1   Title: Evaluating species distribution models with discrimination accuracy is

2   uninformative for many applications

3

4   Running title: Evaluating niche models with occurrences

5

6   Authors: Dan L. Warren[1,2,3*], Nicholas J. Matzke[4], and Teresa L. Iglesias[3,5]

7

8   * To whom correspondence should be addressed: dan.l.warren@gmail.com

9   1. Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany

10 2. Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and

11 Technology, Okinawa, Japan

12 3. Department of Biology, Macquarie University, Sydney, Australia

13 4. Department of Biological Sciences, University of Auckland, Auckland, New

14 Zealand

15 5. Physics and Biology Unity, Okinawa Institute of Science and Technology

16

17

18

22 **Abstract**

23 *Aim*

24 Species distribution models are used across evolution, ecology, conservation,

25 and epidemiology to make critical decisions and study biological phenomena,

26 often in cases where experimental approaches are intractable. Choices regarding

27 optimal models, methods, and data are typically made based on discrimination

28 accuracy: a model's ability to predict subsets of species occurrence data that

29 were withheld during model construction. However, empirical applications of

30 these models often involve making biological inferences based on continuous

31 estimates of relative habitat suitability as a function of environmental predictor

32 variables. We term the reliability of these biological inferences "functional

33 accuracy." We explore the link between discrimination accuracy and functional

34 accuracy.

35

36 *Methods*

37 Using a simulation approach we investigate whether models that make good

38 predictions of species distributions correctly infer the underlying relationship

39 between environmental predictors and the suitability of habitat.

40

41 *Results*

42 We demonstrate that discrimination accuracy is only informative when models

43 are simple and similar in structure to the true niche, or when data partitioning is

44 geographically structured. However, the utility of discrimination accuracy for

45 selecting models with high functional accuracy was low in all cases.

46

47   *Main conclusions*

48   These results suggest that many empirical studies and decisions are based on

49   criteria that are unrelated to models' usefulness for their intended purpose. We

50   argue that empirical modeling studies need to place significantly more emphasis

51   on biological insight into the plausibility of models, and that the current

52   approach of maximizing discrimination accuracy at the expense of other

53   considerations is detrimental to both the empirical and methodological

54   literature in this active field. Finally, we argue that future development of the

55   field must include an increased emphasis on simulation; methodological studies

56   based on ability to predict withheld occurrence data may be largely

57   uninformative about best practices for applications where interpretation of

58   models relies on estimating ecological processes, and will unduly penalize more

59   biologically informative modeling approaches.

60

61   Species distribution models (SDM, alternatively environmental niche

62 models or ENM) use data on species occurrences in conjunction with

63 environmental data to generate statistical models of species' ecological

64 tolerances, environmental limits, and potential to occupy different geographic

65 areas. These methods have been used since the 1920s (Cook 1925, Sutherst

66 2014), but recent years have seen rapid growth in the number of studies

67 employing SDM in fields including ecology, conservation biology, evolutionary

68 biology, and epidemiology (Peterson, Soberón et al. 2011, Coro, Pagano et al.

69 2013, Allen and Lendemer 2016, Gutierrez-Tapia and Palma 2016, Lezama

70 Ochoa, Murua et al. 2016, Raghavan, Goodin et al. 2016, Guisan, Thuiller et al.

71 2017). The primary appeal of SDMs is their tractability; estimating

72 environmental tolerances experimentally is expensive and time-consuming at

73 best, and impractical for many species. In contrast, SDMs can be constructed

74 with minimal investment of resources, using freely available data and software

75 (Hijmans, Cameron et al. 2005, Phillips, Anderson et al. 2006, Thuiller,

76 Lafourcade et al. 2009, Hijmans, Phillips et al. 2012, Kriticos, Webber et al. 2012).

77 For many species of conservation concern, they are one of the only tractable

78 means of estimating habitat suitability, often in cases where stakeholders need

79 these estimates urgently (Keith, Mahony et al. 2014, Warren, Wright et al. 2014).

80   SDM construction involves many decisions which may affect model

81 predictions. These include choice of modeling algorithm, required sample size,

82 optimal model complexity, choice of study area from which data are drawn, the

83 exclusion of outliers, and selection of environmental predictors, among others

84 (Guisan, Graham et al. 2007, Wisz, Hijmans et al. 2008, Acevedo, Jimenez-

4

85   Valverde et al. 2012, Domisch, Kuemmerlen et al. 2013, Boria, Olson et al. 2014,

86   Varela, Anderson et al. 2014, Garcia-Callejas and Araujo 2016, Soley-Guardia,

87   Gutierrez et al. 2016, van Proosdij, Sosef et al. 2016). The literature surrounding

88   these decisions is large and growing rapidly, as is the suite of associated software

89   tools. Decisions about how best to model species are typically made using

90   metrics that test discrimination accuracy on subsets of species occurrence data

91   that have been withheld during model construction (Elith, Graham et al. 2006,

92   Radosavljevic and Anderson 2014). However, the binary prediction of withheld

93   occurrence data is rarely the intended application of SDMs; they are more

94   frequently used to make continuous estimates of habitat suitability, and to make

95   predictions outside of the training conditions both in space and in time. These

96   applications often implicitly assume that there is biological meaning to the

97   continuous suitability scores produced by the model, or to the functional

98   relationship between environmental gradients and habitat suitability. However,

99   it is often not clear which (if any) measurable biological phenomena should be

100   correlated with suitability estimates from SDMs. Many of the measurable

101   phenomena that are potentially related to suitability (e.g., population density

102   (Carrascal, Aragon et al. 2015), upper limit of local abundance (VanDerWal, Shoo

103   et al. 2009, Gomes, IJff et al. 2018)) have not been quantified in detail for many

104   real species and as such are unavailable for model validation.

105   This impracticality of studying environmental suitability experimentally

106   makes it difficult to measure the ability of SDMs to correctly make continuous

107   estimates of habitat suitability. As such, modeling decisions are typically

108   predicated on an assumed relationship between a model's ability to make

109   continuous estimates of relative habitat suitability (hereafter referred to as

5

110 "functional accuracy") and its ability to predict withheld occurrence data

111 (discrimination accuracy). This assumption has been questioned before (Lobo,

112 Jimenez-Valverde et al. 2008), but its importance and validity for SDM studies is

113 largely untested.

114 Discrimination accuracy is known to be a potentially misleading measure

115 for many applications; it is known to be a poor indicator of model calibration

116 (Reineking and Schröder 2006, Jimenez-Valverde, Acevedo et al. 2013), and may

117 even be negatively correlated with calibration and functional accuracy under

118 some conditions (Murphy and Winkler 1992). This general statistical problem

119 may be exacerbated by attributes of the SDM process in a number of ways. First,

120 spatial autocorrelation present in species distributions and in the environment

121 can generate spurious correlations that a model might treat as biological truth,

122 resulting in models that produce high discrimination accuracy even when

123 occurrence data is random (Raes and ter Steege 2007) or the predictors are

124 biologically meaningless (Bahn and McGill 2007, Bahn and McGill 2013,

125 Fourcade, Besnard et al. 2018). Second, there are phenomena other than the

126 suitability of habitat that shape species distributions (e.g., historical

127 biogeography, dispersal, biotic interactions, Figure 1) (Soberon and Peterson

128 2005, Kearney 2006, Anderson 2012, Warren 2012, Warren 2013, Warren,

129 Cardillo et al. 2014). Although it is possible to include these processes as

130 predictors for SDMs, this is not often done in practice. Failure to explicitly

131 consider these processes introduces spurious correlations between species

132 occurrences and the environment into the estimate of the environmental niche.

133 Similarly, the collection of occurrence data often shows spatial biases

134 (Figure 1, panel F), which may be correlated with spatially autocorrelated

6

135   predictors (Phillips, Dudik et al. 2009).   All of these phenomena can lead to poor

136   niche estimates (Figure 2) that still have high discrimination accuracy in

137   geographic space. Since these non-target phenomena are shared between

138   training and test data, a model that parameterizes the environmental correlates

139   of these processes may have higher discrimination accuracy than a model that

140   accurately estimates the species' environmental tolerances, and yet may produce

141   pathological behavior in applications where model transferability or continuous

142   estimates of habitat suitability are desired (Lobo, Jimenez-Valverde et al. 2008,

143   Veloz 2009, Radosavljevic and Anderson 2014, Torres, Sutton et al. 2015, Huang

144   and Frimpong 2016).

