Testing the ability of species distribution models to infer variable importance

Adam B. Smith[1] & Maria J. Santos[2]

[1] Center for Conservation and Sustainable Development, Missouri Botanical Garden, Saint Louis MO 63116 USA. Email: adam.smith@mobot.org

[2] University Research Priority Program in Global Change and Biodiversity and Department of Geography, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

Abstract [350 word limit]

Identifying biophysical factors that define species' niches and influence geographical ranges is a fundamental pursuit of ecology. Frequently, models of species' distributions or niches are used to infer the importance of range- and niche-defining variables. However, very few—if any—studies examine how reliably distribution and niche models can be used for inference. Here we use a simulation approach to understand the conditions under which species distribution models reliably measure variable importance. Using a set of scenarios of increasing complexity, we explore how well models can be used to 1) discriminate between variables that vary in importance and 2) calibrate the effect of variables relative to an "omniscient" model used to simulate the species. Variable importance was assessed using a sensitivity test in which each predictor was permuted in turn. Importance was inferred by comparing model performance between permuted and unpermuted predictions and by calculating the correlation between permuted and unpermuted predictions. Of five metrics of importance (correlation statistic and AUC each calculated with presences/absences or presences/background sites, plus the Continuous Boyce Index), only the Continuous Boyce Index was capable of indicating absolute (versus relative) variable importance. In simple scenarios with one influential environmental variable with a linear spatial gradient and one uninfluential randomly-distributed variable, models were unable to discriminate reliably between variables under conditions that are typically challenging (low sample size, high prevalence, small spatial extent, coarse spatial data resolution with low spatial autocorrelation, and high collinearity between variables). In more complex scenarios with two influential environmental variables, models successfully discriminated between variables when they acted unequally, but overestimated the importance of the stronger variable and underestimated the importance of the weaker variable. When variables had equal influence, models underestimated importance when niche breadth was narrow. Generalized additive models and Maxent had better discrimination accuracy than boosted regression trees. Our work demonstrates that permutation tests can reliably discriminate between variables with different levels of influence but cannot accurately measure the magnitude of influence. The frequency with which distribution and niche models are used to identify influential variables begs further research into methods for assessing variable importance.

Introduction

What environmental factors determine species' ranges and environmental tolerances? The answer to this question remains elusive for most species on Earth despite being crucial for addressing long-standing issues in theoretical ecology (e.g., Francis & Currie 2003; Smith 2013a; Barnes & Clark 2017; Angert et al. 2018; Hoban et al. in press) and conservation (e.g., Albrecht et al. 2016; Smith et

al. 2016; Stralberg et al. 2018; Svancara et al. 2019). Although manipulative field- and laboratory-based studies can be used to shed light on the factors that shape ranges and niches (e.g., Hargreaves et al. 2016; Lee-Yaw et al. 2016), the logistical difficulties inherent in executing these studies preclude studying range limits across wide geographic scales for many species. Alternatively, environmental limits can be inferred using sensitivity analysis of geographic range and niche models. These ecological niche models (ENMs) and species distribution models (SDMs) are constructed by correlating observations of occurrence with data on environmental conditions at occupied sites.

Niche and distribution models can be evaluated along several dimensions, one of which is the ability to identify the most important variables and estimate their influence accurately (Araújo et al. 2019). Indeed, identification of important variables is one of the most common uses of ENMs and SDMs (cf. 227 studies analyzed by Bradie & Leung 2017 using just the Maxent algorithm). Knowing which variables shape species' ranges and niches is critical to understanding evolutionary limitations (Araújo et al. 2013; Quintero & Wiens 2013) and modes of speciation (Pyron et al. 2015; Reino et al. 2017), delineating the scale of species' responses to the environment (Graf et al 2005; Franklin et al. 2013; Pradervand et al. 2013; Mertes & Jetz 2017), predicting responses to past and future global change (Quintero & Wiens 2013; Smith 2013a; Feng & Papeş 2014; Morelli et al. 2012 and 2016; Santos et al. 2017), designing appropriate conservation strategies (Smith et al. accepted; Morelli et al. in review), and predicting invasions (Callen & Miller 2015; Petitpierre et al. 2017). Given the frequency with which SDMs and ENMs are used in this regard and the impact these studies have on our understanding of the natural world, there is a pressing need to understand the conditions under which they can identify influential variables. However, we are aware of no studies that evaluate how well SDMs and ENMs actually measure variable importance (i.e., model "realism" *sensu* Araújo et al. 2019).

The ability of an ENM or SDM to infer variable importance will likely be affected by many of the same factors that influence model accuracy such as sample size, study region extent, species' prevalence, spatial autocorrelation in predictors, resolution of environmental data, correlations between variables on the landscape, and interactions between variables in how they define the niche. Although the effect of these factors on variable importance has received some attention (e.g., Menke et al. 2009; Franklin et al. 2013; Pradervand et al. 2013; Mertes & Jetz 2017; Acevedo et al. 2017), studies using real species are often complicated by the fact that the true importance of variables is unknown. Moreover, in real-world situations factors like spatial extent, prevalence, and correlation between variables are often interdependent and so cannot be manipulated in isolation.

The goal of this article is to begin to fill the gap in our knowledge about how ENMs and SDMs can be used to measure variable importance. We utilize an experimental approach based on virtual species. Simulating species' distributions allows us to know the true importance of variables and thus determine the efficacy of our methods (Austin et al. 2006; Zurell et al. 2009; Meynard & Kaplan 2013; Thibaud et al. 2014; Moudrý 2015). The virtual species approach also enables us to isolate specific aspects of study design and the modeling situation that otherwise would be difficult or impossible to change in a real-world setting. In the following section we provide a brief overview of some aspects of modeling practices that influence the ability of models to measure variable

importance. We then present results from 10 "experiments" in which we manipulate practices and common ecological conditions we hypothesized would affect the ability of models to measure variable importance. Finally, we end with a discussion emphasizing future directions. Compared to the abundance of studies assessing the predictive accuracy of models of niches and distributions (e.g., Guisan et al. 1999; Araújo et al. 2005; Elith et al. 2006; Dobrowski et al. 2011; Rapacciuolo et al 2012; Smith et al. 2013; Guevara et al. 2018; Yates et al. 2018; Norberg et al. in press), the lack of research on the efficacy of these same models for identifying important variables is a striking oversight.

Variable selection, niches, and distributions

The present work builds upon yet is distinct from the critical problem of selecting ecologically relevant variables (Austin 2002; Austin & Van Neill 2011; Mod et al. 2016; Petitpierre et al. 2017; Araújo et al. 2019; Gardener et al. 2019). Similarly, this work is related to yet distinct from the practice of model selection (i.e., choosing amongst different models, often containing different sets of variables; e.g., Burnham & Anderson 2002; Warren & Siefert 2011; La Rest et al. 2014; Gobeyn et al. 2017). Variables that are excluded by the modeler initially or during the model process are assumed to be of no importance with zero uncertainty. Thus, both practices are highly relevant to measuring variable importance. However, our focus here is on measuring the influence of variables that are provided to a model, whether or not they "survive" the model selection process or are otherwise downweighted by the model parameterization routine.

The present work also intrinsically relates to yet is generally unaffected by the distinction between niche and distribution models (Peterson et al. 2011). Even though we use the term "landscape" to refer to the space from which training and test samples are obtained, our simulations are typically constructed such that this space could represent accessible environmental (niche) space or accessible geographic space and thereby yield models of the niche or distribution, respectively. Obviously, a key focus for further research is distinguishing under what conditions models can be used to infer the importance of niche-limiting versus distribution-limiting variables (Warren et al. bioRxiv).

Variable importance

We assume a variable is "important" if predictions from a model are highly sensitive to changes in values of the predictor (Smith & Santos in prep.). Sensitivity can be summarized using a test statistic such as a metric of model performance (e.g., the area under the receiver-operator characteristic curve, or AUC). Robust tests of variable importance should have a high degree of *discrimination* (qualitatively differentiate variables with different influence) and *calibration*, the latter of which can be divided into "accuracy" (the central tendency of importance should be similar to the central tendency of importance from a model created with perfect knowledge of a species' requirements) and "precision" (variation around central tendency matches that of a perfect model). In addition, ideal test statistics should be *bounded* within a known range without unknown limits and *proportionate*, meaning that magnitude connotes effect size (e.g., halving the effect of a variable should halve the value of the test statistic; Smith & Santos in prep.). Test statistics can assess absolute importance (i.e., how much information does the variable contribute to predicting the

distribution or niche of the species?) or relative importance (i.e., how much does the variable influence predictions relative to other variables in the model?). Absolute measures can be compared across species and modeling scenarios while relative measures can only be compared within species in the same scenario.

Of the many different test statistics of variable importance (Smith & Santos in prep.), here we focus on the permute-after-calibration test (Breiman 2001). The permute-after-calibration test is algorithm-independent so can be used with any SDM or ENM. To implement the procedure, a model is trained on training data then used to make "control" predictions to test sites. Then, each variable is permuted in turn and the scrambled predictors are used to create "permuted" predictions. Control and permuted predictions are then compared. If the permuted variable has a strong influence on the model, then permuted and control predictions will be very different (i.e., model performance will be substantially lower for the permuted predictions).

The test statistic used to compare permuted and control predictions determines whether the permute-after-calibration test reflects absolute or relative importance. In the main text we use the Continuous Boyce Index (CBI) measure of a model's calibration accuracy as a metric of absolute importance. CBI is the Pearson rank correlation between the ratio of binned frequencies of predictions at test occurrence sites and actual occurrences, and the ranks of the bins (Boyce et al. 2002; Hirzel et al. 2006). CBI reflects a model's calibration accuracy, or how well its predictions serve as an index of the actual probability of presence. Being a correlation coefficient, CBI has the range [-1, 1], with values >0 indicating performance better than random, values ~0 no better than random, and <0 worse. Thus, to conclude that a variable is important in an absolute sense two conditions need to occur. First, the unpermuted "control" predictions must yield a high value of CBI since it makes little sense to claim a variable is influential if models using the (unmanipulated) variable fails to predict the probability of occurrence reliably (Meinshausen & Bühlmann 2010; Smith & Santos in prep.). Second, permuting the variable of interest must substantially reduce CBI. Thus, the degree of absolute importance is given by the difference in performance between the control and permuted model predictions (Meinshausen & Bühlmann 2010). We also calculated variable importance using the permute-after-calibration test with three other test statistics: AUC calculated with presences and absences (AUC_{pa}) or with presences and background sites (AUC_{bg}), and the correlation between predictions of the control model and permuted model calculated with presences and absences (COR_{pa}) or with presences and background sites (COR_{bg}). Results of these test statistics are presented in Appendices 1 through 4.

General approach

To simplify scenarios and help isolate factors affecting assessment of variable importance, unless otherwise noted, we assume an ideal modeling situation: the species is in statistical equilibrium with its environment; sampling is unbiased with perfect detectability; environmental variables are measured without error and the species interacts with the environment at the spatial and temporal scales at which the environmental data is represented; biotic interactions and disturbances do not influence occupancy of the range or niche; and there is no process-based spatial autocorrelation in the distribution of the species.

The lack of process-based autocorrelation means that there is no hidden process in the "generative" model that would mislead a correlative model attempting to estimate the generative model from data. For example, the simulations include no source/sink dynamics that would cause an erstwhile unsuitable site to appear suitable simply because it is populated (Pulliam 2000). Similarly, there are no dispersal lags or disturbances, so there are no regions of suitable habitat that remain unoccupiable (aside from sampling probabilistically from the true probability of presence to generate presence/absence for training and testing the models; Moudrý 2015). Thus, in our simulations the lack of process-based autocorrelation in the species' range and predictors obviates the need to use geographically distinct training and test sites. Our simulations do have incidental spatial autocorrelation: many of the environmental layers follow a linear gradient across the landscape, and as a result the species also has a spatially smoothed distribution that is spatially autocorrelated. However, we could just as well have simulated the environment such that variables and the probability of occurrence had the same univariate and multivariate frequency distribution of values as their smooth versions but distributed in a spatially random manner. In this case neither the environment nor the species' distribution would have had spatial autocorrelation above that expected by random chance, but the results would have been exactly the same (except for *Experiment 5: Spatial resolution and spatial autocorrelation of environmental data* where we explicitly evaluate the effects of spatial autocorrelation). The lack of process-based spatial autocorrelation in our simulations is a convenience that is unlikely to be met in real-world situations where robust inference of variable importance requires using evaluation data that is geographically or temporally as independent as possible from training data (La Rest et al. 2014; Fourcade et al. 2018; Warren et al. in prep.).

We evaluated three modeling algorithms: generalized additive models (GAMs; Wood 2006), Maxent (MAX; Phillips et al. 2006; Phillips & Dudík 2008), and boosted regression trees (BRTs; Elith et al. 2008). We focus on these three algorithms because they are commonly employed, tend to have high accuracy (Elith et al. 2006), have very different approaches to recreating species' response functions (GAMs: smoothed hypervolumes; MAX: smoothed and step-function hypervolumes; BRTs: overlapping rectangular hypervolumes), and represent three very different approaches to modeling (GAMs: semi-parametric likelihood maximization, MAX: maximization of information entropy/deterministic machine learning; and BRTs: stochastic machine learning). Detailed methods on model parameterization are described in the Appendix 5. Briefly stated, we used an AIC_c-based approach to identify optimal GAM models; a grid search and AIC_c to find the optimal master regularization parameter and combination of feature types for MAX (Warren & Siefert 2011); and a grid search for optimal tree depth and learning rate for BRTs.

To simulate species' distributions, we defined a generative function of one or more variables on the landscape, then used it to produce a raster of the probability of occurrence. For each cell, occupancy was determined by a random draw from [0, 1] and designating it as occupied if the value was ≤ the probability of presence (Meynard & Kaplan 2013). Using a probabilistic representation of the range is superior to applying a threshold above which every cell is assumed occupied because it allows us to characterize uncertainty related just to variation of probability of presence in sampling of sites, and because applying a fixed threshold above which every cell is assumed occupied converts smooth responses into step functions (Meynard & Kaplan 2013; Moudrý 2015). We then selected

203 distinct sets of training presences, test presences, and test absences, plus two sets of randomly
204 located background sites for training and testing. For model training we used 200 presences plus
205 10,000 (GAMs and MAX) or 200 (BRTs) randomly selected background sites (Barbet-Massin et al.
206 2012). For model testing we used 200 test presences plus either 10,000 background sites (for CBI,
207 AUC_{bg}, and COR_{bg}) or 200 absences (for AUC_{pa} and COR_{pa}). We generated 100 sets of presences,
208 absences, and background sites for each combination of factors we wished to manipulate (e.g.,
209 differences in prevalence, magnitude of correlation between variables, etc.). Models were trained
210 and tested using each of the data 100 sets. For any given iteration each algorithm was trained and
211 tested against the same set of calibration and evaluation sites. Permutation tests were repeated 30
212 times apiece then averaged for a given treatment (e.g., a specific level of prevalence), set of training
213 and test sites, and algorithm. Landscapes were either square or circular, depending on the scenario.

214 To provide an objective benchmark against which to judge the results from the three SDM
215 algorithms, we used an "omniscient" (OMNI) model which uses the same generative model and
216 parameter values used to create the species' distribution (Meynard et al. In press). We evaluated
217 the OMNI models with the same set of test sites used to evaluate the SDMs. Since test sites are
218 drawn probabilistically from the underlying probability distribution describing the species' actual
219 range, there is variability in the test statistics across iterations using OMNI models even though
220 their predictions of the probability of occurrence are deterministic and thus always "correct."
221 Variability in test statistics of OMNI thus reflects only variation in test sites across 100 data
222 iterations. Variability in test statistics of other algorithms reflects variation in test and training
223 sites across the same 100 iterations plus differences in structural forms of the models. For
224 shorthand, we will use "SDM" to refer to GAMs, MAX, and BRTs together (not including OMNI).

225 We adopted a reductionist approach isolate factors that affect assessment of variable importance by
226 conducting simulation experiments focusing on one aspect of interest at a time. Thus, our first
227 scenario begins with the simplest possible test of methods for assessing variable importance: (1) a
228 species' range is determined by a single "TRUE" variable, but the SDMs are presented with data on
229 this variable plus an uncorrelated "FALSE" variable that has no effect on distribution. The second
230 experiment (2) explores the effects of training sample size (number of presences). We then
231 investigate the effects of spatial scaling on measuring variable importance by varying: (3) the
232 prevalence of the species; (4) study region extent; and (5) spatial resolution and autocorrelation of
233 environmental data. In the next experiment, we explore the outcomes of (6) presenting models with
234 a FALSE variable that is correlated with a TRUE variable. We then consider cases where the
235 species' distribution is shaped by two factors in experiments where (7) variables are uncorrelated
236 on the landscape; (8) correlated on the landscape; (9) interact to define the niche but remain
237 uncorrelated; and (10) interact and are correlated. Finally, we consider two cases where (11) a
238 modeler mistakenly excludes a TRUE variable from the model but includes a correlated FALSE
239 variable; and (12) overparameterizes the model with correlated TRUE and FALSE variables.

240 We created the R package *enmSdmPredImport* for generating virtual species and evaluating
241 methods for assessing variable importance (available at
242 www.github.com/adamlilith/enmSdmPredImport). Code for generating the scenarios and
243 illustrations used in this article are available at

www.github.com/adamlilith/enmSdmPredImport_scenarios. The package and scenario code depend primarily on the *dismo* (Hijmans et al. 2017), *raster* (Hijmans 2018), and *enmSdm* (Smith 2019) packages for the R Statistical Environment (R Core Team 2018).

