Fauxcurrence: simulating multi-species occurrences for null models in species distribution modelling and biogeography

Authors

Owen G. Osborne[1*], Henry G. Fell[2], Hannah Atkins[3], Jan van Tol[4], Daniel Phillips[1], Leonel Herrera-Alsina[5], Poppy Mynard[5], Greta Bocedi[5], Cécile Gubry-Rangin[5], Lesley T. Lancaster[5], Simon Creer[1], Meis Nangoy[6], Fahri Fahri[7], Pungki Lupiyaningdyah[8], I Made Sudiana[9], Berry Juliandi[10], Justin M.J. Travis[5], Alexander S.T. Papadopulos[1], Adam C. Algar[2,11†]

Author affiliations

[1] School of Natural Sciences, Bangor University, Environment Centre Wales, Deiniol Road, Bangor, LL57 2UW, UK

[2] School of Geography, Sir Clive Granger Building, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

[3] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK

[4] Naturalis Biodiversity Center, P.O. Box 9517, 2300 RA Leiden, Netherlands

[5] School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UK

[6] Faculty of Animal Husbandry, Sam Ratulangi University, Kampus Bahu Street, Manado, 95115 Indonesia

[7] Department of Biology, Tadulako University, Palu, Indonesia

[8] Museum Zoologicum Bogoriense, Research Center for Biology, Indonesian Institute of Sciences, Cibinong, 16911, Indonesia

[9] Microbial Ecology Research Group, Research Center for Biology, Indonesian Institute of Sciences, Cibinong 19611, Indonesia

[10] Department of Biology, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia

[11] Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada

Corresponding authors

[*]Owen G. Osborne: o.osborne@bangor.ac.uk

[†]Adam C. Algar: aalgar@lakeheadu.ca

ABSTRACT

Defining appropriate null expectations for species distribution hypotheses is important because sampling bias and spatial autocorrelation can produce realistic, but ecologically meaningless, geographic patterns. Generating null species occurrences with similar spatial structure to observed data can help overcome these problems, but existing methods focus on single or pairs of species and do not incorporate between-species spatial structure that may occlude comparative biogeographic analyses. Here, we describe an algorithm for generating randomised species occurrence points that mimic the within- and between-species spatial structure of real datasets and implement it in a new R package - *fauxcurrence*. The algorithm can be implemented on any geographic domain for any number of species, limited only by computing power. To demonstrate its utility, we apply the algorithm to two common analysis-types: testing the fit of species distribution models (SDMs) and evaluating niche-overlap. The method works well on all tested datasets within reasonable timescales. We found that many SDMs, despite a good fit to the data, were not significantly better than null expectations and identified only two cases (out of a possible 32) of significantly higher niche divergence than expected by chance. The package is user-friendly, flexible and has many potential applications beyond those tested here, such as joint SDM evaluation and species co-occurrence analysis, spanning the areas of ecology, evolutionary biology and biogeography.

KEYWORDS

Environmental niche model; Joint species distribution modelling; Niche conservatism; Niche divergence; Niche overlap; Null biogeographical model; Species distribution model