145   A further issue with discrimination accuracy is the lack of true absence

146   data. One of the primary reasons that SDM methods are so tractable is that they

147   can be used without true absence data, which is often difficult and expensive to

148   obtain. SDMs deal with the lack of true absences by sampling "pseudoabsence"

149   or "background" points which are ideally intended to represent the set of

150   environmental conditions that are potentially accessible to the species. This

151   requires users to make decisions about the size of the appropriate study area for

152   background samples (Acevedo, Jimenez-Valverde et al. 2012), as well as the

153   nature of sampling (e.g., random points or points from closely related species

154   (Phillips, Dudik et al. 2009)). These decisions are often somewhat arbitrary (e.g.,

155   background areas chosen using political boundaries or poorly-justified

156   assumptions about dispersal), and can affect both the inferred model (Acevedo,

157   Jimenez-Valverde et al. 2012) and the performance of metrics used to evaluate

158   models (Acevedo, Jimenez-Valverde et al. 2012, Hijmans 2012, Jimenez-Valverde,

159   Acevedo et al. 2013). The lack of real absence data results in models that are

160   incapable of accurately predicting prevalence, and that incorrectly treat some

161   suitable conditions as unsuitable.

162   Finally, the usefulness of discrimination accuracy as a criterion for

163   selecting SDMs may also be negatively impacted by model complexity.

164   Discrimination accuracy only measures whether a model assigns higher

165   suitability values to presence points than it does to background or absence

166   points, and highly flexible algorithms may produce a broad range of marginal

167   suitability functions that have similar, or even identical, discrimination accuracy

168   (Figure 3). This phenomenon is likely compounded by the frequent use of large

169   numbers of predictors that are highly collinear; as the number of predictors and

170   the complexity of marginal suitability functions increase, the number of potential

171   models with similar discrimination accuracy grows very rapidly.

172   Although many of these problems with discrimination accuracy have been

173   noted before, the utility of discrimination metrics for SDM studies has not been

174   examined in a system where the true niche and habitat suitability are known. As

175   a result, we have little information on how useful these metrics are for empirical

176   studies where the goal is to estimate the relative suitability of habitat, despite

177   the ubiquity of discrimination metrics in SDM model selection.

178   Here we adopt a simulation approach to explore the relationship between

179   discrimination and functional accuracy using virtual species for which the true

180   niche is known. We build models using a number of different algorithms, study

181   area sizes, and methods of partitioning training and test data. However, these

182   simulations are not intended to represent all possible modeling approaches. The

183   goal of these simulations is not to determine which method produces the best

184   niche or distribution estimates, but rather to evaluate commonly used methods

185 for model selection across a broad range of models in a system where we know

186 the underlying true habitat suitability.

187

188 **Methods**To examine how model selection was affected by modeling

189 algorithm, sampling bias, and non-target spatial phenomena, we conducted four

190 sets of simulation experiments:

191 1. High complexity: Artificial niches were generated using three

192   randomly chosen variables and models were constructed using all

193   nineteen bioclimatic variables. Background data were drawn from

194   100 km circular buffers around occurrence points.

195 2. Geographic partitioning: Same conditions as (1) but presence and

196   background points for each species were split into four quadrants

197   using ENMEval (Muscarella, Galante et al. 2014).   Presence and

198   background data from one randomly selected quadrant were withheld

199   for model evaluation.

200 3. Large background: Same conditions as (1), but background data were

201   drawn from 1000 km circular buffers around occurrence points.

202 4. Low complexity: Niches were based on two randomly chosen

203   variables and models were built using those two variables plus two

204   more chosen at random.

205

206 All analyses used CliMond data for 19 bioclimatic variables typically

207 employed in SDM studies (Nix 1986) ENREF_1, including data for the present

208 and for 24 combinations of future emissions scenario (A1B and A2), year (2030,

209 2050, 2070, 2080, 2090, and 2100), and climate model (CSIRO-Mk3.0 and

210   MIROC-H) (Kriticos, Webber et al. 2012). Simulations and analyses were

211   restricted to Australia, including Tasmania.

212   Simulated niches were created using the generateRandomSp function in

213   the virtualspecies R package (Leroy, Meynard et al. 2015). Simulated species

214   with fewer than 400 suitable grid cells in the initial presence/absence raster

215   showed a strong tendency to produce models that showed no suitable habitat on

216   future climate scenarios, rendering comparisons that were uninformative for

217   model selection. As a result, these simulations were discarded.

218   To simulate the effects of non-target spatial processes (e.g., historical

219   biogeography, biotic interactions, dispersal limitation), the initial

220   presence/absence raster from virtualspecies was converted to point data, which

221   was partitioned into allopatric regions using k-means clustering. Solutions

222   ranging from 2 to 10 clusters were considered, and an algorithm was used to

223   maximize the minimum distance between clusters. One region was assigned at

224   random to be the range of the species, and converted back into a raster. We

225   recorded the total proportion of suitable cells that fell within this range, to

226   measure the extent to which species distributions departed from the distribution

227   of available suitable habitat across the entire study area.

228   Spatial sampling bias was modeled using data from 5,969,252 collection

229   records for 28,286 Australian plant species. These records were harvested from

230   GBIF (GBIF.org 2015) using the rgbif (Chamberlain, Boettiger et al. 2013)

231   package, and converted to a raster representing the number of observations per

232   grid cell at the same extent and resolution as the environmental data. All values

233   were then divided by the maximum cell value, resulting in a range of sampling

234   intensity from 0 to 1.

235   Species occurrence data was sampled from a raster with values in each

236   grid cell *x* calculated as:

237

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

239

240   Where probability of sampling is a function of *p(x)*, g_b is a parameter that

241   controls the magnitude of spatial sampling bias, *b(x)* is the relative strength of

242   spatial sampling bias in cell *x*, *s(x)* is the suitability of habitat in the grid cell, and

243   *r(x)* is a binary variable taking the value 1 inside the species range and 0

244   everywhere else. For each species, we drew 100 simulated occurrence points by

245   selecting grid cells at random and sampling occurrences as a Bernoulli trial with

246   probability of success equal to *p(x)*. Presence points were drawn with

247   replacement so that we could study the effects of sampling bias. We simulated

248   data across eleven levels of spatial sampling bias, with the bias strength

249   parameter ranging from 0 to 1 in increments of 0.1. We performed 20

250   simulations for each of the 11 levels of spatial sampling bias. Each of the four

251   simulation conditions (experiments 1-4, above) therefore consisted of 220

252   simulations, for a total of 880 total simulated species across all experiments.

253   As noted in a recent review (Meynard, Leroy et al. 2019), simulation

254   studies need to choose both the simulated niches and sampling regimes that are

255   appropriate for the question involved. Since our goal here is to test which

256   metrics select models that accurately estimate the niche, it was essential for us to

257   generate data that would be capable of producing accurate niche estimates in

258   ideal conditions. Due to these concerns we chose not to apply a threshold

259   minimum suitability score below which the organism could not possibly occur;

260 the application of such a threshold would truncate the response functions we are

261 trying to estimate (Meynard, Leroy et al. 2019), resulting in lower expected

262 functional accuracy. Additionally, prior work with virtual species has

263 demonstrated that the application of thresholds results in discrimination

264 accuracy metrics that are overly optimistic (Meynard and Kaplan 2013).

265 We built models using seven algorithms; Bioclim, Domain, generalized

266 linear models (GLM), generalized additive models (GAM), Maxent, random

267 forests, and boosted regression trees. Models were built using the dismo R

268 package (Hijmans, Phillips et al. 2012) and Maxent (Phillips, Anderson et al.

269 2006). This resulted in 6160 inferred models, seven for each of the 880

270 simulated niches. Algorithm settings were left at their default values. For each

271 model, 25 occurrence points were withheld from model construction and used

272 for model evaluation. Each model's discrimination accuracy was evaluated

273 using three statistics: the area under the receiver operating characteristic curve

274 (AUC) (Fielding and Bell 1997), the true skill statistic (TSS) (Allouche, Tsoar et al.

