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

Bazzichetto Manuele[1*], Lenoir Jonathan[2], Da Re Daniele[3], Tordoni Enrico[4], Rocchini Duccio[5,6], Malavasi Marco[6,7], Barták Vojtech[6↓] & Sperandii Marta Gaia[1↓]

[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--Suchdol, Czech Republic; [7] Department of Chemistry, Physics, Mathematics and Natural Sciences, University of Sassari, Sassari, Italy.

[*] Corresponding author: manuele.bazzichetto@gmail.com

[↓]Joint senior authors.

Abstract

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

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

Significance statement

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

1. Introduction

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

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

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

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):

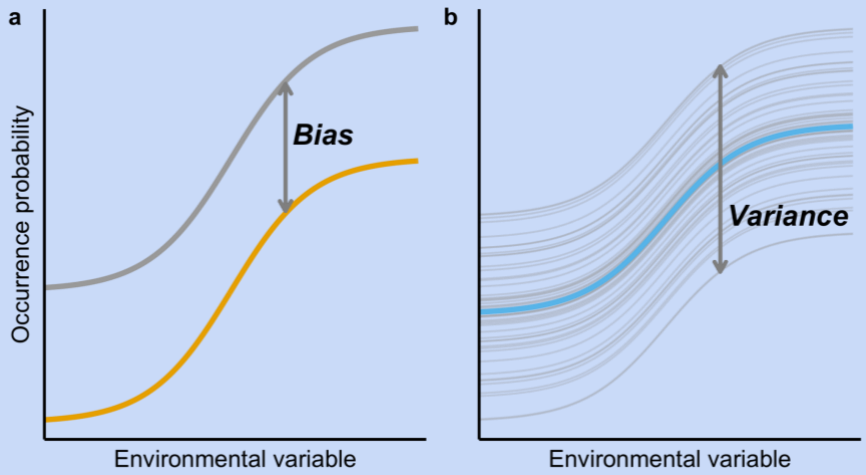
$$Bias = \mathrm{E}[(\widehat{\theta} - \theta)]$$

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):

$$Variance = \mathrm{E}[(\widehat{\theta} - E[\widehat{\theta}])^2]$$

***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:

$$MSE = \mathrm{E}[(\widehat{\theta} - \theta)^2]$$



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.

