Received: 13 December 2018 | Revised: 15 August 2019 | Accepted: 17 August 2019

DOI: 10.1111/jbi.13705

RESEARCH PAPER

Journal of Biogeography | WILEY

Evaluating presence-only species distribution models with discrimination accuracy is uninformative for many applications

Dan L. Warren[1,2,3] | Nicholas J. Matzke[4] | Teresa L. Iglesias[3,5]

[1]Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany

[2]Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology, Okinawa, Japan

[3]Department of Biology, Macquarie University, Sydney, NSW, Australia

[4]Department of Biological Sciences, University of Auckland, Auckland, New Zealand

[5]Physics and Biology Unity, Okinawa Institute of Science and Technology, Okinawa, Japan

Correspondence
Dan L. Warren, Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany.
Email: dan.l.warren@gmail.com

Funding information
Macquarie University; Australian Research Council

Handling Editor: Dr. Wilfried Thuiller

Abstract

Aim: Species distribution models are used across evolution, ecology, conservation and epidemiology to make critical decisions and study biological phenomena, often in cases where experimental approaches are intractable. Choices regarding optimal models, methods and data are typically made based on discrimination accuracy: a model's ability to predict subsets of species occurrence data that were withheld during model construction. However, empirical applications of these models often involve making biological inferences based on continuous estimates of relative habitat suitability as a function of environmental predictor variables. We term the reliability of these biological inferences 'functional accuracy.' We explore the link between discrimination accuracy and functional accuracy.

Methods: Using a simulation approach we investigate whether models that make good predictions of species distributions correctly infer the underlying relationship between environmental predictors and the suitability of habitat.

Results: We demonstrate that discrimination accuracy is only informative when models are simple and similar in structure to the true niche, or when data partitioning is geographically structured. However, the utility of discrimination accuracy for selecting models with high functional accuracy was low in all cases.

Main conclusions: These results suggest that many empirical studies and decisions are based on criteria that are unrelated to models' usefulness for their intended purpose. We argue that empirical modelling studies need to place significantly more emphasis on biological insight into the plausibility of models, and that the current approach of maximizing discrimination accuracy at the expense of other considerations is detrimental to both the empirical and methodological literature in this active field. Finally, we argue that future development of the field must include an increased emphasis on simulation; methodological studies based on ability to predict withheld occurrence data may be largely uninformative about best practices for applications where interpretation of models relies on estimating ecological processes, and will unduly penalize more biologically informative modelling approaches.

1 | INTRODUCTION

Species distribution models (SDM, alternatively environmental niche models or ENM) use data on species occurrences in conjunction with environmental data to generate statistical models of species' ecological tolerances, environmental limits and potential to occupy different geographic areas. These methods have been used since the 1920s (Cook, 1925; Sutherst, 2014), but recent years have seen rapid growth in the number of studies employing SDM in fields including ecology, conservation biology, evolutionary biology and epidemiology (Allen & Lendemer, 2016; Coro, Pagano, & Ellenbroek, 2013; Guisan, Thuiller, & Zimmermann, 2017; Gutierrez-Tapia & Palma, 2016; Lezama-Ochoa et al., 2016; Peterson et al., 2011; Raghavan et al., 2016). The primary appeal of SDMs is their tractability; estimating environmental tolerances experimentally is expensive and time-consuming at best and impractical for many species. In contrast, SDMs can be constructed with minimal investment of resources, using freely available data and software (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005; Hijmans, Phillips, Leathwick, & Elith, 2012; Kriticos et al., 2012; Phillips, Anderson, & Schapire, 2006; Thuiller, Lafourcade, Engler, & Araújo, 2009). For many species of conservation concern, they are one of the only tractable means of estimating habitat suitability, often in cases where stakeholders need these estimates urgently (Keith et al., 2014; Warren, Wright, Seifert, & Shaffer, 2014).

SDM construction involves many decisions which may affect model predictions. These include choice of modelling algorithm, required sample size, optimal model complexity, choice of study area from which data are drawn, the exclusion of outliers and selection of environmental predictors, among others (Acevedo, Jimenez-Valverde, Lobo, & Real, 2012; Boria, Olson, Goodman, & Anderson, 2014; Domisch, Kuemmerlen, Jahnig, & Haase, 2013; Garcia-Callejas & Araujo, 2016; Guisan, Graham, Elith, Huettmann, & Distri, 2007; van Proosdij, Sosef, Wieringa, & Raes, 2016; Soley-Guardia et al., 2016; Varela, Anderson, Garcia-Valdes, & Fernandez-Gonzalez, 2014; Wisz et al., 2008). The literature surrounding these decisions is large and growing rapidly, as is the suite of associated software tools. Decisions about how best to model species are typically made using metrics that test discrimination accuracy on subsets of species occurrence data that have been withheld during model construction (Elith et al., 2006; Radosavljevic & Anderson, 2014); for a recent literature review and summary see Appendices S1 and S2. However, the binary prediction of withheld occurrence data is rarely the intended application of SDMs; they are more frequently used to make continuous estimates of habitat suitability, and to make predictions outside of the training conditions both in space and in time. These applications often implicitly assume that there is biological meaning to the continuous suitability scores produced by the model, or to the functional relationship between environmental gradients and habitat suitability. However, it is often not clear which (if any) measurable biological phenomena should be correlated with suitability estimates from SDMs. Many of the measurable phenomena that are potentially related to suitability (e.g., population density [Carrascal, Aragon, Palomino, & Lobo, 2015], upper limit of local abundance [VanDerWal, Shoo, Johnson, & Williams, 2009, Gomes et al., 2018]) have not been quantified in detail for many real species and as such are unavailable for model validation.

This impracticality of studying environmental suitability experimentally makes it difficult to measure the ability of SDMs to correctly make continuous estimates of habitat suitability. As such, modelling decisions are typically predicated on an assumed relationship between a model's ability to make continuous estimates of relative habitat suitability (hereafter referred to as 'functional accuracy') and its ability to predict withheld occurrence data (discrimination accuracy). This assumption has been questioned before (Lobo, Jimenez-Valverde, & Real, 2008), but its importance and validity for SDM studies is largely untested.

Discrimination accuracy is known to be a potentially misleading measure for many applications; it is known to be a poor indicator of model calibration (Jimenez-Valverde, Acevedo, Barbosa, Lobo, & Real, 2013; Reineking & Schröder, 2006), and may even be negatively correlated with calibration and functional accuracy under some conditions (Murphy & Winkler, 1992). This general statistical problem may be exacerbated by attributes of the SDM process in a number of ways. First, spatial autocorrelation present in species distributions and in the environment can generate spurious correlations that a model might treat as biological truth, resulting in models that produce high discrimination accuracy even when occurrence data are random (Raes & ter Steege, 2007) or the predictors are biologically meaningless (Bahn & McGill, 2007, 2013; Fourcade, Besnard, & Secondi, 2018). Second, there are phenomena other than the suitability of habitat that shape species distributions (e.g., historical biogeography, dispersal, biotic interactions, Figure 1) (Anderson, 2012; Kearney, 2006; Soberon & Peterson, 2005; Warren, 2012, 2013; Warren, Cardillo, Rosauer, & Bolnick, 2014). Although it is possible to include these processes as predictors for SDMs, this is not often done in practice. Failure to explicitly consider these processes introduces spurious correlations between species occurrences and the environment into the estimate of the environmental niche.

Similarly, the collection of occurrence data often shows spatial biases (Figure 1, panel F), which may be correlated with spatially autocorrelated predictors (Phillips et al., 2009). All of these phenomena can lead to poor niche estimates (Figure 2) that still have high discrimination accuracy in geographic space. Since these non-target phenomena are shared between training and test data, a model that parameterizes the environmental correlates of these processes may have higher discrimination accuracy than a model that accurately estimates the species' environmental tolerances, and yet may produce pathological behaviour in applications where model transferability or continuous estimates of habitat suitability are desired (Huang & Frimpong, 2016; Lobo et al., 2008; Radosavljevic & Anderson, 2014; Torres et al., 2015; Veloz, 2009).