Simple scenarios

Experiment 1: A simple scenario

In this scenario, the probability of occurrence of the species has a logistic response to a single TRUE variable:

Eq. 1 $$\Pr(occ) = \frac{exp(\beta \times TRUE)}{1 + exp(\beta \times TRUE)}$$

where β represents the strength of the response (we used $\beta = 2$). The TRUE variable has a linear gradient in geographic space ranging from -1 to 1 across a square landscape 1024 cells on a side (Fig. 1). Given this arrangement, the generative function (Eq. 1) creates a probability of occurrence with an inflection point midway across the landscape (Fig. 1). For each of the 100 iterations, each SDM algorithm (BRT, GAM, and MAX) was trained using values of TRUE at presence and background sites plus values of a randomly distributed FALSE variable with the range [-1, 1] at the same sites. This FALSE variable does not appear in the generative model, so represents a variable "mistakenly" presented to the SDMs.

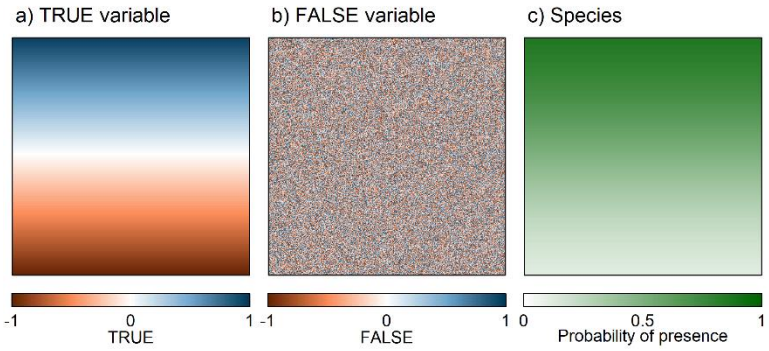


Figure 1. Spatial arrangement of landscape and species for *Experiments 1* and *2*. (a) The "TRUE" variable has a linear trend across the landscape with a range of [-1, 1]. (b) The "FALSE" variable has randomly distributed values, also in the range [-1, 1]. (c) The species responds in a logistic fashion to the TRUE variable. Training and test occurrences are sampled from the species' distribution proportional to cells' probability of occurrence (the "species" map). The landscape is 1024 cells on a side. Scales along the bottom display the range of the values shown in the maps.

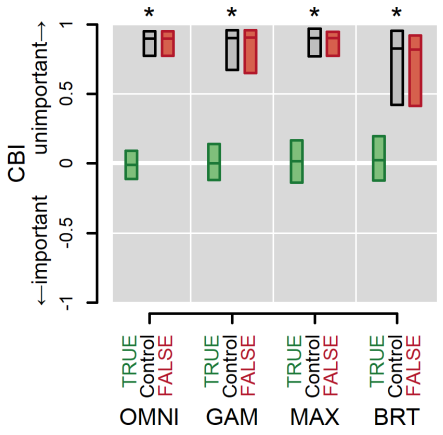


Figure 2. Results for *Experiment 1* examining a simple scenario with an influential TRUE variable and an uninfluential FALSE variable that was "mistakenly" presented to the model. Models successfully discriminated between the two variables (no overlap between green and red bars, indicated by asterisks). If a variable is important then permuting its values will lower model performance, here measured using the Continuous Boyce Index (CBI). "Control" values represent CBI of the unpermuted model. Bars represent the inner 95th-quantile of the distribution of CBI from 100 models across 100 data iterations of training and test sites. Horizontal lines within each bar represent the median value across the 100 iterations. The OMNI ("omniscient") model serves as a benchmark against which to assess the ability of the other models.

When presenting results, we plot the median and the inner 95% of the values of the test statistics across the 100 data iterations. Since we are utilizing simulations, we know a priori that treatments (different SDM algorithms, etc.) are different, so tests of statistical significance are unwarranted (White et al. 2014). Thus, our interest is in the effect size of the treatments (White et al. 2014), which can be discerned by eye. We reiterate that variation in the responses represents variability arising from random sampling of training (GAMs, MAX, and BRTs) and test sites (all algorithms) across 100 models (i.e., 100 OMNI models, GAMs, Maxent models, or BRTs). Thus, they represent the range of variation expected across a set of data samples. In real-world situations there are typically just one set of samples. Thus, if we were to implement the procedures for a real species with just one set of data, the outcome would fall somewhere within the range of variation displayed here. Hence, even if a model cannot reliably discriminate between variables in all of 100 data iterations in our scenarios, it may still be able to do so for some of them. However, in real-world situations the expected distribution of the test statistic will be unknown, so it will not be possible to know if the test statistic robustly indicates the importance of a variable. Hence, define successful discrimination between variables as complete lack of overlap between the distribution of CBI where the TRUE and FALSE variables are permuted in turn. We define successful calibration by comparing the distribution of the test statistic from the SDM and OMNI: the difference between the upper 97.5th and 2.5th percentile of the SDM's distribution of CBI had to be within ±10% of the same range of OMNI's distribution of CBI, and the SDM's median CBI had to be within the inner-most 20th percentile distribution of CBI (i.e., between the 40th and 60th percentile). For a given case (e.g., set of test statistics for a control model plus the two sets of predictions) each distribution

had to be successfully calibrated compared to OMNI for the case as a whole to be considered successfully calibrated. We found no cases using CBI, COR_{pa}, or COR_{bg} had successful calibration accuracy, although using AUC_{pa} and AUC_{bg} sometimes yielded well-calibrated outcomes when the species was modeled with GAMs or MAX (Appendices 1 and 2).

Results for the OMNI model serve as a benchmark for the other SDMs, so we discuss them first. CBI for OMNI control was slightly <1 (inner 95% quantile range: 0.77 to 0.95; Fig. 2). Even though OMNI correctly recreates the generative model, maximum CBI for OMNI is <1 because of the probabilistic assignation of presence/background sites. Thus, the scenario is in theory not too challenging for SDMs to model accurately. Compared to OMNI, MAX had a comparable median CBI (high accuracy) and comparable distribution of CBI (high precision). GAMs also had high accuracy but exhibited lower precision. BRTs had the lowest accuracy and precision.

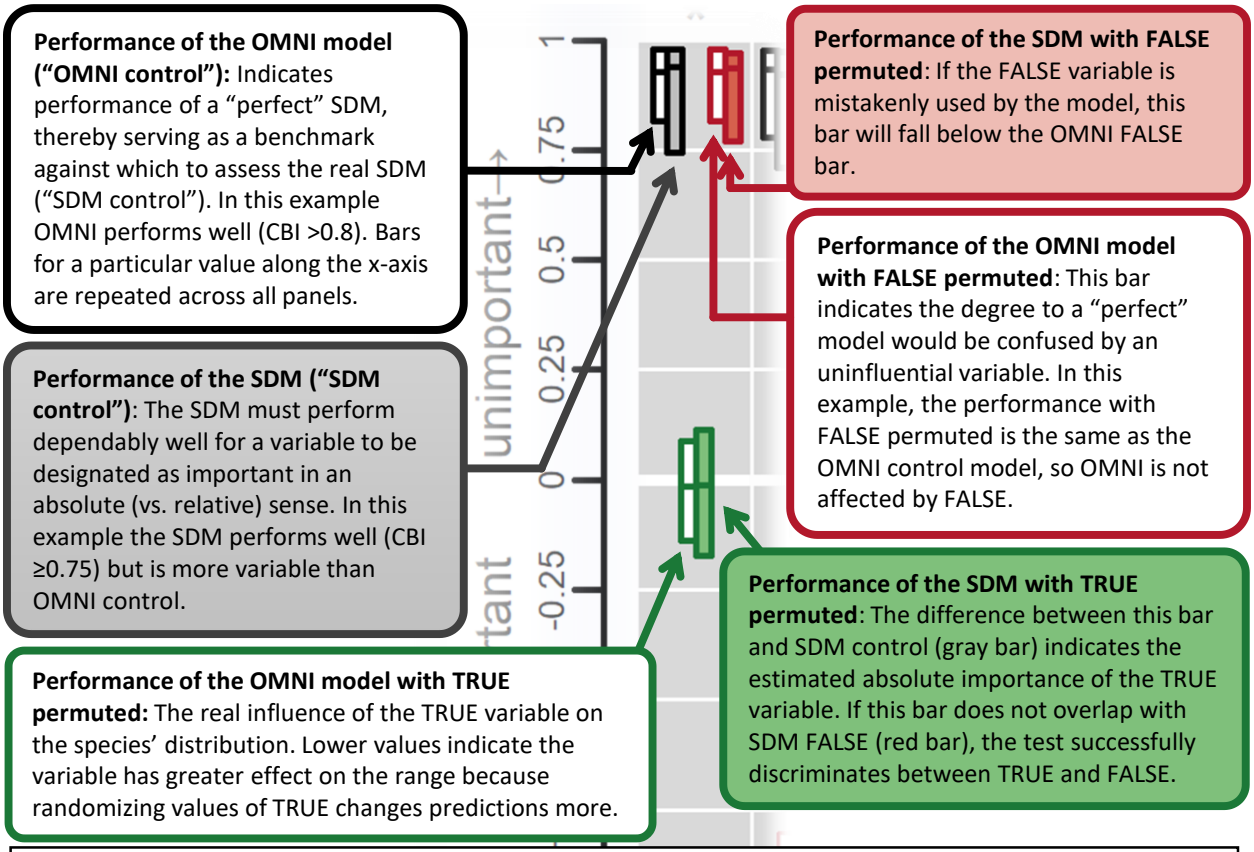
All algorithms had high discriminatory capacity (ability to qualitatively differentiate between TRUE and FALSE variables). In each case, permuting TRUE yielded CBI values centered on 0, which is indicative of a model that performs no better than random. When TRUE was permuted, the SDMs had slightly less precision (greater variability) than OMNI. When FALSE was permuted, only MAX was precise (same distribution as OMNI). BRTs also slightly overestimated the importance of FALSE (lower accuracy indicated by a lower median CBI value compared to OMNI; Fig. 2). BRTs are highly flexible, "local" learners (Elith et al. 2008), which makes them more likely to use information from the FALSE variable. This was evident from examining the BRT-specific measure of variable importance, which is the number of times a variable appears in a model times the deviance reduction associated with its use, standardized to sum to 1 (Elith et al. 2008). For a case with one influential variable and one uninfluential variable, these values should approach 1 and 0, respectively. In this experiment the average value across 100 iterations for TRUE was 0.81 (inner 95% quantile range: 0.57-0.95) and for FALSE 0.19 (0.05-0.43). Thus, BRTs were liable to mistakenly use information in the FALSE, randomly distributed variable.

This scenario demonstrated that even very simple situations can be somewhat challenging for measuring variable importance, especially for BRTs. Although it may seem odd that models occasionally used the FALSE variable, employing variables with no actual relationship to the species' occurrence can yield models as accurate as models trained using "real" variables (Buklin et al. 2015; Fourcade et al. 2018).

Experiment 2: Training sample size

Here we test the effect of the number of presences used in the training sample on estimates of variable importance using the same species' response and landscape as in *Experiment 1* (Eq. 1 and Fig. 1). The number of training occurrences was varied across the series 8, 16, 32, … , 1024. We used the same number of test presences (200) and test background sites (10,000) as in the previous experiment. Presentation of results for *Experiments 2* through *5* and *7* use a similar graphical motif explained in Box 1.

Box 1: Interpreting the permute-after-calibration test of variable importance



Q: Could an SDM perform well for a given situation (a particular value along the x-axis)?
A: Yes, if **OMNI control** has good performance.

Q: Does the SDM (without perturbations) perform as well as it could?
A: Yes, if **SDM control** and **OMNI control** largely overlap.

Q: Could the SDM reliably estimate the absolute importance of the TRUE variable?
A: Yes, if **OMNI control** has a distribution near 1 and **OMNI TRUE permuted** has a distribution near 0.

Q: Does the SDM reliably estimate the importance of the TRUE variable?
A: Yes, if **SDM TRUE permuted** and **OMNI TRUE permuted** are both centered around 0.

Q: Could a SDM reliably estimate the absolute importance of the FALSE variable?
A: Yes, if **OMNI control** has a distribution approaching 1 and **OMNI FALSE permuted** has a distribution approaching 1.

Q: Does the SDM reliably estimate importance of the FALSE variable?
A: Yes, if **OMNI FALSE permuted** and **SDM FALSE permuted** are both centered around 0.

Q: Could a SDM reliably discriminate between TRUE and FALSE?
A: Yes, if **OMNI TRUE permuted** does not overlap with **OMNI FALSE permuted**.

Q: Does the SDM reliably discriminate between TRUE and FALSE?
A: Yes, if **SDM TRUE permuted** does not overlap with **SDM FALSE permuted**.

Q: Is the SDM well-calibrated?
A: Yes, if the central tendency and range of distributions of performance by the SDM (control and permuted) match their OMNI counterparts.

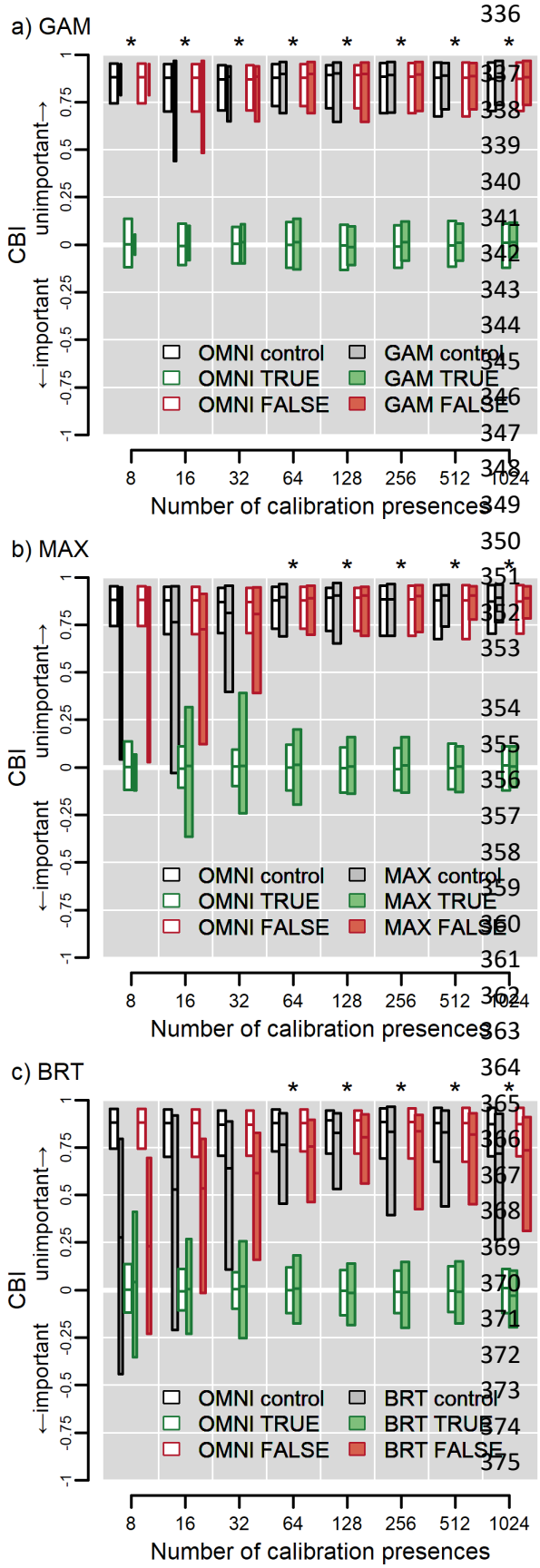


Figure 3. In *Experiment* 2 GAMs can reliably discriminate between an influential and uninfluential variable when training sample size is ≥8 occurrences, but MAX requires ≥32 and BRTs ≥64 occurrences. Asterisks indicate both OMNI and the SDM successfully discriminate between TRUE and FALSE at the given number of presences. Bars represent the inner 95% of values of CBI across 100 models on 100 data iterations of training and test sites at the given sample size. The horizontal line in each bar represents the median value. Bar width is proportional to the number of iterations in which the respective SDM algorithm converged (<100 for small sample size for GAMs and BRTs). Results for OMNI are repeated across all panels. See Box 1 for further guidance on interpretation.

Regardless of sample size, OMNI was always able to generate reliable predictions (OMNI control) and correctly discriminate between TRUE from FALSE (OMNI permuted; Fig. 3). GAM performed very similarly to OMNI at all sample sizes, but sometimes could not converge when sample size was ≤32 (Fig. 3a; the greater variation in GAM controls for n = 16 is likely due to small sample size effects arising from the limited number of models that converged). MAX control performed as well as OMNI control when sample size was ≥64 (Fig. 3b), but had lower precision and accuracy when sample size was ≤32. MAX was also unable to reliably discriminate between TRUE and FALSE when sample size was ≤32. MAX also had trouble "converging" when sample size was 8 (which in this context means it produced an intercept-only model that could not be used to calculate CBI, which requires variation in predictions for calculation). BRT control always had the lowest accuracy and precision across all sample sizes (Fig. 3c). BRTs were unable to

reliably discriminate TRUE from FALSE and had trouble converging when sample size was ≤32.

In summary we found that small sample size challenged the ability of SDMs to differentiate influential from uninfluential variables. In our scenario all SDMs could reliably discriminate between TRUE and FALSE when sample size was ≥64, although we expect that this is a lower limit for real-world applications. Hereafter we use 200 occurrences for model training.

Spatial scale: Prevalence, extent, and spatial resolution

Experiment 3: Prevalence

Here we explore the effects of prevalence (mean probability of occurrence; i.e., range size) on assessments of variable importance. As in the previous scenarios we used a square landscape with a TRUE variable with a linear gradient with the range [-1, 1] and a randomly distributed FALSE variable also with the range [-1, 1]. The species responds to TRUE as per a logistic function

Eq. 2 $$\Pr(occ) = \frac{exp(\beta_1 \times (TRUE - \beta_{11})}{1 + exp(\beta_1 \times (TRUE - \beta_{11})}.$$

The generative function includes an offset parameter β_{11} which shifts the range across the landscape, thereby altering prevalence (Fig. 4). This allows us to alter prevalence while not changing study region extent, which is almost impossible in real-world situations. We chose values of β_{11} that altered prevalence in 9 steps across a range spanning ~0.05 to 0.95. The strength of the species' response to TRUE, β_1, was set equal to 2 as in *Experiments 1* and *2*.