2. Materials and Methods

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

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

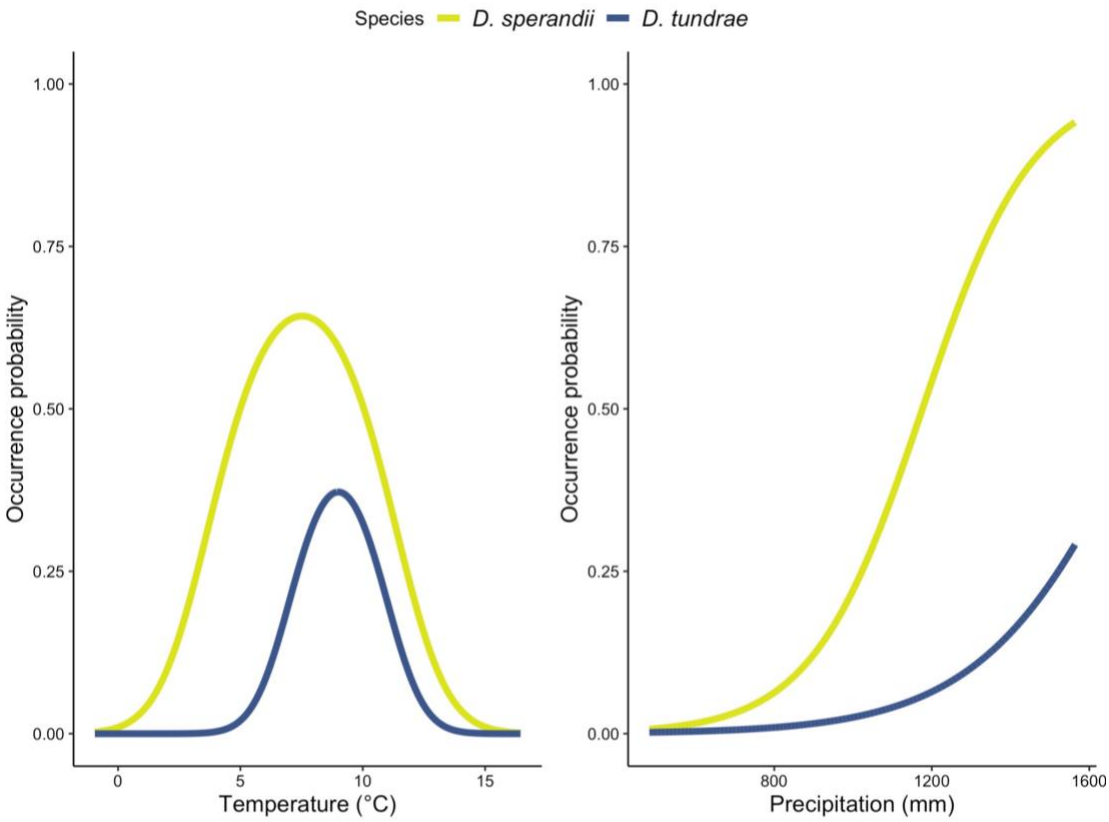


Figure 1 - Simulated response curves of *Dianthus sperandii* (in lime) and *D. tundrae* (in blue) along the temperature (left panel) and precipitation (right panel) gradients.

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:

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

where logit(·) is the natural logarithm of the odds $p_i/(1-p_i)$, 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: -4.614 (α), 0.007 (β_{pr}), 1.384 (β_{tm}), and -0.092 (β_{tmq}). For *D. tundrae*, they were set to: -17.528 (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).

