectively isolates the ecological process of interest (Gotelli & Ulrich, 2012). The approaches available span a spectrum from classical hypothesis testing based on predefined theoretical distributions, to mechanistically derived ones that require explicit specification of particular ecological processes (Gotelli & Ulrich, 2012). The latter approach is often impractical because of the amount of data required to parameterize relevant ecological processes. Conversely, the former restricts the validity of the test to the assumptions of the theoretical distributions. Unfortunately, some of the most common measures of performance for niche models have underlying assumptions that are unreasonable in many situations (including area under the curve and omission rate; Appendix S1 in Supporting Information). For example, violation of a test's assumptions can be caused by sampling bias, spatial autocorrelation in the distribution of the species and/or the environmental variables, and unequal proportions of the various environmental conditions available (Beale et al., 2008; Dormann et al., 2007; Peterson et al., 2011, pp 176–181; Raes & ter Steege, 2007; Veloz, 2009).

1.1 | Null models

Null models offer a useful intermediate approach between theoretical and mechanistically derived statistical distributions. Typically, null model analysis is based on randomization of observed data to emulate a pattern expected by chance in the absence of a particular ecological process (Gotelli & McGill, 2006; Gotelli & Ulrich, 2012). Null models do not require estimating or even specifying all the ecological processes necessary to explain the data. Instead, they create a null distribution by holding some features of the data constant (related to the constraints of the system) while allowing others to vary stochastically—namely, those related to the ecological process of interest (Gotelli & McGill, 2006; Gotelli & Ulrich, 2012).

Null models have been applied to niche modelling to disentangle the association between the distribution of a species and spatial patterns of environmental features (Beale et al., 2008; Raes & ter Steege, 2007). In these analyses, randomization is applied to the pattern of presences (or presences and absences) that are used to calibrate the models, with all other aspects of the data and modelling algorithm held constant. Specifically, null distributions are obtained by calibrating models using randomly sampled pixels in the study area instead of real species records. Depending on the null hypothesis desired, more or less elaborate constraints can be imposed. For example, Raes and ter Steege (2007) opted for an unconstrained approach where null calibration pixels are sampled randomly across the study area. In contrast, Beale et al. (2008) imposed a sampling regime aimed at preserving the same degree of spatial autocorrelation for null calibration pixels as for the real species records. Additional constraints could be used to account for other aspects of the species' distribution (such as the dispersal capabilities of the species) or to impose a specific degree of sampling bias (in either geographic or environmental space).

Hence, null models offer a promising way to obtain appropriate null hypotheses in niche modelling. Importantly, they can be made subject to the same data-related and algorithmic constraints as the real model (e.g. same environmental dataset, modelling algorithm, and model settings/complexity). In this way, they account for the effects of many variables (including latent and confounding ones) without the need to quantify them explicitly. In particular, the spatial structure of environmental conditions (e.g. spatial autocorrelation and unequal proportions of various environmental conditions available) can be difficult to estimate and tend to produce inflated estimates of performance (Merckx, Steyaert, Vanreusel, Vincx, & Vanaverbeke, 2011; Raes & ter Steege, 2007). Subjecting null models to the same spatio-environmental constraints as the real model yields a null distribution that takes into account such heterogeneity, without explicitly quantifying it.

Nevertheless, currently available implementations correspond to null hypotheses that have limited relevance to most predictive tasks in niche modelling (Beale et al., 2008; Raes & ter Steege, 2007). This limitation originates in the way these null models are evaluated. Specifically, current approaches do not calculate model performance based on the same evaluation records that are used for the real model. Instead, null distributions are generated with the following steps:

1. A set of null pixels is selected by random sampling without replacement within the study region (in Beale et al. [2008], the sampling is constrained to preserve the same spatial autocorrelation as the real species records);
2. A null model is generated with these null pixels (in Raes and ter Steege [2007], all of them are used for model calibration; in Beale et al. [2008], a random subset of 70% of the null pixels is used for calibration with the remaining 30% reserved for evaluation);
3. Measures of model performance are calculated (in Raes and ter Steege [2007], this is done using the same null calibration pixels; the procedure in Beale et al. [2008] uses the 30% of the null pixels that were withheld for evaluation);
4. Steps (1–3) are replicated *i* times to create null distributions of the measures of performance.

These methods produce null distributions in which the performance values of each of the *i* null replicates quantify predictive ability on a set of "evaluation" pixels that originates from that null replicate's dataset. The performance of the real model, calibrated and evaluated with real species records, is then compared to such null distributions. Hence, these implementations involve comparisons of statistics calculated using different datasets. Additionally, the approach implemented by Raes and ter Steege (2007) does not use distinct evaluation data (but rather quantifies the model's ability to predict calibration data; i.e. model "verification" *sensu* Araújo & Guisan, 2006). In contrast, approaches that use the same evaluation data for both the real model and the null replicates would allow more direct comparisons and hence likely lead to more realistic inferences.

Therefore, we propose a simple modification that combines aspects of existing approaches but differs in the way null models are evaluated. We suggest evaluating the performance of all models (whether calibrated based on real or null data) with the same set of withheld real species records of the species (Figure 1). With this approach, the null distributions of the performance measures calculated on models calibrated using null pixels are directly comparable to those for the model calibrated with real species records. In doing so, we test a null hypothesis that differs from those of existing null model options (Appendix S2). Specifically, this approach corresponds to the null hypothesis that a model calibrated with a subset of real species records is no better at predicting the remaining (withheld) real records than a model calibrated with null pixels. In principle, any methodology could be used to select the set of real species records to be withheld for model evaluation (random or structured partitioning, whether single or *k*-fold). However, in this particular implementation we use spatially independent evaluation records, which test the model's predictive ability under spatial transfer and can provide a more realistic quantification of performance than randomly partitioned records under such circumstance (Hijmans, 2012; Radosavljevic & Anderson, 2013; Roberts et al., 2017; Veloz, 2009). Furthermore, as determining the magnitude of performance of real model evaluations in comparison to those of null models is also of high importance, in addition to measuring significance we also calculate effect size, which to date has received little emphasis in niche modelling studies.

