

Effect of sampling strategies on the response curves estimated by plant species distribution models

Bazzichetto Manuele^{1*}, Lenoir Jonathan², Da Re Daniele³, Tordoni Enrico⁴, Rocchini Duccio^{5,6}, Malavasi Marco^{6,7}, Barták Vojtech^{6^l} & Sperandii Marta Gaia^{1^l}

¹ Centro de Investigaciones sobre Desertificación (CSIC-UV-GV), Valencia, Spain; ² UMR CNRS 7058 « Ecologie et Dynamique des Systèmes Anthropisés » (EDYSAN), Université de Picardie Jules Verne, 1 rue des Louvels, 80000 Amiens, France; ³ Georges Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Place Louis Pasteur 3, 1348 Louvain-la-Neuve, Belgium; ⁴ Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia; ⁵ BIOME Lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Bologna, Italy; ⁶ Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha--Suchbát, Czech Republic; ⁷ Department of Chemistry, Physics, Mathematics and Natural Sciences, University of Sassari, Sassari, Italy.

* Corresponding author: manuele.bazzichetto@gmail.com

^lJoint senior authors.

1 Abstract

2 Species distribution models (SDMs) rely on species presence/absence or abundance data and environmental
3 variables to estimate species response curves. Therefore, the quality (and quantity, i.e., sample size) of the data to
4 describe the species distribution determines the quality of the estimate of the species-environment relationship.
5 However, SDMs are seldom fitted on high-quality data collected strictly for that purpose. Usually, SDMs rely on
6 a collection of opportunistic datasets sampled from previous projects or public repositories with different
7 objectives. Here, we aim at assessing how the sampling strategy capturing the geographic distribution of a species
8 affects the accuracy and precision of its response curves along environmental gradients, as estimated by parametric
9 SDMs. We simulated the occurrence of two virtual plant species across the Abruzzo region (Italy). We assumed
10 that the two virtual plants were similarly affected by precipitation, but one had a wider realised niche for
11 temperature (i.e., higher thermal tolerance), and, as a result, a wider distribution extent. Then, we sampled
12 occurrence data for the two species following five different sampling strategies: random, stratified, systematic,
13 topographic, and uniform (the latter performed within the environmental space). In addition, we simulated a
14 spatially biased sampling design by collecting presence/absence data close to roads. To account for sample size,
15 we also repeated our simulations along a gradient of increasing sampling effort, i.e., number of sampled locations.
16 In total, we ran 500 replicates for each combination of sampling design and effort. For each replicate, we fitted
17 SDMs using binomial generalised linear models and extracted the model coefficients for precipitation and
18 temperature to be compared with the true coefficients from the virtual species' model. We evaluated the quality
19 of the estimated response curves by computing the following measures: bias (accuracy), variance (precision), and
20 mean squared error (accuracy and precision). Our results suggest that a proper estimate of the species response
21 curve can be obtained when the choice of the sampling strategy is guided by the species' ecology. In particular,
22 species with wide tolerances to environmental drivers may be better modelled using data uniformly collected
23 within the environmental space, while none of the tested sampling designs seemed to substantially outperform the
24 others for modelling species with a narrow realised niche.

25 **Keywords:** virtual species, simulation, mean squared error, bias, environmental space, ecological niche breadth,
26 sampling bias.

27 Significance statement

28 The choice of the most appropriate strategy to sample presence/absence data for plant species distribution models
29 depends on the species' ecology, with generalist species being more sensitive to the sampling strategy used.

30 1. Introduction

31 Species distribution models (SDMs) rely on species observations (presence/absence, abundance) and spatially
32 explicit variables (e.g., climatic, edaphic, topographic, anthropogenic) to estimate the relationship between living
33 organisms and their environment. Specifically, SDMs allow deriving species' response curves along chosen
34 environmental gradients, which define how species respond to the environmental conditions they experience.
35 Being based on statistical models fitted to field-collected observations, SDMs are sensitive to the quality (and
36 quantity) of data used for model calibration (Hirzel & Guisan 2002; McPherson & Jetz 2007; Lobo 2008;
37 Tessarolo et al. 2021). Species presence and absence data, both ideally collected in-situ, would be modelled as a
38 function of environmental variables sampled at the same geographic locations of the species' records. Very often,
39 however, absences are created *in-silico* (e.g., pseudo-absences, background points) to overcome the logistic
40 difficulty of confirming them in the field (Lobo 2008). In any case, SDMs are seldom fitted using species (and
41 environmental) data collected strictly for that purpose. Instead, biodiversity data used as input in SDMs are mostly
42 opportunistic and sampled for different purposes (Hirzel & Guisan 2002, Gábor et al. 2020). Examples include
43 opportunistic data from museum collections or herbaria (Newbold 2010), citizen science (Leandro et al. 2020;
44 Feldman et al. 2021), vegetation surveys (Bazzichetto et al. 2021), or a combination of these (Wasof et al. 2015).
45 The use of data not specifically collected for species distribution modelling can be an issue, as the sampling
46 strategy determines the quality of the species response curves estimated by SDMs (Beck et al. 2014; Baker et al.
47 2022).

48 In principle, species distribution data should be collected in a way that allows for answering our ecological
49 questions. Specifically for SDMs, which are rooted in the niche theory (*sensu* Hutchinson, see Jackson &
50 Overpeck 2000), species distribution data should be sampled so that an adequate description of the realised niche
51 of the species can be achieved (Guisan & Zimmermann 2000). Typically, in vegetation science, SDMs rely on
52 presence-absence data from pre-existing vegetation surveys recorded by botanists and phytosociologists mostly
53 to describe plant communities (co-occurrence data). Such data, not initially collected to model a single species
54 distribution, should be used cautiously, as they might lead to a poor estimate of the relationship between the
55 species and the environment. In this respect, there is a vast scientific literature on the effect of sampling design
56 (and sampling bias) on SDMs, but nearly the totality of these studies evaluated models' predictive performance,
57 i.e., they compared SDMs' predictions to independent observations using accuracy measures such as AUC, True
58 Skill Statistics, Kappa, Sensitivity, Specificity, and the Continuous Boyce Index (see Kadmon et al. 2003; Hirzel
59 et al. 2006; Tessarolo et al. 2014; Varela et al. 2014; Guisan et al. 2017). Instead, and this is not to downplay the
60 importance of assessing models' predictive performance, we argue that SDMs should also be evaluated in terms
61 of their capacity of estimating the true species' response curves and thus the mechanisms generating species
62 distribution. Indeed, measures of predictive accuracy are known to be affected by several factors, including sample
63 prevalence and size (Jiménez-Valverde 2021), which may confound the comparison of SDMs fitted under
64 different circumstances (e.g., different sampling strategies or intensity of sampling bias). Still worse, some
65 accuracy metrics can score high even in the case of poorly defined SDMs (Lobo et al. 2008). Hence, accounting
66 for the performance of coefficients' estimators derived from parametric SDMs, rather than focusing solely on
67 their predictive performance, is important. In this regard, simulations, together with specific measures of accuracy
68 (i.e., bias) and precision (i.e., variance), can provide an alternative for evaluating the influence of different factors
69 on SDMs' capacity of estimating the true coefficients defining species response curves (Gu & Swihart 2004;
70 Fernandes et al. 2018).