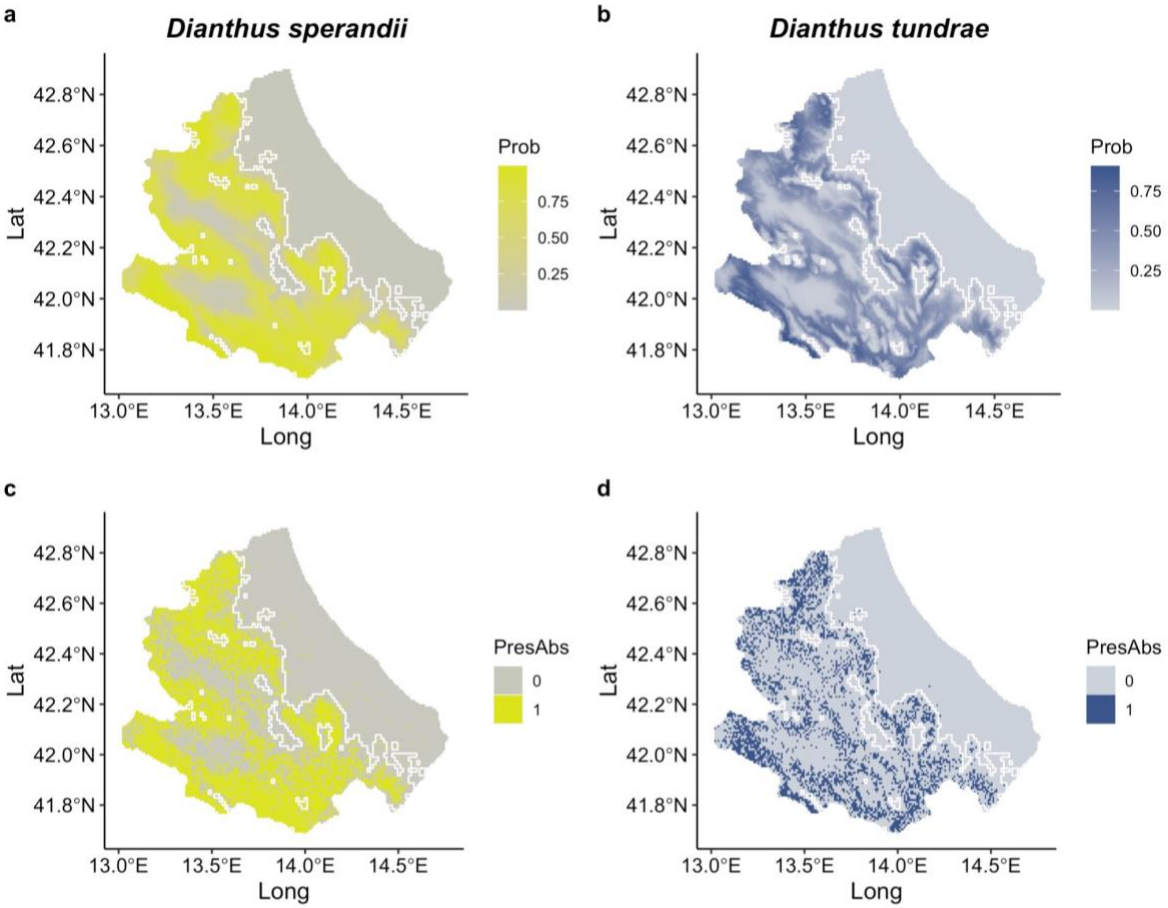


Figure 2 - Simulated occurrence probability and presence/absence data of *Dianthus sperandii* (a, c) and *D. tundrae* (b, d). The white line in the plots delineates the area of interest for the study (i.e., all lands approx. between 500 and 1800 m a.s.l.).

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

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*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).

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

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

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

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

3. Results

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

Results for D. sperandii

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

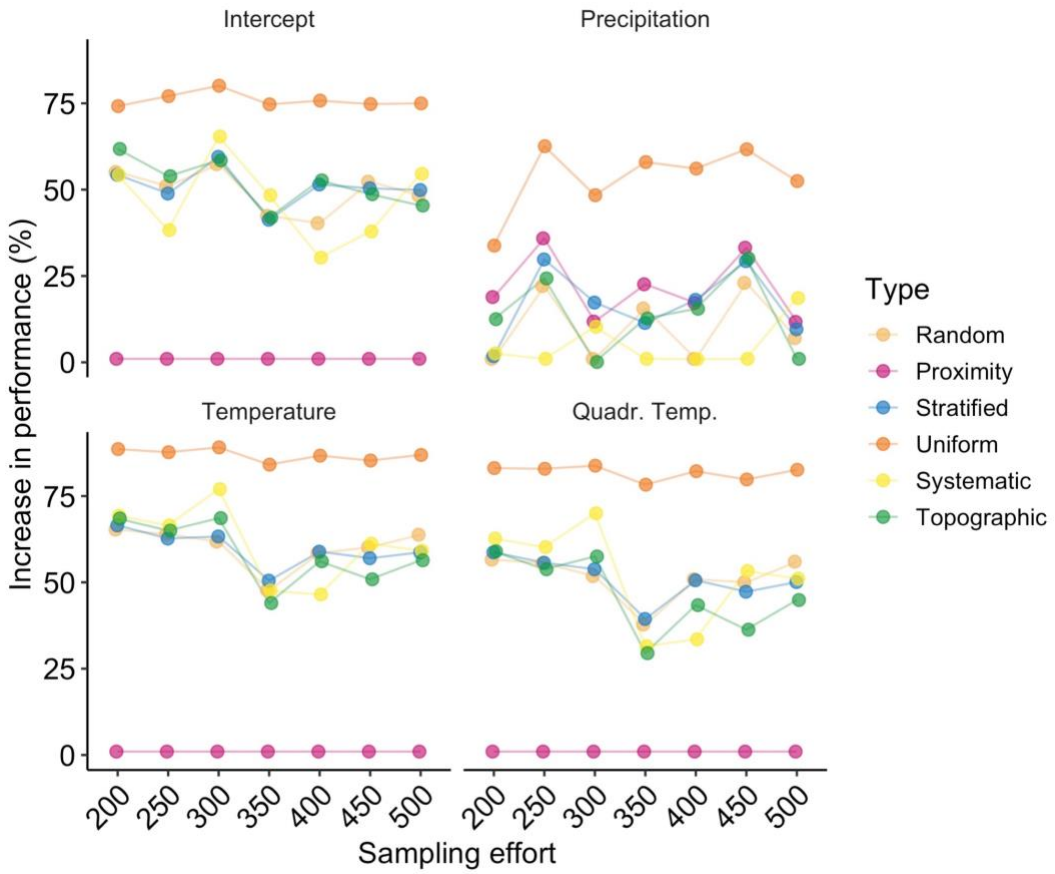


Figure 3 - Performance (expressed by percentage decrease in MSE values with respect to the worst performing approach) obtained for the different sampling strategies used to record the presence/absence of *D. sperandii*. Values are reported for increasing sampling effort. Quadr. Temp.: quadratic term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