275 2006), and Cohen's kappa (Cohen 1960). While AUC can be calculated using

276 continuous suitability scores, TSS and kappa require binary predictions of

277 presence and absence, so the values for these models that were used in model

278 assessment corresponded to their maximum value across all potential

279 thresholds, i.e., the best possible performance of a thresholded model (Fielding

280 and Bell 1997).

281 Models were projected onto the current distribution of the environmental

282 variables used for model construction. Models were also projected onto the 24

283 future climate scenarios. We used the simulated niche from the virtualspecies

284 object to project the true suitability of habitat across those same set of future

12

285   environments, to assess whether discrimination accuracy is a useful predictor of

286   models' ability to extrapolate to new environmental conditions. To measure

287   functional accuracy, we compared geographic projections of habitat suitability

288   from the true niche and the inferred models using Spearman rank correlation.

289   Correlations between projected and true suitability scores for the present and

290   for future climate scenarios were measured separately within the species range

291   (areas where $r(x) = 1$) and across the entire study area (Australia and Tasmania).

292   Spearman rank correlation was chosen as a measure of functional accuracy for

293   this study due to the structural differences between models produced by

294   different algorithms, and in consideration of how SDMs are often applied; any

295   two models that assign identical rankings to a set of habitat patches are

296   effectively interchangeable for applications where models are thresholded, or

297   where suitability scores are used to prioritize one habitat patch over another.

298   Rank correlation will reflect this when models produce similar rankings of

299   relative habitat suitability (e.g., $\rho = 1$ when predictions made from one model are

300   a monotonically increasing function of predictions made from another model).

301   In contrast, Pearson product moment correlation will only assign a value of 1 if

302   the relationship between suitability scores for the two models is linear, which

303   may serve to exaggerate differences that are not relevant to many empirical

304   studies. In order to test the sensitivity of our results to this choice, we also

305   conducted a separate set of analyses using Pearson product moment correlation

306   as a measure of functional accuracy (Appendix S4), but results from these

307   analyses were effectively the same as those seen in tables 1 and 2.

308   We choose to focus on functional accuracy here instead of calibration for

309   several reasons; first, the application of SDMs more often relies on the relative

13

310   suitability of habitat than estimating the exact probability of observing a species

311   in a particular place, and functional accuracy more directly estimates this aspect

312   of model behavior. Second, it is already known that discrimination accuracy may

313   be poorly correlated to calibration even when the model gets the relative

314   ranking of habitat right. For example, if the estimated suitability of habitat is a

315   transformation of the true suitability of habitat that preserves relative rankings

316   but not the magnitude of differences in suitability scores (Reineking and

317   Schröder 2006), both discrimination and functional accuracy would be high, but

318   calibration would be poor. Finally, the link between discrimination accuracy and

319   calibration is known to be severely affected by prevalence (Reineking and

320   Schröder 2006, Elith and Graham 2009), but the link between discrimination

321   accuracy and functional accuracy as measured here would not be so affected.

322   To summarise the relationships between discrimination and functional

323   accuracy for all algorithms considered together (Table 1 and Appendix S3 and

324   S4), we used generalized linear mixed models, and evaluated correlations using

325   McFadden's pseudo-r^2 (McFadden 1974). For the remainder of the regressions,

326   we used linear models and the standard coefficient of determination, r^2. We

327   applied Bonferroni corrections to compensate for problems arising from

328   multiple testing. For these purposes we defined four families of test that we

329   consider independently. Those examining the relationship between

330   discrimination and functional accuracy at each combination of algorithm and

331   complexity level (Tables 1 and 2, n = 12 comparisons per set), and the

332   remainder, which are intended primarily to examine which factors impact

333   overall model quality and as a check to establish that expected relationships

334   between metrics are seen in the simulation results (Appendix S3, n = 11).

14

335   Based on results from experiments 1-4, we performed a fifth set of

336   simulation experiments to examine more thoroughly the effects of niche

337   complexity and the number of predictor variables on the relationship between

338   discrimination and functional accuracy (Figure 4). Due to the computational

339   intensity of some SDM algorithms, we restricted analyses to a simpler set of

340   conditions for these simulations. Presence data was generated with no non-

341   target spatial biological processes and no spatial sampling bias, so occurrence

342   points were sampled across the entirety of the suitable habitat. We restricted the

343   modeling process to GLM and Maxent, and only used AUC for evaluating

344   predictions on randomly withheld test data. Simulated niches were built using a

345   number of environmental variables ranging between 1 and 19, and models were

346   inferred with between 1 and 19 variables (2 to 19 for Maxent, due to issues with

347   the software implementation), subject to the constraint that variables that were

348   included within the species' niche were selected first during model construction.

349   For each combination of number of niche axes and environmental predictors, we

350   performed 300 separate simulations, resulting in 108,300 total simulations per

351   modeling method. For each simulation we recorded the test AUC and the rank

352   correlation between the inferred and true suitability of habitat. For each

353   combination of number of niche axes and predictors, we then measured the rank

354   correlation between discrimination accuracy and functional accuracy across the

355   set of 300 models. This resulted in a metric ranging from 1, where test AUC was

356   a perfect indicator of functional accuracy, to -1, where test AUC was negatively

357   associated with functional accuracy. We fit a linear model to these results which

358   included the number of variables used for the simulated niche, the number of

359   variables used for model construction, and an interaction term.

15

360

361 **Results**

362   Regression outputs for experiments 1-4 are summarized in tables 1

363 (algorithms pooled) and 2 (algorithms analyzed separately), and also in

364 appendices S3, S4, and S5. TSS, kappa, and AUC were all highly correlated with

365 each other, so we will not discuss them separately. We found that discrimination

366 accuracy on training and test data were correlated, and that functional accuracy

367 in the training region was correlated with functional accuracy outside the

368 training region. This indicates that models that perform well at discrimination

369 accuracy tend to do so regardless of whether it is measured on training or test

370 data, and the same is true of models that perform well at functional accuracy.

371 Functional accuracy was generally fairly good; a majority of models produced

372 estimates of habitat suitability that were positively correlated with the true

373 suitability of habitat, whether measured in the training region (73.8%), or

374 projected to the continental scale (80.8%). However, models performed

375 somewhat worse when projected into future climate scenarios (65.7% were

376 positively correlated with true suitability within the species range, 71.0% at the

377 continental scale).

378   When all algorithms were analyzed together in a single GLM,

379 discrimination accuracy was a very poor predictor of functional accuracy in all

380 cases (Table 1). Although 31/48 regressions were statistically significant, five

381 were negative correlations, and none had an r^2 value greater than 0.2. This

382 indicates poor performance of discrimination accuracy at selecting models when

383 comparing between algorithms.

384   Results of regressions conducted for each algorithm separately are

385   presented in Table2. For all experiments, we find that discrimination accuracy is

386   uninformative or actively misleading about models' functional accuracy in a

387   majority of cases (significant positive correlations were seen for less than half of

388   comparisons in any simulation experiment). A majority of these correlations

389   were also quite weak; the average r^2 value was 0.08 (range -.15 to .26). This

390   indicates poor performance of discrimination accuracy at selecting models with

391   high functional accuracy even when comparing models that were built using the

392   same methods. Discrimination accuracy had no negative correlations with

393   functional accuracy when test data were chosen based on a geographic partition

394   of the species' range, but was still a poor predictor of functional accuracy

395   (average r^2 = 0.12).

396   We note an interesting phenomenon here with respect to the size of the

397   buffer regions used to draw background data for model fitting and evaluation;

398   the models built using the largest (1000 km) buffers around occurrence points

399   performed very well, with the highest levels of functional accuracy and

400   discrimination accuracy (Appendix S5). These differences were most prominent

401   for discrimination accuracy, reinforcing previous findings showing that

402   discrimination accuracy is sensitive to study area (Lobo, Jimenez-Valverde et al.

403   2008). Some previous studies have suggested the models perform best when

404   constructed using fairly small study regions, however those studies have largely

405   assessed model quality via discrimination accuracy within the species' native

406   range (Acevedo, Jimenez-Valverde et al. 2012, Zhu, Rédei et al. 2014). These

407   results indicate that the relationship between study area size and model

17

408   performance may be more complex than previously reported, and optimal

409   choices may depend on the applications for which models are designed.