71 Here, we use simulations of virtual plant species and data collection to answer the following questions: how does
72 sampling strategy affect the quality of the species response curves derived from SDMs? And more specifically:
73 to what extent are the coefficients' estimators of the species response curves simulated using different sampling
74 designs accurate and/or precise? To quantify accuracy and precision, we use bias, variance, and mean squared
75 error (see Box 1 for definitions).

76 Box 1. Definitions of bias, variance and mean squared error.

Bias: expected difference between an estimator and the parameter. Bias is used to assess accuracy (i.e., quality

of the answer we can get from the analyses of ecological data, Bolker 2008):

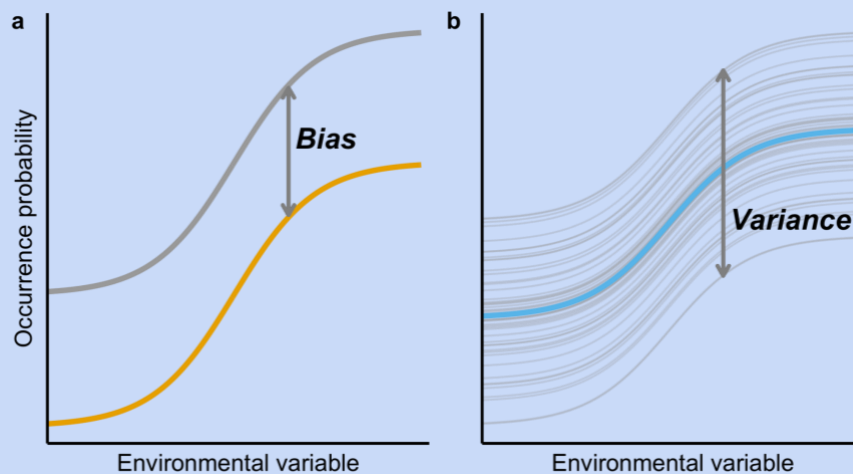
$$\text{Bias} = E[(\hat{\theta} - \theta)]$$

Variance: expected squared difference between an estimator and its expected value (notice that the expected value of the estimator is different from the parameter if the estimator is biased). Variance is used to assess precision (i.e., how largely the estimator fluctuates around its mean on the long-run, Bolker 2008):

$$\text{Variance} = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

Mean Squared Error (MSE): expected squared difference between the estimator and the parameter. The mean squared error can be partitioned in (squared) bias plus variance. Therefore it combines precision and accuracy, and, for this reason, is generally used as a measure of quality of an estimator:

$$\text{MSE} = E[(\hat{\theta} - \theta)^2]$$

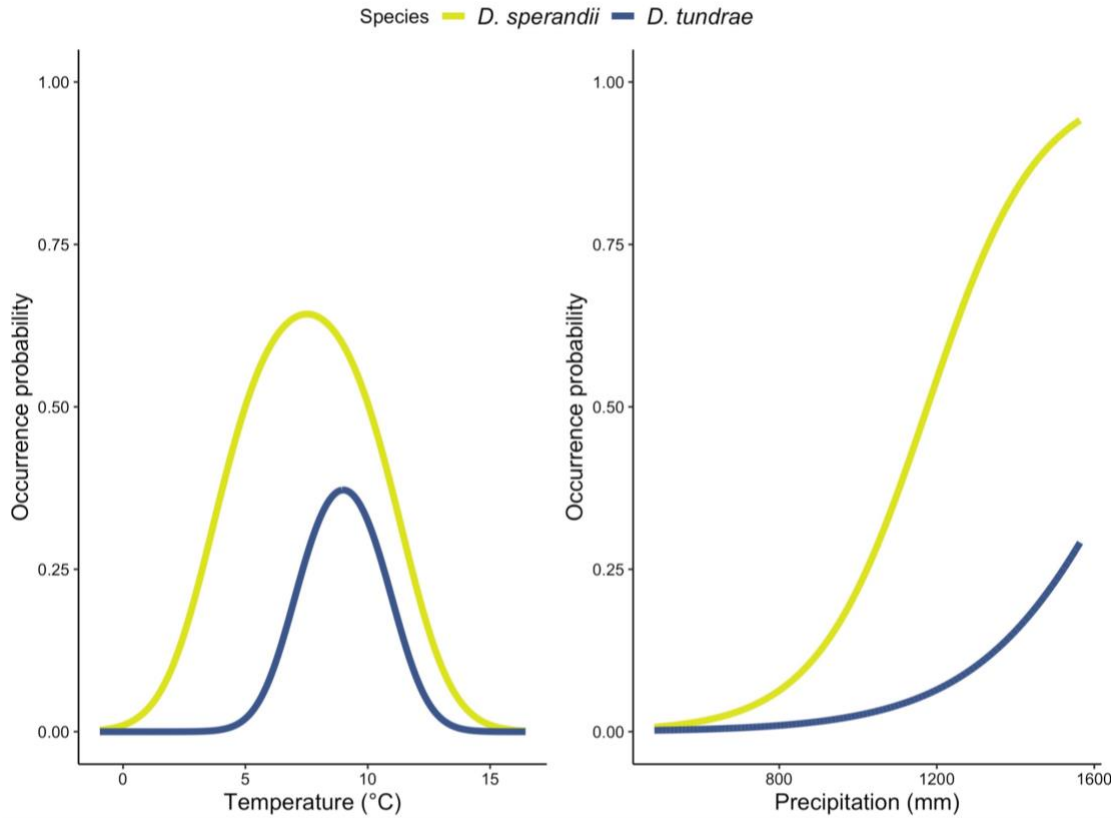


Graphical representation of bias (a) and variance (b). In panel a, the gold logistic function shows the true response curve of the species for a given environmental variable, while the grey function is the long-run average of multiple simulated response curves. The difference between the gold and the grey logistic functions is the bias. In panel b, the blue logistic function represents the long-run average of multiple simulated response curves (in grey). Note that the blue line only represents the true logistic function (in gold in panel a) when the bias equals zero (in which case the variance is the mean squared error). Also note that the figure above provides a ‘simplified’ representation of bias and variance of species’ response curves, as bias does not necessarily produce vertical shifts of the true logistic function, and, similarly, variance may not lead to systematic oscillations around the true response curve.

77 2. Materials and Methods

78 To assess the impact of vegetation sampling on parametric SDMs we used binomial generalised linear models
79 (GLMs). Binomial GLMs are widely used among SDMs practitioners, and their statistical properties are well-
80 known (see McCullagh & Nelder 1989).