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

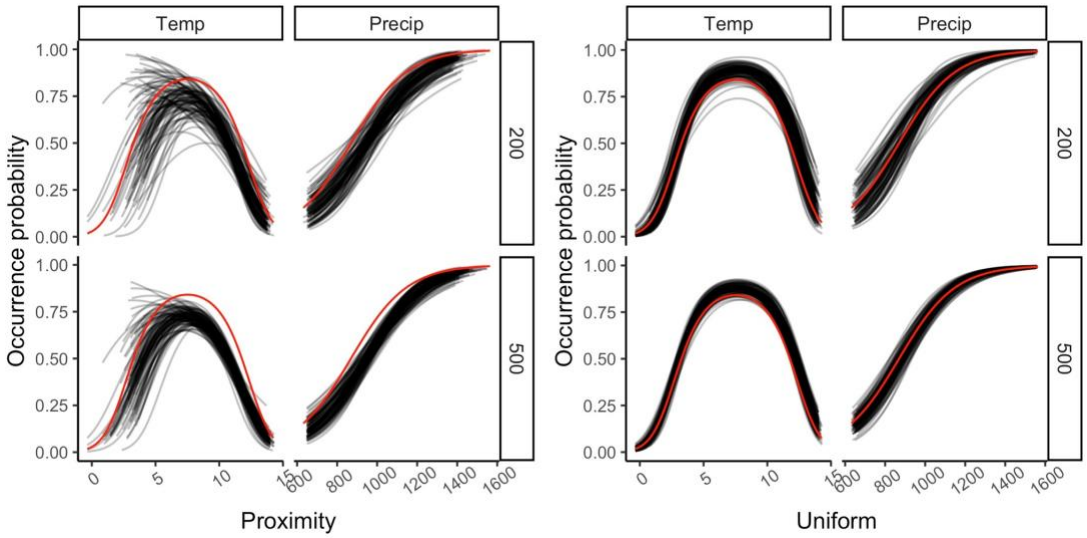


Figure 4 - Comparison between the response curves for *D. sperandii* as estimated by data collected through the proximity-to-road approach (left panel) and the uniform sampling of the environmental space (right panel). Each panel grid combines two sampling efforts (N = 200 and 500) and the two predictors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