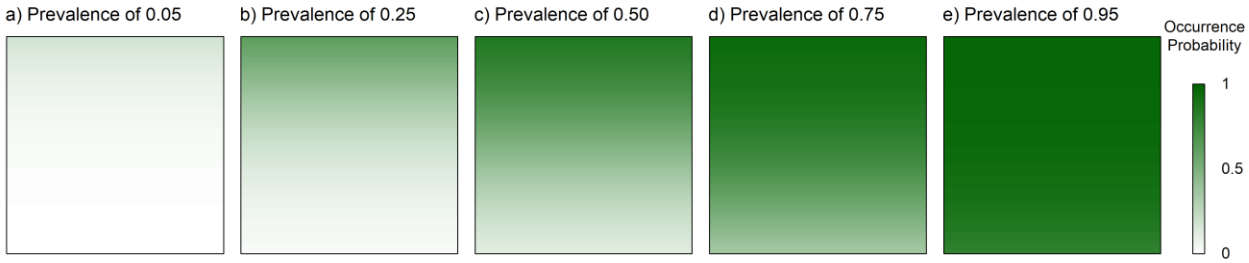


Figure 4. Varying prevalence of the species on the landscape for *Experiment 3*. Prevalence (mean probability of occurrence) was varied from 0.05 to 0.95 along a series of 9 steps (only 5 are shown here). The distribution of the TRUE and FALSE variables remained unchanged and are as in Fig. 1.

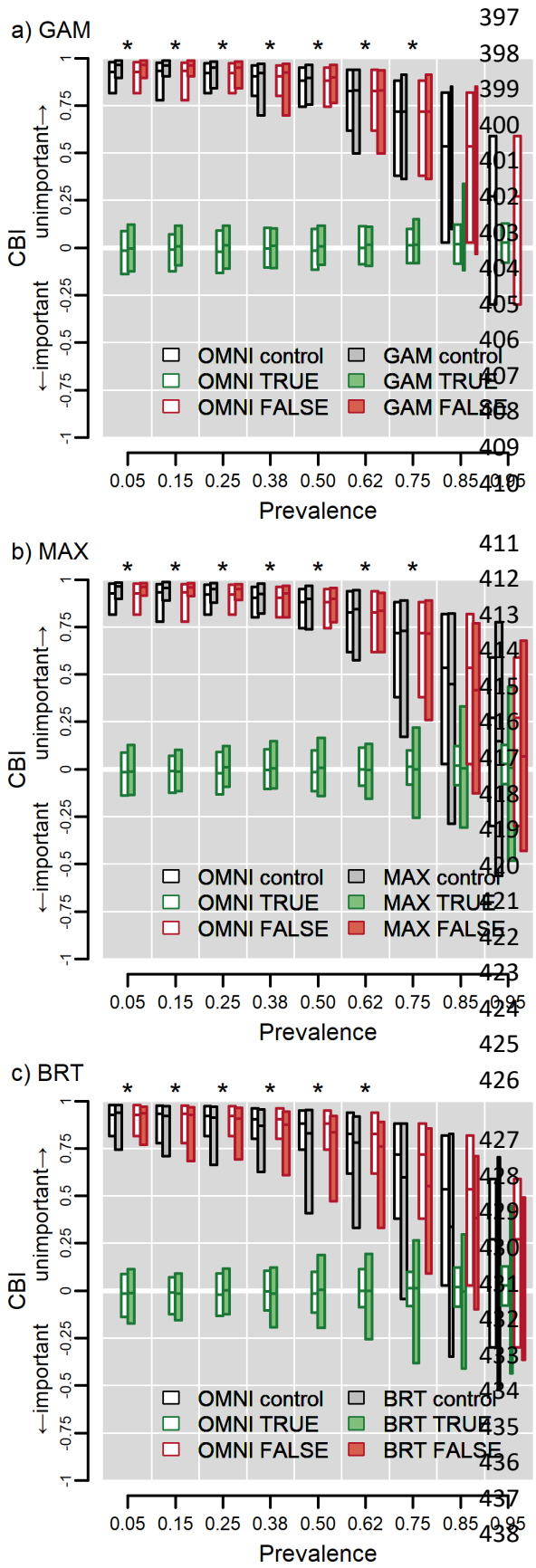


Figure 5. In *Experiment 3*, models measure variable importance most accurately when a species occupies not more than 62-75% of the study region. Asterisks indicate both OMNI and the SDM successfully discriminate between TRUE and FALSE at the given level of prevalence. Bars represent the inner 95% of values of CBI across 100 models for 100 data iterations of training and test sites at the given prevalence. The horizontal line in each bar represents the median value. Bar width is proportional to the number of iterations in which the respective SDM algorithm converged. Results for OMNI are repeated across all panels. See Box 1 for further guidance on interpretation.

Increasing prevalence had only a slight effect on the performance of OMNI control when the species occupied less than half of the landscape (any panel in Fig. 5). When prevalence was ≥0.5 mean performance of OMNI control decreased while variability increased, and approached or overlapped 0 when prevalence was ≥0.85. Thus, even a "perfect" model is not able to reliably estimate a species' range when the species occupies most of the landscape. For OMNI the distributions of permuted TRUE and FALSE overlapped when prevalence was ≥0.85 (white bars outlined in red or green in any panel of Fig. 5). Since OMNI is not able to distinguish TRUE and FALSE at high prevalence, SDMs should also not be able to do so.

Increasing prevalence also decreased the performance of the control SDMs, especially when prevalence was ≥0.5 (any panel in Fig. 5). GAM controls roughly matched OMNI controls for all levels of prevalence, but MAX and BRT controls performed increasingly worse than OMNI controls as prevalence increased. Both GAMs and BRTs struggled to converge at high levels of prevalence, and no GAMs converged when prevalence was 0.95. BRTs were slightly less robust than GAMs and MAX to high levels of prevalence indicated by overlap between

permuted TRUE and FALSE when prevalence was >0.62 (BRTs) versus ≥0.75 (GAMs and MAX).

In summary, all models (including OMNI) were less accurate when prevalence was ≥0.5. For a given level of prevalence, GAMs had better calibration accuracy and sometimes better discrimination accuracy than MAX and BRTs. However, even OMNI could not differentiate TRUE from FALSE when the species was very common (prevalence ≥0.85).

Experiment 4: Study region extent

In real-world situations enlarging the study region typically decreases prevalence (e.g., Acevedo et al. 2017) while increasing the range variability of environmental variables to which models are exposed (VanDerWal et al. 2009). In this experiment we isolate the effect of increasing the extent of the study region on the range environmental variation of the TRUE variable. Again, we used a TRUE variable varying along a linear gradient and a FALSE, randomly distributed variable. Landscape size was varied along the series 128, 256, 512, 1024, 2048, 4096, and 8192 cells on a side, each matched with an increasing range of the TRUE variable from -0.125 to 0.125 for the smallest landscape to -8 to 8 for the largest. The range of FALSE was kept at [-1, 1]. The species responded to TRUE as per a logistic function (Eq. 1) with the inflection of suitability placed at the mid-line of the landscape (Fig. 6). As a result, prevalence was kept constant even though extent was increased.

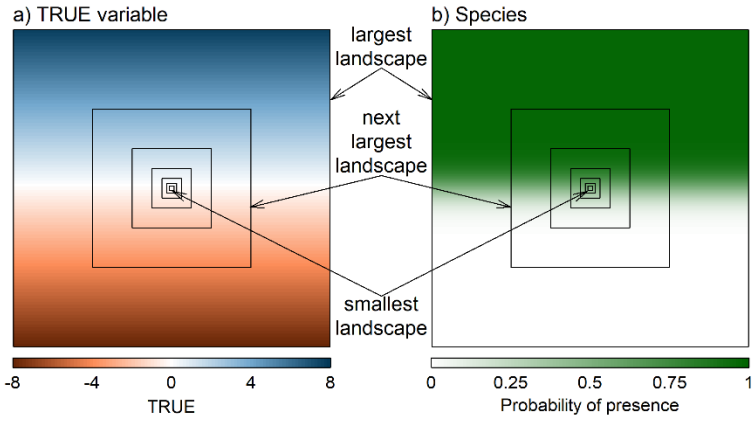


Figure 6. For the landscape in *Experiment 4*, the study region extent was increased while keeping the species' prevalence unchanged. To test the effects of the relationship between extent and the range of environmental variables on assessing variable importance we simulated seven landscapes of increasing size (128 cells on a side to 8192 cells on a side) with concomitant increases in the overall range of the TRUE variable (from 0.25 at the smallest extent to 16 at the largest). The species had a logistic response to the TRUE variable. Prevalence (mean suitability) of the species was kept constant across the landscapes. Increasing extent increased the range of the probability of presence across the landscape.

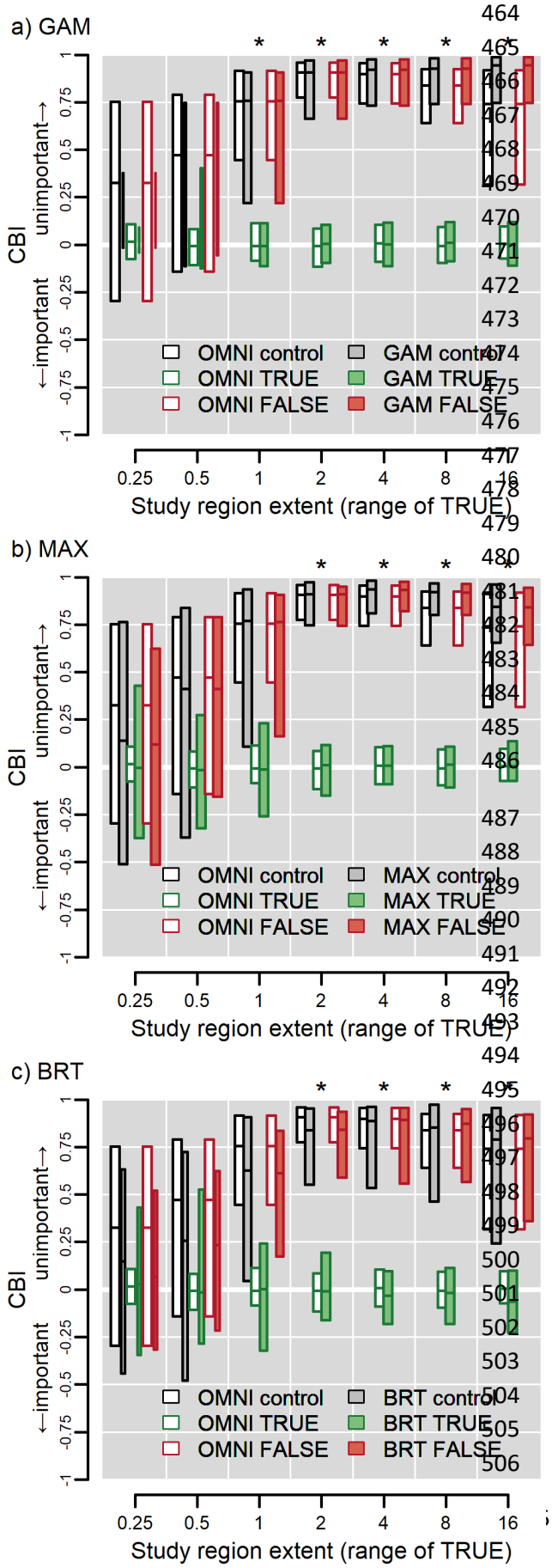


Figure 7. In *Experiment 4* models cannot reliably measure variable importance when study region extent is too small to encompass suitable variation in driving variables, even if the model is "omniscient." The ordinate in each panel displays the range of the TRUE variable (0. 25 for the smallest landscape and 16 for the largest) and thus reflects the effect of increasing the spatial extent of the study area (ranging 128 to 8192 cells on a side). Asterisks indicate OMNI and the SDM both successfully discriminate between TRUE and FALSE at the given extent (range of TRUE). Bars represent the inner 95% of values of CBI from 100 models for 100 data iterations of training and test sites at the given spatial extent. The horizontal line in each bar represents the median value. Bar width is proportional to the number of iterations in which the model successfully converged (<100 for GAMs and BRTs at ranges ≤1). Results for OMNI are repeated across all panels. See Box 1 for further guidance on interpretation.

Performance of the OMNI model peaked at intermediate extents (where the range of TRUE was from 2 to 4 where the landscape was 1024 to 2048 cells on a side), declined slowly above this range, and dramatically below this range (any panel in Fig. 7). OMNI control overlapped with 0 when extent was small (range of TRUE ≤0.5), indicating some models performed no better than or even worse than random. Thus, even the generative model could not robustly recreate the species' distribution when extent was very small. As a result, OMNI could not dependably discriminate between TRUE from FALSE when the range of TRUE was ≤1. Small extents challenge models because there is very little variance in the actual probability of presence across the landscape, causing occurrences to intermingle with background sites approximately at random. Hence, test metrics of model performance find no

difference between slightly unsuitable and slightly more suitable areas even if the model perfectly recreates the probability of presence. Surprisingly, increasing extent such that the range of TRUE was ≥8 increased variability in OMNI control, although OMNI could still reliably discriminate between TRUE and FALSE at large extents. The worsening performance of OMNI controls at large extents is likely the result of the sensitivity of CBI to test presences that are located in areas with a very low probability of presence (Appendix 6). Unlike OMNI control, performance of the control GAMs and MAX did not notably decline at large extents. We believe this is due to GAMs and MAX actually underfitting the data, wherein areas of very low probability of occurrence are predicted to be disproportionately higher in suitability, thereby mistakenly predicting improbable presences belong to the set of more probable presences.

Like OMNI, the SDM algorithms were also unable to discriminate between TRUE and FALSE at small extents. Discriminatory capacity was unreliable when the range of TRUE was ≤0.5 (GAMs) or ≤1 (MAX and BRTs; Fig. 7). GAMs and BRTs often did not converge at small extents.

In summary, small spatial extents that span shallow environmental gradients (Fig. 6) engender little variation the probability of presence. As a result, models have great difficulty differentiating TRUE from FALSE variables.

Experiment 5: Spatial resolution and spatial autocorrelation of environmental data

In this experiment we explore the effects of spatial autocorrelation and spatial resolution (grain size) of environmental variables on estimates of predictor importance. As before, we use a square landscape with a TRUE variable with a linear gradient and a FALSE variable with random values. We assumed the species responded to the environment at a scale represented by dividing the landscape into 1024 cells on a side (the "native" resolution). As before, we constructed the TRUE variable so it had a linear gradient spanning the landscape. When unperturbed, this linear gradient has a high degree of spatial autocorrelation because cells close to one another have similar values. Thus, we manipulated spatial autocorrelation by randomly swapping a set proportion of cell values (no cells, or one third, two thirds, or all of the cells). Swapping values diminished spatial autocorrelation in TRUE because cells with dissimilar values were more likely to be close to one another (Fig. RESOLUTION). (By design, FALSE had a level of spatial autocorrelation no different from random).

To explore the effects of spatial resolution, we applied bilinear interpolation to resample cells to a finer resolution so that the landscape had 16,384 cells on a side or to a coarser resolution with 64 cells on a side. Resampling to a finer resolution caused spatial autocorrelation in TRUE and FALSE to increase slightly because smaller cells on the edge of large cells were spatially averaged with fine cells on the edge of adjacent large cells (Fig. 8). Resampling to a coarser resolution also caused spatial autocorrelation in TRUE and FALSE to increase because aggregation of smaller cells into larger ones "smooths" over values of the smaller cells. This effect was especially pronounced when a large proportion of coarse cells had been swapped (Fig. 8).

The species' probability of presence was generated using the landscape at its native resolution (possibly with cells swapped). Training and test sites were then selected using the landscape at the

native resolution but assigned cell values based on the landscape at the (possibly) resampled resolution. This recreates a realistic situation where the occurrences representing a species' response to the environment are indicative of the scale of the true response but environmental data used to predict the response are available at a (potentially) different resolution. We created 12 scenarios representing all possible combinations of grain size (cell size of 1/64th, 1/1024th, and 1/16,384th of the landscape's linear dimension) and spatial autocorrelation (swapping no cells, or one third, two thirds, or all of the cells; Fig. 8).