1 INTRODUCTION

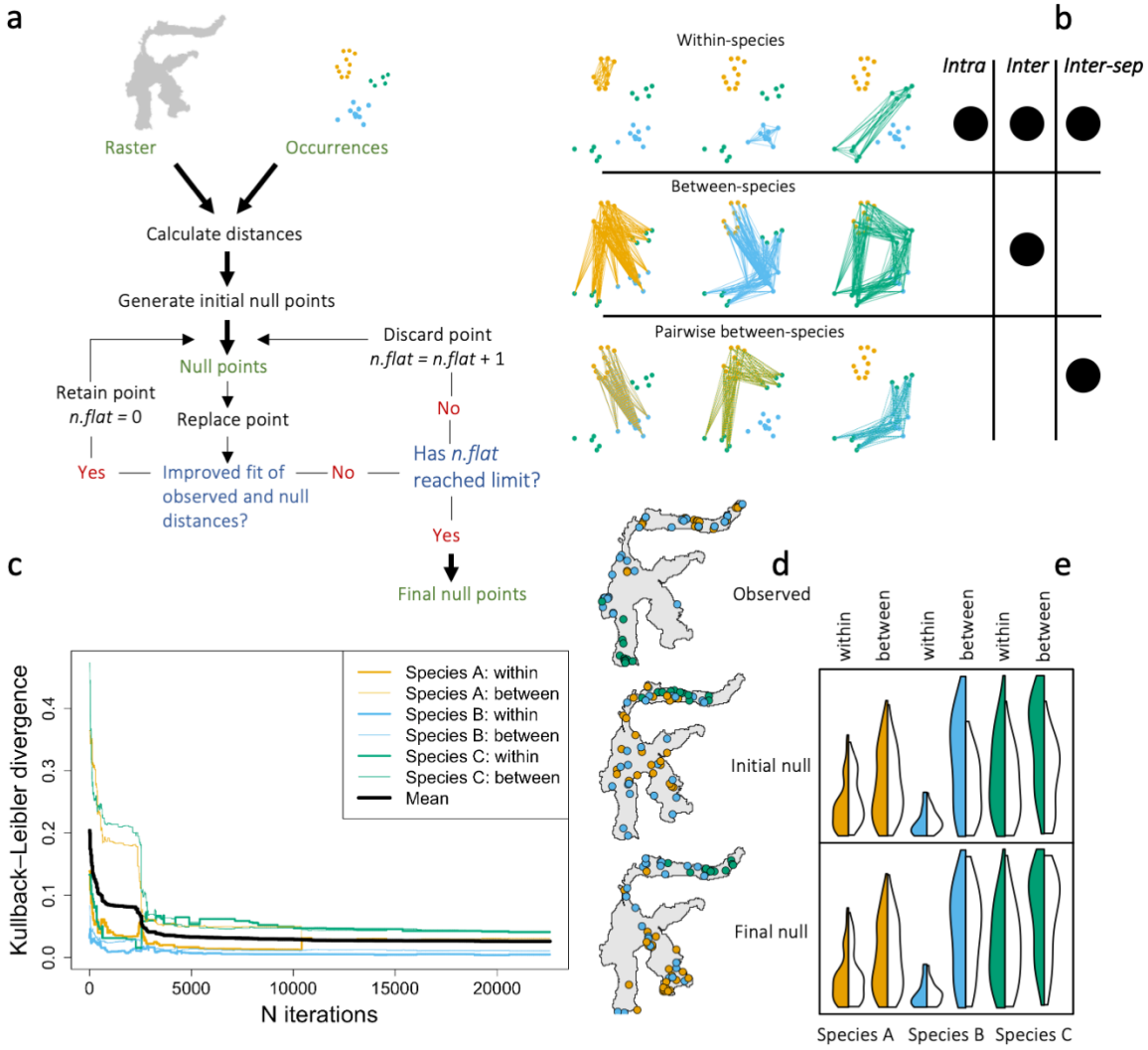
Eco-geographical hypothesis testing using species occurrence data can be hampered by spatial autocorrelation and the difficulty of defining appropriate null expectations (Bahn & Mcgill, 2007; Beale, Lennon, & Gimona, 2008; Chapman, 2010; Fourcade, Besnard, & Secondi, 2018; Moore, Bagchi, Aiello-Lammens, & Schlichting, 2018). A major issue is that spatial clustering of conspecific, or separation of heterospecific, occurrence records can be affected by multiple factors, which are often difficult to disentangle. These include: i) habitat suitability (Phillips, Anderson, & Schapire, 2006), ii) dispersal limitation (Glor & Warren, 2010), iii) interactions between individuals of the same or different species such as conspecific attraction, competitive exclusion or mutualism (Mielke et al., 2020), or iv) sampling bias, where occurrence records are more likely to be collected from more easily accessible or intensively studied areas (Phillips et al., 2009).