Results for D. tundrae

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

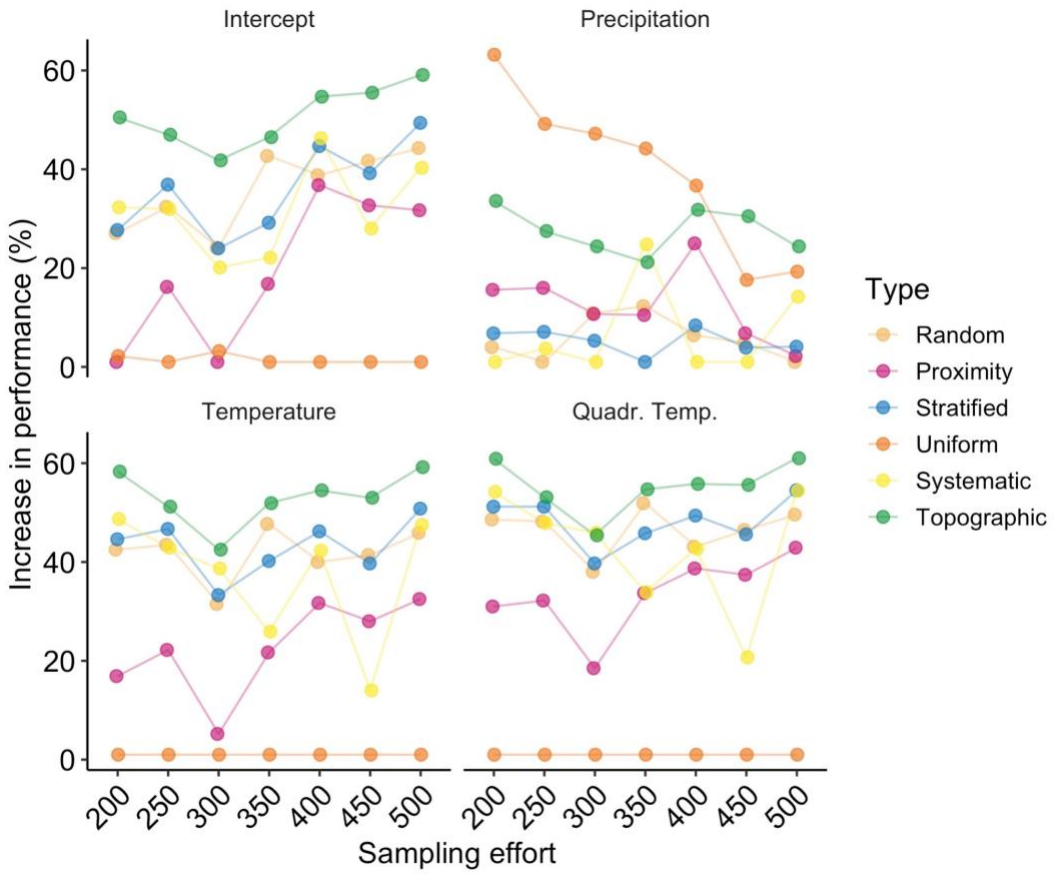


Figure 5 - Performance (expressed by percentage decrease in MSE values with respect to the worst performing approach) obtained for the different sampling strategies used to record the presence/absence of *D. tundrae*. Values are reported for increasing sampling effort. Quadr. Temp.: quadratic term for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

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.

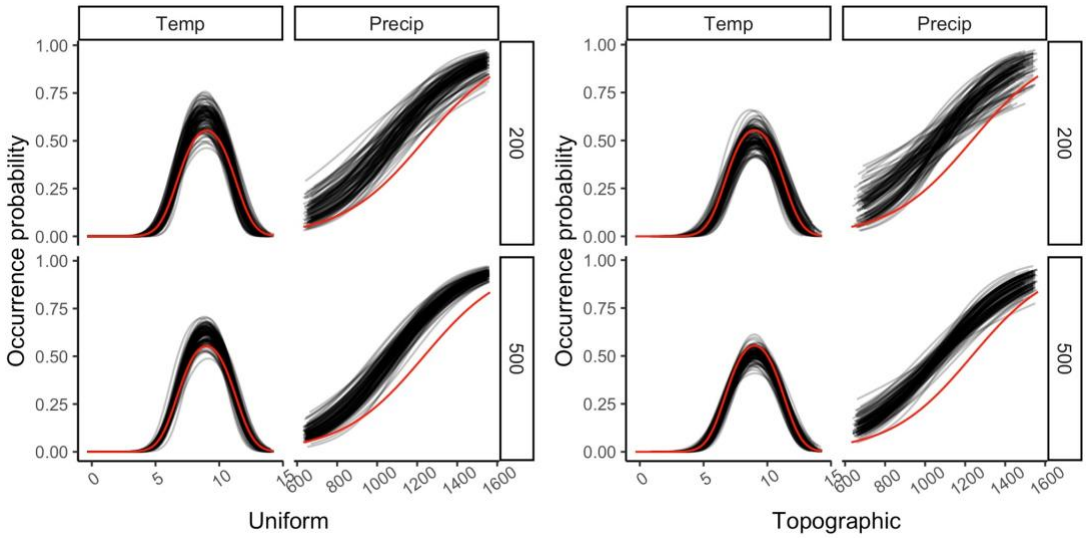


Figure 6 - Comparison between the response curves for *D. tundrae* as estimated by data collected through the uniform sampling of the environmental space (left panel) and the topographic approach (right panel). Each panel grid combines two sampling efforts (N = 200 and 500) and the two predictors used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

4. Discussion

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

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

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

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

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

5. Authors' contribution

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

6. Acknowledgments

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

7. Data availability statement

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

References

Baker, D. J., Maclean, I. M., Goodall, M., & Gaston, K. J. (2022). Correlations between spatial sampling biases and environmental niches affect species distribution models. *Global Ecology and Biogeography*, 31(6), 1038-1050.

Bazzichetto, M., Massol, F., Carboni, M., Lenoir, J., Lembrechts, J. J., Joly, R., & Renault, D. (2021). Once upon a time in the far south: Influence of local drivers and functional traits on plant invasion in the harsh sub-Antarctic islands. *Journal of Vegetation Science*, 32(4), e13057.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19, 10-15.

Bolker, B. M. (2008). Ecological models and data in R. Princeton University Press.