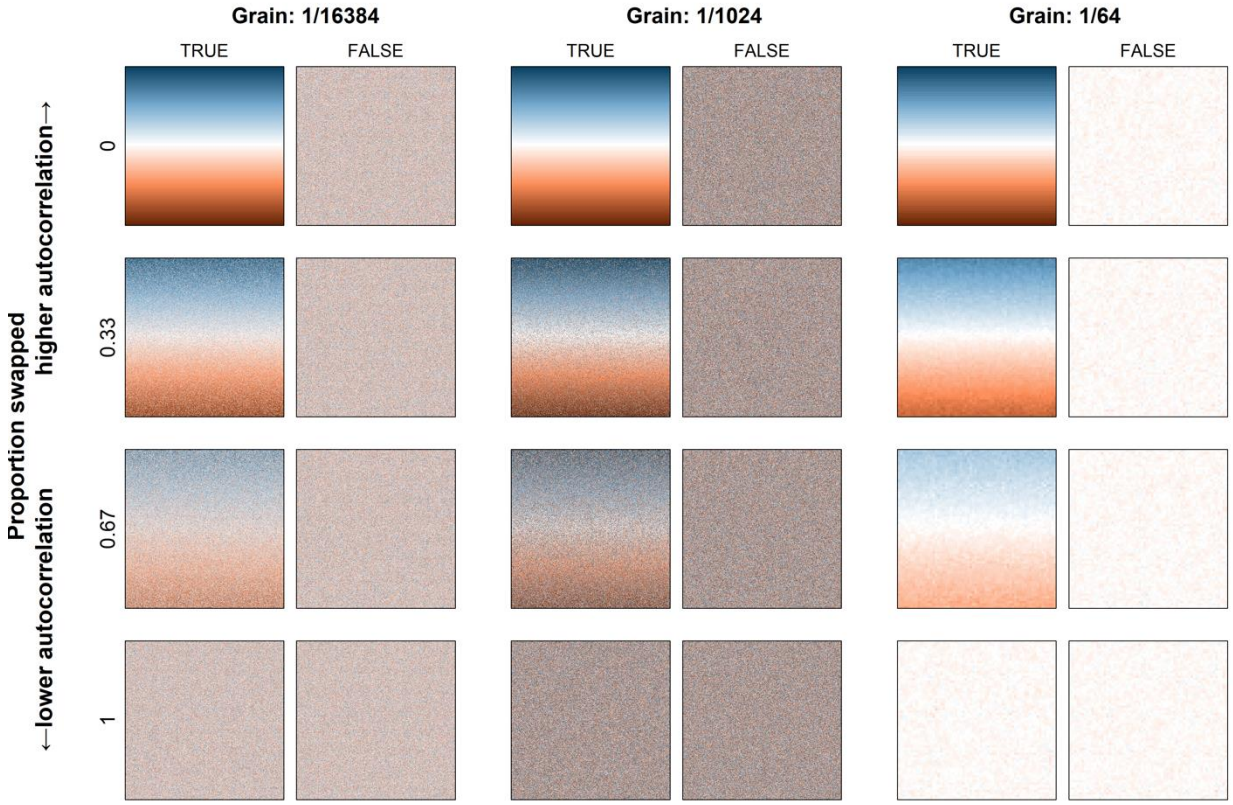


Figure 8. Landscapes for testing the effects of grain size (spatial resolution) of data and spatial autocorrelation in *Experiment 5*. The species was assumed to respond to the environment at a "native" grain size of 1/1024th of the linear dimension of the landscape. The environmental data was resampled to a resolution of 1/16,384th (smaller cells) or 1/64th (large cells) of the landscape's linear dimension. Spatial autocorrelation in the environment mediates the effect of spatial resolution, so we varied autocorrelation for each variable by randomly swapping cell values between none, 1/3, 2/3, and all of the cells. Increasing the proportion of cells swapped reduces spatial autocorrelation in the TRUE variable because cells of similar value are less likely to be close to one another. The left three columns represent the TRUE variable and right three columns the FALSE variable. Each TRUE was paired with the FALSE with the same level of spatial autocorrelation and grain (e.g., the TRUE in the upper left was paired with the FALSE to the immediate right of it). Values range from -1 (brown) to 1 (blue).

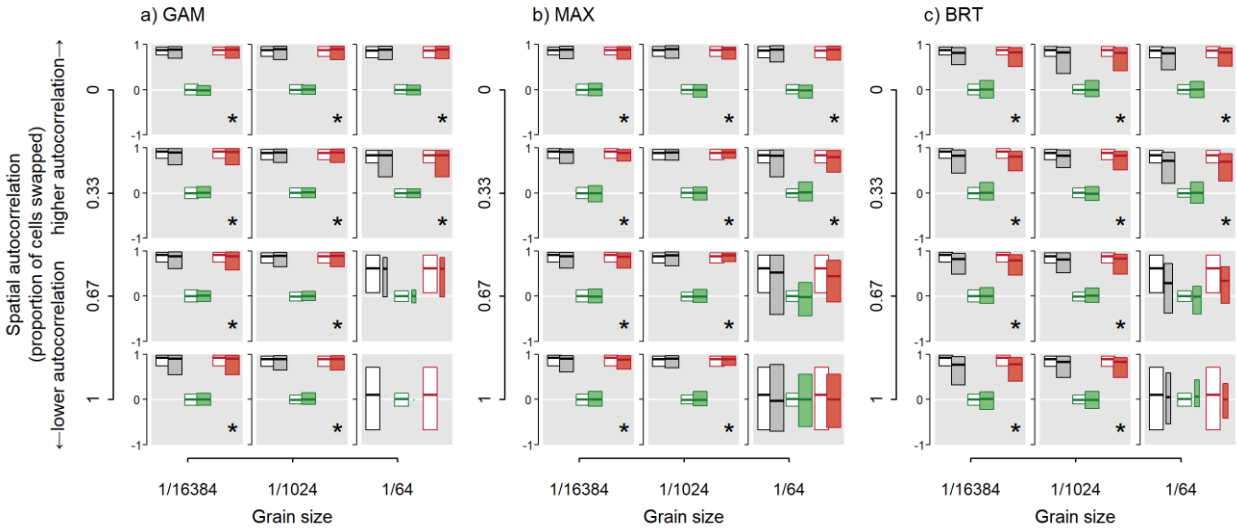


Figure 9. In *Experiment* 5, models successfully discriminate between TRUE and FALSE variables regardless of spatial resolution of environmental data and spatial autocorrelation except when resolution is coarser than the perceptual scale of the species and autocorrelation is low. The resolution at which the species perceives the landscape is at 1/1024 (middle column in each panel). Bars in each subplot of each panel represent the inner 95% of values of CBI from 100 models for 100 data iterations of training and test sites at the given spatial resolution and magnitude of spatial autocorrelation. The horizontal line in each bar represents the median value. Bar width is proportional to the number of iterations in which the model successfully converged. Asterisks in the lower right corner of each subplot indicates both OMNI and the SDM successfully discriminate between the TRUE and FALSE variables. Results for OMNI are repeated across all panels. See Box 1 for further guidance on interpretation of the subplots in each panel.

In this experiment the "true" importance of TRUE and FALSE is indicated by results from OMNI at the "native" resolution of 1/1024th (middle column in any panel in Fig. 9). Results for OMNI for the other resolutions represent the outcome that would be obtainable if a modeler had perfect knowledge of the species' response to the environment but only had data available at finer or coarser resolutions. At the native resolution results for OMNI and SDMs were comparable, although BRTs tended to have lower accuracy and precision compared to GAMs and MAX (Fig. 9c).

Regardless of the level of spatial autocorrelation, downscaling to the finer resolution of 1/16,384th (left column in each panel of Fig. 9) yielded results very similar to results at the native resolution. Similarly, upscaling to 1/64th resolution when spatial autocorrelation was highest (proportion of cells swapped = 0; upper right of any panel in Fig. RESOLUTION_6) was very similar to modeling at the native resolution. However, reducing the level of spatial autocorrelation (going from upper right to lower right in any panel in Fig. RESOLUTION_6) decreased performance of OMNI controls. When spatial autocorrelation was lowest, OMNI control performance was on average no different from random. Permuting either TRUE or FALSE with OMNI reduced CBI to ~0 in both cases, meaning that OMNI could not reliably discriminate between the variables in this situation. Likewise, none of the SDMs dependably discriminated TRUE from FALSE when spatial resolution

was coarser than the scale of the perception of the environment and spatial autocorrelation was low (proportion swapped 2/3 or 1). Tellingly, very few GAM and BRT models converged under these circumstances.

In summary, we found downscaling did not affect the ability of SDMs to measure variable importance, and this was not affected by the level of spatial autocorrelation in the environment. Splitting cells into smaller units has negligible effect on cells values, so we did not expect a great difference between results at the native and downscaled resolutions. Similarly, upscaling when spatial autocorrelation was high also did not challenge models. In this situation, the species responds to the environment at finer scale, but amalgamating cells to yield larger cells does not change their values much because cells with similar environments occur close to one another. Thus, larger cells are nearly just as representative of the environment experienced by the species when spatial autocorrelation high (Mertes & Jetz 2017). However, when the environment varies rapidly across space coarsening resolution tends to "blend" cells with different values, thereby creating mismatches between the scale at which the species responds to the environment and the values assigned to training and test locations (Mertes & Jetz 2017). As a result, SDMs cannot reliably measure variable importance when grain size is coarser than the scale at which the species responds to the environment and environmental spatial autocorrelation is low (McInery & Purvis 2011; Mertes & Jetz 2017).

Collinearity and interactions between predictors

Experiment 6: Collinearity between environmental variables

In this experiment we explore the effects of model misspecification when a FALSE predictor is collinear (correlated) with a TRUE predictor. As before, the species has a logistic response (Eq. 1) to TRUE, which has a linear gradient across the landscape. FALSE also has a linear trend which is rotated relative to the gradient in TRUE to affect the correlation between the variables. To ensure no change in the univariate frequencies of TRUE and FALSE with rotation we used a circular landscape (Fig. 10). The two variables are uncorrelated when their relative rotation is 90° (Fig. TRUEvsFALSEc), and positive or negative if less than or more than 90°, respectively (Fig. 10). We used rotations of FALSE relative to TRUE from 22.5 to 157.5° in steps if 22.5°, which produced correlations between the two variables ranging from -0.91 to 0.91. Both variables had the range [-1, 1].

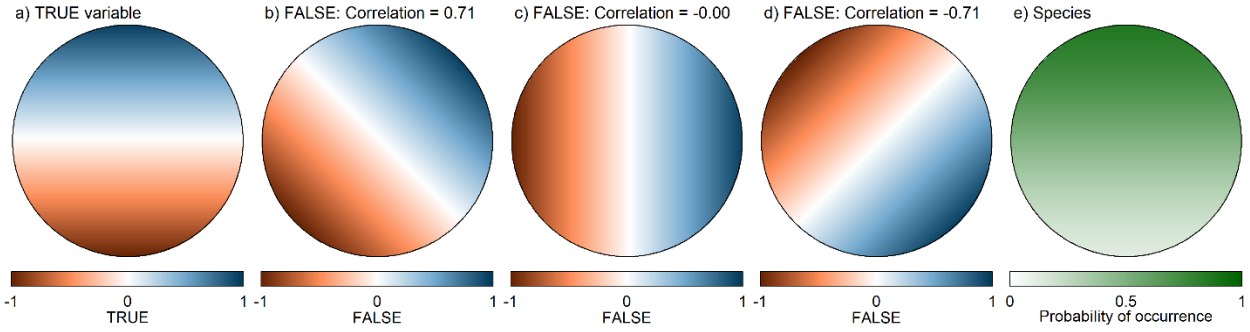


Figure 10. In *Experiment* 6 we varied collinearity between environmental variables. In this scenario we created a circular landscape with (a) a TRUE variable that influenced the species' distribution. The model was also presented with data on (b through d) a FALSE variable that did not influence distribution. The FALSE variable was rotated relative to the gradient of the TRUE variable from 22.5° to 157.5° in steps of 22.5°. The rotations changed the correlation with TRUE from 0.92 to -0.92. Rotations of 45° (b), 90° (c), and 135° (d) are shown here. Since the species only responds to the TRUE variable its distribution remains constant regardless of the rotation of FALSE relative to TRUE (e). However, the SDM algorithm is provided information on both TRUE and FALSE.

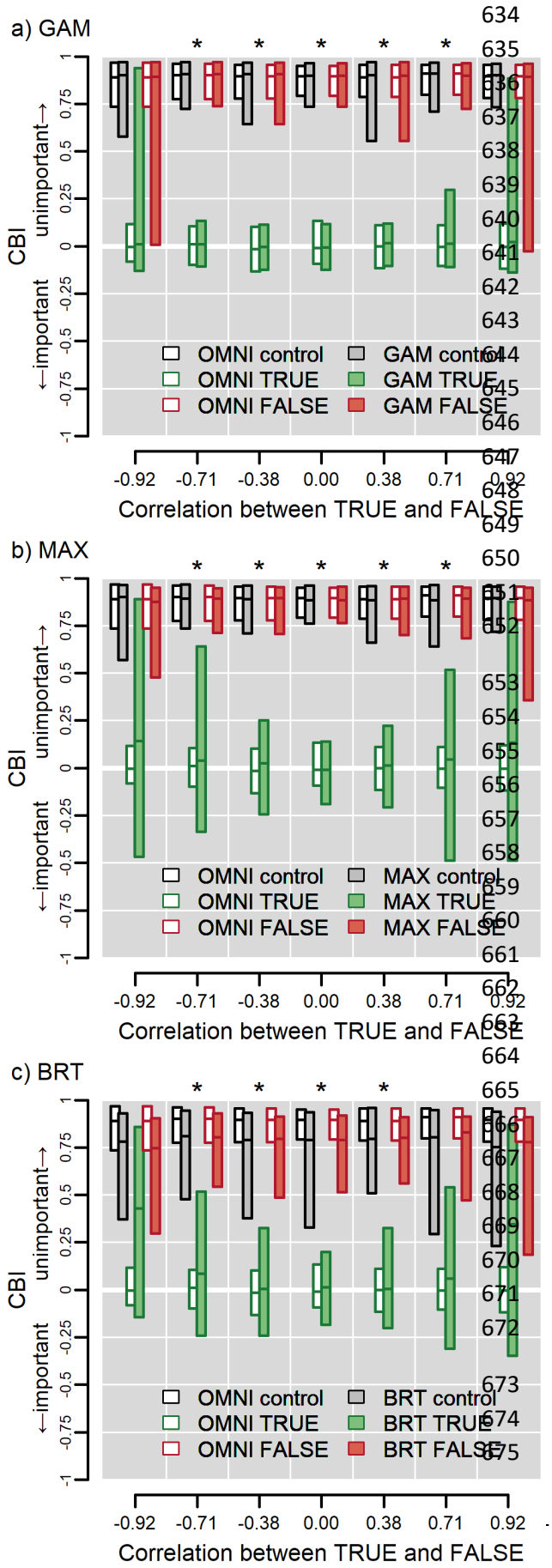


Figure 11. In *Experiment 6*, GAMs are fairly robust to correlations between an influential and uninfluential predictor, but MAX and BRTs have increasingly worse accuracy as the magnitude of correlation increases, especially when $|r| \geq 0.7$. In each panel the ordinate reflects the correlation between the FALSE and TRUE variables which changes as the gradient in FALSE is rotated on the landscape relative to the gradient in TRUE. Asterisks indicate both OMNI and the SDM successfully discriminate between TRUE and FALSE at the given level of correlation. Bars represent the distribution of the inner 95% of values of CBI from 100 models for 100 data iterations of the species and landscape at the given level of correlation. Results from OMNI are repeated across panels. See Box 1 for further guidance on interpretation.

OMNI completely ignores the FALSE variable and thus had high performance regardless of correlation between TRUE and FALSE (Fig. 11). Although slightly less accurate than OMNI control, GAM control had fairly high performance across the entire range of correlation between TRUE and FALSE. GAMs also had high discrimination and calibration capacity except when correlation was very high ($|r| > 0.71$). MAX and BRTs were much more sensitive to collinearity, although both models could usually discriminate between TRUE and FALSE at when $|r| \leq 0.71$. MAX control performed well at all levels of correlation, but BRT control had the poorest performance at all levels of correlation. For both MAX and BRTs the accuracy and precision of the permute test for TRUE decreased rapidly even when the magnitude of correlation was small ($|r| \geq 0.38$; Fig. 11b and c).

Overall GAMs were more robust to collinearity and model misspecification than MAX which was in turn slightly more robust than BRTs.

There was no strong trend in differences between positive and negative correlation (cf. Mela & Kapolle 2002), although responses were somewhat asymmetric for each model. Our results support the common practice of retaining only variables with pairwise correlations $|r| <\sim 0.7$ (cf. Dorman et al. 2013; Júnior & Nóbrega 2018), although even lower thresholds may be required for robust estimates of variable importance. However, discarding variables is problematic since it effectively assumes removed variables have zero importance with no uncertainty.

Experiment 7: Two influential, uncorrelated variables acting independently

Here we explore the ability of SDMs to differentiate the influence of two uncorrelated TRUE variables. The species responds to the two TRUE variables, T1 and T2, as per a Gaussian bivariate function where T1 and T2 act additively (independently) to determine the probability of occurrence:

Eq. 3 $$\Pr(occ) = \frac{1}{2\pi\sigma_1\sigma_2} exp\left(-\frac{(T1-\overline{T1})^2}{2\sigma_1^2} - \frac{(T2-\overline{T2})^2}{2\sigma_2^2}\right).$$

Niche width in T1 and T2 is determined by σ_1 and σ_2. Importantly, decreasing σ_i increases the degree to which a variable restricts distribution, meaning that σ_i and variable importance are inversely related. Changing σ_1 and σ_2 thus also alters the distribution of the species (Fig. 12b). We explored all possible combinations of values of 0.1, 0.3, and 0.5 for σ_1 and σ_2. Altering niche width has the unfortunate side effect of changing prevalence. Based on results from *Experiment 3*, we do not expect to see substantial biases until prevalence is $\geq\sim 0.5$ (Fig. 5), but in none of the current scenarios did prevalence fall above 0.570 (which occurs when niche width is widest, at $\sigma_1 = \sigma 2 = 0.5$). We used a circular landscape and allowed T1 and T2 to vary in a linear fashion across the range [-1, 1].