One approach to overcome these issues has been to use null species occurrences to define the expectations if only the inherent spatial structure within species has shaped their distributions (Beale et al., 2008; Algar, Mahler, Glor, & Losos, 2013). To define the null expectation, these approaches use an iterative procedure to produce null species distributions with similar spatial structure to observed occurrences, excluding any consideration of environment or specific geographic location. While these methods are well-suited to testing habitat-suitability hypotheses for single species, they do not take spatial structure between species into account, making them unsuitable for testing multispecies hypotheses involving, for example, niche overlap or range boundaries. Some methods designed to test niche overlap hypotheses employ null models for pairs of species, but these either do not take spatial structure into account (Warren, Glor, & Turelli, 2008), or simply translocate the entire set of occurrence

82 points. This preserves spatial structure but limits their application to species which are range-
83 restricted relative to the study region (Nunes & Pearson, 2017).
84
85 Here, we present a method to fill this gap, which is implemented in a new *R* package –
86 *fauxcurrence* version 1.0 (available at https://github.com/ogosborne/fauxcurrence). The
87 package can produce null species occurrences which preserve the spatial structure within and
88 between an arbitrary number of species, and provides many options to tailor these
89 occurrences to the user's needs. We demonstrate the utility of the package using a dataset of
90 22 species of plants, vertebrates and arthropods. We use the resulting null occurrence points
91 to test the significance of species distribution models (SDMs) and to test for significant
92 deviations from null expectations of niche overlap between species.
93
94 **2 MATERIALS AND METHODS**
95 **2.1 Method description**
96 Our method (Fig. 1a) has three main modes of operation, distinguished by how inter-point
97 distances are used to define spatial structure. Within-species distances (divided into one
98 subset per species) are always included and can also be used alone (which we refer to here as
99 the "*Intra*" null model; Fig. 1b). The total set of between-species distances can be divided
100 into subsets in two ways: either as sets of general between-species distances per species (i.e.,
101 the distances from a species' occurrence points to all heterospecific occurrence points; the
102 "*Inter*" null model; Fig. 1b) or as a separate set of distances between each pair of species in
103 the dataset (the "*Inter-sep*" model; Fig. 1b; Appendix 1.1).
104
105 The user provides a set of species occurrence points and a raster defining the study area. The
106 algorithm begins by randomly generating one simulated occurrence point per species. It then

adds occurrence points for each species by drawing each new point *D* distance away from a random existing conspecific point (where *D* is sampled from the empirical distribution function of observed within-species distances) until each species has the same number of occurrences as in the observed dataset (Appendix 1.7).

Once the initial set of simulated points are generated, the fit of their spatial structure to that of the observed points is iteratively improved. For each iteration, one point is replaced and the match between null and observed interpoint distances is evaluated using discrete Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), where smaller values indicate a better match between the simulated and observed points. Since there are multiple interpoint distance distributions (i.e., within- and between-species distances for multiple species or pairs of species), a weighted mean of KL-divergence across all distributions is used, weighted such that within- and between-species distances contribute equally. The point replacement is only retained if it improves the match, and this procedure is repeated until either no improvement has been made for a set number of iterations or a maximum iteration limit has reached (Fig. 1c-e; see Appendix 1 for full details).



130

131 Figure 1. Overview of the method. A flowchart (a) shows the basic functioning of the
132 algorithm: initial null points are generated based on observed inter-point distances, these are
133 then iteratively improved. The algorithm finishes when the number of iterations with no
134 improvement (*n.flat*) reaches a user-defined limit. There are three classes of inter-point
135 distance sets in the algorithm (b), within-species (used in all models), general between-
136 species (used in the *Inter* model) and pairwise between-species (used in the *Inter-sep* model).
137 The circles to the right of each indicate which null models they are used in. Panels (c-e) show
138 an example run of the *Inter* model. Kullback-Leibler divergence decreases across iterations
139 (c) and this improvement can be clearly seen when comparing the initial and final null
140 occurrence points (d; map panels). The left of each split violin plot (e) shows the density of

observed interpoint distances for each of the distance sets in the model and the right shows those of the null, which are more similar to the observed distances in the final null (bottom) than the initial null (top).