410   Our fifth experiment examined the effects of complexity across a broader

411   range of complexity values, and found that the ability of AUC to select GLM

412   models with high functional accuracy was negatively correlated with the

413   complexity of the simulated niche and the number of predictor variables. For

414   Maxent models the relationship between discrimination accuracy and functional

415   accuracy (Figure 4) was positively but weakly correlated (r^2 = .09) with the

416   number of variables in the true niche, but uncorrelated with the number of

417   variables included in the modeling process.

418

419   **Discussion**

420   SDM methods are used for many applications in which niche estimates

421   are needed but experimental approaches are impractical. Results for

422   experiments 1-4 demonstrate that many of these models can provide useful

423   estimates of the relative suitability of habitat, the ability of species to invade new

424   areas, and the effects of climate change. However, one of the key steps in any

425   modeling study is the identification of which models from a candidate set

426   perform well and which perform poorly. Our results indicate that the most

427   widely used methods for selecting models are largely uninformative for studies

428   where the goal is to make continuous estimates of habitat suitability, or to

429   estimate the species' response to an environmental gradient. When algorithms

430   were analyzed separately, 15/149 statistically significant correlations between

431   discrimination and functional accuracy were negative. In these cases

432   discrimination metrics were not just uninformative, but in fact positively

433   misleading for applications where the goal of SDM is to predict the relative

434   suitability of habitat.

435   In our fifth experiment we examined the effects of niche and model

436   complexity for GLM models across a broader set of conditions, and found that

437   discrimination accuracy predicts functional accuracy only when both the niche

438   and the environmental space it is being modeled in are far simpler than those

439   seen typically in the empirical literature (Figure 4). Even at low levels of

440   complexity, the relationship between discrimination and functional accuracy for

441   GLM is fairly weak (Spearman rank correlation = 0.31 for a single niche axis and

442   predictor variable), and declines rapidly as models become more complex,

443   becoming minimally informative as models approach levels of complexity that

444   are often seen in the empirical literature.

445   For Maxent models the number of predictors used to model the niche had

446   no effect on the utility of discrimination accuracy for model selection, but there

447   was a weak positive effect of the number of variables used to simulate the true

448   niche (β = .006, p < .05). We hypothesize that the lack of effect of the number of

449   predictors for Maxent is due to its ability to automatically penalize

450   overparameterization; many of the predictors supplied to the algorithm may

451   ultimately have little or no weight in the model. We also note that the most

452   reliable correlations between discrimination accuracy and functional accuracy

453   seen in our simulation results were for Maxent models (Table 2), as would be

454   expected if model complexity and number of predictors were partly responsible

455   for driving the poor performance of discrimination accuracy.

456   Discrimination accuracy was generally a better predictor of functional

457   accuracy for GLM, GAM, and Maxent models than for the other methods of model

458   construction explored in this study. This is likely due to the internal structure of

459   these models. The simulation approach taken here uses a logistic function to

460   generate sampling probabilities based on a simulated niche, which is composed

461   of smooth (linear or quadratic) responses to a set of environmental variables.

462   As such, the function underlying habitat suitability lies within the set of functions

463   that may be exactly estimated by GLM, GAM, or Maxent, so that estimation of

464   simulated niches is considerably more tractable for those methods. We

465   therefore caution users to refrain from interpreting these results as an

466   endorsement of any particular method when constructing SDMs using empirical

467   data. Rather, we suggest that these results indicate that choice of modeling

468   methods should ideally include intuition or data regarding the potential

469   functional relationship between the environmental predictors and the suitability

470   of habitat. If the functional relationships that may be estimated by an algorithm

471   differ significantly from the true functional relationship, discrimination accuracy

472   is largely uninformative or misleading about models' ability to predict habitat

473   suitability. This does not necessarily imply that models built using different

474   functional shapes from the true niche are poor estimates of habitat suitability;

475   rather it indicates that discrimination accuracy is uninformative for selecting

476   models with high functional accuracy under these conditions.

477   Our results clearly indicate that most empirical studies using SDM

478   methods should ideally not rely solely on prediction of withheld occurrence data

479   to assess model quality. However, they also indicate a much more systemic

480   problem for the SDM literature: decades of methodological work in this field

481   have resulted in a set of widely-adopted "best practices", but a great majority of

482   these studies have focused on optimizing models' discrimination accuracy on

20

483   withheld occurrence data from real species distributions (Guisan, Graham et al.

484   2007, Wisz, Hijmans et al. 2008, Domisch, Kuemmerlen et al. 2013, Boria, Olson

485   et al. 2014, Radosavljevic and Anderson 2014, Moreno-Amat, Mateo et al. 2015,

486   Garcia-Callejas and Araujo 2016, Huang and Frimpong 2016, Kuebler,

487   Hildebrandt et al. 2016, Lopatin, Dolos et al. 2016, Rovzar, Gillespie et al. 2016,

488   Soley-Guardia, Gutierrez et al. 2016). Given the disconnect seen here between

489   discrimination and functional accuracy, it is entirely possible that the "best

490   practices" advocated in these studies have negligible, or even detrimental, effects

491   on model quality for applications where functional accuracy is the goal.

492   In order to accurately assess the ability of different methods to achieve

493   useful levels of functional accuracy, we argue that the methodological literature

494   must reevaluate its "best practices" via simulations where true habitat suitability

495   and niche parameters are known. While some simulation studies are already

496   being conducted (Meynard, Leroy et al. 2019), these have typically been done in

497   the context of optimizing discrimination accuracy, and as such may also be

498   largely uninformative about estimating habitat suitability as a function of

499   environmental gradients. There are many common practices and assumptions in

500   the field that may need to be reevaluated based on their ability to estimate

501   habitat suitability; choice of algorithm, methods for choosing predictor variables,

502   choice of study area, rarefaction of data, and optimal model complexity are

503   obvious candidates.

504   In addition, we argue that practitioners must recognize that favoring

505   models based strictly on their spatial predictions is simply inappropriate for

506   many applications. In studies where the goal is to estimate the niche (i.e.,

507   maximize functional accuracy), users must become comfortable with the idea

508   that a biologically accurate model may produce relatively poor estimates of

509   species' current spatial distributions. This is not simply a methodological point

510   brought to light by the current simulation study; it is necessarily true given the

511   existence of non-target phenomena that themselves have spatial structure (e.g.,

512   biotic interactions, dispersal). This has been known for years (Jackson and

513   Overpeck 2000, Soberon and Peterson 2005, Anderson 2012, Warren 2012,

514   Warren 2013), yet has been largely ignored in the continued pursuit of methods

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

516   Investigators familiar with SDM methods will no doubt wish to critically

517   examine the methods used here to infer models; there are other algorithms

518   available, and there are many modeling choices that we did not explore in great

519   depth. However, these criticisms are largely irrelevant to the primary results of

520   this study; while it is certainly possible that greater effort in exploring the space

521   of model choices might improve the accuracy of models, we note that (1)

522   evaluation metrics on randomly withheld test data for the models generated

523   here are not unusual for the range seen in the empirical SDM literature (e.g.,

524   Appendix S5), (2) the overall performance of SDM methods is irrelevant to

525   whether or not discrimination accuracy is a valid indicator of functional

526   accuracy, and (3) most SDM users' methodological preferences are currently

527   chosen based on studies that seek to maximize the very performance metrics

528   that the current study demonstrates are not useful for estimating functional

529   accuracy.

530   We acknowledge the possibility that there is some subset of modeling

531   approaches not addressed here for which discrimination and functional accuracy

532   are highly correlated. It would be both gratifying and very useful to find such a

533   set of conditions, and that topic deserves to be examined in great depth.

534   However, even if such a set can be found it does not invalidate the conclusion

535   presented here; that there is a large range of modeling algorithms and

536   approaches for which the correlation between discrimination accuracy and

537   functional accuracy is not strong enough to be useful in model selection for many

538   purposes. Similarly, we acknowledge that the disconnect between functional

539   accuracy and discrimination seen here may be affected by sample size, but the

540   sample sizes used here (75 training, 25 test) are not atypical for the ENM

541   literature.

542   In summary, we demonstrate that, under a broad range of conditions, the

543   ability of a model to successfully predict withheld occurrence data within the

544   training region does not reliably measure its ability to estimate the relationship

545   between environmental gradients and habitat suitability. Discrimination

546   accuracy may be a reasonable metric when the goal is to guide further sampling

547   of occurrences within a species' current range, without regard for whether the

548   model estimates the true environmental niche or the relative suitability of

549   habitat well. However, this is not often the goal of empirical model construction

550   in the SDM literature.