81 We focus on the Abruzzo region, located in Central Italy and covering different climates and habitat types (see
82 Figure A1 in Appendix 1). We started by generating two virtual plant species: *Dianthus sperandii* and *D. tundrae*
83 (species’ names are invented and do not relate to the species’ ecological preferences). For the sake of simplicity,
84 we assumed the occurrence of the two virtual species to be only driven by temperature and precipitation. As shown
85 in Figure 1, *D. sperandii* has a thermal optimum at approx. 7.5 °C, and its probability of occurrence increases
86 linearly with precipitation. Similarly, *D. tundrae* has an optimum at approx. 9 °C, and its occurrence probability
87 also increases with precipitation. However, *D. tundrae* has a much more constrained thermal tolerance and a lower
88 prevalence (i.e., the ratio between number of presences and absences). As a result, *D. tundrae* has a narrower
89 distribution than *D. sperandii*. By generating virtual species sharing similar ecological preferences, but different
90 distribution extents, we tested the effect of sampling strategy on SDMs for generalist vs specialist species.



91

92
93

Figure 1 - Simulated response curves of *Dianthus sperandii* (in lime) and *D. tundrae* (in blue) along the temperature (left panel) and precipitation (right panel) gradients.

94
95
96
97

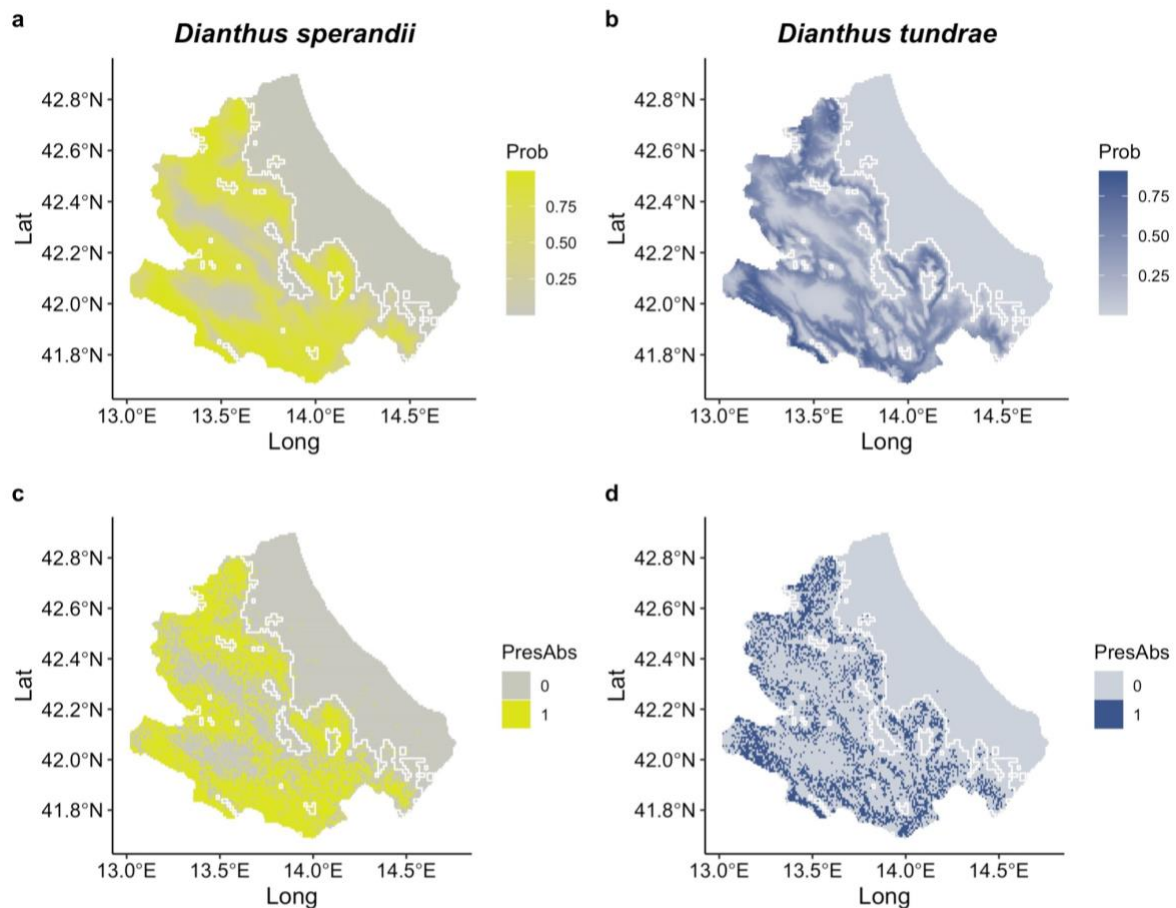
Once we defined the true relationships between the two virtual species and the two climatic variables (by setting the parameters determining the species' response curves, see Equation 1), we computed, for each cell of a raster layer spanning the study area (spatial resolution: ca. 1 km), the true probability of occurrence (p) of the species across the Abruzzo using the following model:

98

$$\text{logit}(p_i) = \alpha + \beta_{pr} * \text{prec}_i + \beta_{tm} * \text{temp}_i + \beta_{tmq} * \text{temp}_i^2 \quad (\text{Equation 1})$$

99
100
101
102
103
104
105
106
107

where $\text{logit}(\cdot)$ is the natural logarithm of the odds $p/(1-p)$, α is the model intercept, β_{pr} is the regression parameter for precipitation, β_{tm} is the parameter for the linear term of temperature, and β_{tmq} is the parameter for the quadratic term of temperature. Regression parameters, here rounded to the third digit, for *D. sperandii* were set to: -11.389 (α), 0.007 (β_{pr}), 1.384 (β_{tm}), and -0.092 (β_{tmq}). For *D. tundrae*, they were set to: -22.173 (α), 0.005 (β_{pr}), 3.779 (β_{tm}), and -0.21 (β_{tmq}). Logit-transformed probabilities were turned to the unit interval [0,1] using the logistic function. Then, we used the true occurrence probability (p) of the two species in a given grid cell of 1 km resolution to simulate their occurrence (presence/absence) across the study region (Figure 2, panels a and b). Specifically, we derived a presence/absence raster layer by drawing at each raster cell a realisation of a Bernoulli trial with probability p . The obtained presence/absence layers are reported in Figure 2 (panels c and d).



108

109 Figure 2 - Simulated occurrence probability and presence/absence data of *Dianthus sperandii* (a, c) and *D. tundrae* (b, d). The white line in
 110 the plots delineates the area of interest for the study (i.e., all lands approx. between 500 and 1800 m a.s.l.).

111 We then simulated what vegetation ecologists would do: go out in the field and collect data! We created six
 112 sampling strategies (Box 2) and fitted SDMs for each of them. Here, our sampling units are the cells of the raster
 113 layer with the presence/absence of the two virtual species (hereafter 'sampling cells', Figure 2, panels c and d). To
 114 keep the simulations as realistic as possible, we carried out the sampling only in a restricted area of the Abruzzo
 115 region: we considered all areas approx. from 500 to 1800 m a.s.l. (90% of the cells included between 518 and
 116 1821 m a.s.l.; minimum elevation: 197 m, maximum elevation: 2791 m) (the perimeter of the area of interest is
 117 marked in white in Figure 2). Indeed, both *D. sperandii* and *D. tundrae* are cold tolerant species, so it wouldn't
 118 make sense to sample their occurrence, e.g., on the coast (where the probability of finding the species is nearly 0,
 119 see Figure 2), or where habitat features are very different from the species' optima. So, by restricting our focus on
 120 a smaller area of interest, we avoided the 'there are no elephants in the Antarctic' paradox (Lobo et al. 2010).