In this worked example, we evaluate the proposed approach using the monk parakeet (*Myiopsitta monachus*; Boddaert, 1783) and the presence-background modelling technique MAXENT (Elith et al., 2011; Phillips, Anderson, Dudík, Schapire, & Blair, 2017; Phillips, Anderson, & Schapire, 2006; Phillips & Dudík, 2008). With a widespread distribution and multiple invasions worldwide, this species provides a study system facilitating scenarios that encompass a range of modelling conditions—including some that are more straightforward and others that are very challenging. Specifically, we

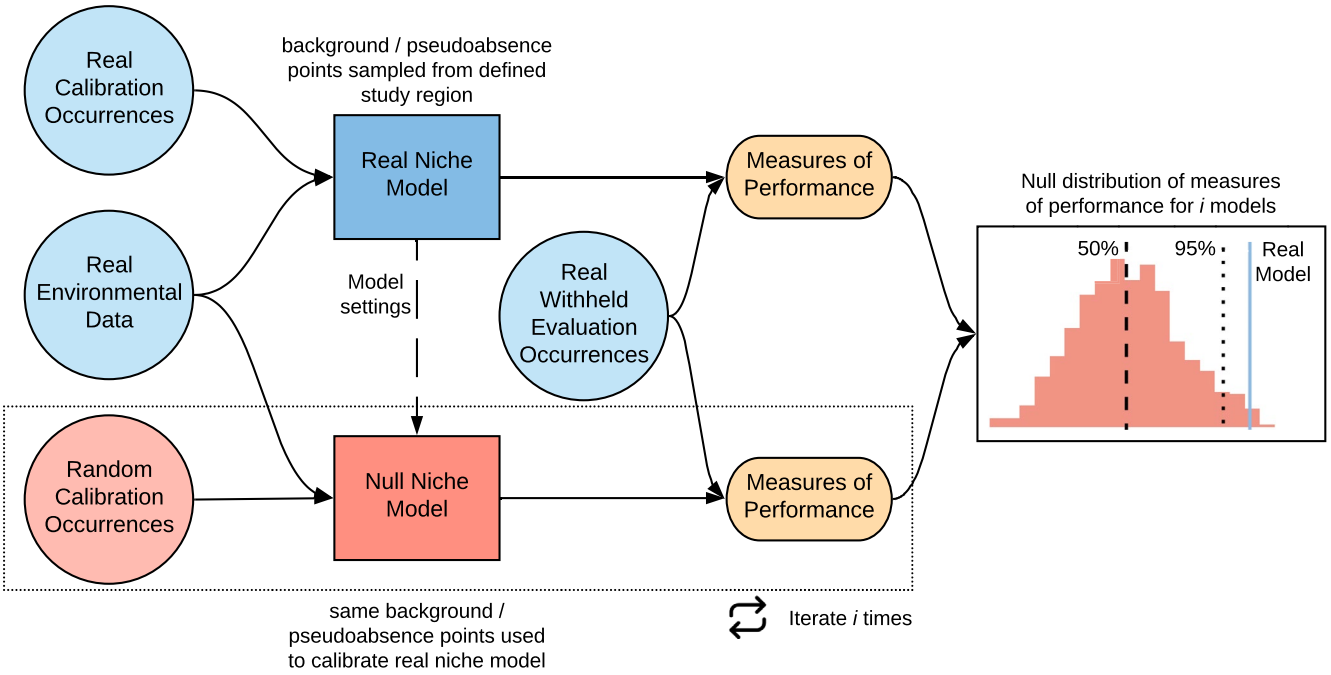

FIGURE 1 Flow chart of the proposed null model approach, including comparison with the niche/distribution model calibrated with real occurrence records of the species. Blue indicates real data and models, and red denotes random/null counterparts. First, a set of random occurrence pixels is sampled for calibrating null models within the same study region as the model that was made using the real occurrence records. Next, a niche/distribution model is constructed with these random calibration pixels and the same environmental predictor variables, background (or pseudo-absence) points and model settings as for the real species' model. The resulting null model is evaluated using the same withheld real occurrences that were used for the evaluation of the real model, and measures of model performance are calculated. This process is iterated *i* times, leading to a null distribution for each measure of performance. Finally, for each measure of performance, the value calculated for the real model is compared with the null distribution [Colour figure can be viewed at wileyonlinelibrary.com]

create models under different modelling scenarios: based on vastly different calibration datasets and varying levels of model complexity. As a test of spatial transfer, we then quantify the ability of these models to predict a single spatially independent dataset of globally distributed real records (based on several measures of performance) and assess significance and effect size with null models. We examine how the different performance measures vary across scenarios and how considerations of significance and effect size influence interpretations. We also compare the results of this null model approach to commonly used significance tests. In doing so, we aim to provide information regarding the behaviour of these evaluation metrics under our proposed null model approach.

2 | MATERIALS AND METHODS

2.1 | Design of analyses

We implement this null model approach across a variety of scenarios (encompassing different environmental and biological conditions as well as levels of model complexity). To do so, we build models with various geographic sets of real calibration records (as well as with null replicates for each set) and algorithmic settings that correspond to three levels of complexity. We then conduct all evaluations using a single withheld, spatially independent dataset of globally distributed records (see Species data and Figure 2). We create models for each of the calibration sets independently as well as for all sets pooled together (see Modelling methods). Here, we concentrate on the behaviour of the metrics, deferring interpretation related to invasion biology for later studies.

When modelling the distribution of an invasive species, various biological factors can affect prediction accuracy on spatially independent datasets, providing variation in the challenges facing any modelling algorithm under spatial transfer. Hence, the species' response to the predictor variables may not be stationary across space (Osborne, Foody, & Suárez-Seoane, 2007; Osborne & Suárez-Seoane, 2002). First, regional sets of populations may experience non-analogue conditions: distinct subsets of environmental space and/or biotic contexts that are different from those experienced in the species' native range (Fitzpatrick & Hargrove, 2009; Williams et al., 2013). Second, regions may differ in the degree of equilibrium the species has with the environment (e.g. depending on how long ago the species arrived). Third, populations may differ functionally, including the possibility of local adaptations (Fitzpatrick & Keller, 2015). Although critical for understanding biological invasions, the aim of this paper is not to delve into methodological issues affecting model transfer (Radosavljevic & Anderson, 2013) or assess niche shifts among populations (e.g. Broennimann et al., 2012; Guisan, Petitpierre, Broennimann,