551   As a result, the applied and methodological literature in this field are

552   largely based on metrics that may be irrelevant to the intended applications of

553   many models. If the field is to continue to attempt to use SDMs to infer species'

554   responses to environmental gradients, we must develop methods for model

555   construction and metrics for model evaluation that are more relevant to the

556   actual goals of the modeling process. While we find that geographically

557   structured partitioning of test data does offer some advantages over randomly

23

558   withheld data, it is clear from this study that even those methods have very

559   limited ability to identify models that accurately estimate the relative suitability

560   of habitat.

561   We would like to particularly highlight the implications of our results for

562   the development of new methods in this field in the coming years. Many

563   investigators are currently developing methods that incorporate more biological

564   and statistical realism into the SDM process, including the integration of

565   physiological and trait data (Pollock, Kelly et al. 2018) and explicit models of bias

566   (Robinson, Ruiz‑Gutierrez et al. 2018), dispersal (Zurell 2017), plasticity (Bush,

567   Mokany et al. 2016), and evolutionary history (Smith, Godsoe et al. 2019). In any

568   system affected by non-target spatial phenomena, these methods will often

569   produce poorer estimates of species' geographic distributions precisely because

570   they provide better estimates of the environmental niche. We hope that the

571   results presented here will compel the field to evaluate these new methods

572   based on their ability to infer the biological phenomena of interest, as

573   demonstrated using simulations or physiological data, rather than simply reject

574   them due to poor discrimination accuracy on misleading occurrence data.

575   We feel it is necessary to specifically address one interpretation of these

576   results that we feel is not appropriate: the work presented here is not intended

577   to suggest that any particular method of SDM construction is inherently better or

578   worse than others. While the relative performance of different methods is a very

579   interesting question and one that deserves further exploration within a

580   simulation framework, this study was not designed to address those questions

581   and it would be inappropriate to interpret these results as such. We emphasize

582   that most of the models built from these simulated species were arguably

24

583 publishable distribution estimates, and were at least somewhat useful as

584 estimates of the species' niche. Rather, this study is intended to examine the

585 performance of widely-used methods of model selection, and it is those methods

586 that are performing poorly. We demonstrate that we can make both good

587 distribution estimates and good niche estimates using common methods, and in

588 fact produced many models that are good for both purposes. However, our

589 results indicate that we have a difficult time distinguishing good models from

590 bad when our goal is functional accuracy.

591 At minimum, our results suggest that any empirical study using

592 discrimination accuracy to assess model quality should start with two crucial

593 steps: (1) use a minimal set of predictor variables for which there is an a priori

594 reason to expect that they limit the suitability of habitat for the species, and (2)

595 select algorithms capable of inferring functional responses that are plausible

596 estimates of the underlying biology (e.g., not using a step function in situations

597 where suitability is expected to be a continuous function of the predictor

598 variable). In a sense, these findings are unsurprising; they recapitulate

599 longstanding best practices in the broader literature regarding statistical

600 modeling (Anderson and Burnham 2004, Burnham and Anderson 2004, Gelman

601 and Hill 2006, Zuur, Ieno et al. 2009). However, here we show that failure to

602 make these choices appropriately does not necessarily lead to poor predictions;

603 instead it means that we are largely unable to distinguish good models from bad

604 using species occurrence data. Under these conditions any preference for a

605 given model based on discrimination accuracy may be little better than choosing

606 a model at random.

607

608

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

Table 1. Results of regressions functional accuracy on discrimination accuracy, all algorithms considered together. Significant positive correlations are represented by "+" and green cell color, negative correlations by "-" and pink cell color. Numbers indicate r^2 values for each regression. Variables accompanied by (F) indicate that they were measured on models projected across 24 future climate scenarios. Variables with (N) and(C) indicated models projected within the species native range or at a continental scale, respectively. Results are presented separately for four model sets: the "simple" set of predictors (2 variables in the true niche, 4 predictors per model,

100km buffer), the "complex" set of predictors (3 variables in the true niche, 19 predictors per model, 100km buffer), the "large background" study region (same simulation settings as "complex" but with a 1000km buffer), and the "geographically structured" model set, for which models were constructed and evaluated using geographically partitioned data (same simulations settings as "complex" but with geographic partitioning of data instead of random holdouts).
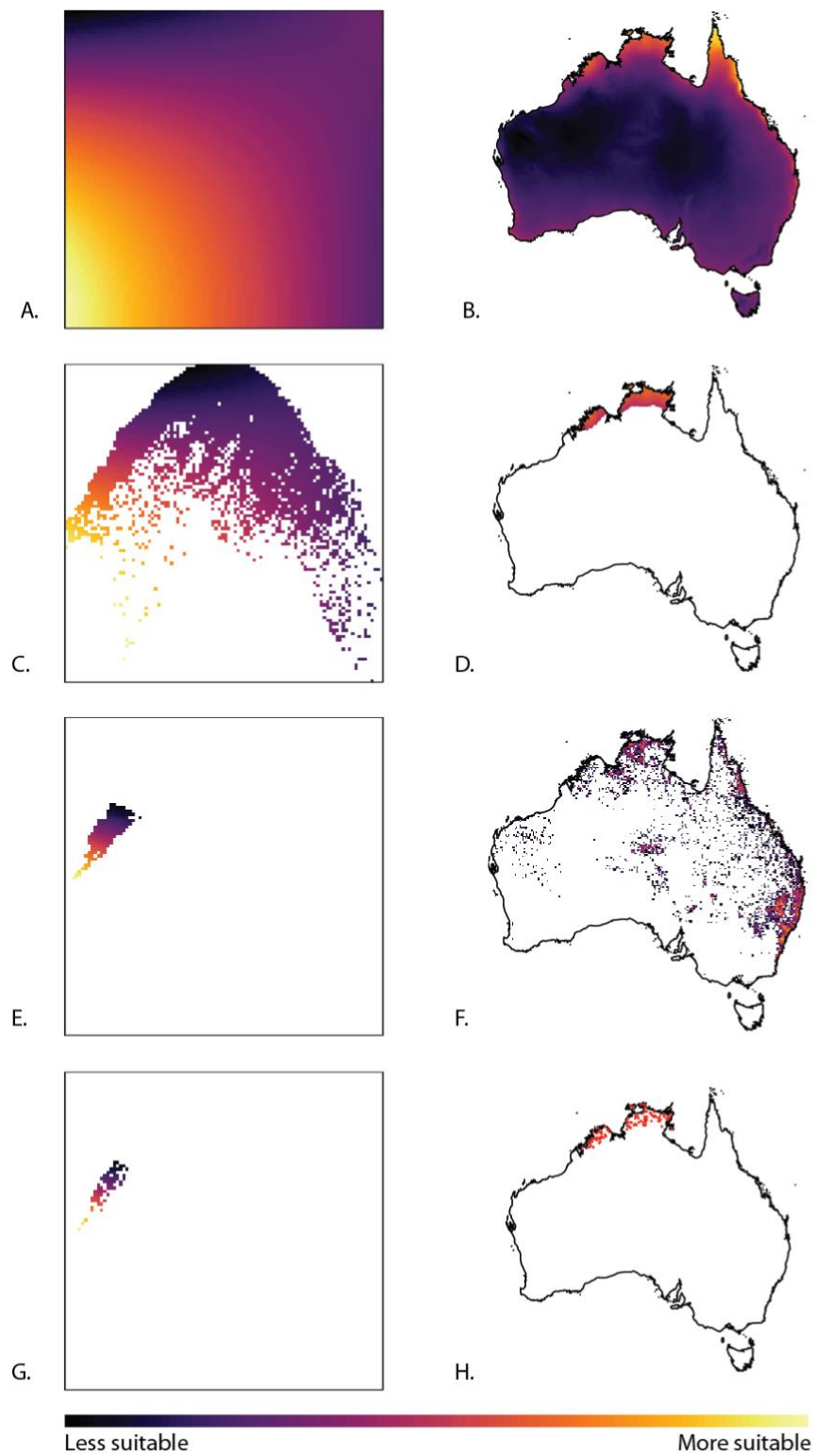
Simple / Complex

