Sampling strategy matters to accurately estimate response curves' parameters in species distribution models

Bazzichetto Manuele[1,2], Lenoir Jonathan[3*], Da Re Daniele[4], Tordoni Enrico[5], Rocchini Duccio[6,1], Malavasi Marco[7,1], Barták Vojtech[1⇂] & Sperandii Marta Gaia[8,2⇂]

[1] Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha--Suchdol, Czech Republic;

[2] Centro de Investigaciones sobre Desertificación (CSIC-UV-GV), Valencia, Spain;

[3] UMR CNRS 7058 « Ecologie et Dynamique des Systèmes Anthropisés » (EDYSAN), Université de Picardie Jules Verne, 1 rue des Louvels, 80000 Amiens, France;

[4] Georges Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Place Louis Pasteur 3, 1348 Louvain-la-Neuve, Belgium;

[5] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia;

[6] BIOME Lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Bologna, Italy;

[7] Department of Chemistry, Physics, Mathematics and Natural Sciences, University of Sassari, Sassari, Italy;

[8] Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic.

[*] Corresponding author: jonathan.lenoir@u-picardie.fr

[⇂]Joint senior authors.

Running title: sampling strategies and species response curves

Abstract

Aim: Assessing how different sampling strategies affect the accuracy and precision of species response curves estimated by parametric Species Distribution Models.

Major taxa studied: Virtual plant species.

Location: Abruzzo (Italy).

Time period: Timeless (simulated data).

Methods: We simulated the occurrence of two virtual species with different ecology (generalist *vs* specialist) and distribution extent. We sampled their occurrence following different sampling strategies: random, stratified, systematic, topographic, uniform within the environmental space (hereafter, uniform), and close to roads. For each sampling design and species, we ran 500 simulations at increasing sampling efforts (total: 42,000 replicates). For each replicate, we fitted a binomial generalised linear model, extracted model coefficients for precipitation and temperature, and compared them with true coefficients from the known species' equation. We evaluated the quality of the estimated response curves by computing bias, variance, and root mean squared error. Additionally, we i) assessed the impact of missing covariates on the performance of the sampling approaches and ii) evaluated the effect of incompletely sampling the environmental space on the uniform approach.

Results: For the generalist species, we found the lowest root mean squared error when uniformly sampling the environmental space, while sampling occurrence data close to roads provided the worst performance. For the specialist species, all sampling designs showed comparable outcomes. Excluding important predictors similarly affected all sampling strategies. Sampling limited portions of the environmental space reduced the performance of the uniform approach, regardless of the portion surveyed.

Main conclusions: 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. Overall, uniformly sampling the environmental space seems more efficient for species with wide environmental tolerances. The advantage of seeking the most appropriate sampling strategy vanishes when modelling species with narrow realised niches.

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

Significance statement

The choice of the most appropriate strategy to sample presence/absence data for 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 for deriving species' response curves along chosen environmental gradients, which define how species respond to the environmental conditions they experience. 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, Lobo, Rangel, Hortal, 2021). Species presence and absence data, both ideally collected in-situ, would be modelled as a function of environmental variables sampled at the same geographical locations where the species was recorded. However, absences are very often 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 primarily opportunistic and sampled for different purposes (Hirzel & Guisan 2002; Gábor, Moudrý, Barták, Lecours, 2020). Examples include opportunistic data from museum collections or herbaria (Newbold 2010), citizen science (Leandro, Jay-Robert, Mériguet, Houard, Renner, 2020; Feldman et al., 2021), vegetation surveys (Bazzichetto et al., 2021); or a combination of these (Wasof et al., 2015). Using data not collected explicitly for species distribution modelling can be an issue, as the sampling strategy determines the quality of the species response curves estimated by SDMs (Beck, Böller, Erhardt, Schwanghart, 2014; Baker, Maclean, Goodall, Gaston, 2022).

In principle, species distribution data should be collected or sampled in a way that helps 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 to describe plant communities (co-occurrence data). Not initially collected to model a single species distribution, such data 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. Still, almost all 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, Farber, Danin, 2003; Hirzel, Le Lay, Helfer, Randin, Guisan, 2006; Tessarolo, Rangel, Araújo, Hortal, 2014; Varela, Anderson, García-Valdés, Fernández-González, 2014; Guisan, Thuiller, Zimmermann, 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 to estimate the actual 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). Even worse, some accuracy metrics can score high in the case of poorly defined SDMs (Lobo, Jiménez-Valverde, Real, 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 to provide accurate estimates of the actual coefficients defining species response curves (Gu & Swihart 2004; Albert et al., 2010; Fernandes, Scherrer, Guisan, 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 root mean squared error (see Box 1 for definitions).