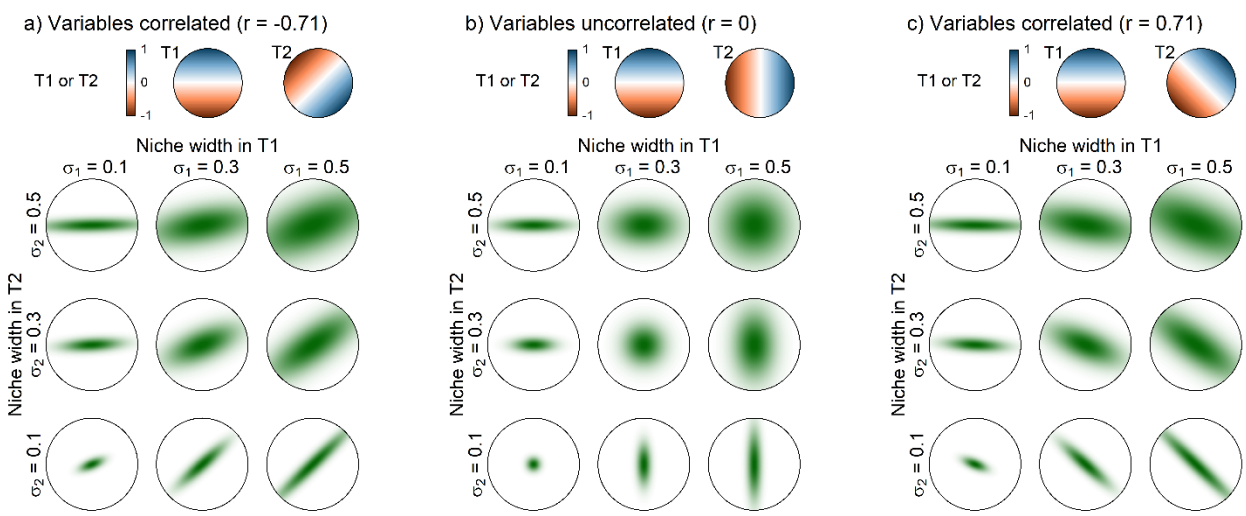


Figure 12. Landscape and species for *Experiments 7* through *9* in which two variables act independently to shape the species' niche and distribution. The top row in each panel displays the distribution of two TRUE variables (T1 and T2) on the landscape. The bottom three rows sow the

species' distribution on the landscapes as a function of niche width in T1 and T2. Larger values of σ connote broader niche width and thus less restriction by the respective variable. As a result, reducing σ_i causes the variable to restrict distribution more (i.e., the variable is more important). In *Experiment 7* we varied niche width while using a 90° rotation between T1 and T2 so they were uncorrelated (panel b). In *Experiment 8* we varied the correlation between T1 and T2 from -0.71 to 0 to 0.72 by rotating T2 relative to T1 on the landscape (panels a and c). The range of values for the species' maps is [0, 1], with darker/greener shades indicating greater probabilities of occurrence. There is no interaction between T1 and T2 in *Experiments 7* through *9* (ρ = 0).

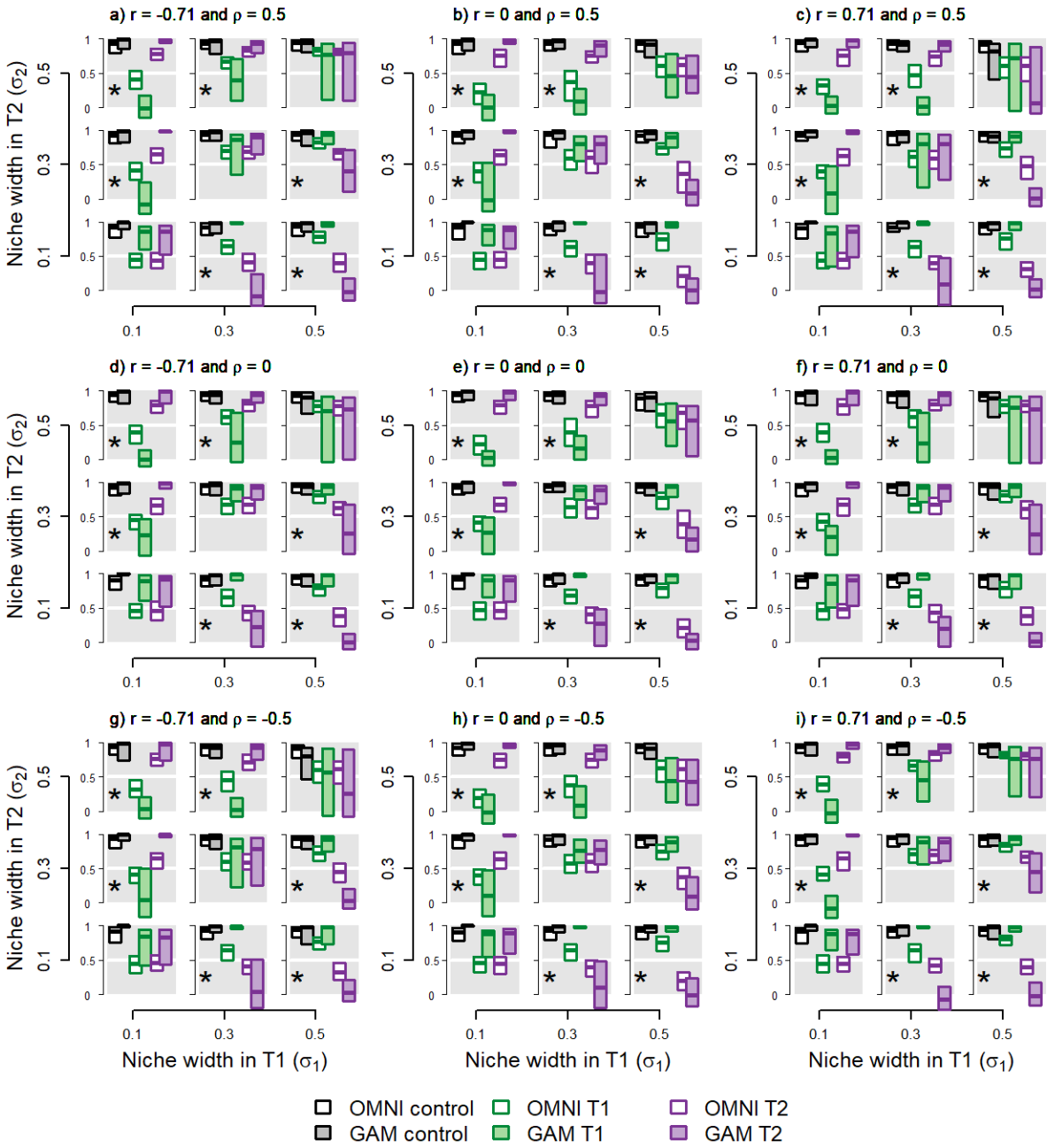


Figure 13. In *Experiments* 7 through *10*, GAMs underestimate importance when variables act equally to shape the niche and niche width is narrow or intermediate. When variables act unequally, GAMs overestimate the importance of the more influential variable (narrow niche width, low σ_i), and underestimate the importance of the less influential variable (broad niche width, high σ_i). Aside from these considerations, neither correlation between predictors (r) nor niche covariance (ρ) have strong effects on estimates of importance. Each panel corresponds to a particular landscape (correlation between predictors) and niche topology (niche covariance structure). Within a panel, subplots contain bar plots of control and permuted predictions from OMNI and GAMs. Smaller niche width corresponds to more restriction by the respective variable (and thus more importance) so should yield lower CBI when the variable is permuted. In each panel a subplot for a given value of σ_1 and σ_2 can be interpreted as explained in Box 1. Asterisks in the lower left of a subplot indicate both OMNI and GAMs successfully discriminate between T1 and T2.

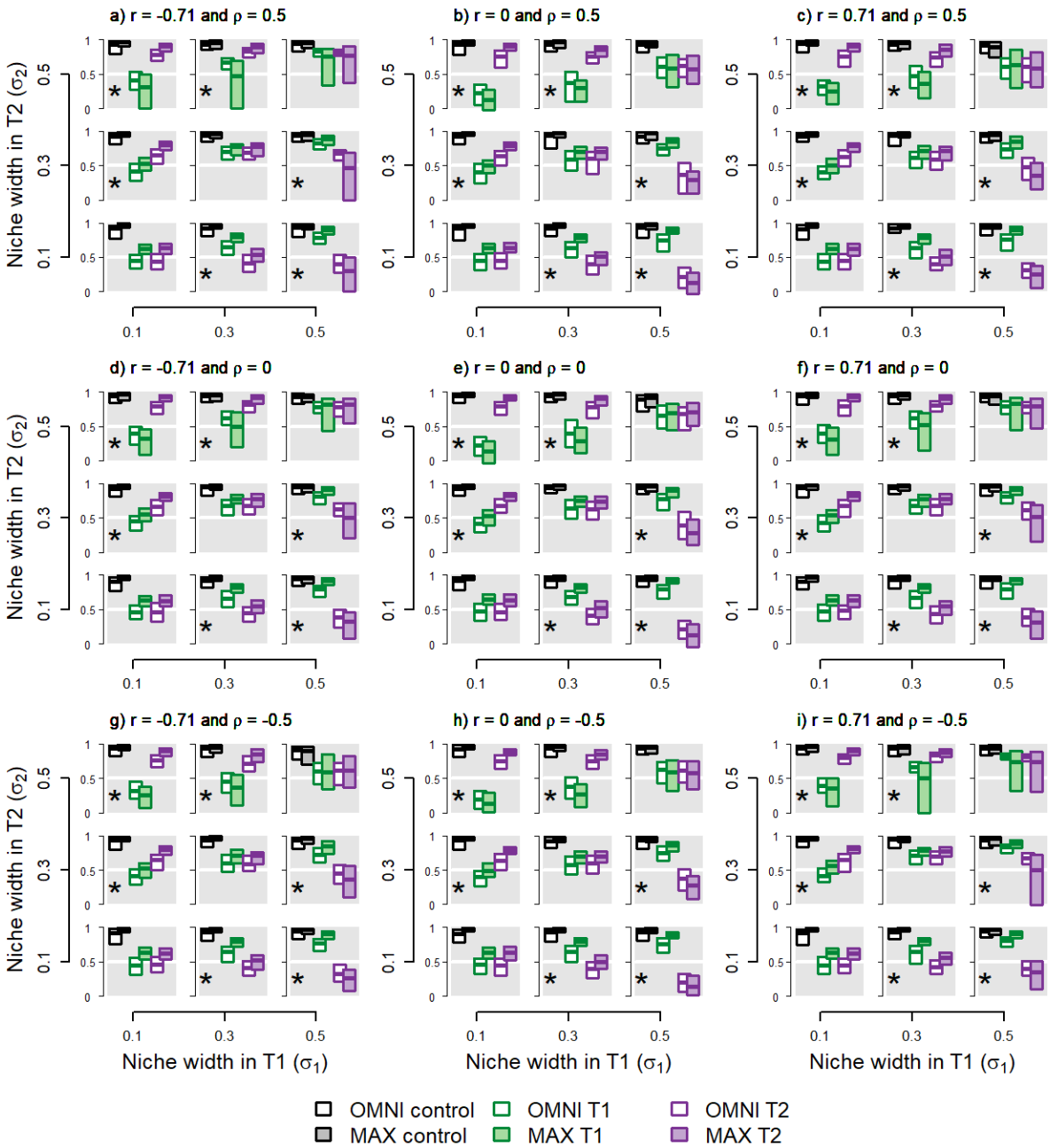


Figure 14. In *Experiments* 7 through *10*, MAX underestimate importance when variables act equally to shape the niche and niche width is narrow or intermediate. When variables act unequally, MAX overestimate the importance of the more influential variable (narrow niche width, low σ_i), and underestimate the importance of the less influential variable (broad niche width, high σ_i). Aside from these considerations, neither correlation between predictors (r) nor niche covariance (ρ) have strong effects on estimates of importance. Each panel corresponds to a particular landscape (correlation between predictors) and niche topology (niche covariance structure). Within a panel, subplots contain bar plots of control and permuted predictions from OMNI and MAX. Smaller niche

width corresponds to more restriction by the respective variable (and thus more importance) so should yield lower CBI when the variable is permuted. In each panel a subplot for a given value of σ_1 and σ_2 can be interpreted as explained in Box 1. Asterisks in the lower left of a subplot indicate both OMNI and MAX successfully discriminate between T1 and T2.

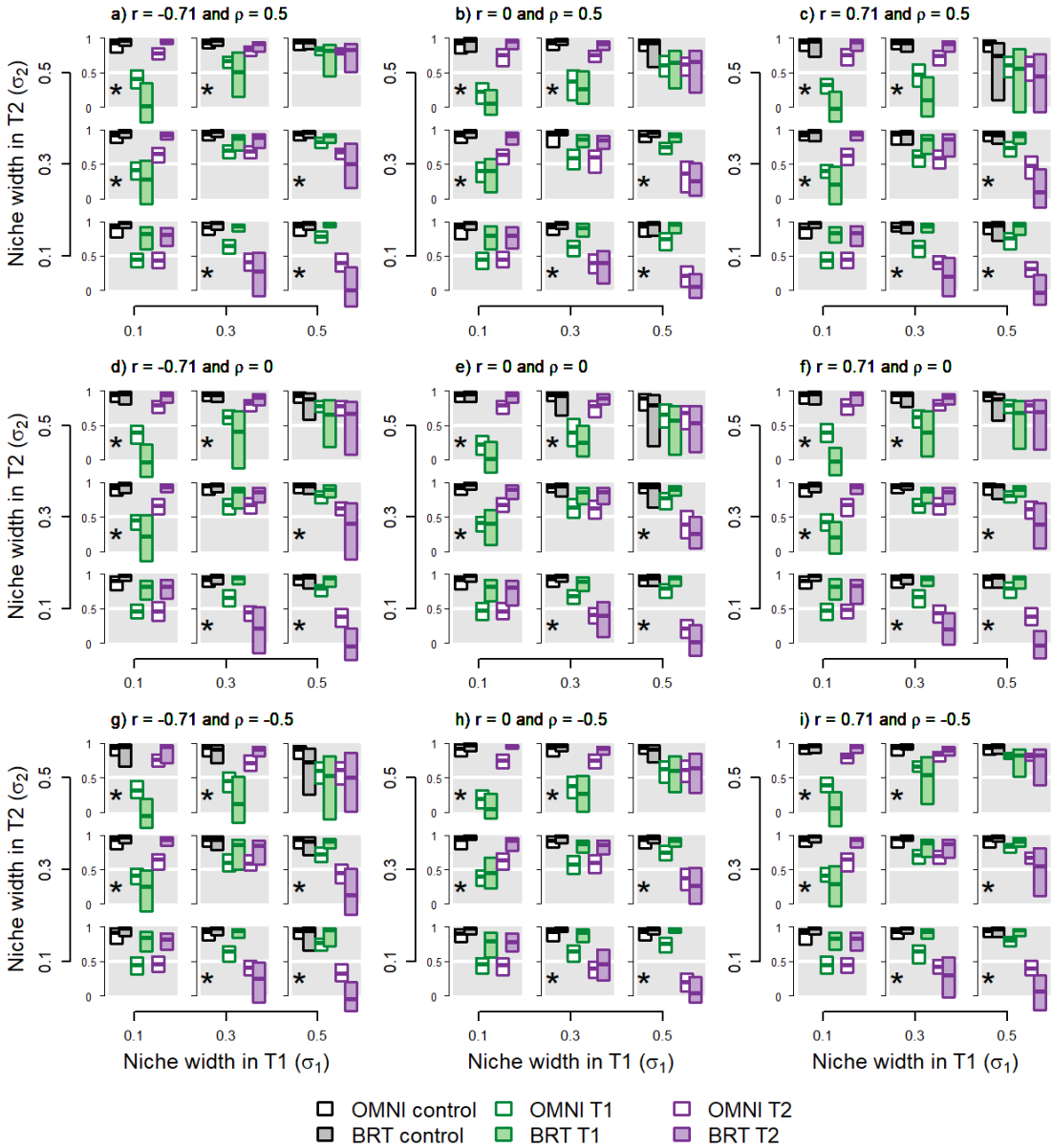


Figure 15. In *Experiments 7* through *10*, BRTs underestimate importance when variables act equally to shape the niche and niche width is narrow or intermediate. When variables act unequally, BRTs overestimate the importance of the more influential variable (narrow niche width, low σ_i), and underestimate the importance of the less influential variable (broad niche width, high σ_i). Aside from these considerations, neither correlation between predictors (r) nor niche covariance (ρ) have strong effects on estimates of importance. Each panel corresponds to a

743 particular landscape (correlation between predictors) and niche topology (niche covariance
744 structure). Within a panel, subplots contain bar plots of control and permuted predictions from
745 OMNI and BRTs. Smaller niche width corresponds to more restriction by the respective variable
746 (and thus more importance) so should yield lower CBI when the variable is permuted. In each panel
747 a subplot for a given value of σ_1 and σ_2 can be interpreted as explained in Box 1. Asterisks in the
748 lower left of a subplot indicate both OMNI and BRTs successfully discriminate between T1 and T2.

749 In the current experiment we examine scenarios in which there is no correlation ($r = 0$) between
750 predictors and variables interact additively to shape the niche ($\rho = 0$; middle panel in each of Figs.
751 13-15). To structure the discussion, we will first focus on cases where T1 and T2 have equal
752 influence on the niche ($\sigma_1 = \sigma 2$, displayed along the upper right-to-lower left diagonal of Figs. 13e,
753 14e, and 15e). Along this sequence niche breadth decreases from broad ($\sigma_1 = \sigma_2 = 0.5$) to moderate
754 ($\sigma_1 = \sigma_2 = 0.3$) to narrow ($\sigma_1 = \sigma_2 = 0.1$). OMNI control had high performance across all combinations
755 of niche width in T1 and T2 (median CBI ranging from 0.89 to 0.93). GAMs and MAX controls
756 performed comparably to OMNI controls across the range of niche breadths. BRTs were also
757 comparable to OMNI control for narrow and moderate niche breadths, but had low precision when
758 niche width was widest ($\sigma_1 = \sigma_2 = 0.5$).

759 Continuing our focus on cases with symmetrical niche breadth ($\sigma_1 = \sigma_2$; Figs. 13e, 14e, and 15e), we
760 recall that narrower breadth increases the restraint imposed by the variables. Thus, permuting a
761 variable for which niche breadth is narrow should reduce CBI more than when breadth is large. The
762 SDMs did not accurately indicate the importance of the variable when niche width was moderate or
763 narrow ($\sigma = 0.1$ or 0.3). In fact, reducing niche width often reduced the estimates of variable
764 importance. We can see this for any of the SDMs in Figs. 13e, 14e, or 15e by going along the
765 diagonal from the upper right to the lower left along which niche changes from broad to narrow.
766 Permuting T1 along this series reduces median CBI for the OMNI model from 0.68 to 0.63 to 0.46,
767 respectively. However, none of the SDMs displayed comparable trends: for GAMs permuted CBI was
768 0.52, 0.90, and 0.90, respectively; for MAX 0.68, 0.76, and 0.64; and for BRTs 0.62, 0.89, and 0.83.
769 (Permuting T2 yields similar discrepancies.) Compared to OMNI treatment models, the SDMs
770 always underestimated the importance of T1 and T2 when niche breadth was moderate or narrow.
771 Despite the failure to accurately estimate importance, MAX had the highest accuracy (median
772 values closest to the respective median values of OMNI) and precision (lowest variability).