Independent variable	Dependent variable	BC	DM	GAM	GLM	MX	RF	BRT	BC	DM	GAM	GLM	MX	RF	BRT
					Simple							Complex			
Test AUC	Spearman (N)				+,.15	+,.06	+,.05	+,.05	-,.05	-,.08		+,.06	+,.06	+,.06	+,.05
Test AUC	Spearman (C)			+,.08	+,.07				-,.10				+,.05		
Test Max TSS	Spearman (N)				+,.14	+,.04			-,.04	-,.1		+,.04	+,.04	+,.04	+,.05
Test Max TSS	Spearman (C)	-,.05		+,.07	+,.07				-,.11				+,.04		
Test Max Kappa	Spearman (N)				+,.11				-,.07	-,.10					
Test Max Kappa	Spearman (C)	-,.06			+,.04				-,.14						
Test AUC	Spearman (F, N)	+,.06		+,.17	+,.14	+,.18	+,.08	+,.15	+,.10	+,.05	+,.14	+,.06	+,.26	+,.11	+,.11
Test AUC	Spearman (F, C)			+,.09	+,.06	+,.04							+,.05		
Test Max TSS	Spearman (F, N)	+,.05		+,.14	+,.14	+,.15	+,.05	+,.12	+,.09	+,.05	+,.14	+,.05	+,.23	+,.08	+,.09
Test Max TSS	Spearman (F, C)			+,.07	+,.05								+,.04		
Test Max Kappa	Spearman (F, N)			+,.10	+,.10	+,.11		+,.06	+,.08	+,.04	+,.07	+,.05	+,.14		
Test Max Kappa	Spearman (F, C)			+,.04											

Large Background / Geographic Partitioning

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

29

Table 2. Relationship between discrimination accuracy and functional accuracy, methods considered separately. Significant positive correlations are represented by "+" and green cell color, negative correlations by "-" and pink cell color. Numbers indicate r^2 values for each regression. Variables accompanied by (F) indicate that they were measured on models projected across 24 future climate scenarios. Variables with (N) and(C) indicated models projected within the species native range or at a continental scale, respectively. Results are presented separately for four model sets: the "simple" set of predictors (2 variables in the true niche, 4 predictors per model, 100km buffer), the "complex" set of predictors (3 variables in the true niche, 19 predictors per model, 100km buffer), the "large background" study region (same simulation settings as "complex" but with a 1000km buffer), and the "geographically structured" model set, for which models were constructed and evaluated using geographically partitioned data (same simulations settings as "complex" but with geographic partitioning of data instead of random holdouts).

Figure 1. Phenomena affecting species distributions and inference of suitability of habitat. Panel A depicts the niche of a simulated species in the first two principal component axes of the 19 Bioclim variables for Australia. Panel B represents the distribution of suitable habitat for the simulated species. The available habitat present across the continent of Australia only represents a subset of the possible niche space (C). The species' current range only encompasses a subset of the suitable habitat (D), which further limits the potential distribution of data in environment space (E). Spatial sampling bias (F, see methods) contributes further bias to the representation of the species both in environment space (G) and geographic space (H). While the geographic distribution of the data (H, red points) may resemble the current range of the species (D), the distribution of that data in environment space (G) is a poor representation of the species' true niche (A). As a result, it may be relatively easy to achieve accurate predictions on randomly withheld occurrence data while still producing a poor estimate of the underlying biology and suitability of habitat.
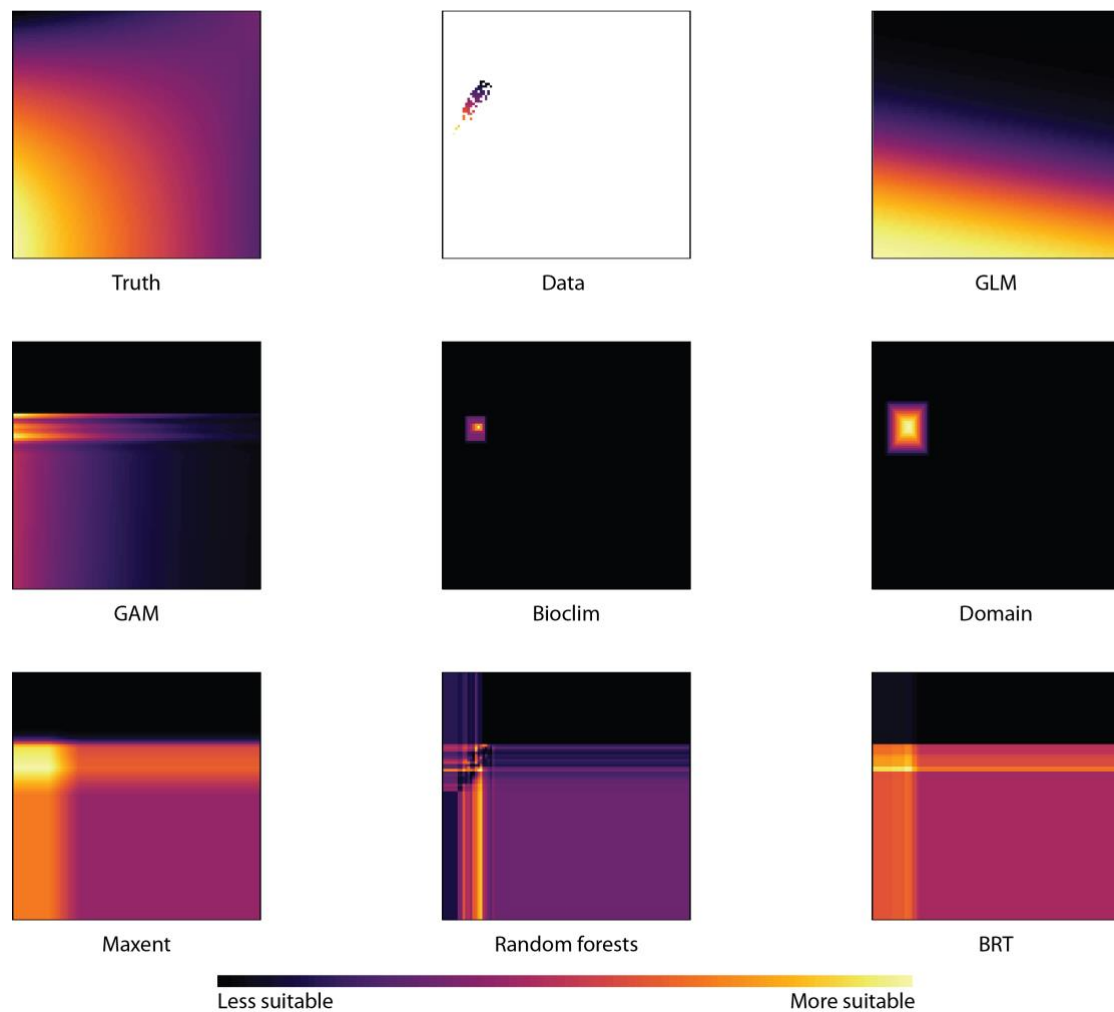
Figure 2. Projection of modeling algorithms in environment space. Using 100 occurrence points for the simulated species in figure 1, we built models using the seven algorithms employed in this study and projected them into the same two dimensional principal component space. The lowest AUC score on 20 randomly withheld data points belonged to random forests (AUC = 0.55), while the highest came from domain (AUC = 0.73). The top left and top center panels show the true niche of the simulated species and the environmental distribution of the data, respectively.
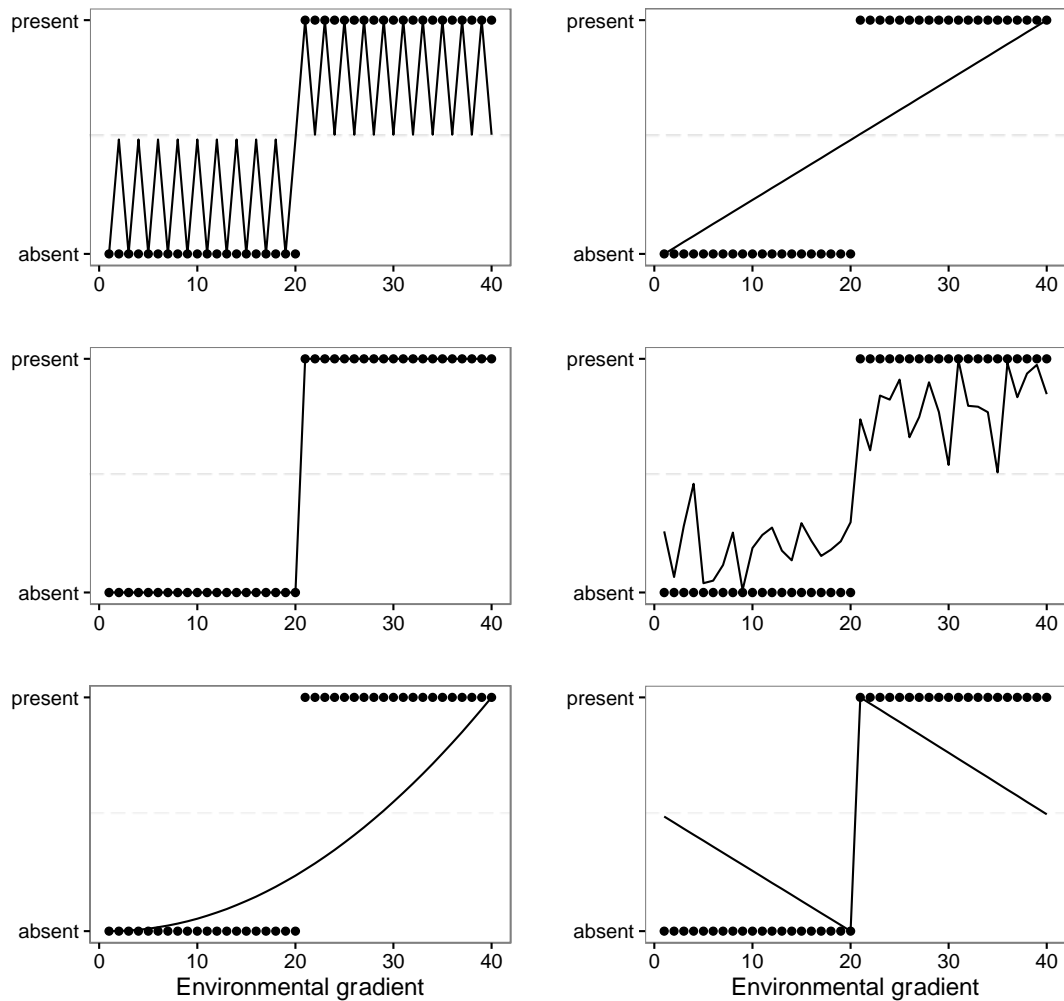
Figure 3. Low information content of discrimination accuracy for inferring functional accuracy. In the above plots, we have simulated 20 absence and 20 presence points along a hypothetical environmental gradient. The six panels represent six hypothetical functions that might be inferred using this data. Each function assigns a higher suitability score (y axis) to all of the presence points than it does to any of the background or absence points. As a result, each function has perfect discrimination accuracy. All six functions are therefore indistinguishable from each other based on discrimination metrics (AUC, TSS,