A further issue with discrimination accuracy is the lack of true absence data. One of the primary reasons that SDM methods are so tractable is that they can be used without true absence data, which is often difficult and expensive to obtain. SDMs deal with the

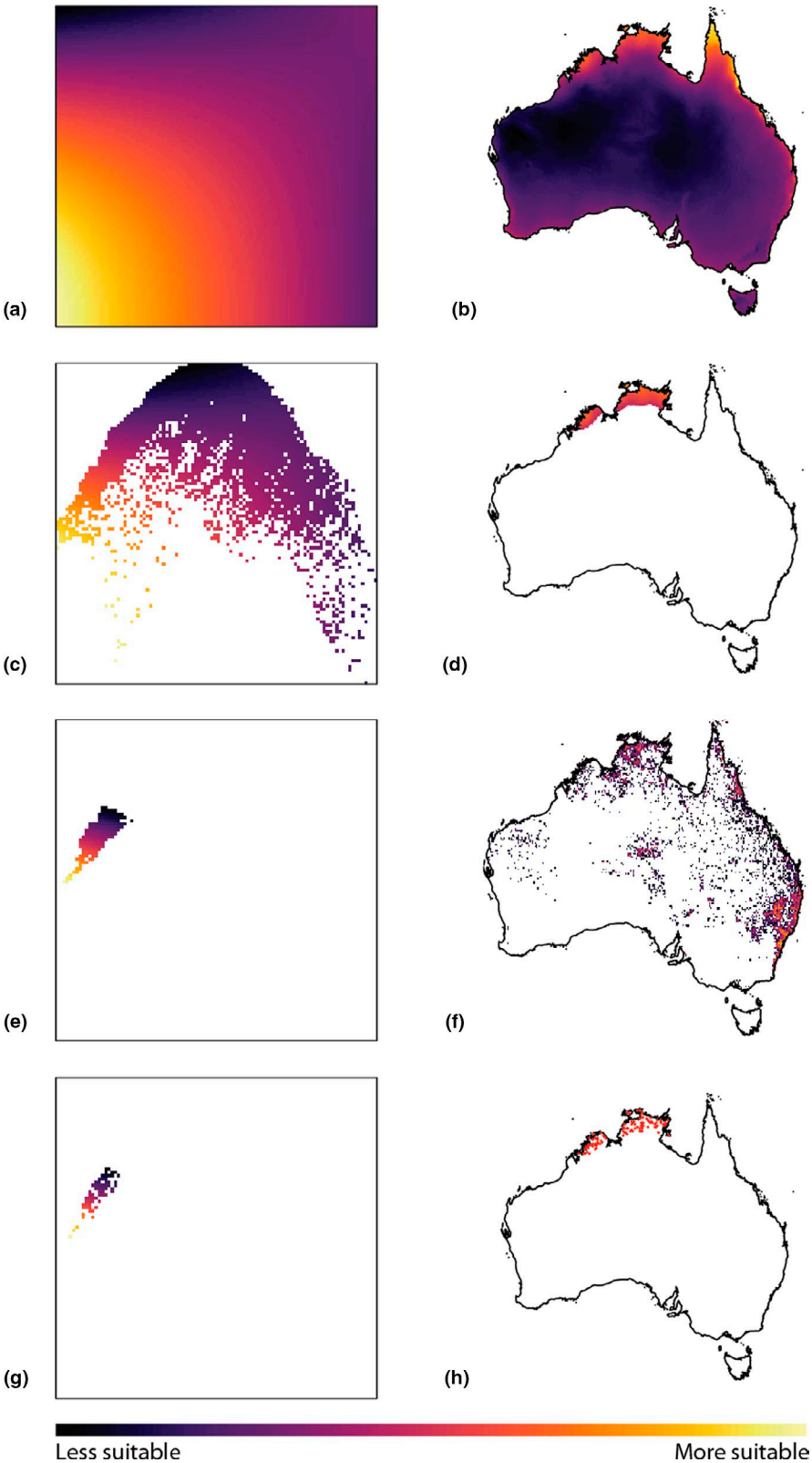

FIGURE 1 Phenomena affecting species distributions and inference of suitability of habitat. Panel a depicts the niche of a simulated species in the first two principal component axes of the 19 Bioclim variables for Australia. Panel b represents the distribution of suitable habitat for the simulated species. The available habitat present across the continent of Australia only represents a subset of the possible niche space (c). The species' current range only encompasses a subset of the suitable habitat (d), which further limits the potential distribution of data in environment space (e). Spatial sampling bias (f, see methods) contributes further bias to the representation of the species both in environment space (g) and geographic space (h). While the geographic distribution of the data (h, red points) may resemble the current range of the species (d), the distribution of that data in environment space (g) is a poor representation of the species' true niche (a). As a result, it may be relatively easy to achieve accurate predictions on randomly withheld occurrence data while still producing a poor estimate of the underlying biology and suitability of habitat

lack of true absences by sampling 'pseudoabsence' or 'background' points which are ideally intended to represent the set of environmental conditions that are potentially accessible to the species. This requires users to make decisions about the size of the appropriate study area for background samples (Acevedo et al., 2012), as well as the nature of sampling (e.g., random points or points from closely related species (Phillips et al., 2009)). These decisions are often somewhat arbitrary (e.g., background areas chosen using political

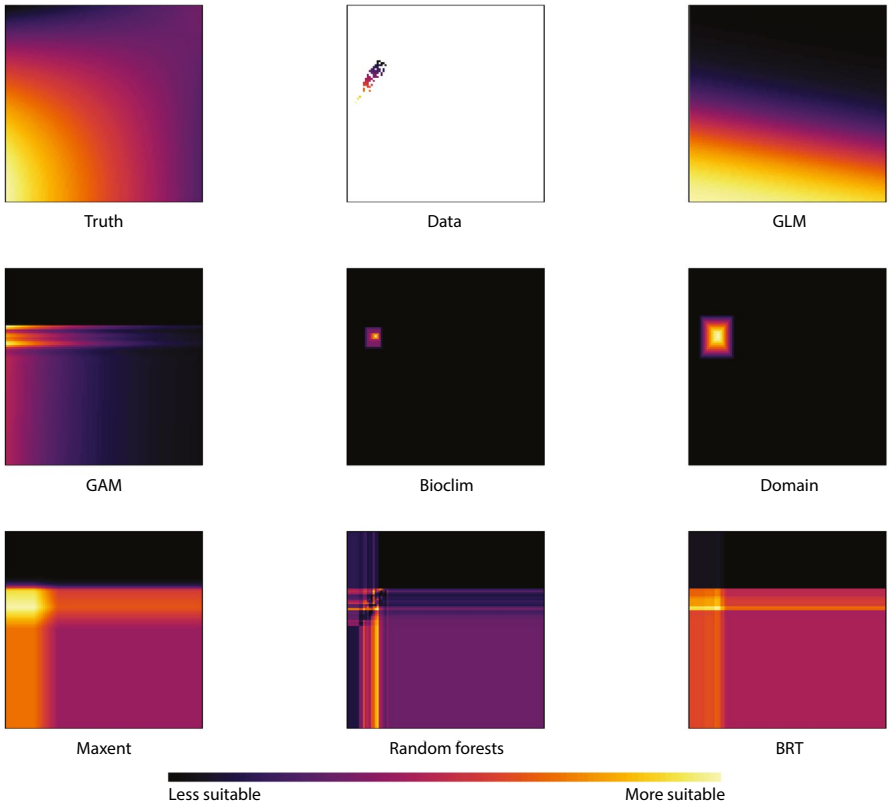

FIGURE 2 Projection of modelling algorithms in environment space. Using 100 occurrence points for the simulated species in Figure 1, we built models using the seven algorithms employed in this study and projected them into the same two dimensional principal component space. The lowest AUC score on 20 randomly withheld data points belonged to random forests (AUC = 0.55), while the highest came from domain (AUC = 0.73). The top left and top centre panels show the true niche of the simulated species and the environmental distribution of the data, respectively

boundaries or poorly justified assumptions about dispersal), and can affect both the inferred model (Acevedo et al., 2012) and the performance of metrics used to evaluate models (Acevedo et al., 2012; Hijmans, 2012; Jimenez-Valverde et al., 2013). The lack of real absence data results in models that are incapable of accurately predicting prevalence, and that incorrectly treat some suitable conditions as unsuitable.