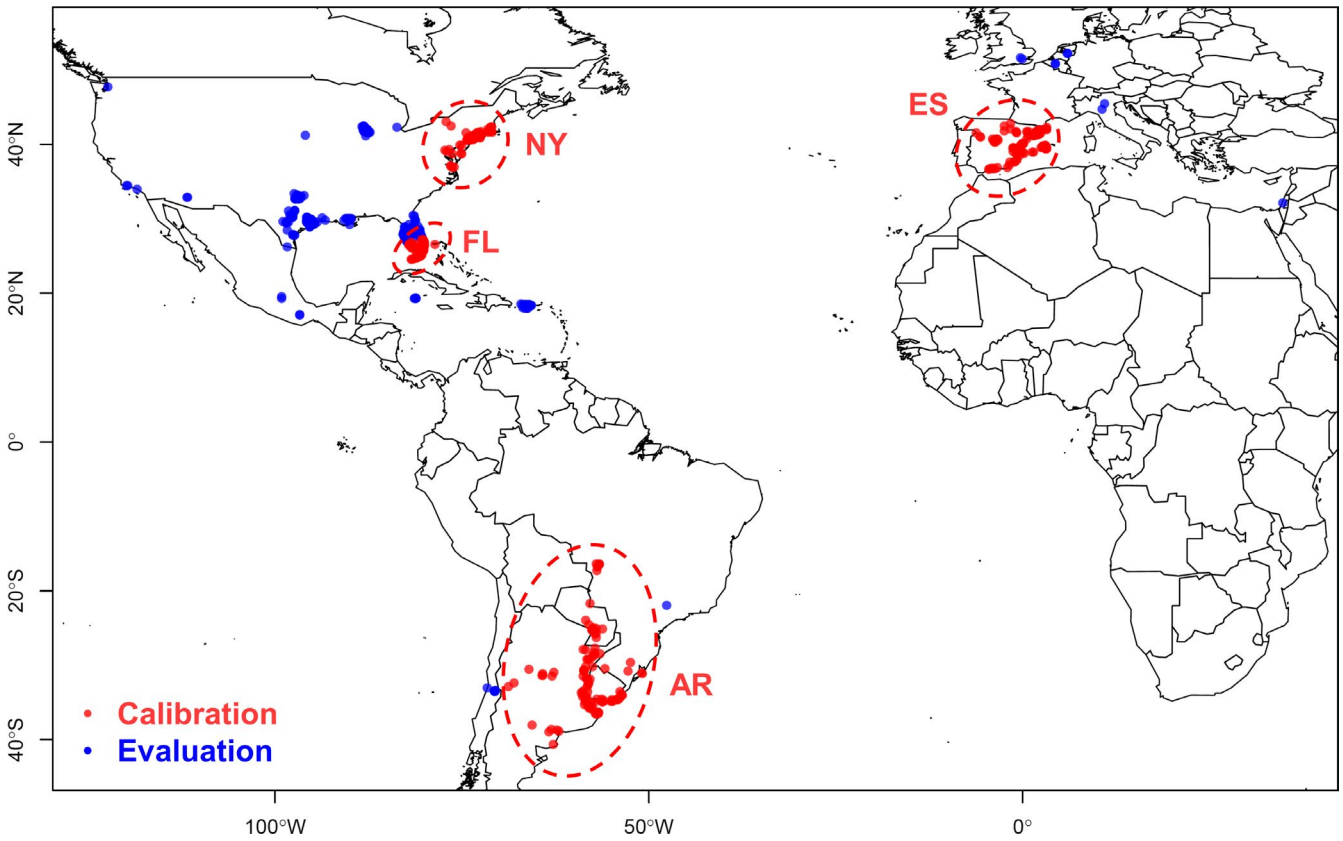

FIGURE 2 Map of monk parakeet presence records. The calibration records are indicated in red and the evaluation records in blue. The four respective regions used for model calibration are circled and labelled as follow: AR = native range (Argentina and surrounding countries), ES = Spain, NY = the wider New York metropolitan area, FL = southern Florida [Colour figure can be viewed at wileyonlinelibrary.com]

Daehler, & Kueffer, 2014). Rather, our intent is to explore the proposed null model approach across a wide variety of modelling scenarios. Hence, these issues that are likely to affect estimates of model quality under transfer increase the biological breadth of our explorations and the likely variability in observed model performance across scenarios. For example, we expect that models built on the native range or the combination of all considered regions to face fewer challenges than those built only using data from one of the invaded regions. As a result, the current scenarios allow exploration of the proposed approach across a wide range of modelling challenges, while still using real occurrence data.

2.2 | Species data

We gathered occurrence records of the behaviourally and morphologically distinctive monk parakeet from various sources (GBIF and Species Link databases, records from the primary literature, and personal observations; 2,997 total records; compiled in 2011). We examined these records in geographic and environmental space, and excluded from the subsequent analyses outliers unlikely to be part of well-established monk parakeet populations (see Appendix S3). We assigned the resulting records to different native and introduced regional sets of populations worldwide (Figure 2 and Appendix S3; 797 records). To provide varied datasets, we selected four of these regions as calibration sets to build models: the native population in Argentina and adjacent countries (AR, n = 121), and the three largest introduced populations (Spain, ES, n = 114; the wider New York City metropolitan area, NY, n = 96; and southern Florida, FL, n = 119). Note that for Florida we only included records from the southern part of the species' range in order to induce an extreme artificial violation of the equilibrium assumption. We also built models for the four calibration regions combined. The remaining records worldwide (347) were reserved as a single spatially independent evaluation set.

2.3 | Modelling methods

We built models using the algorithm Maxent (Elith et al., 2011; Phillips & Dudík, 2008; Phillips et al., 2006), which has been shown effective yet sensitive to differences in modelling settings (Shcheglovitova & Anderson, 2013; Syfert, Smith, & Coomes, 2013; Warren & Seifert, 2011). For all models, we considered a combination of 24 climatic, anthropogenic, and land cover predictor variables at 5 arc-minutes resolution (see Appendix S3). Except for the modifications described in the next paragraph, we used the default settings of Maxent 3.3.3k. The background data for each model corresponded to all the pixels within a 106 km buffer of the respective input calibration records (VanDerWal, Shoo, Graham, & Williams, 2009), rather than a random sample of pixels from this area. This distance is the maximum reported dispersal distance for monk parakeets (Gonçalves da Silva, Eberhard, Wright, Avery, & Russello, 2010) and thus constitutes a good estimate of the potentially accessible areas for this species (Anderson & Raza, 2010; Barve et al., 2011; Anderson, 2013; additionally, see nearest neighbour distances, Appendix S3).

We examined models made with three different levels of complexity by varying the types of feature classes (Linear, Quadratic, Product, Threshold and Hinge) and the value of the regularization multiplier (RM) used in Maxent. Feature classes determine the flexibility of the shape of the response that can be modelled. Higher RM values correspond to stronger penalties for model complexity (Elith et al., 2011; Merow, Smith, & Silander, 2013; Phillips & Dudík, 2008). At one end of the spectrum, we considered settings that should lead to very simple models, made using linear features only and a RM of 4.00. At the other extreme, we used all the feature classes and a RM of 0.25, which should produce highly complex models. Finally, as an intermediate level of complexity, we also made models using the default settings (which, for the sample sizes in the regional occurrence datasets, uses all the feature classes and a RM of 1.00).