121 Box 2. Description of the simulated sampling strategies.

Random: probably one of the most common sampling strategies. It is used for several purposes, including the description of vegetation patterns across space, and it is usually adopted to ensure independence among sampling units (Lájer 2007).

Systematic: also very common, the systematic strategy collects data from regularly spaced grids to maximise the sampling effort for any number of sampling units. Our systematic approach is similar to the 'grid approach' implemented in Hirzel & Guisan (2002).

Proportional random-stratified: (hereafter, stratified): this survey is one step forward of the random approach. It accounts for the fact that habitat types (i.e., abiotic conditions) are not homogeneously distributed across the geographic space. So, the sampling is performed within *strata* covering many (if not all) combinations of abiotic conditions, including rare habitats (Roleček et al. 2007). In our case, as we only focus on temperature and precipitation (climatic data gathered from CHELSA; Karger et al. 2017), the stratification provides an exhaustive sampling of combinations of the two climatic variables within the geographic space. As the strata were not evenly distributed (i.e., some strata were more widely spread than others), in each stratum we sampled a number of cells proportional to the area of the stratum. The strata were generated as 16 classes combining temperature and precipitation conditions. Notice that the proportional random-stratified converges to the random design when sample size (N) gets large (Hirzel & Guisan 2002).

Topographic: this sampling strategy is commonly used by ecologists to capture a large amount of variability along a given transect. It reproduces the idea of collecting data across transects located in areas with high topographic (and potentially climatic) heterogeneity. To generate this sampling design, we used 4 topographic layers: elevation, slope, northness, and eastness. The last three were derived from the elevation layer, which, in turn, was retrieved at a spatial resolution of approx. 48 m x 65 m using the R package *elevatr* (Hollister 2021; for elevation data sources see <https://github.com/tilezen/joerd/blob/master/docs/data-sources.md#what-is-the-ground-resolution>). To identify areas with highly heterogeneous terrain conditions, we first standardised each topographic layer to have mean value zero and unit variance, and aggregated its spatial resolution to match that of the bioclimatic layers (approx. 1 km). Specifically, each 1 km cell was assigned the standard deviation of the aggregated cells. Then we summed the 4 resulting layers to derive a single one. Finally, to focus the sampling only on those areas featuring high heterogeneity, we retained (and then randomly sampled) only those cells with a standard deviation larger than the median standard deviation of the final layer (all other cell values were set to NAs, and were, therefore, not sampled).

Proximity to roads (hereafter, proximity-to-road): this sampling design reflects the reality of logistic constraints during fieldwork. Specifically, to account for the fact that sampling activities are sometimes preferentially carried out in the most accessible places (e.g., citizen science), we simulated a sampling strategy that maximises access through proximity to roads. The resulting bias has been widely investigated in analyses of species distribution data (Kadmon et al. 2004; Tessarolo et al. 2014). To generate this sampling scenario, we downloaded from OpenStreetMap a layer comprising all major roads in the Abruzzo (using the *osmdata* R package, Padgham et al. 2017). Then we derived a raster layer reporting, for each cell, the corresponding (Euclidean) distance from the closest road. Finally, we transformed the value of each cell (i.e., distance from the closest road) to the corresponding negative exponential (e.g., $\exp(-1 \cdot \text{road distance})$), so that the probability of sampling a given cell decayed (exponentially) as the distance from the closest road increased.

Uniform sampling of the environmental space (hereafter, uniform): this sampling strategy is conceptually similar to the stratified sampling, while, practically, it is implemented as the systematic approach. Indeed, the uniform sampling of the environmental space aims at collecting data from as many habitat types as possible by regularly selecting sampling units within a (here, 10 x 10) grid overlaid to a 2-dimensional (environmental) space. This way, the uniform sampling allows, at the same time, to maximise information on environmental variability and minimise sampling bias (e.g., it avoids over-sampling habitat types that are more widely distributed within the geographic space). In this study, the environmental space was defined as the 2-dimensional plane spanned by temperature and precipitation (see Varela et al. 2014, Hattab et al. 2017; see also Figure A2 in Appendix 2, which shows the portion of the environmental space occupied by the two virtual species).

Maps of design-specific sampling effort are reported in Figure A3 (Appendix 3).

122 The data collected through the 6 sampling approaches (see Box 2) were then used to fit binomial GLMs (link
123 logit), which always included the following terms as predictors: precipitation + temperature + temperature². Each
124 model was fitted to the sampled data using the exact same model formula as in Equation 1, i.e., the one used to

125 generate the occurrence pattern of *D. sperandii* and *D. tundrae*. This allowed quantifying: (i) how much - on
126 average - the estimated coefficients deviated from the true parameters (i.e., bias), (ii) how much - on average -
127 they fluctuated around the average of the coefficient estimator (i.e., variance), and (iii) how much - on average -
128 they fluctuated around the true parameters (i.e., mean squared error) (see Box 1). Note that our measures of bias,
129 variance, and mean squared error (hereafter, MSE) are estimators of these quantities, which we computed
130 replacing expectations by averages (computed over multiple simulations). The simulated sampling activities were
131 replicated 500 times for each of the six sampling strategies we tested, thereby fitting 3000 GLMs. Because
132 regression coefficients of GLMs are estimated by maximum likelihood, they feature desirable properties such as
133 asymptotic unbiasedness and efficiency (i.e., decreasing bias, variance, and therefore MSE, with increasing
134 sample size). As a consequence, a comparison of the impact of different sampling strategies on the bias and
135 variance (and MSE) of the species' response curve cannot be undertaken without accounting for the effect of
136 sample size (i.e., total number of presence/absence records used to fit our GLMs). Therefore, we repeated the 500
137 sampling-specific simulations for an increasing number of sampling cells (i.e., sampling effort): from 200 to 500
138 cells using an increment of 50 cells between both limits. As a result, for each sampling strategy, we obtained 500
139 values of the regression coefficients as estimated by GLMs fitted to datasets of sizes from 200 to 500 (by 50). All
140 datasets contained at least 30 presences, which means 10 presences for each regressor included in the model, i.e.,
141 precipitation, linear and second order polynomial term for temperature (intercept excluded). Correlation among
142 predictor variables (here, temperature and precipitation) was checked at each iteration to avoid its impact on the
143 variance of the coefficients.

144 We compared the sampling approaches, as simulated for the different sampling efforts, in terms of the relative
145 difference among their MSE values. To this aim, we computed the sampling type-specific drop in MSE from the
146 worst performing approach (i.e., the one associated with the highest MSE). We considered an approach as the best
147 performing (at a given sampling effort) when it was associated with the lowest MSE. We then used bias and
148 variance to assess their impact on the species' response curves. It should be noted that, although statistical power
149 calls for big numbers, sample size is one of the most important limiting factors when planning actual sampling
150 campaigns. In this sense, sampling strategies providing high performance at low sampling effort should be
151 preferred for their efficiency, as they represent the best trade-off between feasibility and accuracy of species
152 response curves.