Finally, the usefulness of discrimination accuracy as a criterion for selecting SDMs may also be negatively impacted by model complexity. Discrimination accuracy only measures whether a model assigns higher suitability values to presence points than it does to background or absence points, and highly flexible algorithms may produce a broad range of marginal suitability functions that have similar, or even identical, discrimination accuracy (Figure 3). This phenomenon is likely compounded by the frequent use of large numbers of predictors that are highly collinear; as the number of predictors and the complexity of marginal suitability functions increase, the number of potential models with similar discrimination accuracy grows very rapidly.

Although many of these problems with discrimination accuracy have been noted before, the utility of discrimination metrics for SDM studies has not been thoroughly examined in a system where the true niche and habitat suitability are known. As a result, we have little information on how useful these metrics are for empirical studies where the goal is to estimate the relative suitability of habitat, despite the ubiquity of discrimination metrics in SDM model selection.

Here, we adopt a simulation approach to explore the relationship between discrimination and functional accuracy using virtual species for which the true niche is known. We build models using a number of different algorithms, study area sizes, and methods of partitioning training and test data. However, these simulations are not intended to represent all possible modelling approaches. The goal of these simulations is not to determine which method produces the best niche or distribution estimates, but rather to evaluate commonly used methods for model selection across a broad range of models in a system where we know the underlying true habitat suitability.

2 | MATERIALS AND METHODS

To examine how model selection was affected by modelling algorithm, sampling bias and non-target spatial phenomena, we conducted four sets of simulation experiments:

1. High complexity: Artificial niches were generated using three randomly chosen variables and models were constructed using all 19 bioclimatic variables. Background data were drawn from 100 km circular buffers around occurrence points.
2. Geographic partitioning: Same conditions as (1) but presence and background points for each species were split into four quadrants using ENMEval (Muscarella et al., 2014). Presence and background data from one randomly selected quadrant were withheld for model evaluation.
3. Large background: Same conditions as (1), but background data were drawn from 1,000 km circular buffers around occurrence points.

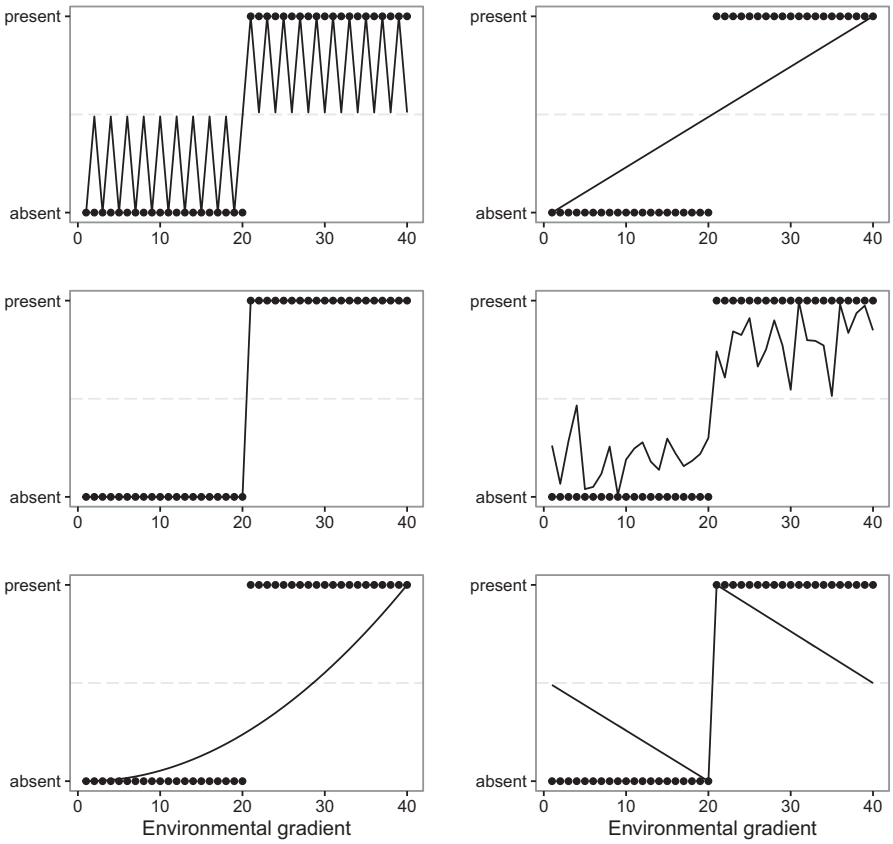

FIGURE 3 Low information content of discrimination accuracy for inferring functional accuracy. In the above plots, we have simulated 20 absence and 20 presence points along a hypothetical environmental gradient. The six panels represent six hypothetical functions that might be inferred using these data. Each function assigns a higher suitability score (*y* axis) to all of the presence points than it does to any of the background or absence points. As a result, each function has perfect discrimination accuracy. All six functions are therefore indistinguishable from each other based on discrimination metrics (AUC, TSS, Kappa), while making very different estimates of the functional relationship of habitat suitability to the environmental predictor variable

4. Low complexity: Niches were based on two randomly chosen variables and models were built using those two variables plus two more chosen at random.

All analyses used CliMond data for 19 bioclimatic variables typically employed in SDM studies (Nix, 1986), including data for the present and for 24 combinations of future emissions scenario (A1B and A2), year (2030, 2050, 2070, 2080, 2090 and 2100), and climate model (CSIRO-Mk3.0 and MIROC-H) (Kriticos et al., 2012). Simulations and analyses were restricted to Australia, including Tasmania.

Simulated niches were created using the generateRandomSp function in the 'virtualspecies' R package (Leroy, Meynard, Bellard, & Courchamp, 2015). Simulated species with fewer than 400 suitable grid cells in the initial presence/absence raster showed a strong tendency to produce models that showed no suitable habitat on future climate scenarios, rendering comparisons that were uninformative for model selection. As a result, these simulations were discarded.

To simulate the effects of non-target spatial processes (e.g., historical biogeography, biotic interactions, dispersal limitation), the initial presence/absence raster from virtualspecies was converted to point data, which was partitioned into allopatric regions using k-means clustering. Solutions ranging from 2 to 10 clusters were considered, and an algorithm was used to maximize the minimum distance between clusters. One region was assigned at random to be the range of the species, and converted back into a raster. We recorded the total proportion of suitable cells that fell within this range, to measure the extent to which species distributions departed from the distribution of available suitable habitat across the entire study area.

Spatial sampling bias was modelled using data from 5,969,252 collection records for 28,286 Australian plant species. These records were harvested from GBIF (GBIF.org,) using the 'rgbif' (Chamberlain, Boettiger, Ram, Barve, & Mcglinn, 2013) package, and converted to a raster representing the number of observations per grid cell at the same extent and resolution as the environmental data. All values were then divided by the maximum cell value, resulting in a range of sampling intensity from 0 to 1.

Species occurrence data was sampled from a raster with values in each grid cell *x* calculated as:

$$p(x) = (1 - g_b)s(x)r(x) + g_b b(x)s(x)r(x)$$

where probability of sampling is a function of $p(x)$, g_b is a parameter that controls the magnitude of spatial sampling bias, $b(x)$ is the relative strength of spatial sampling bias in cell *x*, $s(x)$ is the suitability of habitat in the grid cell and $r(x)$ is a binary variable taking the value 1 inside the species range and 0 everywhere else. For each species, we drew 100 simulated occurrence points by selecting grid cells at random and sampling occurrences as a Bernoulli trial with probability of success equal to $p(x)$. Presence points were drawn with replacement so that we could study the effects of sampling bias. We simulated data across 11 levels of spatial sampling bias, with the bias strength parameter ranging from 0 to 1 in increments of 0.1. We performed 20 simulations for each of

the 11 levels of spatial sampling bias. Each of the four simulation conditions (experiments 1–4, above) therefore consisted of 220 simulations, for a total of 880 total simulated species across all experiments.