2.4 | Model evaluation

We assessed the ability of each model to predict the records in the single evaluation dataset, quantifying model performance with measures of discriminatory ability and overfitting. For discriminatory ability, we considered the area under the receiver operating characteristic curve (AUC) calculated on the evaluation data (AUC_{TEST}; Phillips et al., 2006), which, despite its criticisms, remains valid for comparisons for the same species over the same study region (Lobo, Jiménez-Valverde, & Real, 2008; Peterson et al., 2011). We measured overfitting first with the omission error rate of the evaluation data (OR), using as a threshold the value of the prediction that leads to omission of 10% of the calibration records. This thresholding rule should, theoretically, give an OR of 10% with an unbiased evaluation sample (Liu, Berry, Dawson, & Pearson, 2005), with overfit models yielding one higher than 10%. We also considered a threshold-independent measure of overfitting: AUC_{DIFF} (Warren & Seifert, 2011), the difference between the AUCs calculated with calibration records (AUC_{TRAIN}) and evaluation records (AUC_{TEST}). Overfit models should yield relatively high values of AUC_{DIFF}.

As outlined in the Introduction, we assessed the effect size and significance of the obtained evaluation statistics with a null model approach that implements an important modification to previously proposed ones (Beale et al., 2008; Raes & ter Steege, 2007). In contrast to those, which evaluated null models using multiple sets of null pixels, we evaluated both the null models and real ones with the same spatially independent real evaluation records. Apart from being calibrated on random pixels, null models were identical to the real models in every way (same number of input records, same background pixels, same model settings, etc.).

In choosing the appropriate constraint(s) on the randomization method, we opted for simplicity to demonstrate the proposed approach. Null models that are too unconstrained can fail to isolate the factors of the desired null hypothesis, and are therefore prone to Type I statistical error. In contrast, imposing too many constraints can reduce statistical power excessively, increasing the risk of Type II error (Araújo, Thuiller, & Yoccoz, 2009; Beale, Lennon, & Gimona, 2009; Gotelli & Ulrich, 2012; Peterson et al., 2009; Thomas, 2010).

Following Gotelli and Ulrich (2012), we favoured the conservative choice of an unconstrained approach, as in Raes and ter Steege (2007). Specifically, we sampled the null model calibration pixels randomly (and without replacement) from the background pixels used for each respective scenario. Nevertheless, we emphasize that other constraints (including the approach of Beale et al. [2008] for sampling that mimics spatial autocorrelation of occurrences) deserve further testing.

For each combination of calibration records and MAXENT settings, we made one model based on the real species data, and 1,000 corresponding null replicate models based on sets of random pixels. To allow direct comparison of the evaluation statistics across the different scenarios, we applied each model to a single larger geographic region, corresponding to the area a 106 km distance buffer around all occurrence records (i.e. both calibration and evaluation records). We evaluated both the real and null models with the single evaluation dataset of withheld real records, calculating AUC_{TRAIN}, AUC_{TEST}, AUC_{DIFF} and OR for the single larger geographic region (see Appendix S3; Radosavljevic & Anderson, 2013). We then calculated standardized effect sizes (as in Ulrich & Gotelli, 2010) and one-tailed *p*-values (see Appendix S3) to compare the real model evaluations against their respective null distributions. Note that although high performance for AUC_{TEST} and AUC_{TRAIN} is represented by high values and thus positive effect sizes, low values and negative effect sizes indicate high performance for AUC_{DIFF} and OR. Furthermore, tests are one-tailed because we are only interested in results that are better than (rather than just different from) the null distribution.

When applied to AUC_{TRAIN}, our approach is equivalent to the null model test proposed by Raes and ter Steege (2007). We therefore assessed the benefits of our approach compared with that one by comparing the effect sizes and significance levels of AUC_{TRAIN} to those of the other metrics. For further comparison, we also calculated *p*-values for OR using a commonly used binomial test (Anderson, Gómez-Laverde, & Peterson, 2002; Peterson et al., 2011, p 168). Finally, we compared the results for all possible pairs of metrics (across all scenarios) using Spearman's rank correlation coefficients, controlling for the inflated risk of Type I error due to multiple comparisons with Holm's (1979) sequential Bonferroni correction. Otherwise, significance was assessed for all analyses with a decision rule of $\alpha = 0.05$. These analyses were performed in R (R Core Team, 2014) and can be replicated with the script included in Appendix S5 (currently, we are working on generalized functions that run these analyses in the R package 'nullENM' under development; https://github.com/ndimhypervol/nullENM).

3 | RESULTS

3.1 | Performance measures

The performance of both real and null models varied greatly across scenarios and evaluation metrics, with clear and informative general trends. For AUC_{TRAIN}, the values obtained for the real models were for the most part very high and tended to increase with increasing model complexity (Figure 3a). The same was also true for this measure in the null models (equivalent to the approach of Raes & ter Steege, 2007). However, the null AUC_{TRAIN} values varied drastically across scenarios, and their range generally tightened with increasing model complexity (Figure 3a). In contrast, AUC_{TEST} values obtained with real models showed no consistent pattern across scenarios, although they were all fairly high (between 0.69 and 0.87) and models based on calibration records from introduced populations generally yielded lower values (Figure 3b). With null models, AUC_{TEST} values were symmetrically centred on the theoretical expectation of 0.5. This observation was consistent across scenarios, although the models with the simplest settings displayed slightly wider ranges of variability (Figure 3b).

Regarding the measures of overfitting (AUC_{DIFF} and OR; Figure 3c,d), few of the models based on real species records performed well (i.e. showed values close to zero). The ones based on introduced populations performed notably worse. For both real and null models, there was a consistent pattern of decreasing performance with increasing model complexity. The null distributions obtained for AUC_{DIFF} showed levels of variability comparable to those of AUC_{TEST} and were generally symmetrical, albeit slightly skewed towards low values (Figure 3c). For OR, the distributions of null model values were highly variable, often asymmetrical, and their variability generally decreased with increasing model complexity (Figure 3d).