2.2 Test data

We tested the method on seven species occurrence datasets from Sulawesi, Indonesia each containing between one and six species from a single genus (Fig. S1a; Appendix 2). We ran the *Intra* model on all datasets, and the *Inter* model on all datasets with over one species. Since the *Inter* and *Inter-sep* models are identical for species pairs, we only ran the *Inter-sep* model for datasets with over two species. The iterative improvement was continued until there had been no improvement for 10,000 iterations. For each dataset/model combination, we produced 1,000 independent, null occurrence replicates, each with similar spatial structure, but differing in the final locations selected by the algorithm.

2.3 SDM model-fitting and niche overlap

We used *Maxent* v. 3.4.1 (Phillips et al., 2006) to build Species Distribution Models (SDMs) for all species using the 19 *BIOCLIM* climate variables and altitude from the *WorldClim1* database (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005). To determine if SDMs for the observed data had a significantly better fit than the null models, we calculated P-values by comparison of area under the receiver operating characteristic curve (AUC) in SDMs built from observed data to those from all null model replicates. For datasets with more than one species, we compared null and observed niche overlap using Schoener's *D* (Schoener, 1968) and Warren's *I* (Warren et al., 2008), between all pairs of congeneric species (See Appendix 3 for full details).

166 **3 RESULTS**

167 **3.1 Method performance**

168 Average time per iteration ranged from 3.2 milliseconds (ms) to 21.5 ms and the mean
169 number of iterations ranged from 48,155 to 341,610 (Fig. S1). Number of occurrences was
170 the best predictor of number of iterations to model completion, although number of species
171 also had an effect (Table S1). Plotting KL divergence across iterations suggested that 10,000
172 iterations without improvement was more than sufficient to minimise the KL-divergence
173 statistic for most datasets (Fig. S2-S4).

174

175 **3.2 Application to SDM model-fitting and niche overlap**

176

177 The AUC values were over 0.8 for 13 of 22 species and over 0.9 for four species, values
178 typically interpreted as "excellent" or "outstanding" discrimination, respectively (Hosmer,
179 Lemeshow, & Sturdivant, 2013). Nevertheless, just six of these were significantly greater
180 than null expectations according to at least one of our null model types (Fig. 2a-c) and where
181 they were significantly greater than those of one null model type, they were often not
182 significantly different to those of the others (Fig. 2a-d). In fact, SDMs for only two species
183 had a significantly better AUC than all applied null models (Fig. 2d).

184

185 Observed niche divergence differed from null expectations in only two species pairs, both of
186 which showed significant niche divergence (Fig. S5-S8). Both of these involved *Cyrtandra*
187 *geocarpa*, which was also one of only two species whose SDM fit significantly better than
188 those from all the null models (Fig. 2). Both *Inter* and *Inter-sep* null models, and both niche
189 overlap statistics found the same two species pairs to be significant (Figs. S5-S8).