Chefaoui, R. M., Lobo, J. M., & Hortal, J. (2011). Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, 34(2), 229-247.

Chiffard, J., Marciau, C., Yoccoz, N., Mouillot, F., Duchateau, S., Nadeau, I., ... & Besnard, A. (2020). Adaptive niche-based sampling to improve ability to find rare and elusive species: Simulations and field tests. *Methods in Ecology and Evolution*, 11(8), 899-909.

Feldman, M. J., Imbeau, L., Marchand, P., Mazerolle, M. J., Darveau, M., & Fenton, N. J. (2021). Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PloS one*, 16(3), e0234587.

Fernandes, R. F., Scherrer, D., & Guisan, A. (2018). How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach. *Ecological Informatics*, 48, 125-134.

Gábor, L., Moudrý, V., Barták, V., & Lecours, V. (2020). How do species and data characteristics affect species distribution models and when to use environmental filtering?. *International Journal of Geographical Information Science*, 34(8), 1567-1584.

Gu, W., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological conservation*, 116(2), 195-203.

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

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135(2-3), 147-186.

Haider, S., Lembrechts, J. J., McDougall, K., Pauchard, A., Alexander, J. M., Barros, A., ... & Seipel, T. (2022). Think globally, measure locally: The MIREN standardized protocol for monitoring plant species distributions along elevation gradients. *Ecology and evolution*, 12(2), e8590.

Hattab, T., Garzón-López, C. X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., ... & Lenoir, J. (2017). A unified framework to model the potential and realized distributions of invasive species within the invaded range. *Diversity and Distributions*, 23(7), 806-819.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological modelling*, 199(2), 142-152.

Hirzel, A., & Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological modelling*, 157(2-3), 331-341.

Hollister, J.W. (2021). elevatr: Access Elevation Data from Various APIs. R package version 0.4.1. https://CRAN.R-project.org/package=elevatr/

Jackson, S.T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, *26*(S4), 194-220.

Jeliazkov, A., Gavish, Y., Marsh, C. J., Geschke, J., Brummitt, N., Rocchini, D., ... & Henle, K. (2022). Sampling and modelling rare species: Conceptual guidelines for the neglected majority. *Global change biology*, 28(12), 3754-3777.

Jiménez-Valverde, A. (2021). Prevalence affects the evaluation of discrimination capacity in presence-absence species distribution models. *Biodiversity and Conservation*, 30(5), 1331-1340.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14(2), 401-413.

Kadmon, R., Farber, O., & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, 13(3), 853-867.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., ... & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific data*, 4(1), 1-20.

Lájer, K. (2007). Statistical tests as inappropriate tools for data analysis performed on non-random samples of plant communities. *Folia Geobotanica*, *42*(2), 115-122.

Leandro, C., Jay-Robert, P., Mériguet, B., Houard, X., & Renner, I. W. (2020). Is my sdm good enough? insights from a citizen science dataset in a point process modeling framework. *Ecological Modelling*, *438*, 109283.

Lobo, J. M., Jiménez-Valverde, A., and Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114.

Lobo, J. M. (2008). More complex distribution models or more representative data?. *Biodiversity informatics*, 5.

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

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models. CRC Press.

McPherson, J., & Jetz, W. (2007). Effects of species' ecology on the accuracy of distribution models. *Ecography*, 30(1), 135-151.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in physical geography*, 34(1), 3-22.

Padgham, M., Lovelace, R., Salmon, M., & Rudis, B. (2017). osmdata. *Journal of Open Source Software*, 2(14).

Roleček, J., Chytrý, M., Hájek, M., Lvončík, S., & Tichý, L. (2007). Sampling design in large-scale vegetation studies: Do not sacrifice ecological thinking to statistical purism!. *Folia Geobotanica*, *42*(2), 199-208.

Tessarolo, G., Lobo, J. M., Rangel, T. F., & Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, 121, 107147.

Tessarolo, G., Rangel, T. F., Araújo, M. B., & Hortal, J. (2014). Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions*, 20(11), 1258-1269.

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11), 1084-1091.

Wasof, S., Lenoir, J., Aarrestad, P. A., Alsos, I. G., Armbruster, W. S., Austrheim, G., ... & Decocq, G. (2015). Disjunct populations of European vascular plant species keep the same climatic niches. *Global Ecology and Biogeography*, 24(12), 1401-1412.