Kappa), while making very different estimates of the functional relationship of

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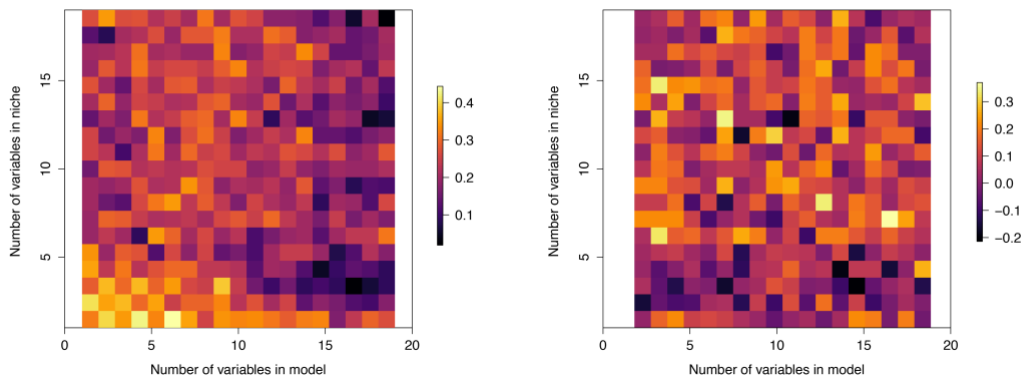
Figure 4. Relationship between number of variables in simulated niche, number of variables in model, and the ability of discrimination to infer functional accuracy for GLM (left) and Maxent (right). Each grid cell represents the output of 300 simulations. The color of each grid cell represents the Spearman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

Supplementary materials:

Appendix S1. Frequency of use of model fit metrics (columns) and partitioning scheme for test data (rows) from a survey of 94 recent applied SDM studies.

Appendix S2. Literature review for metrics of model fit.

Appendix S3. Relationship between evaluation metrics and simulation settings, all algorithms considered together.

Appendix S4. Relationship between discrimination accuracy and functional accuracy using Pearson product moment correlations.

Appendix S5. Discrimination and functional accuracy performance for each simulation experiment.

Data Availability:

Sample code is available on github here:

https://github.com/danlwarren/sim-code-Warren-et-al-2019

Acknowledgements

Acevedo, P., A. Jimenez-Valverde, J. M. Lobo and R. Real (2012). "Delimiting the geographical background in species distribution modelling." Journal of Biogeography **39**(8): 1383-1390.

Allen, J. L. and J. C. Lendemer (2016). "Climate change impacts on endemic, high-elevation lichens in a biodiversity hotspot." Biodiversity and Conservation **25**(3): 555-568.

Allouche, O., A. Tsoar and R. Kadmon (2006). "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)." Journal of Applied Ecology **43**(6): 1223-1232.

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

Anderson, R. P. (2012). "Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions." Ann N Y Acad Sci **1260**: 66-80.

Bahn, V. and B. J. McGill (2007). "Can niche-based distribution models outperform spatial interpolation?" Global Ecology and Biogeography **16**(6): 733-742.

Bahn, V. and B. J. McGill (2013). "Testing the predictive performance of distribution models." Oikos **122**(3): 321-331.

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

Burnham, K. P. and D. R. Anderson (2004). "Multimodel inference: understanding AIC and BIC in model selection." Sociological methods & research **33**(2): 261-304.

Bush, A., K. Mokany, R. Catullo, A. Hoffmann, V. Kellermann, C. Sgrò, S. McEvey and S. Ferrier (2016). "Incorporating evolutionary adaptation in species distribution modelling reduces projected vulnerability to climate change." Ecology letters **19**(12): 1468-1478.

Carrascal, L. M., P. Aragon, D. Palomino and J. M. Lobo (2015). "Predicting regional densities from bird occurrence data: validation and effects of species traits in a Macaronesian Island." Diversity and Distributions **21**(11): 1284-1294.

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

Cohen, J. (1960). "A coefficient of agreement for nominal scales." Educational and Psychological Measurement **20**(1): 37-46.

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

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

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

Elith, J. and C. H. Graham (2009). "Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models." Ecography **32**(1): 66-77.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G.

Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz and N. E. Zimmermann (2006). "Novel methods improve prediction of species' distributions from occurrence data." Ecography **29**(2): 129-151.

Fielding, A. H. and J. F. Bell (1997). "A review of methods for the assessment of prediction errors in conservation presence/absence models." Environmental Conservation **24**(1): 38-49.

Fourcade, Y., A. G. Besnard and J. Secondi (2018). "Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics." Global Ecology and Biogeography **27**(2): 245-256.

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

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

Gelman, A. and J. Hill (2006). Data analysis using regression and multilevel/hierarchical models, Cambridge university press.

Gomes, V. H., S. D. IJff, N. Raes, I. L. Amaral, R. P. Salomão, L. de Souza Coelho, F. D. de Almeida Matos, C. V. Castilho, D. de Andrade Lima Filho and D. C. López (2018). "Species Distribution Modelling: Contrasting presence-only models with plot abundance data." Scientific reports **8**(1): 1003.

Guisan, A., C. H. Graham, J. Elith, F. Huettmann and N. S. Distri (2007). "Sensitivity of predictive species distribution models to change in grain size." Diversity and Distributions **13**(3): 332-340.