3.2 | Effect sizes and significance

As with performance, effect sizes and significance varied greatly across scenarios and evaluation metrics (Figure 4 and Appendix S4), with the following clear trends. For AUC_{TRAIN} (equivalent to the approach of Raes & ter Steege, 2007), the models based on all calibration records yielded extremely high effect sizes and were highly significant. Lower effect sizes were observed for all the other models, although a few of them were marginally significant or approached significance. Overall, effect sizes decreased with increasing model complexity. For AUC_{TRAIN}, the null hypothesis was rejected for 6 of the 15 scenarios (Figure 4a). For AUC_{TEST}, most models produced fairly high effect sizes, especially those with default and complex settings for the models based on all calibration records and the ones made using native records only. In contrast to AUC_{TRAIN}, effect size generally increased with increasing model complexity. The null hypothesis was rejected for 12 of the 15 scenarios (the three exceptions being for models built with simple settings; Figure 4b).

Regarding the measures of overfitting, the effect sizes were generally low, especially for OR (Figure 4c,d). In fact, several of the models for introduced populations performed worse than random (i.e. positive effect size). For both AUC_{DIFF} and OR, only the model using native records and default settings reached significance (Figure 4c,d). In contrast, significance calculated via the binomial test yielded very different results (Appendix S4). For most scenarios, the results of the binomial test (Anderson et al., 2002) were highly significant, and only four of them did not reach significance. Specifically, the null hypothesis was rejected in 73% of the cases, as opposed to a 7% rejection

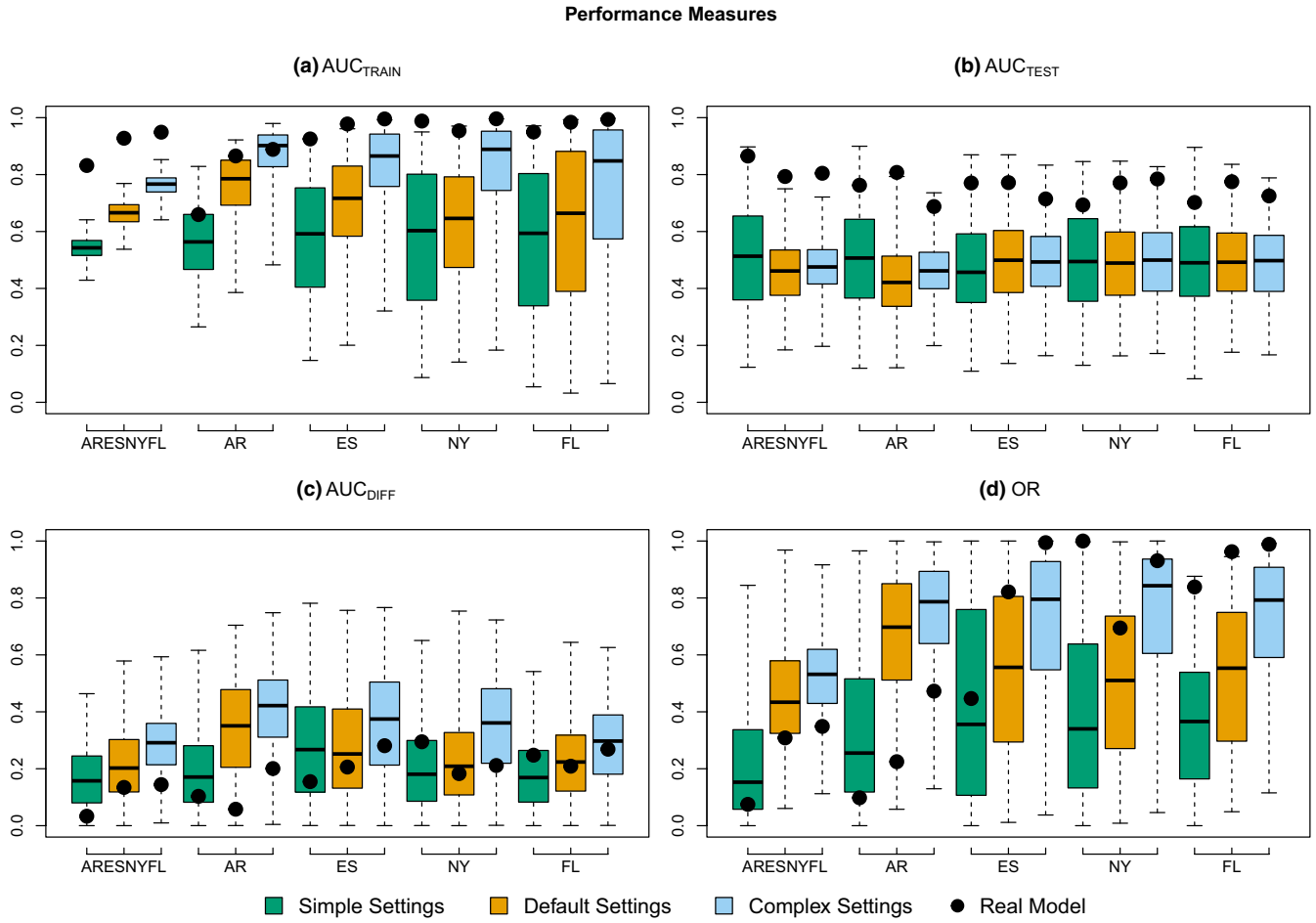

FIGURE 3 Performance of 15 scenarios of MAXENT models for the monk parakeet created with five different sets of calibration records (see Figure 2) across three levels of model complexity (simple, default and complex, shown in green, orange and light-blue respectively). Four measures of model performance are presented in different panels: (a) AUC_{TRAIN} (AUC calculated on calibration records), (b) AUC_{TEST} (AUC calculated on evaluation records), (c) AUC_{DIFF} (=AUC_{TRAIN}–AUC_{TEST}) and OR (the omission error rate with a threshold of 10% of calibration presences). In each panel and for each scenario, the performance of the model built with real species records is shown with a black circle, while the performance of 1,000 replicate null models based on sets of random geographic pixels is summarized with a boxplot showing the range and quartiles of the distribution [Colour figure can be viewed at wileyonlinelibrary.com]

rate with the null model approach applied to OR. These two methods led to the same conclusion regarding the decision to retain or reject the null hypothesis for only five scenarios (33% agreement).