773 Now, we turn attention to cases with asymmetrical influence by T1 and T2 ($\sigma_1 \neq \sigma_2$). Again, OMNI
774 control had reliable performance in these cases, and permuting T1 or T2 caused CBI to decrease
775 monotonically with decreasing niche breadth as the respective variable became narrower. All three
776 SDMs successfully discriminated between T1 and T2 (i.e., no overlap between distributions of
777 permuted CBI for T1 and T2). However, all three SDMs tended to overestimate the importance of
778 the more restrictive variable while underestimating the importance of the less restrictive variable.
779 Again, MAX had greater relative accuracy and precision than GAMs and BRTs. Estimates of
780 importance tended to be more uncertain for the more influential variable compared to the less
781 influential variable.

In summary, SDMs successfully discriminated between two variables of unequal influence. However, SDMs had poor calibration accuracy. When variables acted equally, increasing niche breadth increased the apparent importance of the variables. When variables acted unequally SDMs overestimated the importance of the more influential variable and underestimated the importance of the less influential variable. Poor calibration accuracy may be the result of regularization/shrinkage used by MAX and GAMs (Tibshirani 1996; Phillips et al. 2006; Wood 2011) and forward-stagewise-like variable selection in BRTs which functions similarly to the LASSO regularization procedure (Hastie et al. 2009, Ch. 16). These techniques are intended to focus the model on the most important variables while down-weighting less influential factors. Although we used standard tuning practices to adjust the optimal level of model fit (Appendix 5), it is possible that these procedures penalized the less influential variable too much.

Experiment 8: Two collinear, influential variables acting independently

This experiment builds upon the last by allowing the two TRUE variables to be correlated. Here we rotated T2 relative to T1 on the landscape to alter the correlation (r) between them from -0.71 to 0 to 0.72. Many modeling studies discard variables among sets that have pairwise correlations with a magnitude >0.7 (cf. Dorman et al. 2013). The species' response to the environment was again bivariate Gaussian with additive effects of T1 and T2 (Eq. 3), so we will constrain our discussion to panels in the middle row of panels in Figs. 13 through 15.

The results for OMNI and the SDMs were generally qualitatively the same as *Experiment 7* in which variables were uncorrelated on the landscape. As before, when niche breadth was equal and either moderate or narrow, the SDMs underestimated importance of the two variables. When niche breadth was equal and broad, the SDMs accurately estimated variable importance, albeit with low precision. When niche width was unequal, the SDMs reliably discriminated between the variables, but overestimated the importance of the more influential variable and underestimated the importance of the less influential variable. Again, MAX had greater accuracy and precision than GAMs and BRTs. In summary, collinearity between variables had little qualitative effect on the SDMs' discriminatory capacity and did not worsen calibration capacity.

Experiment 9: Two uncorrelated, interacting variables

Here we modify the bivariate niche described in Eq. 3 by considering interactions between variables in how they determine environmental preferences (probability of occurrence), a phenomenon we call "niche covariance." The modified model is bivariate Gaussian and includes a term ρ specifying the sign and strength of niche covariance between the variables:

Eq. 4 $$\Pr(occ) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\left(-\frac{(T1-\overline{T1})^2}{2\sigma_1^2(1-\rho^2)} + \rho\frac{(T1-\overline{T1})(T2-\overline{T2})}{\sigma_1\sigma_2(1-\rho^2)} - \frac{(T2-\overline{T2})^2}{2\sigma_2^2(1-\rho^2)}\right).$$

Again, σ_1 and σ_2 represent the degree of limitation imposed by T1 and T2, the effect of which is mediated by the magnitude of ρ.

In this experiment we restrict our investigation to cases when there is no correlation between the variables on the landscape (r = 0). Varying the covariance parameter ρ alters the orientation of the

species' distribution in niche space and on the landscape even when variables are uncorrelated. Negative covariance (ρ <0; Fig. 16a) causes an increase in one variable and a decrease in the other to increase the probability of presence. Positive covariance (ρ>0; Fig. 16c) causes an increase in both variables or a decrease in both variables to increase the probability of presence. Lack of niche covariance (ρ = 0; Fig. 16b) causes variables to shape the niche only additively so favorable values of one variable can offset unfavorable values of another.

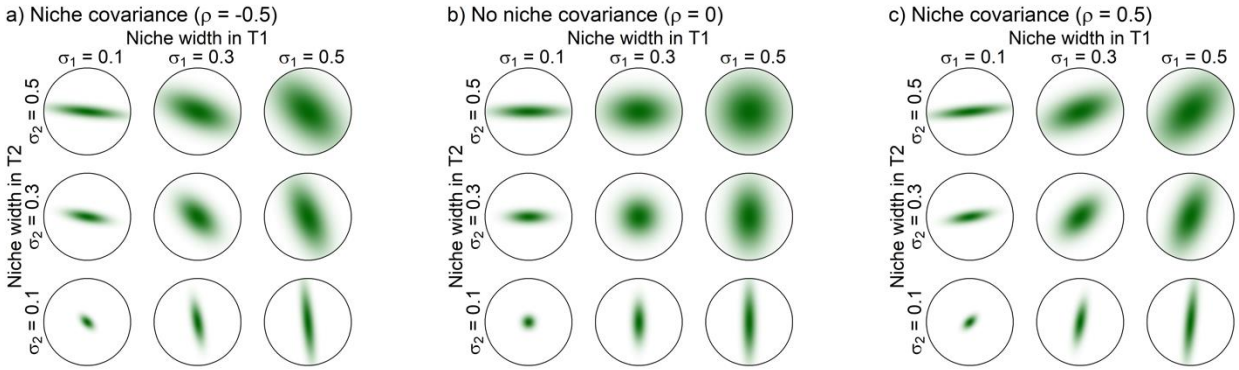


Figure 16. Range maps for *Experiments 9* and *10* in which two TRUE variables (T1 and T2) interact to define the niche. In all cases shown here T1 and T2 were distributed orthogonally to one another on the landscape so were uncorrelated even though they interact to shape the niche (cf. Fig. TRUEvsFALSEa and c). Parameter ρ determines the interaction between T1 and T2 and was varied from -0.5 to 0 to 0.5. The degree of limitation imposed by T1 and T2 are determined by σ_1 and σ_2 plus the magnitude and sign of ρ. Panel (b) is the same as Fig. 12b but repeated here to aid visual comparison. The predicted probability of presence ranges from 0 (white) to 1 (dark green).

Results for this experiment are displayed across the middle panels of Figs. 13-15. OMNI and the SDMs responded in a manner qualitatively similar to that in *Experiments 7* and *8*. Namely, SDMs reliably discriminated between variables when they had unequal influence, but otherwise had low calibration accuracy.

Experiment 10: Niche covariance with collinear variables

This experiment combines the scenarios in *Experiments 7* through *9* varying niche breadth, correlation between predictors, and niche covariance (corner panels of Figs. 13-15). Again, we found that OMNI control and the SDM controls usually performed well and had reliable discriminatory accuracy, but estimated variable importance poorly when niche breadth was equal. When variables acted unequally the SDMs underestimated the importance of the weaker variable and overestimated the stronger variable.

Results using other test statistics

Test statistics other than CBI can be used with permute-after-calibration test to measure variable importance, including AUC calculated with presences and absences (AUC_{pa}) or presences and background sites (AUC_{bg}), and the correlation between control and treatment predictions calculated with presences and absences (COR_{pa}) or with presences and background sites (COR_{bg}). Commonly

used distribution and niche modeling software like Maxent (Phillips et al. 2006; Phillips & Dudík 2008) and the BIOMOD2 (Thuiller et al. 2009) and sdm (Naimi & Araújo 2016) packages for the R Statistical System (R Core Team 2018) allow users the option to estimate variable importance using AUC and/or COR. Results using each of these test statistics are presented in Appendices 1 through 4 so are briefly summarized here.

Compared to CBI, we found that AUC_{pa}, AUC_{bg}, COR_{pa}, and COR_{bg} had generally comparable discriminatory power. However, both versions of AUC generally had somewhat better calibration capacity than CBI when the species was modeled with GAMs or MAX. In some circumstances AUC was more robust to challenging situations. For example, MAX and BRTs, could successfully discriminate between TRUE and FALSE when training sample size was ≥32 AUC_{pa} and AUC_{bg}, versus ≥64 when using CBI. In contrast, GAMs required ≥128 occurrences when using either version of AUC, but required only 8 occurrences when using CBI (Figures S2 in both Appendices 1 and 2). Nevertheless, in most situations AUC performed equivalently to CBI. Both versions of COR had equivalent discrimination compared to CBI but worse accuracy and precision, especially under challenging situations (e.g., low sample size, high collinearity, etc.).

These results suggest AUC might be a viable measure of variable importance. However, neither AUC nor COR is capable of assessing absolute importance of variables. Both variants of AUC suffer from unknowable maximum expected values related to non-deterministic sampling (Jímenez-Valverde et al. 2013) and prevalence (Smith 2013b) which makes it impossible to know if a model performs arbitrarily well (see Figs. S3 in both Appendices 1 and 2). Similarly, both variants of COR are calculated from control and permuted predictions so cannot by themselves indicate whether the control model is accurate. As a result, these tests statistics should be only be used to assess relative importance of variables within the same scenario (i.e., same species in the same region and time period).

Discussion

In this study we used a reductionist approach to test the ability of species distribution models and ecological niche models to measure the influence of variables (results summarized in Table 1). In general, situations that are challenging for generation of accurate models are also challenging for generating well-calibrated measures of variable importance (i.e., estimates that have the same central tendency and distribution as an "omniscient" model with the same form used to simulate the species). Challenging conditions include cases with small sample size, high prevalence, low spatial extent (low environmental variability), and using environmental data that is coarser than the perceptual scale of the species when fine-scale variation in the environment is high. We also found that the permute-after-calibration test (Breiman 2001) for SDMs cannot produce well-calibrated estimates of importance when variables act simultaneously to shape distribution. However, tests were often able to discriminate when variables acted unequally even when the tests had low calibration accuracy. In summary, the ability to measure variable importance reliably was affected by aspects related to 1) study design, 2) spatial scale, 3) landscape, and 4) the niche. We structure the discussion of results around these four themes.

Study design and variable importance

Difference between algorithms.—The primary distinction between SDM and ENM algorithms is the manner in which they attempt to delineate the multi-dimensional environmental space inhabited by a species (Guisan & Zimmermann 2000). Algorithms differ in their ability to fit smooth (GAMs) versus "step-like" (BRTs) functions or a mixture thereof (MAX). Algorithms also differ in their ability to capture interactions between variables (MAX: two-way, BRTs: potentially highly complex, GAMs: typically two-way), manage collinearity (MAX: regularization; GAMs: shrinkage; BRTs: forward-stagewise-like variable selection; Hastie et al. 2009, Ch. 16), and perform model and variable selection (MAX as implemented here: AIC_c, GAMs: automated shrinkage; BRTs: variables most correlated with response at each step). Owing to these differences, the choice of modeling algorithm is one of the largest contributors to variation in predictions (Dorman et al. 2008; Diniz-Filho et al. 2009; Buisson et al. 2010; Barbet-Massin et al. 2012; Rapacciuolo et al. 2012; Buklin et al. 2015; Norberg et al. Eco Monog in press), with no one algorithm being best for all species (Qiao et al. 2015). As a result, the manner in which an algorithm estimates the niche will affect its capacity to measure variable importance.

We found that in simple scenarios with one influential and one uninfluential variable (*Experiments 1* through *6*) GAMs often had better discrimination and somewhat calibration than MAX or BRTs, even though they sometimes did not converge under challenging situations (e.g. low sample size Fig. 3a and b). In scenarios with two influential variables (*Experiments 7* through *10*), the three SDMs had similar discrimination accuracy but MAX had somewhat better calibration capacity (mean values closer to mean values of the OMNI model and lower variability around the central tendency). BRTs had the poorest discrimination and calibration accuracy across all scenarios despite the use of standard parameterization procedures to generate well-calibrated models (Appendix 5). Troublingly, some algorithms (especially BRTs) invariably used information from the uninfluential variable. This is particularly concerning because in many cases these same models (when using unpermuted variables) often yielded values of CBI that were still within the range indicative of "reliable" performance (i.e., were still seemingly accurate). In some cases, use of the uninfluential variable even enhanced model performance. For example, in the simplest scenario (*Experiment* 1), the correlation between CBI of the unpermuted control predictions and the BRT-specific importance variable metric (deviance reduction times frequency of variable use; Elith et al. 2008) was -0.19 for the influential variable and 0.19 for the FALSE variable. Thus, using the uninfluential variable actually increased model performance. This is alarming but supports findings that using variables with no actual relationship to a species' occurrence can yield models as accurate as models trained using "real" variables (Buklin et al. 2015; Fourcade et al. 2018). Thus, performance of a control model is not a reliable guide to the capacity of the model to measure variable importance.

Ultimately the ability of a model to measure variable importance accurately depends on its capacity to estimate the functional relationship between a species and its environment. In our case, models had to recreate either a logistic or a bivariate Gaussian function, both of which are similar to functional forms we expect to see in the niches of real species (Whittaker 1970?; but see Austin et al. 1990). Each of the algorithms we used were able to reproduce good approximations of these functions under the simplest situation (Fig. 2). However, small sample size (Fig. 3) or insufficient differentiation between occurrences and background sites (Fig. 6) quickly eroded models' ability to

931 estimate the real functional response. In some cases, even the omniscient model was not able to
932 effectively differentiate between influential and uninfluential variables (e.g., high prevalence in Fig.
933 5). The omniscient model does not use calibration data, so in this case the failure of the model is
934 due solely to lack of differentiation between presence and background sites (or presence and
935 absence sites) used to evaluate the model. SDMs are even more challenged than the omniscient
936 model since they must estimate the response to the environment from potentially inadequate data
937 then assess importance using similarly inadequate data.

938 Obviously, more research needs to be conducted on how well different SDM algorithms assess
939 variable importance. Additionally, a small but growing number of methods that do not rely on SDMs
940 can be used for assessing the effects of variables on distributions (Greenberg et al. 2015; Guisande
941 et al. 2017). We expect that no single algorithm or method is best for all species in all scenarios
942 (Qiao et al. 2015) and thus recommend methodological plurality when estimating variable
943 importance.

944 *Sample size*.—Rare species or species with few available occurrences present special challenges
945 because a) occurrences may not adequately capture the full extent of the species' environmental
946 tolerances (Sax new article on islands in NCC?); b) small samples size diminishes confidence in
947 coefficient estimates (Lomba et al. 2010); and c) small samples may contain spurious correlations
948 between predictors (Ashcroft et al. 2011). As a result, models calibrated on small samples are likely
949 to have highly variable performance (Kadmon et al 2003; Hernandez et al. 2006; Thibaud et al.
950 2014; Soultan & Safi 2017), and thus by inference, less ability to measure variable importance.

951 In our case SDMs reliably discriminated between influential and uninfluential variables when
952 sample size was ≥8 (GAMs), or ≥64 (BRTs and MAX), although we expect that this is a lower limit
953 for real-world applications. Recent work demonstrates that very small sample sizes may be
954 adequate for generating accurate models in some situations (Soultan & Safi 2017), especially when
955 coupled with novel techniques (Lomba et al. 2010; Breiner et al. 2015). However, we found that
956 across scenarios performance of the control models was not always a reliable guide to the ability of
957 the same models to differentiate between variables. If we make the reasonable assumption that
958 increasing sample size would likely improve measurement of variable importance in challenging
959 situations, then we can conclude generating accurate estimates of variable importance probably
960 requires larger sample sizes than needed for generating accurate models.

961 *Spatial scale and variable importance*

962 For real landscapes decisions around spatial extent, prevalence, spatial resolution of environmental
963 data, and environmental autocorrelation almost always affect one another (e.g., Thuiller et al. 2003;
964 Suárez-Seoane et al. 2014; Connor et al. 2017; Mertes & Jetz 2017). For example, increasing study
965 region extent nearly always reduces prevalence because species will be absent from larger areas
966 (Anderson & Raza 2010). Study region extent may also make locating fine-resolution
967 environmental data difficult (Record et al. 2018), thereby necessitating use of coarse-grained data.
968 In turn, using cells with coarse resolution generally increases prevalence because a species needs
969 only occur in a portion of a cell to be scored as "present" in the entire cell (Seo et al. 2008; Lauzeral
970 et al. 2013; Connor et al. 2017). Cell size typically affects the magnitude of spatial autocorrelation

apparent in environmental data because smaller cells emphasize fine-scale differences while larger cells average across them (Mertes & Jetz). Spatial autocorrelation interacts with the spatial scale of the species' perception of the environment and niche breadth to shape a model's ability to reconstruct/predict the species' response (Soultan & Safi 2017; Connor et al. 2017). Adding to these complications, many metrics of model performance are confounded by extent and prevalence (Santika 2011; Smith 2013b; Somodi et al. 2017; Sor et al. 2017), which compromises their use as metrics of variable importance (Figs. 5 and 7). Owing to the interdependencies of aspects of scale, it is extremely difficult to disentangle how scale affects the ability of models to measure variable importance.