As noted in a recent review (Meynard, Leroy, & Kaplan, 2019), simulation studies need to choose both the simulated niches and sampling regimes that are appropriate for the question involved. Since our goal here is to test which metrics select models that accurately estimate the niche, it was essential for us to generate data that would be capable of producing accurate niche estimates in ideal conditions. Due to these concerns we chose not to apply a threshold minimum suitability score below which the organism could not possibly occur; the application of such a threshold would truncate the response functions we are trying to estimate (Meynard et al., 2019), resulting in lower expected functional accuracy. Additionally, prior work with virtual species has demonstrated that the application of thresholds results in discrimination accuracy metrics that are overly optimistic (Meynard & Kaplan, 2013).

We built models using seven algorithms; Bioclim, Domain, generalized linear models (GLM), generalized additive models (GAM), Maxent, random forests and boosted regression trees. Models were built using the 'dismo' R package (Hijmans et al., 2012) and Maxent (Phillips et al., 2006). This resulted in 6,160 inferred models, seven for each of the 880 simulated niches. Algorithm settings were left at their default values, and models were provided with all 19 predictor variables. While much has been written about the desirability of reducing the set of predictor variables and minimizing multicollinearity, we forego that procedure for experiments 1–3 because practitioners in the field frequently do use all 19 variables, relying on the modelling algorithm to determine their relative importance. We explore the effects of this procedure in experiment 4 and in a fifth experiment, outlined below. For each model, 25 occurrence points were withheld from model construction and used for model evaluation. Each model's discrimination accuracy was evaluated using three statistics: the area under the receiver operating characteristic curve (AUC) (Fielding & Bell, 1997), the true skill statistic (TSS) (Allouche, Tsoar, & Kadmon, 2006), and Cohen's kappa (Cohen, 1960). While AUC can be calculated using continuous suitability scores, TSS and kappa require binary predictions of presence and absence, so the values for these models that were used in model assessment corresponded to their maximum value across all potential thresholds, i.e., the best possible performance of a thresholded model (Fielding & Bell, 1997).

Models were projected onto the current distribution of the environmental variables used for model construction. Models were also projected onto the 24 future climate scenarios. We used the simulated niche from the virtualspecies object to project the true suitability of habitat across those same set of future environments, to assess whether discrimination accuracy is a useful predictor of models' ability to extrapolate to new environmental conditions. To measure functional accuracy, we compared geographic projections of habitat suitability from the true niche and the inferred models using Spearman rank correlation. Correlations between projected and true suitability scores for the present and for future climate scenarios were measured separately within the species range (areas where $r(x) = 1$) and across the entire study area (Australia and Tasmania). Spearman rank correlation was chosen as a measure of functional accuracy for this study due to the structural differences between models produced by different algorithms, and in consideration of how SDMs are often applied; any two models that assign identical rankings to a set of habitat patches are effectively interchangeable

TABLE 1 Results of regressions functional accuracy on discrimination accuracy, all algorithms considered together

Independent variable	Dependent variable	Simple	Complex	Large BG	Geographic
Test AUC	Spearman (*N*)	+,.01		+,.02	+,.08
Test AUC	Spearman (C)				+,0.02
Test Max TSS	Spearman (*N*)	+,.01		+,.02	+,.07
Test Max TSS	Spearman (C)		−,.01		+,.01
Test Max Kappa	Spearman (*N*)		−,.01	+,.01	+,.06
Test Max Kappa	Spearman (C)	−,.01	−,.01	−,.01	+,.01
Test AUC	Spearman (F, *N*)	+,.10	+,.11	+,.11	+,.12
Test AUC	Spearman (F, C)	+,.01		+,.08	+,.01
Test Max TSS	Spearman (F, *N*)	+,.09	+,.10	+,.11	+,.12
Test Max TSS	Spearman (F, C)				+,.01
Test Max Kappa	Spearman (F, *N*)	+,.20	+,.05	+,.11	+,.10
Test Max Kappa	Spearman (F, C)				+,.01

Note: Significant positive correlations are represented by '+' and green cell colour, negative correlations by '−' and pink cell colour. Numbers indicate r^2 values for each regression. Variables accompanied by (F) indicate that they were measured on models projected across 24 future climate scenarios. Variables with (N) and(C) indicated models projected within the species native range or at a continental scale, respectively. Results are presented separately for four model sets: the 'simple' set of predictors (two variables in the true niche, four predictors per model, 100 km buffer), the 'complex' set of predictors (3 variables in the true niche, 19 predictors per model, 100 km buffer), the 'large background' study region (same simulation settings as 'complex' but with a 1,000 km buffer) and the 'geographically structured' model set, for which models were constructed and evaluated using geographically partitioned data (same simulations settings as 'complex' but with geographic partitioning of data instead of random holdouts).

for applications where models are thresholded, or where suitability scores are used to prioritize one habitat patch over another. Rank correlation will reflect this when models produce similar rankings of relative habitat suitability (e.g., $\rho = 1$ when predictions made from one model are a monotonically increasing function of predictions made from another model). In contrast, Pearson product moment correlation will only assign a value of 1 if the relationship between suitability scores for the two models is linear, which may serve to exaggerate differences that are not relevant to many empirical studies. In order to test the sensitivity of our results to this choice, we also conducted a separate set of analyses using Pearson product moment correlation as a measure of functional accuracy (Appendix S4), but results from these analyses were effectively the same as those seen in Tables 1 and 2.

We choose to focus on functional accuracy here instead of calibration for several reasons; first, the application of SDMs more often relies on the relative suitability of habitat than estimating the exact probability of observing a species in a particular place, and functional accuracy more directly estimates this aspect of model behaviour. Second, it is already known that discrimination accuracy may be poorly correlated with calibration even when the model gets the relative ranking of habitat right. For example, if the estimated suitability of habitat is a transformation of the true suitability of habitat that preserves relative rankings but not the magnitude of differences in suitability scores (Reineking & Schröder, 2006), both discrimination and functional accuracy would be high, but calibration would be poor. Finally, the link between discrimination accuracy and calibration is known to be severely affected by prevalence (Elith & Graham, 2009; Reineking & Schröder, 2006), but the link between discrimination accuracy and functional accuracy as measured here would not be so affected. Further studies exploring calibration measurements as an alternative to discrimination accuracy for model selection are already underway.

To summarize the relationships between discrimination and functional accuracy for all algorithms considered together (Table 1 and Appendices S3 and S4), we used generalized linear mixed models, and evaluated correlations using McFadden's pseudo-r^2 (McFadden, 1974). For the remainder of the regressions, we used linear models and the standard coefficient of determination, r^2. We applied Bonferroni corrections to compensate for problems arising from multiple testing. For these purposes we defined four families of test that we consider independently. Those examining the relationship between discrimination and functional accuracy at each combination of algorithm and complexity level (Tables 1 and 2, $n = 12$ comparisons per set), and the remainder, which are intended primarily to examine which factors impact overall model quality and as a check to establish that expected relationships between metrics are seen in the simulation results (Appendix S3, $n = 11$).

Based on results from experiments 1–4, we performed a fifth set of simulation experiments to examine more thoroughly the effects of niche complexity and the number of predictor variables on the relationship between discrimination and functional accuracy (Figure 4). Due to the computational intensity of some SDM algorithms, we restricted analyses to a simpler set of conditions for these simulations. Presence data were generated with no non-target spatial biological processes and no spatial sampling bias, so occurrence points were sampled across the entirety of the suitable habitat. We restricted the modelling process to GLM and Maxent, and only used AUC for evaluating predictions on randomly withheld test data. Simulated niches were built using a number of environmental variables ranging between 1 and 19, and models were inferred with between 1 and 19 variables (2–19 for Maxent, due to issues with the software implementation), subject to the constraint that variables that were included within the species' niche were selected first during model construction. For each combination of number of niche axes and environmental predictors, we performed 300 separate simulations, resulting in 108,300 total simulations per modelling method. For each simulation we recorded the test AUC and the rank correlation between the inferred and true suitability of habitat. For each combination of number of niche axes and predictors, we then measured the rank correlation between discrimination accuracy and functional accuracy across the set of 300 models. This resulted in a metric ranging from 1, where test AUC was a perfect indicator of functional accuracy, to −1, where test AUC was negatively associated with functional accuracy. We fit a linear model to these results which included the number of variables used for the simulated niche, the number of variables used for model construction, and an interaction term.