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

Bias: the 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):

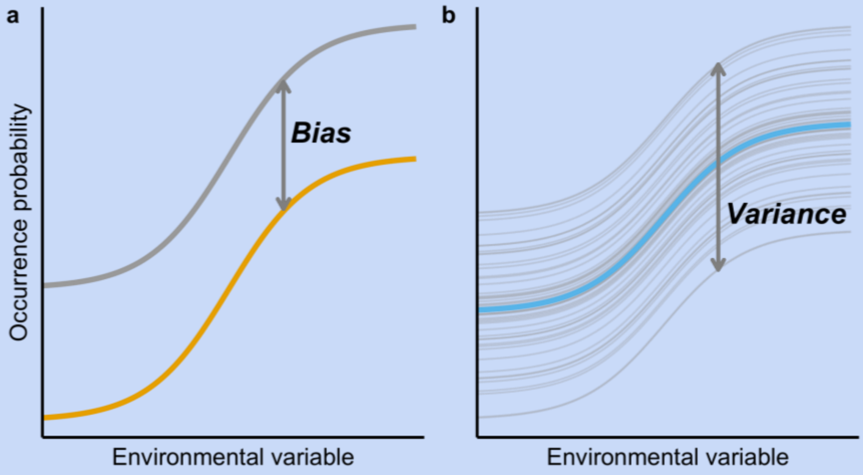
$$Bias = \mathrm{E}[(\hat{\theta} - \theta)]$$

Variance: expected squared difference between an estimator and its expected value (notice that the estimator’s expected value 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 in the long-run, Bolker 2008):

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

Root Mean Squared Error **(RMSE)**: square root of the Mean Squared Error (MSE), which is the 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 the quality of an estimator. We here use the RMSE as it is expressed on the same scale as the data:

$$RMSE = \sqrt{}\mathrm{E}[(\hat{\theta} - \theta)^2]$$



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 evenly spread oscillations around the true response curve.

2. Materials and Methods

2.1. Simulations of plant virtual species and their sampling

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

We focus on the Abruzzo region in Central Italy, which covers different climates and habitat types (see Figure S1.1, Appendix S1 in Supporting Information). 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 thus 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 more or less similar ecological preferences, but different thermal niche breadth and thus different distribution extents, we tested the effect of sampling strategy on SDMs for generalist *vs* specialist species.