We used simulations to assess the independent effects of prevalence, study region extent, spatial resolution of predictors, and spatial autocorrelation in environmental variables in a manner that isolates the effects of each. We found that models were most reliable when prevalence was <0.5 (Fig. 5) and when study region extent was "intermediate" or "large" relative to the size of the geographic gradient across which the species responded to its environment (Fig. 7). We also found that models when trained on data that had a spatial resolution different from the perceptual scale of the species, except when the spatial resolution of predictors was coarse and spatial autocorrelation in the environment was low (Fig. 9).

As noted, nearly all of these aspects of scale are under direct control of the modeler or at least directly affected by decisions made by the modeler, but nearly all are tightly interdependent. Thus, our recommendations need to be interpreted with regards to how decisions over one aspect of scale affect the others. We recommend using study regions that encompass the entire species' range plus enough extra area to ensure the species resides in less than half the region. Current best practices in distribution modeling recommend that the region from which background/absence sites are drawn should only represent the area that is "accessible" to the species through dispersal (Barve et al. 2011; Peterson et al. 2011; Araújo et al. 2019). Depending on the scale of dispersal, the accessible area may or may not be large enough to ensure the species occupies less than half the area. Thus, modelers may be faced with a difficult decision. We expect that the "accessible area" rule will be more applicable when the goal of a study is to delineate a niche and/or predict its potential future geographic distribution. But we expect that our "less than half" rule will be more applicable when the goal of a study is to measure variable importance. We also recommend delineating study regions that are large enough to cover broad environmental gradients or at least selecting predictors that display sufficient variation across the study region. Smaller extents and predictors with low variability across samples do not offer sufficient variation to reliably differentiate occurrences from background/absences and therefore confound attempts to measure variable importance (Fig. 7). Finally, we recommend using environmental data at a resolution at the same grain size or smaller than the scale of species' perception of the environment (Mertes & Jetz 2017; Araújo et al. 2019). When this is not possible, problems arising from mismatching perceptual and data scales might be ameliorated by high levels of spatial autocorrelation. We did not explore situations where the species perceives different environmental factors at different scales (Graf et al. 2005; Hawkins et al. 2012; Mertes & Jetz 2017). However, it seems reasonable to assume that our results would apply to individual variables regardless of the level of spatial autocorrelation and scale of perception of other variables. Likewise, we did not explore how niche width interacts with

spatial resolution and spatial autocorrelation. However, we expect that scale mismatches might also be ameliorated if the species has broad environmental tolerances and so is indifferent to fine-grained variation (Connor et al. 2017; Mertes & Jetz 2017). Thus, the scale of environmental data may not always need to be matched with the species' scale of environmental perception to reliably estimate variable importance.

Niche dimension, collinearity, and niche covariance

Niches and distributions are shaped by a multitude of factors acting independently and interactively. Environmental factors can act with different intensities, spatial scales (Mertes & Jetz 2017; Smith et al. in review), and life stages (Doak & Morris 2010), and/or in different seasons or years (Garroway et al. 2011). Environmental limits are mediated by biotic interactions (Stanton-Geddes et al. 2012; Anderson 2017; Godsoe et al. 2017), disturbance (Conlisk et al. 2013; Liang et al. 2018), and dispersal and population dynamics (Holt 2009; Pagel & Schurr 2012; Conlisk et al. 2013; Fordham et al. 2013). Moreover, environmental factors are nearly always spatially autocorrelated and therefore statistically collinear (Jiménez-Valverde et al. 2009; Braunisch et al. 2013). Finally, environmental factors can interact additively, multiplicatively, or in more complex ways to shape the niche (Smith 2013a). For example, a species may be able to tolerate hot conditions if ample water is available, but otherwise thrive under colder conditions even if water is scarce. Taken together, these considerations make identifying important variables and measuring their influence extremely difficult.

We found that models were unable to produce accurately calibrated estimates of variable importance when variables acted equally. In this same situation models tended to underestimate importance as the variables exerted more influence (*Experiments 8* through *10*; Figs. 13-15). When variables acted unequally, models overestimated the importance of the more influential variable and underestimated the importance of the less influential variable. Nonetheless, models were usually able to discriminate between variables of different influence. Surprisingly, correlations between variables and interactions between variables in how they shape the niche (niche covariance) did not quantitatively affect these conclusions, at least at the levels of correlation (±0.71) and covariance (±0.5) investigated here. However, we did find that collinearity between predictors reduced discrimination and calibration accuracy for MAX and especially BRTs (Fig. 11). For both of these algorithms even low levels of collinearity (i.e., $|r| \geq 0.38$) increased variability around estimates of importance. Although results varied somewhat by the sign of the relationship between variables, we did not find strong differences between negative and positive correlations (cf. Mela & Kapolle 2002; Smith et al. 2009).

Our results support the common practice of retaining only variables with pairwise correlations $|r| < \sim 0.7$ (but see Braunisch et al. 2013), although even lower thresholds may be required for high calibration accuracy (Fig. 11; Dorman et al. 2013; Júnior & Nóbrega 2018). However, discarding variables is problematic since it effectively assumes excluded variables have zero importance with no uncertainty. Our results highlight the problem of including uninfluential variables among the variables used to train SDMs (Braunisch et al. 2013; Fourcade et al. 2018). The most effective solution to this dilemma is to ensure variables used in the models are indeed at least moderately influential as evidenced by field or laboratory studies (Mod et al. 2016; Gardener et al. 2019).

However, in most cases we lack complete knowledge of relevant variables, so another solution may be to construct multiple models for different sets of uncorrelated variables (e.g., Petitpierre et al. 2017) then compare performance.

Given that most niches are almost surely defined by more than one variable, that environmental variables are typically correlated, and that variables likely interact to determine distribution and niches, it is challenging to make general recommendations. The good news is that algorithms were able to discriminate between variables when they acted unequally despite moderate levels of correlations and interactions between variables. The bad news is that models could not produce accurate calibrations of importance. Notably, the effect size (mean tendency) of changes in CBI (and other metrics) did not reflect the magnitude of restriction imposed by the niche variables. For example, reducing niche width by a factor of five did not reduce CBI for permuted models by a factor of five (it sometimes even increased it). Thus, neither CBI, AUC, nor the COR are "proportionate" measures of variable importance (Smith & Santos in prep.) when coupled with the permute-after-calibration test. In summary, interpreting the actual numerical value of a measure of importance as indicative of effect size is unwarranted. Rather, we strongly recommend relying on discrimination (i.e., ranking of variables) over calibration (i.e., measures of central tendency and variability) of variable influence, especially when variables have seemingly equal contributions.

Model performance as an indicator of inferential power

We found that performance of the control SDMs was a reliable indicator of discrimination accuracy but not of calibration accuracy. That is, when the control model performed well it was usually able to discriminate between variables even if it was unable to calibrate the magnitude of their influence. This is somewhat surprising given that the choice of predictors has a direct effect on model performance (Peterson & Cohoon 1999; Petitpierre et al. 2017; Bosh et al. 2018; Guevara et al. 2018; Barbet-Massin & Jetz 2014; for exceptions see Synes & Osborne 2011 and Buklin et al. 2015). Presumably sets of predictors that produce reliable models do so because they have more direct bearing on the niche or range (Austin 2002). However, we found this not to be the case. Thus, we strongly advise against using the permute-after-calibration procedure coupled with any of the test statistics employed here (CBI, AUC_{pa}, AUC_{bg}, COR_{pa}, and COR_{bg}) in studies requiring calibrated estimates of variable importance (e.g., testing a hypothesis that states one variable is twice as important as another). However, the permute-after-calibration test is capable of differentiating between variables under most conditions where models perform well (using unpermuted data). Discriminating between variables is a less informative test than calibrating their influence (Barbet-Massin & Jetz 2014), but more reliable under common situations.

Further work

Studies using real species are often complicated by the fact that the true importance of variables is unknown. We used a reductionist approach in which only a few factors were varied at a time while all other potentially confounding issues were assumed constant or non-extant. Owing to the many simplifications inherent in the simulations, we expect that the conditions under which models did succeed are too lenient for real-world application. Likewise, we also expect many of the simplifying assumptions we made such as perfect detectability (Lohaz-Manafort et al. 2014) or lack of biotic

interactions (Godsoe et al. 2017) affect inferences about variable importance (Meynard et. al in press). Indeed, in reviewing the literature on the effects of factors investigated here (scale, correlation, etc.) on SDM performance, we often noted differences between studies using real species and studies using virtual species. Part of the discrepancy may be due to the fact that for virtual species complicating factors can be controlled, meaning simulations may be a better reflection of a reality that is difficult to observe (Meynard et. al in press). For example, simulations do not suffer issues related to detectability (Lohaz-Manafort et al. 2014) and uncertainty in coordinate locations (Naimi et al. 2014), unless these are explicitly incorporated into the scenarios. On the other hand, discrepancies between empirical and simulation studies may be due to the fact that virtual species and landscapes are necessarily much simpler and therefore a poor reflection of reality (Meynard et. al in press). An alternative strategy to the reductionist one used here is to simulate niches or distributions as realistically as possible (e.g., Zurell et al. 2016; Norberg et al. in press Eco Monog), then apply a battery of procedures and test statistics and compare them against an omniscient model to identify situations that are conducive to measuring variable importance accurately.

We examined just one procedure, the permute-after-calibration test, and focused on just five test statistics (CBI, plus two versions of AUC and COR) to ascertain absolute variable importance. Many other procedures and test statistics deserve attention (Warren et al. in prep; Smith & Santos in prep.). Outstanding methodological questions include: Which tests are most reliable? What conditions are most challenging for disentangling the importance of variables? Are algorithm-specific tests as reliable as algorithm-independent tests? Does data type (abundance, presence/background, presence/absence, detection/non-detection) affect inferences for the same species? How well do tests of performance perform in high-dimensional niche spaces?

Ultimately, correlative models can only establish association, not causality (Austin 2002; Fourcade et al. 2018). Even many "mechanistic" models of niches and distributions rely on regressions for estimating parameters, and so cannot by themselves establish causality. Obviously, either type of model can only identify variables considered by modelers in the first place (Mod et al. 2014; Gardener et al. 2019). This is a critical point because the importance of a variable—even if measured using a test reflecting "absolute" importance—will always be relative to the set of candidate variables or processes considered in the modeling process.

The fundamental outcome we desire from the present work is to inspire further research into how and under what conditions particular procedures and test statistics can be used with SDMs and ENMs to identify and measure the most important variables shaping distributions and niches. Ecologists have a robust (yet still growing) appreciation for factors that affect the predictive accuracy of SDMs and ENMs (e.g., Guisan et al. 1999; Araújo et al. 2005; Elith et al. 2006; Dobrowski et al. 2011; Rapacciuolo et al 2012; Smith et al. 2013; Zurell et al. 2016; Guevara et al. 2018; Yates et al. 2018; Smith et al. 2019)… and yet we have scant knowledge of how reliably these same models assess variable importance despite the enormous popularity of this particular application (Bradie & Leung 2017). What environmental factors determine species' ranges and environmental tolerances, and how can we identify them?

Acknowledgements

ABS was supported by the Alan Graham Fund in Global Change.

Online appendices

Appendix 1: Using AUC calculated with presences and absences (AUC_{pa}) as a test statistic for the permute-after-calibration test of variable importance

Appendix 2: Using AUC calculated with presences and absences (AUC_{bg}) as a test statistic for the permute-after-calibration test of variable importance

Appendix 3: Using the correlation test statistic calculated with presences and absences (COR_{pa}) as a test statistic for the permute-after-calibration test of variable importance

Appendix 4: Using the correlation test statistic calculated with presences and absences (CORbg) as a test statistic for the permute-after-calibration test of variable importance

Appendix 5: Model parameterization

Appendix 6: Sensitivity of the Continuous Boyce Index to improbable test presences

Literature cited

Acevedo, P., Jiménez-Valverde, A., Lobo, J.M., and Real, R. 2017. Predictor weighting and geographical background delimitation: Two synergistic sources of uncertainty when assessing species sensitivity to climate change. Climatic Change 145:131-143.

Albrecht, M.A., Becknell, R.E., and Long, Q. 2016. Habitat change in insular grasslands: Woody encroachment alters the population dynamics of a rare ecotonal plant. Biological Conservation 196:91-102.

Anderson, R.P. 2017. When and how should biotic interactions be considered in models of species niches and distributions? Journal of Biogeography 44:8-17.

Anderson, R.P. and A. Raza. 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 37:1378-1393.

Angert, A.L., Bayly, M., Sheth, S.N., and Paul, J.R. 2018. Testing range-limit hypotheses using range-wide habitat suitability and occupancy for the Scarlet Monkeyflower (*Erythranthe cardinanis*). American Naturalist 191:E76-E89.

Araújo, M.B, Anderson, R.P., Barbosa, A.M., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., and Rahbek, C. 2019. Standards for distribution models in biodiversity assessments. Science Advances 5:eaat4858.

Araújo, M.B., Ferri-Yáñez, F., Bozinovic, F., Marquet, P.A., Valladares, F., and Chown, S.L. 2013. Heat freezes niche evolution. Ecology Letters 16:1206-1219.

Araújo, M.B., R.J. Whittaker, R.J. Ladle, and M. Erhard. 2005. Reducing uncertainty in projections of extinction risk from climate change. Global Ecology and Biogeography 14:529-538.

Ashcroft, M.B., French, K.O., and Chisholm, L.A. 2011. An evaluation of environmental factors affecting species distributions. Ecological Modeling 222:524-531.

Austin, M.P. 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical modeling. Ecological Modeling 157:101-118. Austin et al. 1990

Austin, M.P. and K.P. Van Niel. 2011. 2011 Improving species distribution models for climate change studies: Variable selection and scale. Journal of Biogeography 38:1-8.

Austin, M.P., L. Belbin, J.A. Meyers, M.D. Doherty, and M. Luoto. 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. Ecological Modelling 199:197-216.

Barbet-Massin, M. and Jetz, W. 2014. A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modeling. Diversity and Distributions 20:1285-1295.

Barbet-Massin, M., F. Jiguet, C.H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution 3:327-338.

Barnes, R., and Clark, A.T. 2017. Sixty-five million years of change in temperature and topography explain evolutionary history in Eastern North American plethodontid salamanders. American Naturalist 190:E1-E12.

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

Bosch, S., Tyberghein, L., Deneudt, K., Hernandez, F., and De Clerck, O. 2018. In search of relevant predictors for marine species distribution modeling using the MarineSPEED benchmark test. Diversity and Distributions 34:144-157.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., and Schmiegelow, F.K.A. 2002. Evaluating resource selection functions. Ecological Modeling 157:281-300.

Bradie, J. and Leung, B. 2017. A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. Journal of Biogeography 44:1344-1361.

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., and Bollmann, K. 2013. Selecting from correlated climate variables: A major source of uncertainty for predicting species distributions under climate change. Ecography 36:971-983.

Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

1201 Breiner, F.T., Guisan, A., Bergamini, A., and Nobis, M.P. 2015. Overcoming limitations of modeling
1202 rare species by using ensembles of small models. Methods in Ecology and Evolution 6:1210-
1203 1218.

1204 Buisson, L., W. Thuiller, N. Casajus, S. Lek, and G. Grenouillet. 2009. Uncertainty in ensemble
1205 forecasting of species distribution. Global Change Biology 16:1145-1157.

1206 Buklin, D.N., Basille, M., Benscoter, A.M., Brandt, L.A., Mazotti, F.J., Romañach, S.S., Speroterra, C., and
1207 Watling, J.I. 2015. Comparing species distribution models constructed with different subsets of
1208 environmental predictors. Diversity and Distributions 21:23-35.

1209 Burnham, K.P. and Anderson, D.R. 2002. Model Selection and Multimodel Inference: A Practical
1210 Information-Theoretic Approach, 2nd ed. Springer, New York.

1211 Callen, S.T. and Miller, A.J. 2015. Signatures of niche conservatism and niche shift in the North
1212 American kudzu (*Pueraria montana*) invasion. Diversity and Distributions 21:853-863.

1213 Conlisk, E., Syphard, A.D., Franklin, J., Flint, L., Flint, A., and Regan, H. 2013. Uncertainty in
1214 assessing the impacts of global change with coupled dynamic species distribution and
1215 population models. Global Change Biology 19:858-869.

1216 Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F., and Liu, J. 2017. Effects of
1217 grain size and niche breadth on species distribution modeling. Ecography 41:1270-1282.

1218 Diniz-Filho, J.A., Bini, L.M., Rangel, T.F., Loyola, R.F., Hof, C., Nogués-Bravo, D., and Araújo, M.B.
1219 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover
1220 under climate change. Ecography 32:897-906.

1221 Doak, D.F. and Morris, W.F. 2010. Demographic compensation and tipping points in climate-
1222 induced range shifts. Nature 467:959-962.

1223 Dobrowski, S.Z., J.H. Thorne, J.A. Greenberg, H.D. Safford, A.R. Mynsberge, S.M. Crimmins, and A.K.
1224 Swanson. 2011. Modeling plant ranges over 75 years of climate change in California, USA:
1225 Temporal transferability and species traits. Ecological Monographs 81:241-257.

1226 Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl G., Carré, G., Marquéz, J.R.G., Gruber, B,.
1227 Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder,
1228 B., Skidmore, A.K., Zurrell, D., and Lautenbach, S. 2013. Collinearity: A review of methods to
1229 deal with it and a simulation study evaluating their performance. Ecography 36:27-46.

1230 Dormann, C.F., Purchke, O., Márquez, J.R.G., Lautenbach, S., and Schroder, B. 2008. Components of
1231 uncertainty in the species distribution analysis: A case study of the Great Gray Shrike. Ecology
1232 89:3371-3386.

1233 Elith, J., C.H. Graham, R.P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R.
1234 Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y.
1235 Nakazawa, J.McC. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E.

Schapire, J. Soberón, S. Williams, M.S. Wisz, and N.E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813. Feng & Papeş 2014

Fordham, D.A., Akçakaya, H.R., Araújo, M.B., Keith, D.A., and Brook, B.W. 2013. Tools for integrating range change, extinction risk and climate change information into conservation management. Ecography 36:956-964. Fourcade et al. 2018

Francis, A.P and D.J. Currie. 2003. A globally consistent richness-climate relationship for angiosperms. American Naturalist 161:523-536.