3 | RESULTS

Regression outputs for experiments 1–4 are summarized in Tables 1 (algorithms pooled) and 2 (algorithms analyzed separately), and also in Appendices S3, S4, and S5. TSS, kappa, and AUC were all highly correlated with each other, so we will not discuss them separately. We found that discrimination accuracy on training and test data were correlated, and that functional accuracy in the training region was correlated with functional accuracy outside the training region (Table S3.1). This indicates that models that perform well at discrimination accuracy tend to do so regardless of whether it is measured on training or test data, and the same is true of models that perform well at functional accuracy.

3.1 | Functional accuracy of models

Functional accuracy was generally fairly good; a majority of models produced estimates of habitat suitability that were positively correlated with the true suitability of habitat, whether measured in the training region (73.8%), or projected to the continental scale (80.8%) (figures given here are for all experiments pooled together; plots for each experiment separately are presented in Appendix S5). Models performed somewhat worse when projected into future climate scenarios (65.7% were positively correlated with true suitability within the species range, 71.0% at the continental scale, Appendix S6).

TABLE 2 Relationship between discrimination accuracy and functional accuracy, methods considered separately

		Simple							Complex						
Independent variable	Dependent variable	BC	DM	GAM	GLM	MX	RF	BRT	BC	DM	GAM	GLM	MX	RF	BRT
Test AUC	Spearman (*N*)				+,.15	+,.06	+,.05	+,.05	−,.05	−,.08		+,.06	+,.06	+,.06	+,.05
Test AUC	Spearman (C)			+,.08	+,.07				−,.10				+,.05		
Test Max TSS	Spearman (*N*)				+,.14	+,.04			−,.04	−,.1		+,.04	+,.04	+,.04	+,.05
Test Max TSS	Spearman (C)	−,.05		+,.07	+,.07				−,.11				+,.04		
Test Max Kappa	Spearman (*N*)				+,.11				−,.07	−,.10					
Test Max Kappa	Spearman (C)	−,.06			+,.04				−,.14						
Test AUC	Spearman (F, *N*)	+,.06		+,.17	+,.14	+,.18	+,.08	+,.15	+,.10	+,.05	+,.14	+,.06	+,.26	+,.11	+,.11
Test AUC	Spearman (F, C)			+,.09	+,.06	+,.04							+,.05		
Test Max TSS	Spearman (F, *N*)	+,.05		+,.14	+,.14	+,.15	+,.05	+,.12	+,.09	+,.05	+,.14	+,.05	+,.23	+,.08	+,.09
Test Max TSS	Spearman (F, C)			+,.07	+,.05								+,.04		
Test Max Kappa	Spearman (F, *N*)			+,.10	+,.10	+,.11		+,.06	+,.08	+,.04	+,.07	+,.05	+,.14		
Test Max Kappa	Spearman (F, C)			+,.04											
		Large background							Geographic partitioning						
		BC	DM	GAM	GLM	MX	RF	BRT	BC	DM	GAM	GLM	MX	RF	BRT
Test AUC	Spearman (*N*)				+,.08	+,.08	+,.05	+,.07	+,.06	+,.02	+,.05	+,.09	+,.14	+,.07	+,.17
Test AUC	Spearman (C)					+,.17								+,.04	
Test Max TSS	Spearman (*N*)				+,.08	+,.08	+,.05	+,.07	+,.06		+,.06	+,.09	+,.12	+,.07	+,.19
Test Max TSS	Spearman (C)	−,.04		−,.05		+,.18									
Test Max Kappa	Spearman (*N*)				+,.05	+,.07		+,.04	+,.05			+,.09	+,.09	+,.06	+,.13
Test Max Kappa	Spearman (C)	−,.12		−,.08		+,.10	+,.07								
Test AUC	Spearman (F, *N*)	+,.06			+,.10	+,.25	+,.16	+,.21			+,.10	+,.11	+,.16	+,.19	+,.21
Test AUC	Spearman (F, C)					+,.25									
Test Max TSS	Spearman (F, *N*)	+,.05			+,.09	+,.26	+,.18	+,.23			+,.12	+,.12	+,.17	+,.20	+,.19
Test Max TSS	Spearman (F, C)					+,.27							+,.04		
Test Max Kappa	Spearman (F, *N*)	+,.04			+,.06	+,.23	+,.13	+,.16			+,.10	+,.12	+,.13	+,.15	+,.15
Test Max Kappa	Spearman (F, C)					+,.17	+,.06								

Note: Significant positive correlations are represented by '+' and green cell colour, negative correlations by '−' and pink cell colour. Numbers indicate r^2 values for each regression. Variables accompanied by (F) indicate that they were measured on models projected across 24 future climate scenarios. Variables with (N) and(C) indicated models projected within the species native range or at a continental scale, respectively. Results are presented separately for four model sets: the 'simple' set of predictors (2 variables in the true niche, 4 predictors per model, 100 km buffer), the 'complex' set of predictors (3 variables in the true niche, 19 predictors per model, 100 km buffer), the 'large background' study region (same simulation settings as 'complex' but with a 1,000 km buffer) and the 'geographically structured' model set, for which models were constructed and evaluated using geographically partitioned data (same simulations settings as 'complex' but with geographic partitioning of data instead of random holdouts).

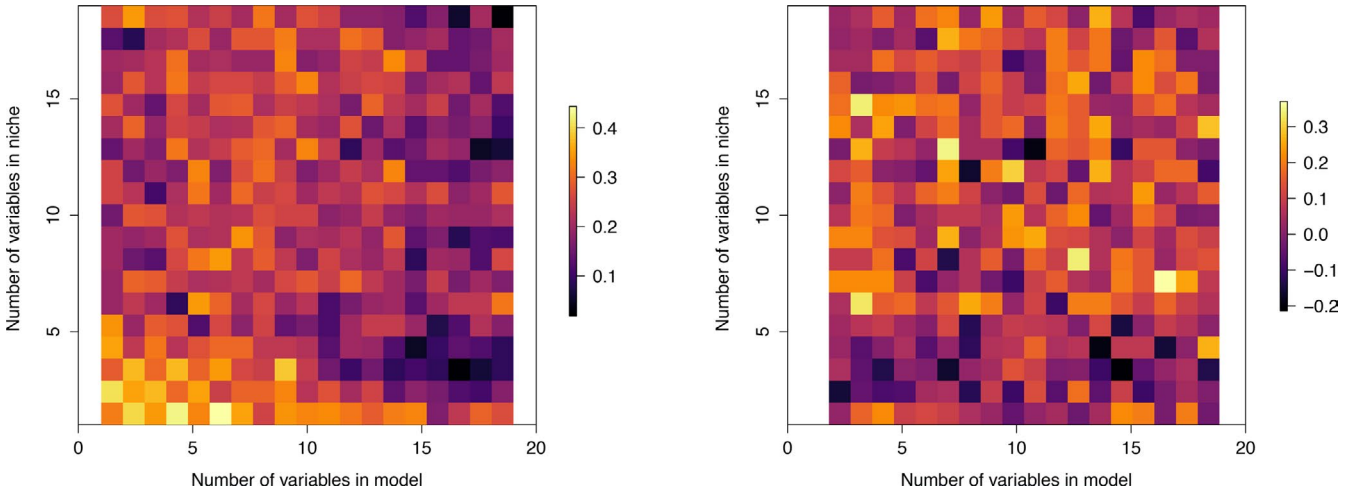

FIGURE 4 Relationship between number of variables in simulated niche, number of variables in model, and the ability of discrimination to infer functional accuracy for GLM (left) and Maxent (right). Each grid cell represents the output of 300 simulations. The colour of each grid cell represents the Spearman rank correlation between test AUC values and functional accuracy

3.2 | Discrimination capacity as a predictor of functional accuracy

When all algorithms were analysed together in a single GLM, discrimination accuracy was a very poor predictor of functional accuracy in all cases (Table 1). Although 31/48 regressions were statistically significant, five were negative correlations, and none had an r^2 value >.2. This indicates poor performance of discrimination accuracy at selecting models when comparing between algorithms.