41

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

Gutierrez-Tapia, P. and R. E. Palma (2016). "Integrating phylogeography and species distribution models: cryptic distributional responses to past climate change in an endemic rodent from the central Chile hotspot." Diversity and Distributions **22**(6): 638-650.

Hijmans, R. J. (2012). "Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model." Ecology **93**(3): 679-688.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones and A. Jarvis (2005). "Very high resolution interpolated climate surfaces for global land areas." International Journal of Climatology **25**(15): 1965-1978.

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

Huang, J. and E. A. Frimpong (2016). "Limited transferability of stream-fish distribution models among river catchments: reasons and implications." Freshwater Biology **61**(5): 729-744.

Jackson, S. T. and J. T. Overpeck (2000). "Responses of plant populations and communities to environmental changes of the late Quaternary." Paleobiology **26**(4): 194-220.

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

Kearney, M. (2006). "Habitat, environment and niche: what are we modelling?" Oikos **115**(1): 186-191.

Keith, D. A., M. Mahony, H. Hines, J. Elith, T. J. Regan, J. B. Baumgartner, D. Hunter, G. W. Heard, N. J. Mitchell, K. M. Parris, T. Penman, B. Scheele, C. C. Simpson, R. Tingley, C. R. Tracy, M. West and H. R. Akcakaya (2014). "Detecting Extinction Risk from Climate Change by IUCN Red List Criteria." Conservation Biology **28**(3): 810-819.

Kriticos, D. J., B. L. Webber, A. Leriche, N. Ota, I. Macadam, J. Bathols and J. K. Scott (2012). "CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling." Methods in Ecology and Evolution **3**(1): 53-64.

Kuebler, D., P. Hildebrandt, S. Guenter, B. Stimm, M. Weber, R. Mosandl, J. Munoz, O. Cabrera, N. Aguirre, J. Zeilinger and B. Silva (2016). "Assessing the importance of topographic variables for the spatial distribution of tree species in a tropical mountain forest." Erdkunde **70**(1): 19-47.

Leroy, B., C. N. Meynard, C. Bellard and F. Courchamp (2015). "virtualspecies, an R package to generate virtual species distributions." Ecography **39**(6): 599–607.

Lezama Ochoa, N., H. Murua, G. Chust, E. Van Loon, J. Ruiz, M. Hall, P. Chavance, A. Delgado de Molina and E. Villarino (2016). "Present and future potential habitat distribution of *Carcharhinus falciformis* and *Canthidermis maculata* by-catch species in the tropical tuna purse-seine fishery under climate change." Frontiers in Marine Science **3**.

Lobo, J. M., A. Jimenez-Valverde and R. Real (2008). "AUC: a misleading measure of the performance of predictive distribution models." Global Ecology and Biogeography **17**(2): 145-151.

43

Lopatin, J., K. Dolos, H. J. Hernandez, M. Galleguillos and F. E. Fassnacht (2016). "Comparing Generalized Linear Models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile." Remote Sensing of Environment **173**: 200-210.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics. P. Zarembka, Academic Press**:** 105-142.

Meynard, C. N. and D. M. Kaplan (2013). "Using virtual species to study species distributions and model performance." Journal of Biogeography **40**(1): 1-8.

Meynard, C. N., B. Leroy and D. M. Kaplan (2019). "Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?" Ecography.

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

Murphy, A. H. and R. L. Winkler (1992). "Diagnostic verification of probability forecasts." International Journal of Forecasting **7**(4): 435-455.

Muscarella, R., P. J. Galante, M. Soley-Guardia, R. A. Boria, J. M. Kass, M. Uriarte and R. P. Anderson (2014). "ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models." Methods in Ecology and Evolution **5**(11): 1198-1205.

Nix, H. A. (1986). Biogeographic Analysis of Australian Elapid Snakes. Atlas of Elapid Snakes of Australia. R. Longmore. Canberra, Australian Government Publishing Service**:** 4-15.

Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura and M. B. Araújo (2011). Ecological niches and geographic distributions (MPB-49), Princeton University Press.

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

Phillips, S. J., M. Dudik, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick and S. Ferrier (2009). "Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data." Ecological Applications **19**(1): 181-197.

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

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

Raes, N. and H. ter Steege (2007). "A null-model for significance testing of presence-only species distribution models." Ecography **30**(5): 727-736.

Raghavan, R. K., D. G. Goodin, G. A. Hanzlicek, G. Zolnerowich, M. W. Dryden, G. A. Anderson and R. R. Ganta (2016). "Maximum Entropy-Based Ecological Niche Model and Bio-Climatic Determinants of Lone Star Tick (*Amblyomma americanum*) Niche." Vector-Borne and Zoonotic Diseases **16**(3): 205-211.

Reineking, B. and B. Schröder (2006). "Constrain to perform: regularization of habitat models." Ecological Modelling **193**(3-4): 675-690.

Robinson, O. J., V. Ruiz‐Gutierrez and D. Fink (2018). "Correcting for bias in distribution modelling for rare species using citizen science data." Diversity and Distributions **24**(4): 460-472.

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

Smith, A. B., W. Godsoe, F. Rodriguez-Sanchez, H. H. Wang and D. Warren (2019). "Niche Estimation Above and Below the Species Level." Trends Ecol Evol **34**(3): 260-273.

Soberon, J. and A. T. Peterson (2005). "Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas." 2005 **2**.

Soley-Guardia, M., E. E. Gutierrez, D. M. Thomas, J. Ochoa-G, M. Aguilera and R. P. Anderson (2016). "Are we overestimating the niche? Removing marginal localities helps ecological niche models detect environmental barriers." Ecology and Evolution **6**(5): 1267-1279.

Sutherst, R. W. (2014). "Pest species distribution modelling: origins and lessons from history." Biological Invasions **16**(2): 239-256.

Thuiller, W., B. Lafourcade, R. Engler and M. B. Araújo (2009). "BIOMOD – a platform for ensemble forecasting of species distributions." Ecography **32**(3): 369-373.

Torres, L. G., P. J. H. Sutton, D. R. Thompson, K. Delord, H. Weimerskirch, P. M. Sagar, E. Sommer, B. J. Dilley, P. G. Ryan and R. A. Phillips (2015). "Poor Transferability of Species Distribution Models for a Pelagic Predator, the Grey Petrel, Indicates Contrasting Habitat Preferences across Ocean Basins." Plos One **10**(3).

van Proosdij, A. S., M. S. Sosef, J. J. Wieringa and N. Raes (2016). "Minimum required number of specimen records to develop accurate species distribution models." Ecography **39**(6): 542-552.

VanDerWal, J., L. P. Shoo, C. N. Johnson and S. E. Williams (2009). "Abundance and the Environmental Niche: Environmental Suitability Estimated from Niche Models Predicts the Upper Limit of Local Abundance." American Naturalist **174**(2): 282-291.

Varela, S., R. P. Anderson, R. Garcia-Valdes and F. Fernandez-Gonzalez (2014). "Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models." Ecography **37**(11): 1084-1091.

Veloz, S. D. (2009). "Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models." Journal of Biogeography **36**(12): 2290-2299.

Warren, D. L. (2012). "In defense of 'niche modeling'." Trends in Ecology & Evolution **27**(9): 497-500.

Warren, D. L. (2013). "'Niche modeling': that uncomfortable sensation means it's working. A reply to McInerny and Etienne." Trends in Ecology & Evolution **28**(4): 193-194.

Warren, D. L., M. Cardillo, D. F. Rosauer and D. I. Bolnick (2014). "Mistaking geography for biology: inferring processes from species distributions." Trends in Ecology & Evolution **29**(10): 572-580.

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

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan and N. P. S. D. W. Group (2008). "Effects of sample size on the performance of species distribution models." Diversity and Distributions **14**(5): 763-773.

Zhu, G.-P., D. Rédei, P. Kment and W.-J. Bu (2014). "Effect of geographic background and equilibrium state on niche model transferability: predicting areas of invasion of *Leptoglossus occidentalis*." Biological Invasions **16**(5): 1069-1081.

Zurell, D. (2017). "Integrating demography, dispersal and interspecific interactions into bird distribution models." Journal of Avian Biology **48**(12): 1505-1516.

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