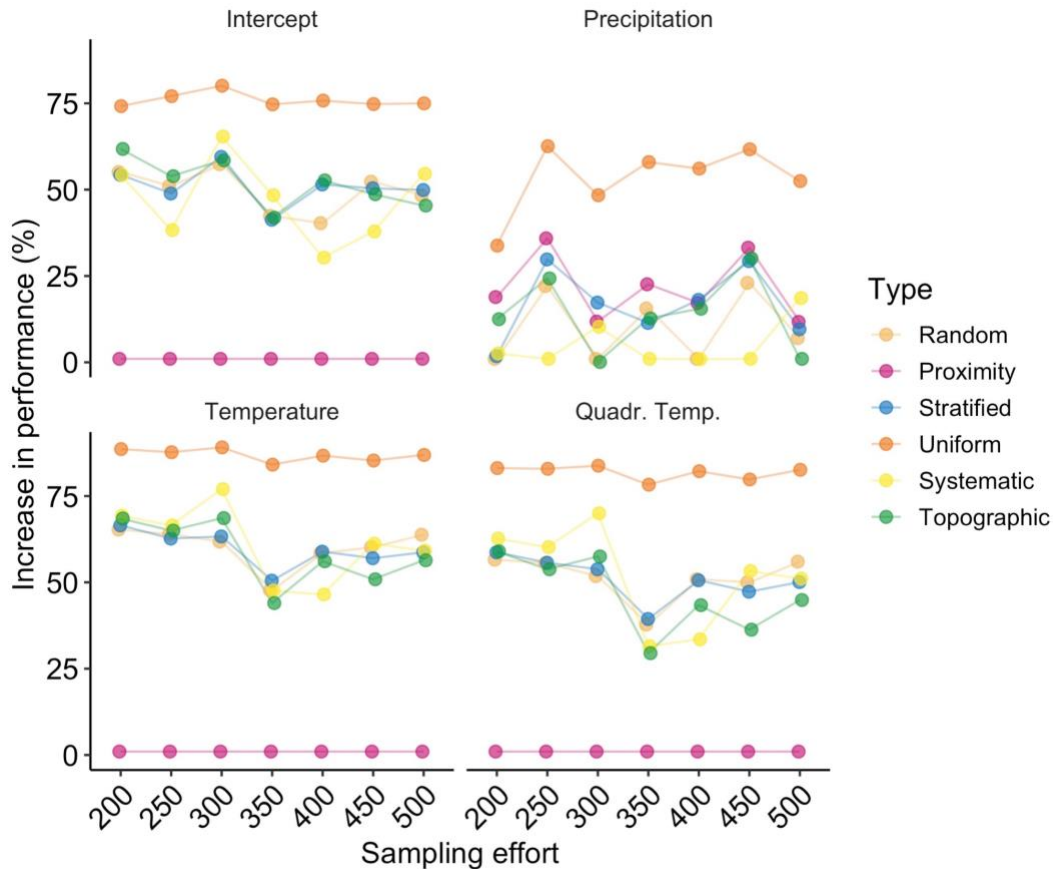
153 The R code of the simulations is available at: <https://github.com/ManueleBazzichetto/SamplingRespCurves>.

154 3. Results

155 As a general result, the MSE of the coefficients' estimators decreased with sampling effort, irrespective of the
156 sampling strategy, and converged towards a similar minimum value (Figures A4a, A5a). This is not surprising, as
157 it reflects the asymptotic unbiasedness and efficiency of the regression coefficients estimated by GLMs. For *D.*
158 *sperandii*, the most important discriminant factor in the performance (i.e., MSE) of the sampling strategies was
159 variance, while, for *D. tundrae*, it was bias (Figures A4b, A5b).

160 *Results for D. sperandii*

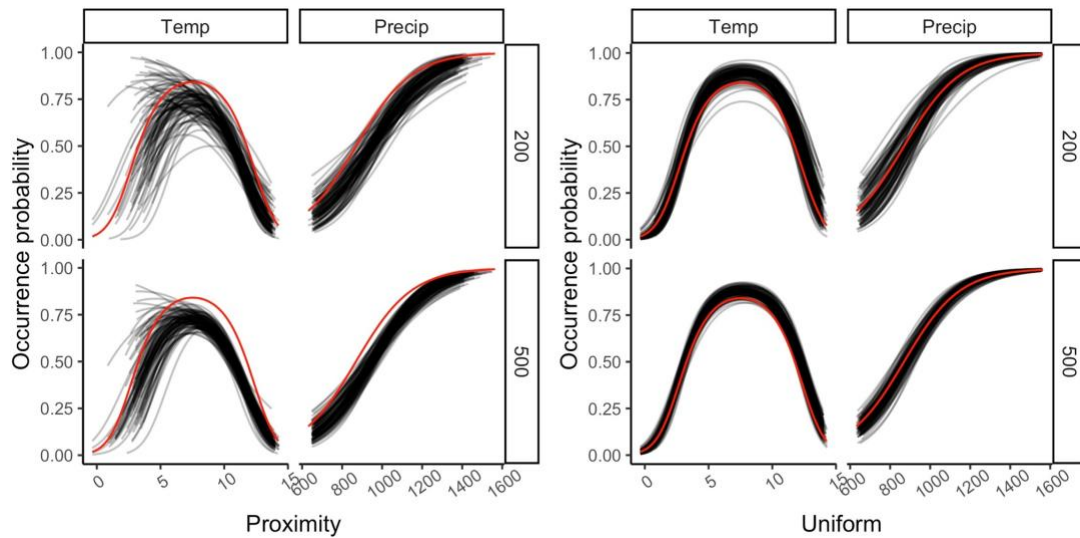
161 The proximity-to-road as a sampling design consistently provided the worst performance in terms of MSE at all
162 sample sizes (Figure 3). The only exception was for the estimation of the precipitation parameter, for which the
163 performance of the proximity-to-road approach was comparable to that of the other sampling designs. On the
164 contrary, the uniform sampling design within the environmental space scored the lowest MSE values at all
165 sampling efforts for all parameters. Specifically, the MSE of the uniform sampling was systematically 75% lower
166 than that of the proximity-to-road sampling for all coefficients but precipitation (Figure 3). The random, stratified,
167 systematic and topographic sampling designs performed similarly, with their MSE values generally included right
168 in between those of the proximity-to-road and uniform approach (Figures 3, A4a).



169

170 Figure 3 - Performance (expressed by percentage decrease in MSE values with respect to the worst performing approach) obtained for the
 171 different sampling strategies used to record the presence/absence of *D. sperandii*. Values are reported for increasing sampling effort. Quadr.
 172 Temp.: quadratic term for temperature.

173 All designs, except for the proximity-to-road approach, overestimated, in the long-run, the partial effect of
 174 precipitation (Figure A4b). In this regard, the largest bias (averaged across all simulations of increasing sampling
 175 effort) was associated with the uniform approach, which predicted a 111% increase in the odds of finding *D.*
 176 *sperandii* for each 100 mm increase in precipitation, in spite of a 105% increase predicted by the true model (see
 177 Figure A4c for the effect of the bias on the response curves). For the linear and quadratic temperature terms, the
 178 estimators derived from the uniform sampling within the environmental space were upwardly and downwardly
 179 biased, respectively (Figures 4, A4b). Concerning the variance, the uniform sampling within the environmental
 180 space provided the most efficient estimators for all coefficients, regardless of sample size (Figure A4b). This
 181 resulted in a more consistent shape of the response curve across simulations (Figures 4, A4c).