190

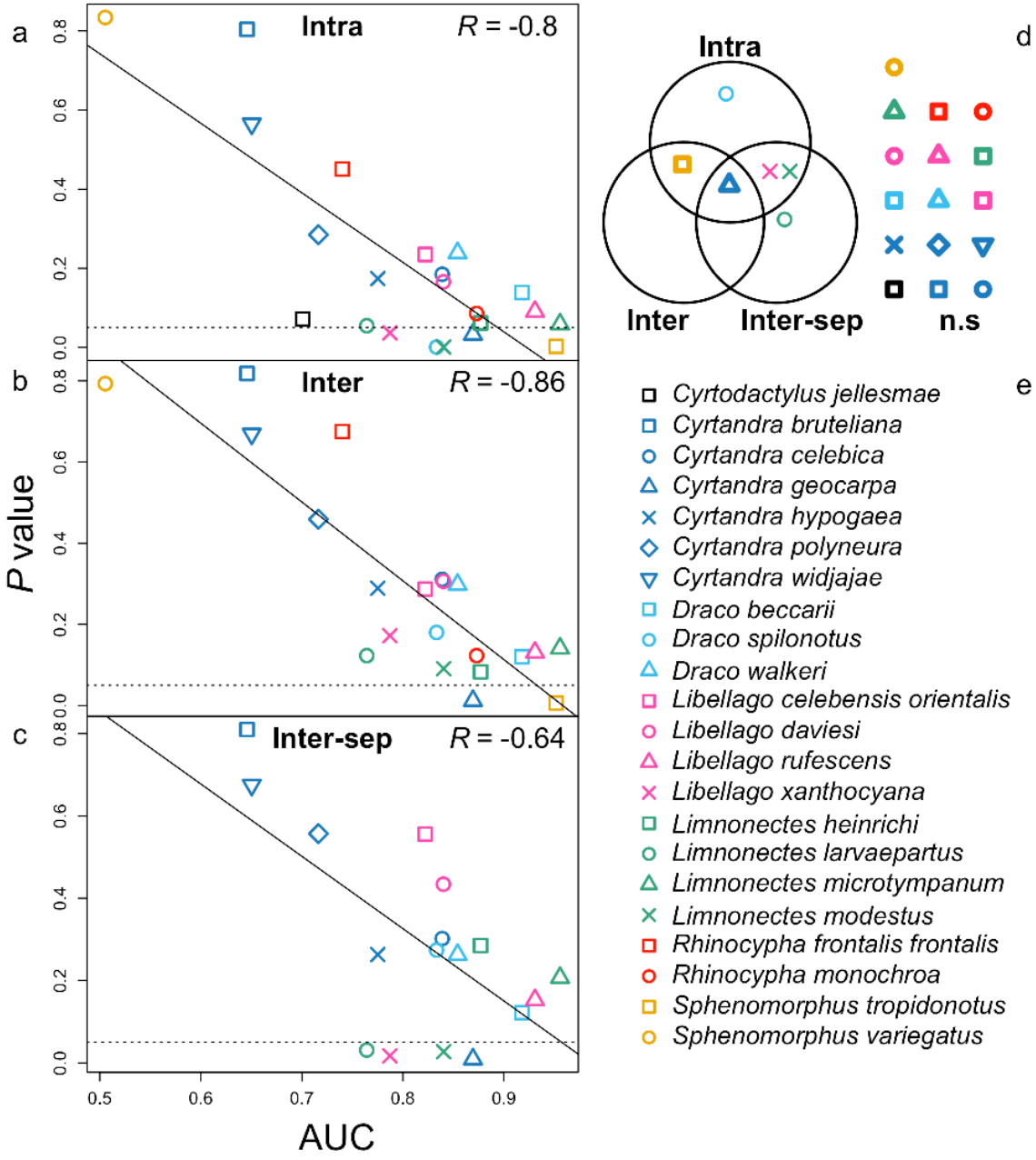


Figure 2. The relationship between Area Under Curve (AUC) of the Species Distribution Models (SDMs) for observed occurrences, and the *P*-values for comparison to AUC of each null model type (a-c). Dotted lines mark 0.05 on the *P*-value axis, the solid line is a linear regression line and Pearson's correlation is shown in the top-right of each plot. A Venn diagram (d) shows the overlap in significance ($P < 0.05$) between the three models. Each labelled circle contains species with significantly higher observed AUC than those of