Franklin, J., Davis, F.W., Ikegami, M., Syphard, A.D., Flint, L.E., Flint, A.L., and Hannah, L. 2013. Modeling plant species distributions under future climates: How fine do climate projections need to be? Global Change Biology 19:473-483.

Gardener, A.S., Mclean, I.M.D., and Gaston, K.J. 2019. Climatic predictors of species distributions neglect biophysiologically meaningful variables. Diversity and Distributions 25:1273-1288.

Garroway, C.J., J. Bowman, G.L. Holloway, J.R. Malcom, and P.J. Wilson. 2011. The genetic signature of rapid range expansion by flying squirrels in response to contemporary climate warming. Global Change Biology 17:1760-1769.

Gobeyn, S., Volk, M., Dominguez-Granda, L., and Goethals, P.L.M. 2017. Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in Ecuador. Environmental Modeling and Software 92:269-316.

Godsoe, W., Franklin, J., and Blanchet, F.G. 2017. Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. Ecology and Evolution 7:645-664.

Graf, R.F., Bollman, K., Suter, W., and Bugmann, H. 2005. The importance of spatial scale in habitat models: Capercaillie in the Swiss Alps. Landscape Ecology 20:703-717.

Greenberg, J.A., Santos, M.J., Dobrowski, S.Z., Vanderbilt, V.C., and Ustin, S.L. 2015. Quantifying environmental limiting factors on tree cover using geospatial analysis. Public Library of Science ONE 10:e0114686.

Guevara, L., Gerstner, B.E., Kass, J.M., and Anderson, R.P. 2018. Toward ecologically realistic predictions of species distributions: A cross-time example from tropical montane cloud forests. Global Change Biology 24:1511-1522.

Guisan, A. and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147-186.

Guisan, A., S.B. Weiss, and A.D. Weiss. 1999. GLM versus CCA spatial modeling of plant species distribution. Plant Ecology 143:107-122.

Guisande, C., García-Roselló, Heine, J., González-Dacosta, González Vilas, L., García Pérez, B.J., and Lobo, J.M. 2017. SPEDInstandR: An algorithm based on a fluctuation index for selecting predictors in species distribution modeling. Ecological Informatics 37:18-23.

Hargreaves, A.L., Samis, K.E., and Eckert, C.G. 2014. Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. American Naturalist 183:157-173.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, New York.

Hawkins, B.A. 2012. Eight (and a half) deadly sins of spatial analysis. Journal of Biogeography 39:1-9.

Hernandez, P.A., Graham, C.H., Master, L.L., and Albert, D.L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773-785.

Hijmans, R.J. 2019. raster: Geographic Data Analysis and Modeling. R package version 2.9-23. https://CRAN.R-project.org/package=raster.

Hijmans, R.J., Phillips, S.J., Leathwick, J., and Elith, J. 2017. dismo: Species Distribution Modeling. R package version 1.1-4. https://CRAN.R-project.org/package=dismo.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., and Guisan, A. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modeling 199:142-152.

Hoban, S., Dawson, A. Robinson, J., Smith, A.B., Strand, A. In press. Inference of biogeographic history by formally integrating distinct lines of evidence: genetic, environmental niche, and fossil. Ecography.

Holt, R.D. 2009. Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. Proceedings of the National Academy of Sciences USA 106 (Suppl. 2):19659-19665.

Jiménez-Valverde, A., Acevedo, P., Barbosa, M., Lobo, J.M., and Real, R. 2013. Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. Global Ecology and Biogeography 22:508-516.

Jiménez-Valverde, A., Nakazawa, Y., Lira-Noriega, A., and Peterson, A.T. 2009. Environmental correlation structure and ecological niche projections. Biodiversity Informatics 6:28-35.

Júnior, P.D.M. and Nóbrega, C.C. 2018. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. Public Library of Science ONE 13:e0202403.

Kadmon, R., Farber, O., Danin, A. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13:853-867. La Rest et al. 2014

1305 Lauzeral, C., Grenouillet, G., and Brosse, S. 2013. Spatial range shape drives grain size effects in
1306 species distribution models. Ecography 36:778-787.

1307 Lee-Yaw, J.A., Kharouba, H.M., Bontrager, M., Mahony, C., Csergö, A.M., Noreen, A.M.E., Li, Q.,
1308 Schuster, R., and Angert, A.L. 2016. A synthesis of transplant experiments and ecological niche
1309 models suggest that range limits are often niche limits. Ecology Letters 19:710-722.

1310 Liang, Y., Duveneck, M.J., Gustafson, E.J, Serra-Diaz, J.M., and Thompson, J.R. 2018. How
1311 disturbance, competition, and dispersal interact to prevent tree range boundaries from keeping
1312 pace with climate change. Global Change Biology 24:e335-e351.

1313 Lohaz-Manafort, J., Guillera-Arroita, G., and Wintle, B.A. 2014. Imperfect detection impacts the
1314 performance of species distribution models. Global Ecology and Biogeography 23:504-515.

1315 Lomba, A., L. Pellissier, C. Randin, J. Vicente, J. Horondo, and A. Guisan. 2010. Overcoming the rare
1316 species modeling complex: A novel hierarchical framework applied to an Iberian endemic plant.
1317 Biological Conservation 143:2647-2657.

1318 McInery, G.J. and Purvis, D.W. 2011. Fine-scale environmental variation in species distribution
1319 modeling: regression dilution, latent variables and neighborly advice. Methods in Ecology and
1320 Evolution 2:248-257.

1321 Meinshausen, N. & Bühlmann, P. 2010. Stability selection. Journal of the Royal Statistical Society:
1322 Statistical Methodology Series B 72:417-473.

1323 Mela, C.F., and Kapolle, P.K. 2002. The impact of collinearity on regression analysis: The
1324 asymmetric effect of negative and positive correlations. Applied Economics 34:667-677.

1325 Menke, S.B., D.A. Holway, R.N. Fisher, and W. Jetz. 2009. Characterizing and predicting species
1326 distributions across environments and scales: Argentine ant occurrences in the eye of the
1327 beholder. Global Ecology and Biogeography 18:50-63.

1328 Mertes, K. and Jetz, W. 2017. Disentangling scale dependencies in species environmental niches
1329 and distributions. Ecography 41:1604-1615.

1330 Meynard, C.N. and D.N. Kaplan. 2013. Using virtual species to study species distributions and
1331 model performance. Journal of Biogeography 40:1-8.

1332 Meynard, C.N., Leroy, B., and Kaplan, D.M. In press. Testing methods in species distribution
1333 modelling using virtual species: what have we learnt and what are we missing? Ecography.

1334 Mod, H.K., Scherrer, D., and Guisan, A. 2016. What we use is not what we know: Environmental
1335 predictors in plant distribution models. Journal of Vegetation Science 27:1308-1322.

1336 Morelli, T.L., Daly, C., Dobrowski, S.Z., Dulen, D.M., Ebersole, J.L., Jackson, S.T., Lundquist, J.D., Millar,
1337 C.I., Maher, S.P., Monahan, W.B., Nydick, K.R., Redmond, K.T., Sawyer, S.C., Stock, S., and
1338 Beissinger, S.R. 2016. Managing climate change refugia for climate adaptation. Public Library
1339 of Science ONE 11:e0159909.

1340 Morelli, T.L., Smith, A.B., Kastely, C.R., Mastroserio, I., Moritz, C., and Beissinger, S.R. 2012.
1341 Anthropogenic refugia ameliorate the severe climate-related decline of a montane mammal
1342 along its trailing edge. Proceedings of the Royal Society of London B 279:4279-4286.

1343 Morelli, T.L., Smith, A.B., Mancini, A.N., Balko, E. A., Borgenson, C., Dolch, R., Farris, Z., Federman, S.,
1344 Golden, C.D., Holmes, S., Irwin, M., Jacobs, R.L., Johnson, S., King, T., Lehman, S., Louis, E.E. Jr.,
1345 Murphy, A., Randriahaingo, H.N.T., Lucien, Randriannarimanana, H.L.L., Ratsimbazafy, J.,
1346 Razafindratsima, O.H., and Baden, A.L. In review. Deforestation and climate change projected to
1347 eliminate Madagascar rainforest habitat by 2080.

1348 Moudrý, V. 2015. Modelling species distributions with simulated virtual species. Journal of
1349 Biogeography 42:1365-1366.

1350 Naimi, B. & Araújo, M.B. 2016. sdm: a reproducible and extensible R platform for species
1351 distribution modelling. Ecography 39:368-375.

1352 Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., and Toxopeus, A.G. 2014. Where is positional
1353 uncertainty a problem for species distribution modeling? Ecography 57:191-203.

1354 Norberg, A., Abrego, N., Blanchet, F.G. Adler, F.R., Anderson, B.J., Anttila, J., Araújo, M.B., Dallas, T.,
1355 Dunson, D., Elith, J., et al. In press. A comprehensive evaluation of predictive performance of 33
1356 species distribution models at species and community levels. Ecological Monographs.

1357 Pagel, J., and Schurr, F.M. 2012. Forecasting species ranges by statistical estimation of ecological
1358 niches and spatial population dynamics. Global Ecology and Biogeography 21:293-304.

1359 Peterson, A.T. and K.P. Cohoon. 1999. Sensitivity of distributional prediction algorithms to
1360 geographic data completeness. Ecological Modelling 117:159-164.

1361 Peterson, A.T., J. Soberón, R.G. Pearson, R.P. Anderson, E. Martínez-Meyer, M. Nakamura, and M.B.
1362 Araújo. 2011. Ecological Niches and Geographic Distributions. Monographs in Population
1363 Biology 49. Princeton University Press, Princeton, NJ.

1364 Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., and Guisan, A. 2017. Selecting predictors
1365 to maximize the transferability of species distribution models: Lessons from cross-continental
1366 plant invasions. Global Ecology and Biogeography 26:275-287.

1367 Phillips, S.J. and Dudík, M. 2008. Modeling species distributions with Maxent: New extensions and
1368 a comprehensive evaluation. Ecography 31:161-175.

1369 Phillips, S.J., Anderson, R.P., and Schapire, R.E. 2006. Maximum entropy modeling of species
1370 geographic distributions. Ecological Modelling 190:231-259.

1371 Pradervand, J-N., Dubuis, A., Pellisier, L., Guisna, A., and Randin, C. 2013. Very high resolution
1372 environmental predictors in species distribution models: Moving beyond topography? Progress
1373 in Physical Geography 38:79-96.

Pulliam, H.R. 2000. On the relationship between niche and distribution. Ecology Letters 3:349-361.

Pyron, R.A., Costa, G.C., Patten, M.A., and Burbrink, F.T. 2015. Phylogenetic niche conservatism and the evolutionary basis of ecological speciation. Biological Reviews 90:1248-1262.

Qiao, H., Soberón, J., and Peterson, A.T. 2015. No silver bullets in correlative ecological niche modeling: Insights from testing among many potential algorithms for niche estimation. Methods in Ecology and Evolution 6:1126-1136.

Quintero, I. and Wiens, J.J. 2013. What determines the climatic niche width of a species? The role of spatial and temporal climatic variation in three vertebrate clades. Global Ecology and Biogeography 22:422-432.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Rapacciuolo, G., Roy, D.B., Gillings, S., Fox, R., Walker, K., and Purvis, A. 2012. Climatic associations of British species distributions show good transferability in time but low predictive accuracy for range change. Public Library of Science ONE 7:e40212.

Record, S., Strecker, A., Tuanmu, M-N., Beaudrot, L., Zarnetske, P., Belmaker, J., and Gerstner, B. 2018. Does scale matter? A systematic review of incorporating biological realism when predicting changes in species distributions. Public Library of Science ONE 13:e0194650.

Reino, L., Ferreria, M., Martínez-Solano, Segurado, P., Xu, C., and Barbosa, A.M. 2017. Favorable areas for co-occurrence of parapatric species: Niche conservatism and niche divergence in Iberian tree frogs and midwife toads. Journal of Biogeography 44:88-98.

Rosenblad, K.C., Perret, D.L., and Sax, D.F. In press. Niche syndromes reveal climate-driven extinction threat to island endemic conifers. Nature Climate Change.

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Global Ecology and Biogeography 20:181-192.

Santos, M.J., Smith, A.B., Thorne, J.H., and Moritz, C. 2017. The relative influence of change in habitat and climate on elevation range limits in small mammals in Yosemite National Park, California, U.S.A. Climate Change Responses 4:7.

Seo, C., Thorne, J.H., Hannah, L., and Thuiller, W. 2008. Scale effects in species distribution models: Implications for conservation planning under climate change. Biology Letters 5:39-43.

Smith, A.B. 2013a. The relative influence of temperature, moisture and their interaction on range limits of mammals over the past century. Global Ecology and Biogeography 22:334-343.

Smith, A.B. 2013b. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. Diversity and Distributions 19:867-872.

Smith, A.B. 2019. enmSdm: Tools for modeling species niches and distributions. R package version 0.3.8.2. URL https://github.com/adamlilith/enmSdm.

Smith, A.B. and Santos, M.J. In prep. Using species distribution models to infer the importance of range-shaping factors: A review of methods and concepts.

Smith, A.B., Beever, E.A., Kessler, A.E. Johnston, A.N., Ray, C., Epps, C.W., Lanier, H.C., Klinger, R.C., Rodhouse, T.J., Varner, et al. In review. Alternatives to genetic affinity as a context for within-species response to climate.

Smith, A.B., Godsoe, W., Rodríguez-Sánchez, F., Wang, H-H., and Warren, D. 2019. Niche estimation above and below the species level. Trends in Ecology and Evolution 34:260-273.

Smith, A.B., Long, Q.G., and Albrecht, M.A. 2016. Shifting targets: spatial priorities for ex situ plant conservation depend on interactions between current threats, climate change, and uncertainty. Biodiversity and Conservation 25:905-922.

Smith, A.B., M.J. Santos, M. S. Koo, K.C. Rowe, K.M.C. Rowe, J.L. Patton, S. Beissinger, and C. Moritz. 2013. Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. Ecography 36:1017-1031.

Smith, A.C., Koper, N., Francis, C.M., and Fahrig, L. 2009. Confronting collinearity: Comparing methods for disentangling the effects of habitat loss and fragmentation. Landscape Ecology 24:1271-1285.

Somodi, I., Lepesi, N., annd Botta-Dukat, Z. 2017. Prevalence dependence in model goodness measures with special emphasis on true skill statistics. Ecology and Evolution 7:863-872.

Sor, R., Park, Y-S., Boets, P., Goethals, P., and Lek, S. 2017. Effects of species prevalence on the performance of predictive models. Ecological Modeling 354:11-19.

Soultan, A. and Safi, K. 2017. The interplay of various sources of noise and reliability on species distribution models hinges on ecological specialization. Public Library of Science ONE 12:e0187906.

Stanton-Geddes, J., Tiffin, P., and Shaw, R.G. 2012. Role of climate and competitors in limiting fitness across range edges of an annual plant. Ecology 93:1604-1613.

Stralberg, D. Carrol, C., Wisley, C.B., McKenney, D.W., and Nielsen, S.E. 2018. Macrorefugia for North American trees and songbirds: Climatic limiting factors and multi‐scale topographic influences. Global Ecology and Biogeography 27:690-703.

Suárez-Seoane, S., Virgós, Terroba, O., Pardavila, X., and Barea-Azcón, J.M. 2014. Scaling species distribution models across spatial resolutions and extents along a biogeographic gradient. The case of the Iberian mole *Talpa occidentalis*. Ecography 37:279-292.

Svancara, L.K., Abatzoglou, J.T., and Waterbury, B. 2019. Modeling current and future potential distributions of milkweeds and the Monarch butterfly in Idaho. Frontiers in Ecology and Evolution 7:168.

Synes, N.W. and Osborne, P.E. 2011. Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. Global Ecology and Biogeography 20:904-914.

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A.C., and Guisan, A. 2014. Measuring relative effect of factors affecting species distribution model predictions. Methods in Ecology ane Evolution 5:947-955.

Thuiller, W., Araújo, M.B., and Lavorel, S. 2003. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. Journal of Vegetation Science 14:669-680.

Thuiller, W., B. Lafourcade, R. Engler, and M.B. Araújo. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. Ecography 32:369-373. (Version XXXXX).

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58:267-288.

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

Warren, D.L. and S.N. Siefert. 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. Ecological Applications 21:335-342.

Warren, D.L., Matzke, N.J., and Iglesias, T.L. In review. Evaluating species distribution models with discrimination accuracy is uninformative for many applications. bioRxiv. doi: https://doi.org/10.1101/684399

White, J.W., Rassweiler, A., Samhouri, J.F., Stier, A.C., and White, C. 2014. Ecologists should not use statistical significance tests to interpret simulation model results. Oikos 123:385-388.

Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30:279–338.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, Boca Raton. 392 pp.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society B 73:3-36.

1475 Yates, K.L., Bouchet, P.J., Caley, M.J., Mengerson, K., Randin, C.F., Parnell, S., Fielding, A.H., Bamford,
1476 A.J., Ban, S., Barbosa, A.M., et al. 2018. Outstanding challenges in the transferability of
1477 ecological models. Trends in Ecology and Evolution 33:790-802.

1478 Zurell, D., F. Jeltsch, C.F. Dormann, and B. Schroder. 2009. Static species distribution models in
1479 dynamically changing systems: How good can prediction be? Ecography 32:733-744.

1480 Zurell, D., Thuiller, W., Pagel, J., Cabral, J.S., Münkenmüller, T., Gravel, D., Dullinger, S., Normand, S.,
1481 Schiffers, K.H., Moore, K.A., and Zimmermann, N.E. 2016. Benchmarking novel approaches for
1482 modeling species range dynamics. Global Change Biology 22:2651-2664.