3.3 | Correlations

The correlation analyses primarily revealed a strong positive association between the measures of overfitting (AUC_{DIFF} and OR). This association was consistently strong and highly significant across all types of comparisons: the performances of real models (Table 1a, below diagonal), the average performances of null models (Table 1a, above diagonal) and the effect sizes (Table 1b). Both of these measures were also strongly and positively correlated with AUC_{TRAIN} for the comparisons involving the performances of both real and null models (Table 1a). However, this was not the case for comparisons involving effect sizes. In fact, the AUC_{TRAIN} effect sizes showed no significant association with any other measure (Table 1b). Similarly, AUC_{TEST} showed no significant association with any of the other measures across all types of comparisons (Table 1), except for a moderate and negative correlation with AUC_{DIFF} when comparing effect sizes (Table 1b). Finally, the association between the obtained performance estimates and their corresponding effect sizes for each statistic varied greatly. It was non-existent for AUC_{TRAIN} ($\rho = -0.06$, $p > 0.05$), moderate but not significant for AUC_{TEST} ($\rho = 0.68$, $p > 0.05$), fair for AUC_{DIFF} ($\rho = 0.70$, $p < 0.05$) and strong for OR ($\rho = 0.81$, $p < 0.001$).

4 | DISCUSSION

This study indicates several ways in which null models can provide useful information regarding model quality. First, the results suggest that our proposed modification to existing null model approaches (i.e. evaluating both real and null models with the same set of withheld evaluation records) is an important step in providing unbiased performance estimates. We observed a strong departure

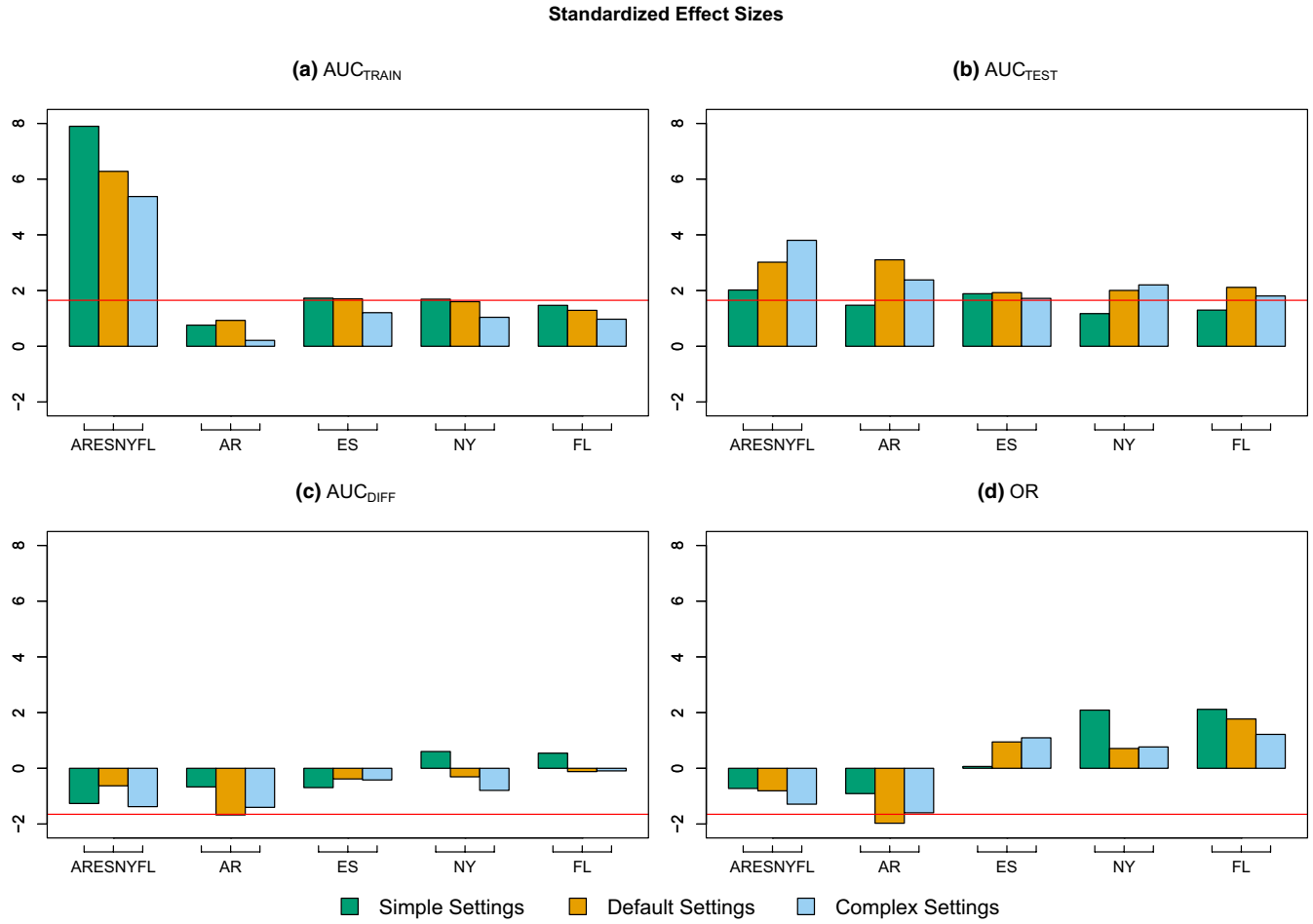

FIGURE 4 Standardized effect size and significance of 15 scenarios of MAXENT models for the monk parakeet created with five different sets of calibration records (see Figure 2) across three levels of model complexity (simple, default and complex, shown in green, orange and light-blue respectively). Four measures of model performance are presented in different panels: (a) AUC_{TRAIN} (AUC calculated on calibration records), (b) AUC_{TEST} (AUC calculated on evaluation records), (c) AUC_{DIFF} (=AUC_{TRAIN}–AUC_{TEST}) and OR (the omission error rate with a threshold of 10% of calibration presences). The height of the bars represents the strength of the effect sizes, expressing (in standard deviations units) how extreme the performance of the model built with real species records is compared with a distribution created with 1,000 replicate null models based on sets of random geographic pixels. The red horizontal line indicates the critical threshold beyond which effect sizes are significant at the 0.05 level. This threshold is positive for AUC_{TEST} and AUC_{TRAIN}, as good performance is represented by high values (and thus positive effect sizes). In contrast, the critical threshold is negative for AUC_{DIFF} and OR, because high performance corresponds to low values (and hence negative effect sizes) [Colour figure can be viewed at wileyonlinelibrary.com]

from theoretical expectations for AUC_{TRAIN} (equivalent to Raes & ter Steege, 2007) but not for AUC_{TEST} (our modification). Random pixels led to highly skewed null distributions for AUC_{TRAIN} values (shifted towards higher values than the theoretical expectation of 0.5). This observation matches that of Merckx et al. (2011), who employed the null model approach of Raes and ter Steege (2007) and found that the methodology used to sample calibration records had a major impact on AUC_{TRAIN} values. In contrast to results for AUC_{TRAIN}, the null model approach presented here yielded distributions of AUC_{TEST} values that were all symmetrically centred near the theoretical expectation of 0.5. In addition, the spread and skewness of these distributions were affected by the modelling settings to a much lesser extent than with AUC_{TRAIN}.