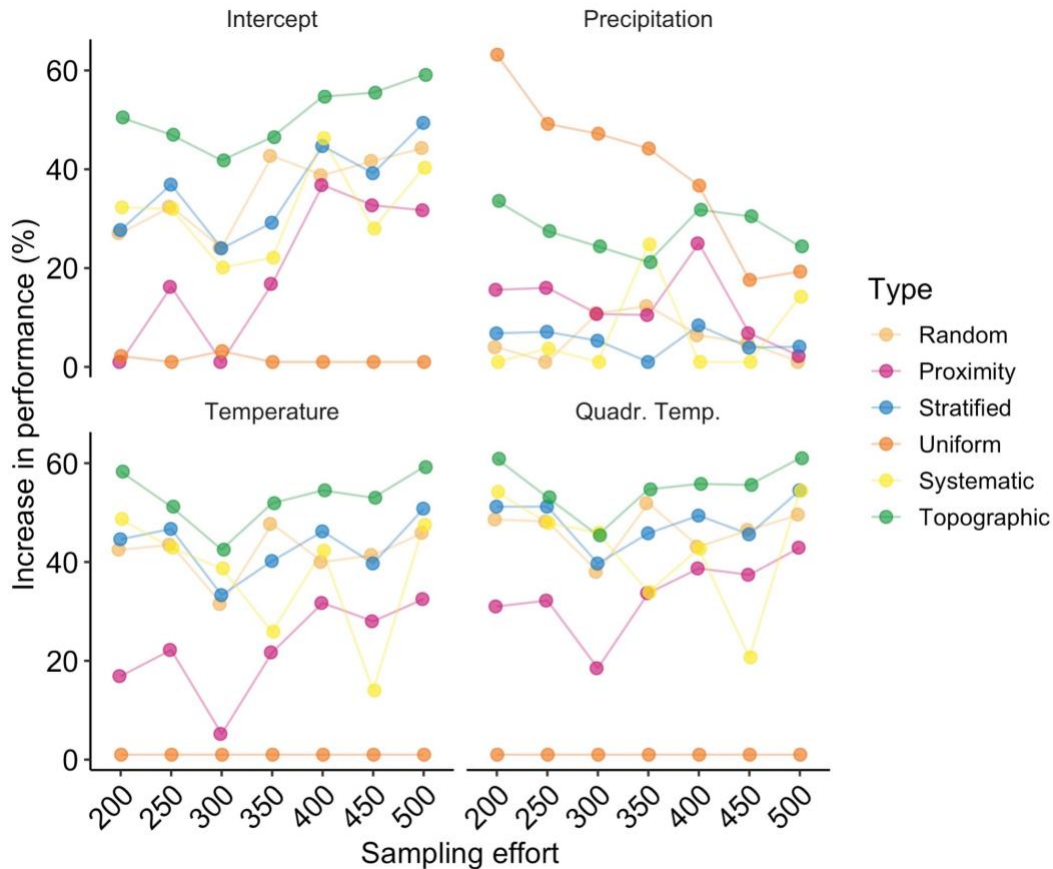


182

183 Figure 4 - Comparison between the response curves for *D. sperandii* as estimated by data collected through the proximity-to-road approach
 184 (left panel) and the uniform sampling of the environmental space (right panel). Each panel grid combines two sampling efforts ($N = 200$ and
 185 500) and the two predictors used in the models (i.e., temperature and precipitation). Red lines represent the true relationship between *D.*
 186 *sperandii* and the predictors. Comparisons between estimated and true response curves for all sampling strategies and efforts are reported in
 187 Appendix 4 (Figure A4c).

188 *Results for D. tundrae*

189 Regardless of sampling effort, the topographic approach scored the highest performance for all regression
 190 coefficients but precipitation (Figure 5). Also, except for precipitation, the topographic approach was always (i.e.,
 191 across sampling effort) followed by the systematic, stratified and random strategies (Figure 5). On the contrary,
 192 the uniform sampling within the environmental space showed the worst performance (i.e., highest MSE) for the
 193 intercept and the temperature (both linear and quadratic terms) at nearly all sampling efforts, whereas it scored
 194 best for the precipitation.



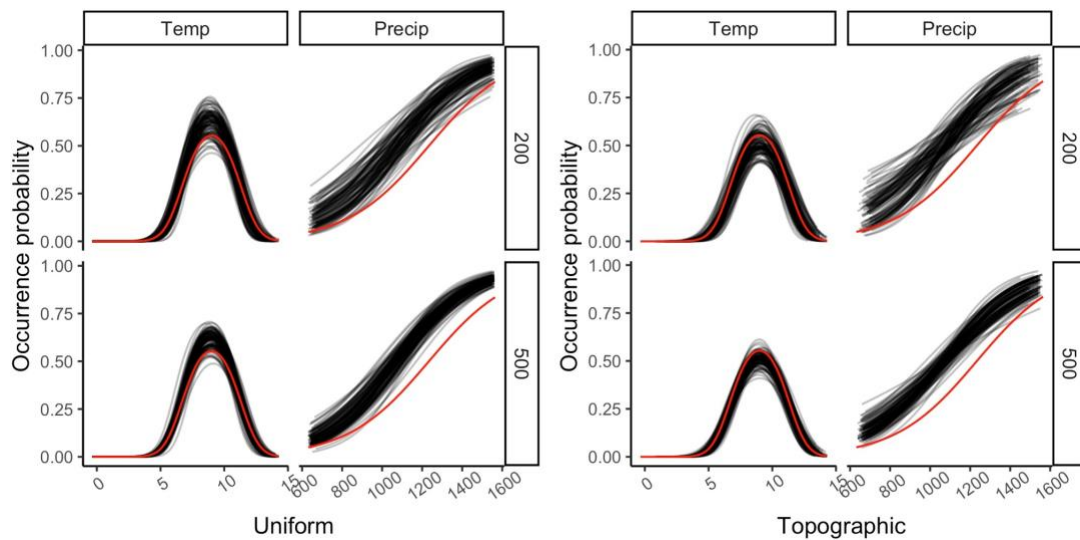
195

196
197
198

Figure 5 - Performance (expressed by percentage decrease in MSE values with respect to the worst performing approach) obtained for the different sampling strategies used to record the presence/absence of *D. tundrae*. Values are reported for increasing sampling effort. Quadr. Temp.: quadratic term for temperature.