198 simulated data of each null model type and those which were significantly higher than
199 simulated data from multiple null models are shown in the appropriate intersection (those
200 outside the circles were not significant: "n.s"). Where all implemented models agree, symbols
201 are in bold. Species symbols are shown in the legend (e).
202
203
204 **4 DISCUSSION**
205 Here, we describe a new tool to help overcome the difficulties of defining null expectations
206 when working with multi-species occurrence data. Our case study demonstrates the utility of
207 the method. Although SDMs were calculated individually for each species, the between-
208 species distances used in the null model had a major effect on their significance. For
209 example, *Draco spilonotus* had a significantly better fit than the *Intra* null model ($P = 0.001$),
210 but was not significantly different from either the *Inter* or *Inter-sep* models ($P = 0.18$ and $P =$
211 0.27, respectively). Such a pattern may be expected if, for example, competitive exclusion
212 between species, rather than climate suitability, was responsible for the geographical
213 separation of species into ranges that happen to have distinct climates (Godsoe, Franklin, &
214 Blanchet, 2017). While AUC and *P*-values were highly correlated, many species with very
215 high AUC scores did not have a significantly better fit than the null models. The lowest AUC
216 score which was significantly higher than any of the null models was 0.76, underlining that
217 SDMs with low AUC scores (e.g., < 0.75) should be treated with great caution. Using our
218 method to demonstrate that SDMs fit significantly better than null expectations will provide
219 much greater certainty than simple inspection of AUC.
220
221 We also demonstrated the applicability of our method to identify significant niche divergence
222 (or niche conservatism). Our method has advantages over existing approaches. The approach

of Nunes & Pearson (2017) creates null replicates by translocating and rotating observed species occurrences. This would clearly be inappropriate for a study region such as Sulawesi, since most rotations and translocations would result in a large proportion of occurrences being translocated to the sea, leading to a high number of similar replicates, as noted by the authors (Nunes & Pearson, 2017). The contorted geography of Sulawesi is not unique, and many intensely studied locations such as Isabela Island in the Galápagos archipelago and Lord Howe Island, Australia, also fall into this category. It is possible our model could have similar issues in extreme cases, where dense occurrence points cover a very large proportion of the study region. However, we expect it to work on a wider range of datasets due to our use of "as similar as possible" rather than identical spatial structure and the capability to include overland distances.

Aside from the two applications shown here, there are many other potential uses for the package. For example, the performance of joint species distribution models (Pollock et al., 2014), which jointly model environmental and community effects on species distributions, could be assessed with our approach in a similar way to our tests of SDM fit. Other potential applications include identifying significant effects of biotic factors (other species) on a focal species' distribution (Algar et al., 2013; Giannini, Chapman, Saraiva, Alves-dos-Santos, & Biesmeijer, 2013) where the comparison of different null models can give insight into the relevance of pairwise and complex biotic interactions, and testing for significant co-occurrence of range-boundaries across clades (Swenson & Howard, 2005). More generally, *fauxcurrence*-generated occurrences could be used in any theoretical biogeographical application where realistic occurrences of species and clades are required. Overall, the method is easy to use, flexible, and can add rigour and insight into investigations of a wide range of problems in ecology, evolution and biogeography.

ACKNOWLEDGEMENTS

The work was funded by Newton Fund (UK)/NERC (UK)/RISTEKDIKTI (Indonesia) grants awarded to JT, BJ, ACA, ASTP, CG-R, GB and LTL (Grant numbers: NE/S006923/1; NE/S006893/1; 2488/IT3.L1/PN/2020; and 3982/IT3.L1/PN/2020). GB and CG-R are funded by Royal Society University Research Fellowships (UF160614 and UF150571 respectively).

AUTHOR CONTRIBUTIONS

Ideas for the study were conceived by ACA, OGO, HGF and ASTP. OGO wrote the package based partly on earlier code by ACA and HGF. HA, JvT and DP collated the species occurrence data. HGF and OGO ran the species distribution modelling and niche overlap analyses. ASTP and ACA supervised the project. OGO drafted the first version of the manuscript and all authors contributed to and approved the final version of the manuscript.