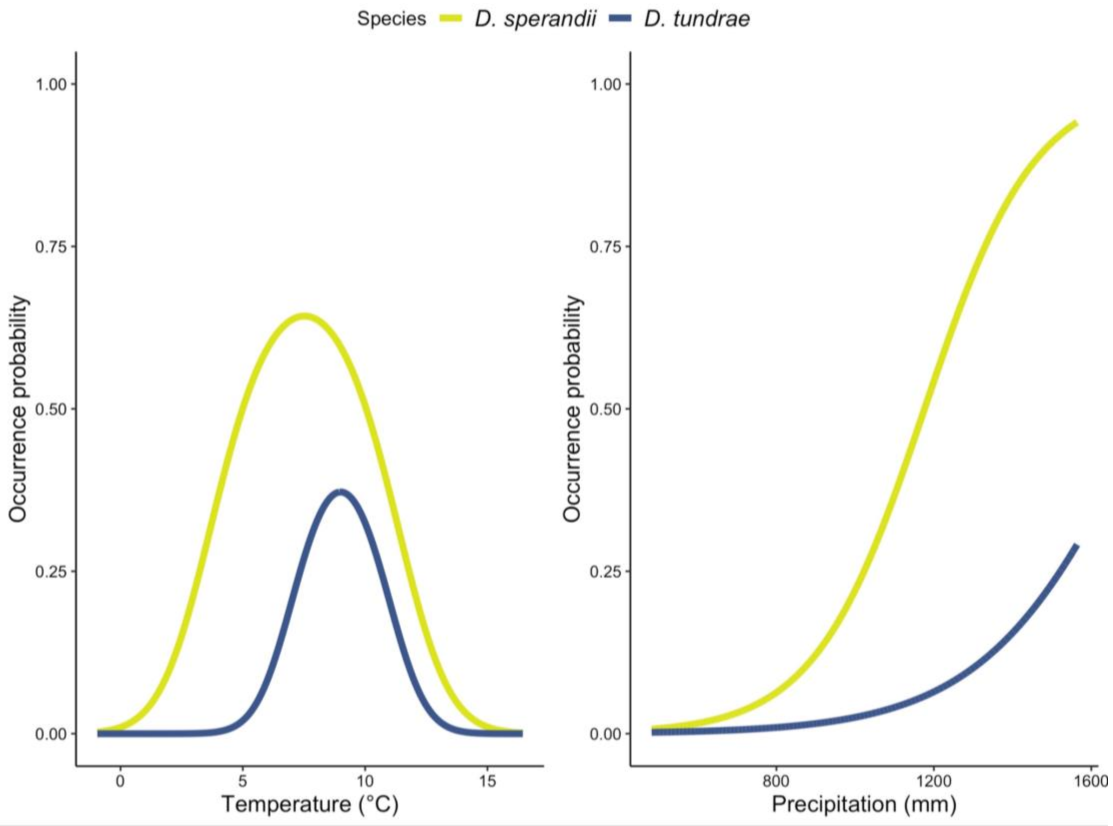


Figure 1. Simulated response curves of *Dianthus sperandii* (in lime) and *D. tundrae* (in blue) along the temperature (left panel) and precipitation (right panel) gradients.

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:

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

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: -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 a random realisation of a Bernoulli trial with probability *p* at each raster cell. The obtained presence/absence layers are reported in Figure 2 (see panels c and d). Note that, by sampling from the Bernoulli distribution, we avoided selecting a fixed threshold to generate the layers of the species presence/absence.