199
200
201
202
203
204
205
206
207

The stratified, systematic, random, and uniform designs, in the long-run, overestimated the partial effect of precipitation, while the estimators derived from the topographic and proximity-to-road approach had low bias (Figures 6, A5b). As for *D. sperandii*, the largest bias was associated with the uniform approach, which predicted a 71% increase in the odds of finding *D. tundrae* for each 100 mm increase in precipitation, whereas the true model predicted a 63% increase (see Figure A5c for the effect of the bias on the response curves). All sampling designs, except for the topographic, underestimated the value of the intercept and provided upwardly biased estimators of the linear term of the temperature and downwardly biased estimators of the quadratic term of the temperature (Figure A5b). Concerning variance, the uniform sampling within the environmental space had the lowest variance for precipitation, while all strategies showed comparable efficiency.



208

209 Figure 6 - Comparison between the response curves for *D. tundrae* as estimated by data collected through the uniform sampling of the
 210 environmental space (left panel) and the topographic approach (right panel). Each panel grid combines two sampling efforts ($N = 200$ and
 211 500) and the two predictors used in the models (i.e., temperature and precipitation). Red lines represent the true relationship between *D.*
 212 *tundrae* and the predictors. Comparisons between estimated and true response curves for all sampling strategies and efforts are reported in
 213 Appendix 5 (Figure A5c).

214 4. Discussion

215 By creating virtual species with different thermal tolerances, and, as a result, different distribution extents across
 216 the Abruzzo (wide for *D. sperandii* and narrow for *D. tundrae*), we tested the impact of data sampling on the
 217 accuracy and precision of species response curves estimated by parametric SDMs. Overall, there seems to be no
 218 'silver bullet' strategy, i.e., a unique sampling approach with optimal performances across species with wide vs
 219 narrow distributions. This suggests that the sampling of presence/absence data should be planned on a case-by-
 220 case basis, i.e., according to the ecological characteristics of the species (span of the niche breadth and distribution
 221 extent) and the environmental heterogeneity of the study area (Chefaoui et al. 2011). We also found that collecting
 222 more data (increasing N) alleviates the impact of the sampling strategy on the variance, and MSE, of the
 223 coefficients, thereby confirming results from previous studies (Chefaoui et al. 2011; Tessarolo et al. 2014; Gábor
 224 et al. 2020). This suggests that, although exhaustive sampling campaigns are time- and cost-consuming, larger
 225 sample sizes successfully improve the estimation of species response curves irrespective of the sampling strategy
 226 used.

227 For generalist species like *D. sperandii*, the uniform sampling strategy within the environmental space seems to
 228 be the best, as well as the most efficient option (i.e., most effective at the lowest sample sizes). Intuitively, species
 229 with low environmental specialisation and wide geographic ranges are better modelled if data are regularly
 230 collected along environmental gradients. Uniformly sampling the environmental space is the best way to achieve
 231 that: data are collected at (generally) spatially aggregated, but environmentally heterogeneous geographic
 232 locations (Figure A3 in Appendix 3; Varela et al. 2014). In this regard, the uniform sampling of the environmental
 233 space was already argued as a suitable strategy for reducing the effect of sampling bias (Varela et al. 2014) or
 234 designing cost-effective, yet highly informative, surveys for species distribution modelling (Hattab et al. 2017).
 235 A key advantage of the uniform approach is certainly the low variance of the estimated coefficients (Figure A4b
 236 in Appendix 4). In this respect, we observed that the correlation between temperature and precipitation in the
 237 datasets generated by the uniform approach was, on average, lower than that associated with other sampling
 238 strategies (average Pearson correlation coefficient computed across sampling efforts: uniform -0.46, proximity
 239 and topographic -0.60, others -0.65), which may partly explain the higher precision of the coefficients estimated
 240 through the uniform approach. On the other hand, the proximity-to-road approach exhibited the worst performance
 241 in terms of MSE. One possible explanation is that, as the probability for *D. sperandii* to occur at a given location
 242 increased with increasing distance from roads, the datasets generated by the proximity approach likely included

243 sub-optimal information on the favourable environmental conditions for that species. For this reason, we feel like
244 warning ecologists against using data gathered through the proximity-to-road strategy for SDMs, except for
245 specific circumstances (e.g., MIREN protocol for species responding to anthropogenic disturbances, see Haider
246 et al. 2022). Indeed, using data collected close to roadsides increases the chance of estimating erroneous species'
247 response curves (Figures 4, A4c), especially when road networks have low environmental coverage (Tessarolo et
248 al. 2014). For instance, in mountain systems, the density of the road network decreases drastically towards higher
249 elevations where accessibility can be a very important constraint. These kinds of side effects should be considered
250 very carefully when calibrating SDMs with empirical data sampled for a completely different purpose.

251 For specialist species (e.g., *D. tundrae*), all sampling designs appear to perform similarly, but the topographic
252 approach scored as best for all regression parameters but precipitation. The similar performance of the sampling
253 strategies might be due to species with a low relative occurrence area (i.e., proportion of area occupied by the
254 species over the whole studied area) being generally easier to model (Lobo 2008). On the other hand, the good
255 performance of the topographic approach could be related to the fact that (i) *D. tundrae* has its thermal optimum
256 close to the mean temperature in the area of interest (i.e., 9 °C; Figure 1) and (ii) by randomly selecting sampling
257 units among those with high topographic heterogeneity (see Box 2), the topographic design mimics the uniform
258 approach, but is likely to over-sample the most common environmental conditions in the study area (e.g., average
259 temperature). As a result, in our study, data collected through the topographic sampling were best for capturing
260 the narrow shape of the response curve of *D. tundrae* along the temperature gradient. Had *D. tundrae*'s thermal
261 optimum lied far from the mean temperature of the study area, the topographic approach would have probably not
262 resulted in such an observed high performance. In this respect, other sampling strategies with similar performances
263 to the topographic approach, e.g., the stratified approach, may be less sensitive to the position of the species'
264 optima. It is worth noticing that, even for *D. tundrae*, the uniform approach provided the best estimators for
265 precipitation, which suggests that as long as a species has a wide tolerance to an environmental driver, this
266 sampling design provides a good estimation of the response curve. As an alternative to the tested approaches,
267 adaptive-sampling strategies could also provide a viable means for modelling species with narrow distributions
268 (Jeliazkov et al. 2022). One example are SDM-guided sampling designs: SDMs are fitted on species and
269 environmental data collected through preliminary sampling. The obtained predictions are then used to identify
270 areas to collect new data on the target species (Chiffard et al. 2020).

271 To sum up, when the aim is to model a widespread and generalist species, choosing an appropriate sampling
272 approach (here: uniformly sampling the environmental space) could represent the most efficient strategy, as it
273 allows obtaining accurate response curves while sparing on resources that would be otherwise allocated to field
274 sampling. As the species' tolerance to environmental drivers shrinks, the advantage of selecting an adequate
275 sampling design vanishes, as all approaches seem to have comparable performances. More specifically, uniformly
276 sampling the environmental space may no longer provide optimal results, while other, equally good approaches
277 (e.g., stratified design), could be chosen.

278 **5. Authors' contribution**

279 MB conceptualised the study with MGS and VB; MB analysed the data with inputs from VB and JL; MB and
280 MGS wrote the first draft of the manuscript. All authors discussed the results, contributed to the improvement of
281 the first manuscript version, and gave their final approval for publication.