Results of regressions conducted for each algorithm separately are presented in Table 2. For all experiments, we find that discrimination accuracy is uninformative or actively misleading about models' functional accuracy in a majority of cases (significant positive correlations were seen for less than half of comparisons in any simulation experiment). A majority of these correlations were also quite weak; the average r^2 value was .08 (range −.15 to .26). This indicates poor performance of discrimination accuracy at selecting models with high functional accuracy even when comparing models that were built using the same methods. Discrimination accuracy had no negative correlations with functional accuracy when test data were chosen based on a geographic partition of the species' range, but was still a poor predictor of functional accuracy (average r^2 = .12).

We note an interesting phenomenon here with respect to the size of the buffer regions used to draw background data for model fitting and evaluation; the models built using the largest (1,000 km) buffers around occurrence points performed very well, with the highest levels of functional accuracy and discrimination accuracy (Appendix S5). These differences were most prominent for discrimination accuracy, reinforcing previous findings showing that discrimination accuracy is sensitive to study area (Lobo et al., 2008). Some previous studies have suggested that the models perform best when constructed using fairly small study regions, however, those studies have largely assessed model quality via discrimination accuracy within the species' native range (Acevedo et al., 2012; Zhu, Rédei, Kment, & Bu, 2014). These results indicate that the relationship between study area size and model performance may be more complex than previously reported, and optimal choices may depend on the applications for which models are designed.

3.3 | Model complexity and the relationship between discrimination capacity and functional accuracy

Our fifth experiment examined the effects of complexity across a broader range of complexity values, and found that the ability of AUC to select GLM models with high functional accuracy was negatively correlated with the complexity of the simulated niche and the number of predictor variables. For Maxent models the relationship between discrimination accuracy and functional accuracy (Figure 4) was positively but weakly correlated (r^2 = .09) with the number of variables in the true niche, but uncorrelated with the number of variables included in the modelling process.

4 | DISCUSSION

4.1 | Implications of the inability of discrimination capacity to select models with high functional accuracy

SDM methods are used for many applications in which niche estimates are needed but experimental approaches are impractical. Results for experiments 1–4 demonstrate that many of these models can provide useful estimates of the relative suitability of habitat, the ability of species to invade new areas, and the effects of climate change. However, one of the key steps in any modelling study is the identification of which models from a candidate set perform well

and which perform poorly. Our results indicate that the most widely used methods for selecting models are largely uninformative for studies where the goal is to make continuous estimates of habitat suitability, or to estimate the species' response to an environmental gradient. When algorithms were analyzed separately, 15/149 statistically significant correlations between discrimination and functional accuracy were negative. In these cases discrimination metrics were not just uninformative, but in fact positively misleading for applications where the goal of SDM is to predict the relative suitability of habitat.

In our fifth experiment we examined the effects of niche and model complexity for GLM models across a broader set of conditions, and found that discrimination accuracy predicts functional accuracy only when both the niche and the environmental space it is being modelled in are far simpler than those seen typically in the empirical literature (Figure 4). Even at low levels of complexity, the relationship between discrimination and functional accuracy for GLM is fairly weak (Spearman rank correlation = .31 for a single niche axis and predictor variable), and declines rapidly as models become more complex, becoming minimally informative as models approach levels of complexity that are often seen in the empirical literature.

For Maxent models the number of predictors used to model the niche had no effect on the utility of discrimination accuracy for model selection, but there was a weak positive effect of the number of variables used to simulate the true niche ($\beta = 0.006$, $p < .05$). We hypothesize that the lack of effect of the number of predictors for Maxent is due to its ability to automatically penalize overparameterization; many of the predictors supplied to the algorithm may ultimately have little or no weight in the model. We also note that the most reliable correlations between discrimination accuracy and functional accuracy seen in our simulation results were for Maxent models (Table 2), as would be expected if model complexity and number of predictors were partly responsible for driving the poor performance of discrimination accuracy.

Discrimination accuracy was generally a better predictor of functional accuracy for GLM, GAM, and Maxent models than for the other methods of model construction explored in this study. This is likely due to the internal structure of these models. The simulation approach taken here uses a logistic function to generate sampling probabilities based on a simulated niche, which is composed of smooth (linear or quadratic) responses to a set of environmental variables. As such, the function underlying habitat suitability lies within the set of functions that may be exactly estimated by GLM, GAM or Maxent, so that estimation of simulated niches is considerably more tractable for those methods. We therefore caution users to refrain from interpreting these results as an endorsement of any particular method when constructing SDMs using empirical data. Rather, we suggest that these results indicate that choice of modelling methods should ideally include intuition or data regarding the potential functional relationship between the environmental predictors and the suitability of habitat. If the functional relationships that may be estimated by an algorithm differ significantly from the true functional relationship, discrimination accuracy is largely uninformative or misleading about models' ability to predict habitat suitability. This does not necessarily imply that models built using different functional shapes from the true niche are poor estimates of habitat suitability; rather it indicates that discrimination accuracy is uninformative for selecting models with high functional accuracy under these conditions.

Our results clearly indicate that most empirical studies using SDM methods should ideally not rely solely on prediction of withheld occurrence data to assess model quality. However, they also indicate a much more systemic problem for the SDM literature: decades of methodological work in this field have resulted in a set of widely-adopted 'best practices', but a great majority of these studies have focused on optimizing models' discrimination accuracy on withheld occurrence data from real species distributions (Boria et al., 2014; Domisch et al., 2013; Garcia-Callejas & Araujo, 2016; Guisan et al., 2007; Huang & Frimpong, 2016; Kuebler et al., 2016; Lopatin, Dolos, Hernandez, Galleguillos, & Fassnacht, 2016; Moreno-Amat et al., 2015; Radosavljevic & Anderson, 2014; Rovzar, Gillespie, & Kawelo, 2016; Soley-Guardia et al., 2016; Wisz et al., 2008). Given the disconnect seen here between discrimination and functional accuracy, it is entirely possible that the 'best practices' advocated in these studies have negligible, or even detrimental, effects on model quality for applications where functional accuracy is the goal.

In order to accurately assess the ability of different methods to achieve useful levels of functional accuracy, we argue that the methodological literature must reevaluate its 'best practices' via simulations where true habitat suitability and niche parameters are known. While some simulation studies are already being conducted (Meynard et al., 2019), these have typically been done in the context of optimizing discrimination accuracy, and as such may also be largely uninformative about estimating habitat suitability as a function of environmental gradients. There are many common practices and assumptions in the field that may need to be reevaluated based on their ability to estimate habitat suitability; choice of algorithm, methods for choosing predictor variables, choice of study area, rarefaction of data, and optimal model complexity are obvious candidates.

In addition, we argue that practitioners must recognize that favouring models based strictly on their spatial predictions is simply inappropriate for many applications. In studies where the goal is to estimate the niche (i.e., maximize functional accuracy), users must become comfortable with the idea that a biologically accurate model may produce relatively poor estimates of species' current spatial distributions. This is not simply a methodological point brought to light by the current simulation study; it is necessarily true given the existence of non-target phenomena that themselves have spatial structure (e.g., biotic interactions, dispersal). This has been known for years (Anderson, 2012; Jackson & Overpeck, 2000; Soberon & Peterson, 2005; Warren, 2012, 2013), yet has been largely ignored in the continued pursuit of methods

that produce tighter and tighter fits to training or test data in geographic space.

4.2 | Generalizability of results

Investigators familiar with SDM methods will no doubt wish to critically examine the methods used here to infer models; there are other algorithms available, and there are many modelling choices that we did not explore in great depth. However, these criticisms are largely irrelevant to the primary results of this study; while it is certainly possible that greater effort in exploring the space of model choices might improve the accuracy of models, we note that (a) evaluation metrics on randomly withheld test data for the models generated here are not unusual for the range seen in the empirical SDM literature (e.g., Appendix S5), (b) the overall performance of SDM methods is irrelevant to whether or not discrimination accuracy is a valid indicator of functional accuracy, and (c) most SDM users' methodological preferences are currently chosen based on studies that seek to maximize the very performance metrics that the current study demonstrates are not useful for estimating functional accuracy.