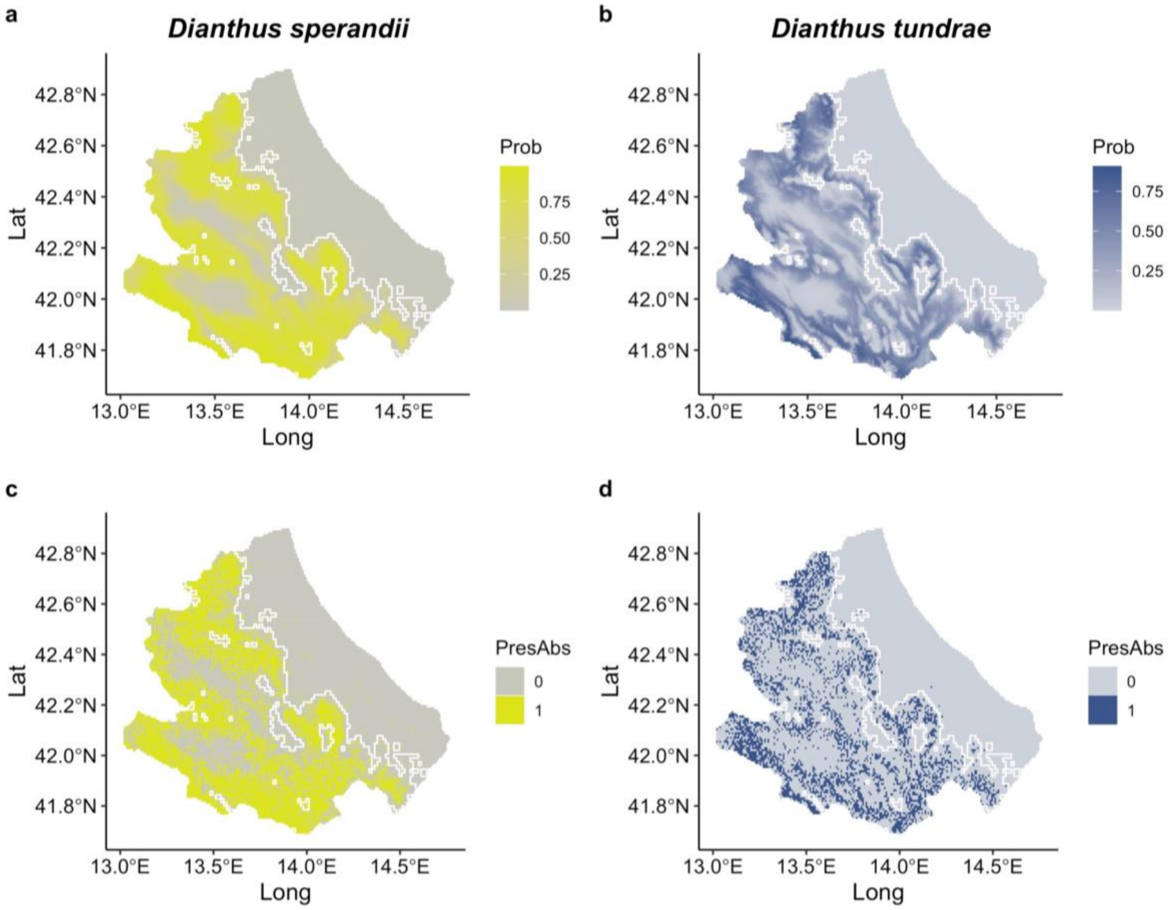


Figure 2. Simulated occurrence probability and presence/absence data of *Dianthus sperandii* (a, c) and *D. tundrae* (b, d). The white line in the plots delineates the area of interest for the study (i.e., all lands approx. between 500 and 1,800 m a.s.l.).

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 conducted the sampling only in a restricted area of the Abruzzo region: we considered all areas approx. from 500 to 1,800 m a.s.l. (90% of the cells included between 518 and

1,821 m a.s.l.; minimum elevation: 197 m, maximum elevation: 2,791 m) (the perimeter of the area of interest is marked in white in Figure 2; the frequency distribution of temperature and precipitation is reported in Figure S1.2, Appendix S1). 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 to a smaller area of interest, we avoided the 'there are no elephants in the Antarctic' paradox (Lobo, Jiménez-Valverde, Hortal, 2010).

Box 2. Description of the simulated sampling strategies.

Random: 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 by 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, Chytrý, Hájek, Lvončik, Tichý, 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 very 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 traditional 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* (*get_elev_raster* function with *zoom* argument set at 10; 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., this is the case for citizen science data), 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, Farber, Danin, 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, Lovelace, Salmon, Rudis, 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 cells × 10 cells) grid overlaid to a 2-dimensional (environmental) space. In practice, the grid is scanned on a cell-by-cell basis and, from each cell, a fixed number of randomly selected sampling units is extracted (see Figure S2.1, Appendix S2). If the amount of sampling units in a cell is lower than the fixed, desired number of units to be collected per cell, then all sampling units are extracted. 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 Albert et al., 2010; Varela et al., 2014, Hattab et al., 2017; see also Figure S2.2 in Appendix S2, which shows the portion of the environmental space occupied by the two virtual species). In Figure S2.3 (Appendix S2), we also show the effect of randomly sampling the environmental space, which results in over-sampling the most widespread environmental conditions encountered in the geographic space. Note that randomly sampling the geographic space leads to the same sampling bias (Elith et al., 2011).

Maps of design-specific sampling effort are reported in Figure S3.1 (Appendix S3).

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 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., root mean squared error) (see Box 1). Note that our measures of bias, variance, and root mean squared error (hereafter, RMSE) are estimators of these quantities, which we computed by replacing expectations with averages (computed over multiple simulations). The simulated sampling realisations were replicated 500 times for each of the six sampling strategies we tested, thereby fitting 3,000 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 RMSE, with increasing sample size). As a consequence, a comparison of the impact of different sampling strategies on the bias and variance (and RMSE) of the species' response curve cannot be undertaken without accounting for the effect of sample size (i.e., the 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 terms 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 root mean squared error (hereafter, RMSE) values. We considered an approach as the best performing approach (at a given sampling effort) when it was associated with the lowest RMSE. 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.

It is worth mentioning that the uniform approach has a limitation depending on the chosen number of sampling units to be extracted from each grid cell overlaid to the environmental space. Indeed, as explained in Box 2, when this number is larger than the amount of sampling units present in a cell, all units are extracted. This circumstance usually happens at the boundary of the environmental space, where the density of sampling units is lower. Although this is not an issue for the sampling strategy itself, having a certain amount of sampling units included in all datasets used to fit the GLMs can downwardly bias the variance of coefficients' estimators, in turn affecting RMSE. To account for this, we repeated another time the simulations for the uniform sampling of *D. sperandii* and *D. tundrae*, this time computing the variance of the estimators as the average (across simulations) of the 'theoretical' variance, i.e., inverse of the Fisher information matrix. We used this variance estimator, which is not affected by the issue of fixed number of sampling units, to re-compute the RMSE for the uniform approach, and we compared it with the one obtained from the original simulations.