282 **6. Acknowledgments**

283 All authors are grateful to Dr. Joaquin Hortal (Department of Biogeography and Global Change of the Museo
284 Nacional de Ciencias Naturales - CSIC), who provided a friendly revision of the manuscript.

285 **7. Data availability statement**

286 The results presented in this manuscript are the product of simulated data. The R code of the simulations is
287 available at: <https://github.com/ManueleBazzichetto/SamplingRespCurves>.

288 **References**

- 289 Baker, D. J., Maclean, I. M., Goodall, M., & Gaston, K. J. (2022). Correlations between spatial sampling biases
290 and environmental niches affect species distribution models. *Global Ecology and Biogeography*, 31(6), 1038-
291 1050.
- 292 Bazzichetto, M., Massol, F., Carboni, M., Lenoir, J., Lembrechts, J. J., Joly, R., & Renault, D. (2021). Once upon
293 a time in the far south: Influence of local drivers and functional traits on plant invasion in the harsh sub-Antarctic
294 islands. *Journal of Vegetation Science*, 32(4), e13057.
- 295 Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on
296 modeling species' geographic distributions. *Ecological Informatics*, 19, 10-15.
- 297 Bolker, B. M. (2008). *Ecological models and data in R*. Princeton University Press.
- 298 Chefaoui, R. M., Lobo, J. M., & Hortal, J. (2011). Effects of species' traits and data characteristics on distribution
299 models of threatened invertebrates. *Animal Biodiversity and Conservation*, 34(2), 229-247.
- 300 Chiffard, J., Marciau, C., Yoccoz, N., Mouillot, F., Duchateau, S., Nadeau, I., ... & Besnard, A. (2020). Adaptive
301 niche-based sampling to improve ability to find rare and elusive species: Simulations and field tests. *Methods in*
302 *Ecology and Evolution*, 11(8), 899-909.
- 303 Feldman, M. J., Imbeau, L., Marchand, P., Mazerolle, M. J., Darveau, M., & Fenton, N. J. (2021). Trends and
304 gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PloS*
305 *one*, 16(3), e0234587.
- 306 Fernandes, R. F., Scherrer, D., & Guisan, A. (2018). How much should one sample to accurately predict the
307 distribution of species assemblages? A virtual community approach. *Ecological Informatics*, 48, 125-134.
- 308 Gábor, L., Moudrý, V., Barták, V., & Lecours, V. (2020). How do species and data characteristics affect species
309 distribution models and when to use environmental filtering?. *International Journal of Geographical Information*
310 *Science*, 34(8), 1567-1584.
- 311 Gu, W., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on
312 wildlife-habitat models. *Biological conservation*, 116(2), 195-203.
- 313 Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with*
314 *applications in R*. Cambridge University Press.
- 315 Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological*
316 *modelling*, 135(2-3), 147-186.
- 317 Haider, S., Lembrechts, J. J., McDougall, K., Pauchard, A., Alexander, J. M., Barros, A., ... & Seipel, T. (2022).
318 Think globally, measure locally: The MIREN standardized protocol for monitoring plant species distributions
319 along elevation gradients. *Ecology and evolution*, 12(2), e8590.
- 320 Hattab, T., Garzón-López, C. X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., ... & Lenoir, J. (2017). A unified
321 framework to model the potential and realized distributions of invasive species within the invaded range. *Diversity*
322 *and Distributions*, 23(7), 806-819.
- 323 Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability
324 models to predict species presences. *Ecological modelling*, 199(2), 142-152.

- 325 Hirzel, A., & Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling.
326 *Ecological modelling*, 157(2-3), 331-341.
- 327 Hollister, J.W. (2021). elevatr: Access Elevation Data from Various APIs. R package version 0.4.1.
328 <https://CRAN.R-project.org/package=elevatr/>
- 329 Jackson, S.T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental
330 changes of the late Quaternary. *Paleobiology*, 26(S4), 194-220.
- 331 Jeliaskov, A., Gavish, Y., Marsh, C. J., Geschke, J., Brummitt, N., Rocchini, D., ... & Henle, K. (2022). Sampling
332 and modelling rare species: Conceptual guidelines for the neglected majority. *Global change biology*, 28(12),
333 3754-3777.
- 334 Jiménez-Valverde, A. (2021). Prevalence affects the evaluation of discrimination capacity in presence-absence
335 species distribution models. *Biodiversity and Conservation*, 30(5), 1331-1340.
- 336 Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced
337 by bioclimatic models. *Ecological Applications*, 14(2), 401-413.
- 338 Kadmon, R., Farber, O., & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic
339 envelope models. *Ecological Applications*, 13(3), 853-867.
- 340 Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., ... & Kessler, M. (2017).
341 Climatologies at high resolution for the earth's land surface areas. *Scientific data*, 4(1), 1-20.
- 342 Lájér, K. (2007). Statistical tests as inappropriate tools for data analysis performed on non-random samples of
343 plant communities. *Folia Geobotanica*, 42(2), 115-122.
- 344 Leandro, C., Jay-Robert, P., Méridet, B., Houard, X., & Renner, I. W. (2020). Is my sdm good enough? insights
345 from a citizen science dataset in a point process modeling framework. *Ecological Modelling*, 438, 109283.
- 346 Lobo, J. M., Jiménez-Valverde, A., and Hortal, J. (2010). The uncertain nature of absences and their importance
347 in species distribution modelling. *Ecography*, 33(1):103–114.
- 348 Lobo, J. M. (2008). More complex distribution models or more representative data?. *Biodiversity informatics*, 5.
- 349 Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of
350 predictive distribution models. *Global ecology and Biogeography*, 17(2), 145-151.
- 351 McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. CRC Press.
- 352 McPherson, J., & Jetz, W. (2007). Effects of species' ecology on the accuracy of distribution models. *Ecography*,
353 30(1), 135-151.
- 354 Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular
355 attention to species distribution models. *Progress in physical geography*, 34(1), 3-22.
- 356 Padgham, M., Lovelace, R., Salmon, M., & Rudis, B. (2017). osmdata. *Journal of Open Source Software*, 2(14).
- 357 Roleček, J., Chytrý, M., Hájek, M., Lvončík, S., & Tichý, L. (2007). Sampling design in large-scale vegetation
358 studies: Do not sacrifice ecological thinking to statistical purism!. *Folia Geobotanica*, 42(2), 199-208.
- 359 Tessarolo, G., Lobo, J. M., Rangel, T. F., & Hortal, J. (2021). High uncertainty in the effects of data characteristics
360 on the performance of species distribution models. *Ecological Indicators*, 121, 107147.

- 361 Tassarolo, G., Rangel, T. F., Araújo, M. B., & Hortal, J. (2014). Uncertainty associated with survey design in
362 Species Distribution Models. *Diversity and Distributions*, 20(11), 1258-1269.
- 363 Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce
364 the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11), 1084-1091.
- 365 Wasof, S., Lenoir, J., Aarrestad, P. A., Alsos, I. G., Armbruster, W. S., Austrheim, G., ... & Decocq, G. (2015).
366 Disjunct populations of European vascular plant species keep the same climatic niches. *Global Ecology and*
367 *Biogeography*, 24(12), 1401-1412.