Second, the null distributions obtained for each evaluation statistic were greatly influenced by both the region from which the calibration records were sampled and the choice of modelling settings. For example, models based on records from some calibration regions produced very tight null distributions, while those based on other regions led to wide variation. This indicates that the availability and/or geographic/spatial structure of environmental conditions in different regions have critical impacts on the null distributions of obtainable performance. Furthermore, modelling choices often influenced both real and null models in a similar way, confirming that an observed effect of model complexity on performance can be an artefact of the settings and study regions selected rather a true species-specific trend. Overall, these results suggest that for any evaluation statistic, there is a unique null distribution for each combination of study region and modelling settings, demonstrating the importance of null model approaches for niche models.

TABLE 1 Spearman rank correlation coefficients for comparisons of measures of performance and effect size for 15 modelling scenarios and spatially independent evaluations of MAXENT models for the monk parakeet. Pairwise comparisons across four evaluation metrics are shown: AUC_{TRAIN} and AUC_{TEST} (AUCs calculated on calibration and evaluation records, respectively), AUC_{DIFF} (= AUC_{TRAIN}–AUC_{TEST}) and OR (the omission error rate with a threshold of 10% of calibration presences). The top portion (a) gives results for the estimates of performance obtained with models based on real species occurrences (below diagonal); and for the average performances of 1,000 null models based on sets of null pixels (above diagonal). The bottom portion (b) provides comparisons for the corresponding effect sizes. Asterisks indicate significance levels (*<0.05, **<0.01, ***<0.001), controlling for multiple comparisons with Holm's sequential Bonferroni correction

	AUC_{TRAIN}	AUC_{TEST}	AUC_{DIFF}	OR
(a) Performance				
AUC_{TRAIN}		−0.40	0.93***	0.93***
AUC_{TEST}	−0.26		−0.43	−0.30
AUC_{DIFF}	0.86***	−0.68		0.91***
OR	0.91***	−0.58	0.98***	
(b) Effect sizes				
AUC_{TRAIN}				
AUC_{TEST}	0.10			
AUC_{DIFF}	0.10	−0.72*		
OR	0.10	−0.69	0.90***	

Third, the importance of comparing real model performance to an appropriate null distribution stresses the utility of considering effect size and significance, rather than just the values of various performance metrics. We found that good performance (high discrimination or low overfitting) did not always accompany high effect sizes or significance, and that measures of performance alone led to an incomplete and biased interpretation of results. For example, when considering performance alone, the model created with all the calibration records and simple settings appeared best because it yielded the highest AUC_{TEST}, as well as the lowest AUC_{DIFF} and OR. Nevertheless, this model had fairly low effect sizes for these measures and failed to reach significance for AUC_{DIFF} or OR. Conversely, when considering performance, effect size and significance together, the model built with native records and intermediate settings yielded the best results overall (and was the only one reaching significance for all three measures).

Fourth, some of the methods available now for producing null distributions and significance estimates seem to be inappropriate. Researchers should be particularly cautious with the binomial test (Anderson et al., 2002) for OR. Here, it yielded very different results than our null model approach. Despite the fact that most models resulted in very poor ORs, the majority of these were highly significant according to the binomial test. This test rests on the assumption that the pixels of the study region are independent observations, which is probably strongly violated in many situations because of spatial structure in the variation of environmental conditions across the study region. It is thus prone to high rates of Type I error. Other methods based on bootstrap replicates of the species occurrence data (e.g. Peterson, Papeş, & Soberón, 2008) may be equally susceptible to this problem because they account for variation associated with sampling of the occurrence data but not for sources of error associated with the environmental structure across the study region.

Finally, the results confirm that despite the fact that AUC_{DIFF} is threshold-independent and OR threshold-dependent, both reflect a similar property that we interpret to be overfitting. For both the real and null models, these statistics were highly correlated for both the actual values and the corresponding effect sizes. As expected, these statistics were also highly correlated with AUC_{TRAIN} (for both the real and null models), reinforcing that models that are too tightly fitted to calibration records tend to predict independent datasets poorly (Radosavljevic & Anderson, 2013). No such relationships existed between AUC_{TEST} (which measures discrimination) and either AUC_{DIFF} or OR (which both reflect overfitting). Consequently, while it may be sufficient to consider only one measure to assess overfitting, model performance should be assessed with both discriminatory ability and overfitting criteria.

5 | CONCLUSIONS AND RECOMMENDATIONS

This study demonstrates the utility of the proposed null model approach, which merits further use and future research. The results reinforce that the estimated performance of ecological niche models is strongly influenced by modelling choices (such as the study region considered and settings related to complexity; Anderson & Raza, 2010; Muscarella et al., 2014), and that measures of overfitting and discriminatory ability account for different aspects of model performance. Additionally, because high performance did not consistently correspond to high effect size and significance, considering measures of performance alone will tend to result in incomplete and likely biased interpretations. Most importantly, because the estimated performance of models evidently was influenced by factors independent from the species' distribution, it is critical to measure performance against an appropriate null distribution that accounts for such factors. The null model approach as implemented here provides a practical way to do so, for either the same or a different region/time period. Obtaining accurate estimates of statistical error rates for this null model approach will ultimately require that similar analyses be undertaken with other species, other constraints on randomization (as in Beale et al., 2008) and with simulated data (as in Ulrich & Gotelli, 2007). We hope that the approach and code presented here facilitate such progress through future studies aimed at reaching more general conclusions regarding these issues, which are critical for the important field of ecological niche modelling of species distributions.

ACKNOWLEDGEMENTS

This work was funded in part by the U.S. National Science Foundation (DEB-1119915 to Anderson). Most analyses were conducted in the laboratory of Jason Munshi-South at Baruch College of CUNY. We thank Robert A. Boria, Peter J. Galante, Valentina Grisales Betancur, Gonzalo Pinilla-Buitrago, Mariano Soley-Guardia, and especially Jason Munshi-South and anonymous reviewers for helpful comments that substantially improved the manuscript.