We acknowledge the possibility that there is some subset of modelling approaches not addressed here for which discrimination and functional accuracy are highly correlated. It would be both gratifying and very useful to find such a set of conditions, and that topic deserves to be examined in great depth. However, even if such a set can be found it does not invalidate the conclusion presented here; that there is a large range of modelling algorithms and approaches for which the correlation between discrimination accuracy and functional accuracy is not strong enough to be useful in model selection for many purposes. Similarly, we acknowledge that the disconnect between functional accuracy and discrimination seen here may be affected by sample size, but the sample sizes used here (75 training, 25 test) are not atypical for the ENM literature.

4.3 | Concluding remarks and future directions

In summary, we demonstrate that, under a broad range of conditions, the ability of a model to successfully predict withheld occurrence data within the training region does not reliably measure its ability to estimate the relationship between environmental gradients and habitat suitability. Discrimination accuracy may be a reasonable metric when the goal is to guide further sampling of occurrences within a species' current range, without regard for whether the model estimates the true environmental niche or the relative suitability of habitat well. However, this is not often the goal of empirical model construction in the SDM literature.

As a result, the applied and methodological literature in this field are largely based on metrics that may be irrelevant to the intended applications of many models. If the field is to continue to attempt to use SDMs to infer species' responses to environmental gradients, we must develop methods for model construction and metrics for model evaluation that are more relevant to the actual goals of the modelling process. While we find that geographically structured partitioning of test data does offer some advantages over randomly withheld data, it is clear from this study that even those methods have very limited ability to identify models that accurately estimate the relative suitability of habitat.

We would like to particularly highlight the implications of our results for the development of new methods in this field in the coming years. Many investigators are currently developing methods that incorporate more biological and statistical realism into the SDM process, including the integration of physiological and trait data (Pollock et al., 2018) and explicit models of bias (Robinson, Ruiz-Gutierrez, & Fink, 2018), dispersal (Zurell, 2017), plasticity (Bush et al., 2016) and evolutionary history (Smith, Godsoe, Rodriguez-Sanchez, Wang, & Warren, 2019). In any system affected by non-target spatial phenomena, these methods will often produce poorer estimates of species' geographic distributions precisely because they provide better estimates of the environmental niche. We hope that the results presented here will compel the field to evaluate these new methods based on their ability to infer the biological phenomena of interest, as demonstrated using simulations or physiological data, rather than simply reject them due to poor discrimination accuracy on misleading occurrence data.

We feel it is necessary to specifically address one interpretation of these results that we feel is not appropriate: the work presented here is not intended to suggest that any particular method of SDM construction is inherently better or worse than others. While the relative performance of different methods is a very interesting question and one that deserves further exploration within a simulation framework, this study was not designed to address those questions and it would be inappropriate to interpret these results as such. We emphasize that most of the models built from these simulated species were arguably publishable distribution estimates, and were at least somewhat useful as estimates of the species' niche. Rather, this study is intended to examine the performance of widely used methods of model selection, and it is those methods that are performing poorly. We demonstrate that we can make both good distribution estimates and good niche estimates using common methods, and in fact produced many models that are good for both purposes. However, our results indicate that we have a difficult time distinguishing good models from bad when our goal is functional accuracy.

At minimum, our results suggest that any empirical study using discrimination accuracy to assess model quality should start with two crucial steps: (a) use a minimal set of predictor variables for which there is an a priori reason to expect that they limit the suitability of habitat for the species, and (b) select algorithms capable of inferring functional responses that are plausible estimates of the underlying biology (e.g., not using a step function in situations where suitability is expected to be a continuous function of the predictor variable). In a sense, these findings are unsurprising; they recapitulate longstanding best practices in the broader literature regarding statistical modelling (Anderson & Burnham, 2004; Burnham & Anderson, 2004; Gelman & Hill, 2006; Zuur, Ieno, Walker, Saveliev,

& Smith, 2009). However, here we show that failure to make these choices appropriately does not necessarily lead to poor predictions; instead it means that we are largely unable to distinguish good models from bad using species occurrence data. Under these conditions any preference for a given model based on discrimination accuracy may be little better than choosing a model at random.

ACKNOWLEDGEMENTS

This work would not have been possible without the financial contributions of the Macquarie University Department of Biology and a DECRA award to D.L.W. from the Australian Research Council. The authors sincerely thank Jorge Lobo and Boris Leroy for their valuable comments on this manuscript, as well as the contributions of two anonymous reviewers.

DATA AVAILABILITY

Sample code is available on github here: https://doi.org/10.5061/dryad.6ft55k9

ORCID

Dan L. Warren https://orcid.org/0000-0002-8747-2451
Nicholas J. Matzke https://orcid.org/0000-0002-8698-7656
Teresa L. Iglesias https://orcid.org/0000-0002-0237-8539

REFERENCES

Acevedo, P., Jimenez-Valverde, A., Lobo, J. M., & Real, R. (2012). Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, *39*(8), 1383–1390. https://doi.org/10.1111/j.1365-2699.2012.02713.x

Allen, J. L., & Lendemer, J. C. (2016). Climate change impacts on endemic, high-elevation lichens in a biodiversity hotspot. *Biodiversity and Conservation*, *25*(3), 555–568. https://doi.org/10.1007/s10531-016-1071-4

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Anderson, D., & Burnham, K. (2004). *Model selection and multi-model inference*. Second (p. 63). NY: Springer-Verlag.

Anderson, R. P. (2012). Harnessing the world's biodiversity data: Promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, *1260*, 66–80.

Bahn, V., & McGill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*(6), 733–742. https://doi.org/10.1111/j.1466-8238.2007.00331.x

Bahn, V., & McGill, B. J. (2013). Testing the predictive performance of distribution models. *Oikos*, *122*(3), 321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Bush, A., Mokany, K., Catullo, R., Hoffmann, A., Kellermann, V., Sgrò, C., … Ferrier, S. (2016). Incorporating evolutionary adaptation in species distribution modelling reduces projected vulnerability to climate change. *Ecology Letters*, *19*(12), 1468–1478. https://doi.org/10.1111/ele.12696

Carrascal, L. M., Aragon, P., Palomino, D., & Lobo, J. M. (2015). Predicting regional densities from bird occurrence data: Validation and effects of species traits in a Macaronesian Island. *Diversity and Distributions*, *21*(11), 1284–1294. https://doi.org/10.1111/ddi.12368

Chamberlain, S., Boettiger, C., Ram, K., Barve, V., & Mcglinn, D. (2013). rgbif: Interface to the Global Biodiversity Information Facility API. R package version 0.4.0. Retrieved from http://cran.r-project.org/package=rgbif

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cook, W. C. (1925). The distribution of the alfalfa weevil (*Phytonomus posticus* Gyll.). A study in physical ecology. *Journal of Agricultural Research*, *30*, 479–491.

Coro, G., Pagano, P., & Ellenbroek, A. (2013). Combining simulated expert knowledge with Neural Networks to produce Ecological Niche Models for *Latimeria chalumnae*. *Ecological Modelling*, *268*, 55–63.

Domisch, S., Kuemmerlen, M., Jahnig, S. C., & Haase, P. (2013). Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. *Ecological Modelling*, *257*, 1–10.

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, *32*(1), 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x

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

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. https://doi.org/10.1017/S0376892997000088

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Garcia-Callejas, D., & Araujo, M. B. (2016). The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling*, *326*, 4–12.

GBIF.org (27th September 2015) GBIF Occurrence Download https://doi.org/10.15468/dl.gsmfd3

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gomes, V. H., IJff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., de Souza Coelho, L., … López, D. C. (2018). Species distribution modelling: Contrasting presence-only models with plot abundance data. *Scientific Reports*, *8*(1), 1003.

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

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge, UK: Cambridge University Press.

Gutierrez-Tapia, P., & Palma, R. E. (2016). Integrating phylogeography and species distribution models: Cryptic distributional responses to past climate change in an endemic rodent from the central Chile hotspot.