2.2. Beyond simulations: accounting for real-life issues associated with species distribution modelling

In real life, the outcome of species distribution modelling is affected by a multitude of potential issues involving all stages of the analysis: from data sampling to model-related factors. Here, we considered the impact of missing covariates, such as potential confounding factors, and the effect of incompletely sampling the environmental space. The former is a common issue in SDMs (Eltih & Leathwick 2009), whereas the latter can impact the performance of, particularly, the uniform approach.

To test for the effect of missing covariates, we generated the distribution of two new virtual subspecies, whose occurrence probability was affected by temperature, precipitation and exposition towards the North (hereafter, northness). Northness was computed as the cosine of the terrain aspect measured in radians. To spatially match the raster of the northness with the climatic layers, we first resampled the layer of the elevation and then computed the northness. The two virtual subspecies were generated to have distribution patterns similar to those of *D. sperandii* and *tundrae*, which again led to compare species with a wide *vs* a restricted distribution. The regression parameters for temperature and precipitation were kept at the same value used for *D. sperandii* and *tundrae*, while the parameter for the northness was set to -1.4 for the widely distributed species (hereafter, *D. sperandii* subsp *thermophilus*) and to -2 for the species with a more restricted distribution (hereafter, *D. tundrae* subsp *thermophilus*). We set negative regression parameters to generate species preferentially occurring in south-oriented areas, seeking topographically warmer expositions (in the Northern Hemisphere). Simulations were repeated for all sampling strategies, excluding northness from the fitted GLMs, and their performance was compared as done for *D. sperandii* and *tundrae*.

To test the effect of missing part of the environmental space used by the virtual species, we repeated the simulations for the uniform approach considering only selected portions of the whole environmental space. Specifically, we performed the uniform sampling within two environmental sub-spaces, including all sampling units located either below or above the mean temperature of the environmental space (Figure S2.4, Appendix S2). This allowed testing the performance of the uniform approach when data on the whole temperature and precipitation gradient were not available. We assumed the impact of incompletely sampling the environmental space on the uniform approach to be the same regardless of species' characteristics and focussed only on *D. sperandii*.

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

3. Results

3.1. Results for D. sperandii and D. tundrae

As a general result, the RMSE of the coefficients' estimators fitted by our parametric SDMs decreased with increasing sampling effort irrespective of the sampling strategy and converged towards a similar minimum value (Figure 3a, 3b). 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 discriminating factor in the performance (i.e., RMSE) of the sampling strategies was variance, while for *D. tundrae*, it was bias (Figures S5.2, S6.2). Having a low proportion of sampling units consistently included in the datasets used to fit the GLMs across simulations did not affect the results for the uniform approach (Figure S4.1, Appendix S4).

Concerning *D. sperandii*, the proximity-to-road, as a sampling design, consistently provided the worst performance in terms of RMSE at all sample sizes (Figure 3a). 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 RMSE values at all sampling efforts for all parameters (Figure 3a). Specifically, the RMSE of the uniform sampling was systematically 50% lower than that of the proximity-to-road sampling for all coefficients but precipitation (Figure S5.1). The random, stratified, systematic and topographic sampling designs performed similarly, with their RMSE values generally included right in between those of the proximity-to-road and uniform approaches (Figures 3a, S5.1). All designs, except for the proximity-to-road approach, overestimated the partial effect of precipitation in the long run, i.e., featuring a positive bias (Figure S5.2). 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 3c and Figure S5.3 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 (Figure S5.2). Concerning the variance, the uniform sampling within the environmental space provided the most efficient estimators for all coefficients, regardless of sample size (Figure S5.2). This resulted in a more consistent shape of the response curve across simulations (Figure 3c, Figure S5.3).

Concerning *D. tundrae*, regardless of the sampling effort, the topographic approach scored the highest performance for all regression coefficients but precipitation (Figures 3b, S6.1). Also, except for precipitation, the topographic approach was always (i.e., across sampling efforts) followed by the systematic, stratified and random strategies (Figures 3b, S6.1). On the contrary, the uniform sampling within the environmental space showed the worst performance (i.e., highest RMSE) for the intercept and the temperature (both linear and quadratic terms) at nearly all sampling efforts, whereas it scored best for precipitation. The stratified, systematic, random, and uniform designs, in the long-run, overestimated the partial effect of precipitation, while the estimators derived from the proximity-to-road and topographic approach had a low and negative bias, respectively (Figure S6.2). Similarly to what was observed 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 3d and Figure S6.3 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 for temperature and downwardly biased estimators of the quadratic term for temperature (Figure S6.2). Concerning variance, the uniform sampling within the environmental space had the lowest variance for precipitation, while all strategies showed comparable efficiency for the other coefficients.