REFERENCES

Algar, A. C., Mahler, D. L., Glor, R. E., & Losos, J. B. (2013). Niche incumbency, dispersal limitation and climate shape geographical distributions in a species-rich island adaptive radiation. *Global Ecology and Biogeography*, *22*(4), 391–402. doi:10.1111/geb.12003

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

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

273 Chapman, D. S. (2010). Weak climatic associations among British plant distributions. *Global*
274 *Ecology and Biogeography*, *19*(6), 831–841. doi:10.1111/j.1466-8238.2010.00561.x
275 Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of
276 species, or the challenge of selecting environmental predictors and evaluation statistics.
277 *Global Ecology and Biogeography*, *27*(2), 245–256. doi:10.1111/geb.12684
278 Giannini, T. C., Chapman, D. S., Saraiva, A. M., Alves-dos-Santos, I., & Biesmeijer, J. C.
279 (2013). Improving species distribution models using biotic interactions: A case study of
280 parasites, pollinators and plants. *Ecography*, *36*(6), 649–656. doi:10.1111/j.1600-
281 0587.2012.07191.x
282 Glor, R. E., & Warren, D. (2010). Testing ecological explanations for biogeographic
283 boundaries. *Evolution*, *65*(3), 673–683. doi:10.1111/j.1558-5646.2010.01177.x
284 Godsoe, W., Franklin, J., & Blanchet, F. G. (2017). Effects of biotic interactions on modeled
285 species' distribution can be masked by environmental gradients. *Ecology and Evolution*,
286 *7*(2), 654–664. doi:10.1002/ece3.2657
287 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high
288 resolution interpolated climate surfaces for global land areas. *International Journal of*
289 *Climatology*, *25*(15), 1965–1978. doi:10.1002/joc.1276
290 Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). 5.2.4 Area Under the Receiver
291 Operating Characteristic Curve. In *Applied Logistic Regression, 3rd ed.* (pp. 173–182).
292 Kullback, S., & Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical*
293 *Statistics*, *22*(1), 79–86.
294 Mielke, K. P., Claassen, T., Busana, M., Heskes, T., Huijbregts, M. A. J., Koffijberg, K., &
295 Schipper, A. M. (2020). Disentangling drivers of spatial autocorrelation in species
296 distribution models. *Ecography*, *43*(12), 1741–1751. doi:10.1111/ecog.05134
297 Moore, T. E., Bagchi, R., Aiello-Lammens, M. E., & Schlichting, C. D. (2018). Spatial

298 autocorrelation inflates niche breadth–range size relationships. *Global Ecology and*
299 *Biogeography*, *27*(12), 1426–1436. doi:10.1111/geb.12818
300 Nunes, L. A., & Pearson, R. G. (2017). A null biogeographical test for assessing ecological
301 niche evolution. *Journal of Biogeography*, *44*(6), 1331–1343. doi:10.1111/jbi.12910
302 Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of
303 species geographic distributions. *Ecological Modelling*, *190*, 231–259.
304 doi:10.1016/j.ecolmodel.2005.03.026
305 Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S.
306 (2009). Sample selection bias and presence-only distribution models: Implications for
307 background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.
308 doi:10.1890/07-2153.1
309 Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., …
310 Mccarthy, M. A. (2014). Understanding co-occurrence by modelling species
311 simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology*
312 *and Evolution*, *5*(5), 397–406. doi:10.1111/2041-210X.12180
313 Schoener, T. W. (1968). The Anolis Lizards of Bimini: Resource Partitioning in a Complex
314 Fauna. *Ecology*, *49*(4), 704–726.
315 Swenson, N. G., & Howard, D. J. (2005). Clustering of contact zones, hybrid zones, and
316 phylogeographic breaks in North America. *American Naturalist*, *166*(5), 581–591.
317 doi:10.1086/491688
318 Warren, D. L., Glor, R. E., & Turelli, M. (2008). Environmental niche equivalency versus
319 conservatism: Quantitative approaches to niche evolution. *Evolution*, *62*(11), 2868–
320 2883. doi:10.1111/j.1558-5646.2008.00482.x
321