Diversity and Distributions, *22*(6), 638–650. https://doi.org/10.1111/ddi.12433

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

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2012). dismo: Species distribution modeling. R package version 0.7-23. Retrieved from http://cran.r-project.org/web/packages/dismo/index.html

Huang, J., & Frimpong, E. A. (2016). Limited transferability of stream-fish distribution models among river catchments: Reasons and implications. *Freshwater Biology*, *61*(5), 729–744. https://doi.org/10.1111/fwb.12743

Jackson, S. T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, *26*(4), 194–220. https://doi.org/10.1017/S0094837300026932

Jimenez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., & Real, R. (2013). Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Global Ecology and Biogeography*, *22*(4), 508–516. https://doi.org/10.1111/geb.12007

Kearney, M. (2006). Habitat, environment and niche: What are we modelling? *Oikos*, *115*(1), 186–191. https://doi.org/10.1111/j.2006.0030-1299.14908.x

Keith, D. A., Mahony, M., Hines, H., Elith, J., Regan, T. J., Baumgartner, J. B., … Akcakaya, H. R. (2014). Detecting extinction risk from climate change by IUCN red list criteria. *Conservation Biology*, *28*(3), 810–819. https://doi.org/10.1111/cobi.12234

Kriticos, D. J., Webber, B. L., Leriche, A., Ota, N., Macadam, I., Bathols, J., & Scott, J. K. (2012). CliMond: Global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods in Ecology and Evolution*, *3*(1), 53–64. https://doi.org/10.1111/j.2041-210X.2011.00134.x

Kuebler, D., Hildebrandt, P., Guenter, S., Stimm, B., Weber, M., Mosandl, R., … Silva, B. (2016). Assessing the importance of topographic variables for the spatial distribution of tree species in a tropical mountain forest. *Erdkunde*, *70*(1), 19–47. https://doi.org/10.3112/erdkunde.2016.01.03

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2015). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Lezama-Ochoa, N., Murua, H., Chust, G., Van Loon, E., Ruiz, J., Hall, M., … Villarino, E. (2016). Present and future potential habitat distribution of *Carcharhinus falciformis* and *Canthidermis maculata* by-catch species in the tropical tuna purse-seine fishery under climate change. *Frontiers in Marine Science 3*. https://doi.org/10.3389/fmars.2016.00034

Lobo, J. M., Jimenez-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

Lopatin, J., Dolos, K., Hernandez, H. J., Galleguillos, M., & Fassnacht, F. E. (2016). Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. *Remote Sensing of Environment*, *173*, 200–210.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

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

Meynard, C. N., Leroy, B., & Kaplan, D. M. (2019). Testing methods in species distribution modelling using virtual species: What have we learnt and what are we missing? *Ecography*. https://doi.org/10.1111/ecog.04385

Moreno-Amat, E., Mateo, R. G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J. C., & Garcia-Amorena, I. (2015). Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, *312*, 308–317.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, *7*(4), 435–455. https://doi.org/10.1016/0169-2070(92)90028-8

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.

Nix, H. A. (1986). Biogeographic analysis of Australian elapid snakes. In R. Longmore (Ed.), *Atlas of elapid snakes of Australia* (pp. 4–15). Canberra: R. Longmore. Australian Government Publishing Service.

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 (MPB-49)*. Princeton, NJ: Princeton University Press.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259.

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

Pollock, L. J., Kelly, L. T., Thomas, F. M., Soe, P., Morris, W. K., White, M., & Vesk, P. A. (2018). Combining functional traits, the environment and multiple surveys to understand semi-arid tree distributions. *Journal of Vegetation Science*, *29*(6), 967–977. https://doi.org/10.1111/jvs.12686

Radosavljevic, A., & Anderson, R. P. (2014). Making better MAXENT models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643.

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

Raghavan, R. K., Goodin, D. G., Hanzlicek, G. A., Zolnerowich, G., Dryden, M. W., Anderson, G. A., & Ganta, R. R. (2016). Maximum entropy-based ecological niche model and bio-climatic determinants of Lone Star Tick (*Amblyomma americanum*) Niche. *Vector-Borne and Zoonotic Diseases*, *16*(3), 205–211.

Reineking, B., & Schröder, B. (2006). Constrain to perform: Regularization of habitat models. *Ecological Modelling*, *193*(3–4), 675–690.

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, *24*(4), 460–472. https://doi.org/10.1111/ddi.12698

Rovzar, C., Gillespie, T. W., & Kawelo, K. (2016). Landscape to site variations in species distribution models for endangered plants. *Forest Ecology and Management*, *369*, 20–28.

Smith, A. B., Godsoe, W., Rodriguez-Sanchez, F., Wang, H. H., & Warren, D. (2019). Niche estimation above and below the species level. *Trends in Ecology & Evolution*, *34*(3), 260–273. https://doi.org/10.1016/j.tree.2018.10.012

Soberon, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species distributional areas. *Biodiversity Informatics*, *2*(0). https://doi.org/10.17161/bi.v2i0.4

Soley-Guardia, M., Gutierrez, E. E., Thomas, D. M., Ochoa-G, J., Aguilera, M., & Anderson, R. P. (2016). Are we overestimating the niche?

Removing marginal localities helps ecological niche models detect environmental barriers. *Ecology and Evolution*, *6*(5), 1267–1279. https://doi.org/10.1002/ece3.1900

Sutherst, R. W. (2014). Pest species distribution modelling: Origins and lessons from history. *Biological Invasions*, *16*(2), 239–256. https://doi.org/10.1007/s10530-013-0523-y

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Torres, L. G., Sutton, P. J. H., Thompson, D. R., Delord, K., Weimerskirch, H., Sagar, P. M., … Phillips, R. A. (2015). Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across ocean basins. *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0120014

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

VanDerWal, J., Shoo, L. P., Johnson, C. N., & Williams, S. E. (2009). Abundance and the environmental niche: Environmental suitability estimated from niche models predicts the upper limit of local abundance. *American Naturalist*, *174*(2), 282–291. https://doi.org/10.1086/600087

Varela, S., Anderson, R. P., Garcia-Valdes, R., & Fernandez-Gonzalez, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

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. (2012). In defense of 'niche modeling'. *Trends in Ecology & Evolution*, *27*(9), 497–500. https://doi.org/10.1016/j.tree.2012.03.010

Warren, D. L. (2013). 'Niche modeling': That uncomfortable sensation means it's working. A reply to McInerny and Etienne. *Trends in Ecology & Evolution*, *28*(4), 193–194. https://doi.org/10.1016/j.tree.2013.02.003

Warren, D. L., Cardillo, M., Rosauer, D. F., & Bolnick, D. I. (2014). Mistaking geography for biology: Inferring processes from species distributions. *Trends in Ecology & Evolution*, *29*(10), 572–580. https://doi.org/10.1016/j.tree.2014.08.003

Warren, D. L., Wright, A. N., Seifert, S. N., & Shaffer, H. B. (2014). Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions*, *20*(3), 334–343. https://doi.org/10.1111/ddi.12160

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A.; N. P. S. D. W. Group (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Zhu, G.-P., Rédei, D., Kment, P., & Bu, W.-J. (2014). Effect of geographic background and equilibrium state on niche model transferability: Predicting areas of invasion of *Leptoglossus occidentalis*. *Biological Invasions*, *16*(5), 1069–1081. https://doi.org/10.1007/s10530-013-0559-z

Zurell, D. (2017). Integrating demography, dispersal and interspecific interactions into bird distribution models. *Journal of Avian Biology*, *48*(12), 1505–1516. https://doi.org/10.1111/jav.01225

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer Science & Business Media.

BIOSKETCHES

Dan L. Warren is a senior scientist at the Senckenberg Biodiversity and Climate Research Centre in Frankfurt, Germany. He is broadly interested in issues at the interface between evolution, ecology and conservation biology. **Nicholas J. Matzke** is a Senior Lecturer in the School of Biological Sciences at the University of Auckland. He works on likelihood and Bayesian methods in biogeography. He is also the author of the R package 'BioGeoBEARS'. **Dr. Teresa L. Iglesias** is currently a postdoc at the Okinawa Institute of Science and Technology (OIST) in Japan. Her research interests include the neurobiology of animal behaviour and the intersection between behaviour, evolution and conservation.

Author contributions: All authors contributed to study design and manuscript preparation. Simulations were coded and run by D.L.W., literature review was performed by T.L.I. Post hoc analyses were performed by T.L.I. and D.L.W.

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: Warren DL, Matzke NJ, Iglesias TL. Evaluating presence-only species distribution models with discrimination accuracy is uninformative for many applications. *J Biogeogr*. 2020;47:167–180. https://doi.org/10.1111/jbi.13705