DATA ACCESSIBILITY

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.cr38pj0

Title: Data from: A new null model approach to quantify performance and significance for ecological niche models of species distributions

DOI: doi:10.5061/dryad.cr38pj0

Journal: Journal of Biogeography

Journal manuscript number: JBI-16-0638

ORCID

Corentin L. Bohl https://orcid.org/0000-0003-1418-8430

Jamie M. Kass https://orcid.org/0000-0002-9432-895X

Robert P. Anderson https://orcid.org/0000-0002-7706-4649

REFERENCES

Anderson, R. P. (2013). A framework for using niche models to estimate impacts of climate change on species distributions. *Annals of the New York Academy of Sciences*, *1297*, 8–28. https://doi.org/10.1111/nyas.12264

Anderson, R. P., Gómez-Laverde, M., & Peterson, A. T. (2002). Geographical distributions of spiny pocket mice in South America: Insights from predictive models. *Global Ecology and Biogeography*, *11*(2), 131–141. https://doi.org/10.1046/j.1466-822X.2002.00275.x

Anderson, R. P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, *37*(7), 1378–1393. https://doi.org/10.1111/j.1365-2699.2010.02290.x

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Araújo, M. B., Thuiller, W., & Yoccoz, N. G. (2009). Reopening the climate envelope reveals macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), E45–E46; author reply E41-3. https://doi.org/10.1073/pnas.0813294106

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*(11), 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Beale, C. M., Lennon, J. J., & Gimona, A. (2008). Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(39), 14908–14912. https://doi.org/10.1073/pnas.0803506105

Beale, C. M., Lennon, J. J., & Gimona, A. (2009). European bird distributions still show few climate associations. *Proceedings of the National Academy of Sciences*, *106*(16), E41–E43. https://doi.org/10.1073/pnas.0902229106

Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., … Guisan, A. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, *21*(4), 481–497. https://doi.org/10.1111/j.1466-8238.2011.00698.x

Dormann, C. F., McPherson, J. M., Araújo, M. B., 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

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

Fitzpatrick, M. C., & Hargrove, W. W. (2009). The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, *18*(8), 2255–2261. https://doi.org/10.1007/s10531-009-9584-8

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

Gonçalves da Silva, A., Eberhard, J. R., Wright, T. F., Avery, M. L., & Russello, M. A. (2010). Genetic evidence for high propagule pressure and long-distance dispersal in monk parakeet (*Myiopsitta monachus*) invasive populations. *Molecular Ecology*, *19*(16), 3336–3350. https://doi.org/10.1111/j.1365-294X.2010.04749.x

Gotelli, N. J., & McGill, B. J. (2006). Null versus neutral models: What's the difference? *Ecography*, *29*(5), 793–800. https://doi.org/10.1111/j.2006.0906-7590.04714.x

Gotelli, N. J., & Ulrich, W. (2012). Statistical challenges in null model analysis. *Oikos*, *121*(2), 171–180. https://doi.org/10.1111/j.1600-0706.2011.20301.x

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., & Kueffer, C. (2014). Unifying niche shift studies: Insights from biological invasions. *Trends in Ecology and Evolution*, *29*(5), 260–269. https://doi.org/10.1016/j.tree.2014.02.009

Hijmans, R. J. (2012). Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology*, *93*(3), 679–688. https://doi.org/10.1890/11-0826.1

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. Retrieved from http://www.jstor.org/stable/4615733

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *3*, 385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., & Vanaverbeke, J. (2011). Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, *222*(3), 588–597. https://doi.org/10.1016/j.ecolmodel.2010.11.016

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

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: An R package

for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, *5*(11), 1198–1205. https://doi.org/10.1111/2041-210X.12261

Osborne, P. E., Foody, G. M., & Suárez-Seoane, S. (2007). Non-stationarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions*, *13*(3), 313–323. https://doi.org/10.1111/j.1472-4642.2007.00344.x

Osborne, P. E., & Suárez-Seoane, S. (2002). Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling*, *157*(2–3), 249–259. https://doi.org/10.1016/S0304-3800(02)00198-9

Peterson, A. T., Barve, N., Bini, L. M., Diniz-Filho, J. A., Jiménez-Valverde, A., & Lira-Noriega, A., … Soberón, J. (2009). The climate envelope may not be empty. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), E47; author reply E41-3. https://doi.org/10.1073/pnas.0809722106

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

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*. In S. A. Levin & H. S. Horn (Eds.), *Monographs in Population Biology* (vol. 49). Princeton, NJ: Princeton University Press. Retrieved from http://press.princeton.edu/titles/9641.html

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, *40*(7), 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S. J., Anderson, R. P., & Schapire, R. (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. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org.

Radosavljevic, A., & Anderson, R. P. (2013). 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

Raes, N., & ter Steege, H. (2007). A null-model for significance testing of presence-only species distribution models. *Ecography*, *30*(5), 727–736. https://doi.org/10.1111/j.2007.0906-7590.05041.x

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Shcheglovitova, M., & Anderson, R. P. (2013). Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling*, *269*, 9–17. https://doi.org/10.1016/j.ecolmodel.2013.08.011

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE*, *8*(2), e55158. https://doi.org/10.1371/journal.pone.0055158

Thomas, C. D. (2010). Climate, climate change and range boundaries. *Diversity and Distributions*, *16*(3), 488–495. https://doi.org/10.1111/j.1472-4642.2010.00642.x

Ulrich, W., & Gotelli, N. J. (2007). Disentangling community patterns of nestedness and species co-occurrence. *Oikos*, *116*(12), 2053–2061. https://doi.org/10.1111/j.2007.0030-1299.16173.x

Ulrich, W., & Gotelli, N. J. (2010). Null model analysis of species associations using abundance data. *Ecology*, *91*(11), 3384–3397. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21141199.

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*(4), 589–594. https://doi.org/10.1016/j.ecolmodel.2008.11.010

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, *36*(12), 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.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*, 29–43, https://doi.org/10.1111/nyas.12226

BIOSKETCHES

Corentin L. Bohl holds a PhD. from the City University of New York. His current research includes the development of novel methods integrating ecological niche modelling, population genetics and landscape genetics. This interdisciplinary investigation aims to quantify patterns of species distributions and genetic variation, and lead to better understanding of underlying behavioural and ecological processes.

Jamie M. Kass holds a PhD from the City College of New York, City University of New York. He currently conducts research on biotic interactions and species distribution models across a variety of ecological systems.

Robert P. Anderson is a Professor of Biology at the City College of New York, City University of New York, specializing on the biogeography of Neotropical mammals. He is interested in spatial patterns of environmental suitability, and their ecological, evolutionary and practical consequences.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Bohl CL, Kass JM, Anderson RP. A new null model approach to quantify performance and significance for ecological niche models of species distributions. *J Biogeogr.* 2019;46:1101–1111. https://doi.org/10.1111/jbi.13573