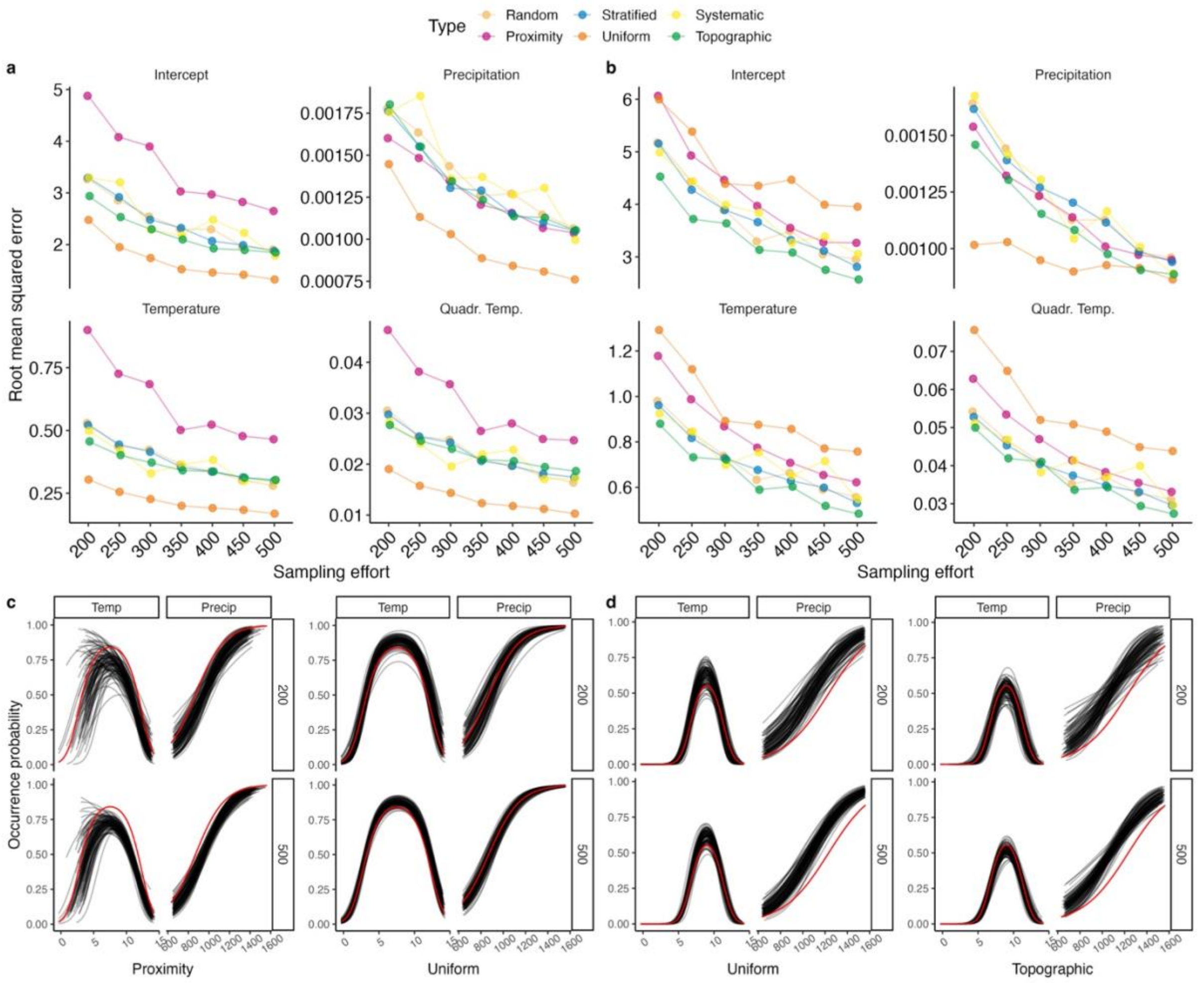


Figure 3. Root mean squared error of regression coefficients for *D. sperandii* (panel a) and *D. tundrae* (panel b). Panel c: Comparison between the response curves for *D. sperandii* as estimated by data collected through the proximity-to-road approach and the uniform sampling of the environmental space. Panel d: Comparison between the response curves for *D. tundrae* as estimated by data collected through the uniform sampling of the environmental space and the topographic approach. Grids of panels c and d combine two sampling efforts (*N* = 200 and 500) and the two predictors used in the models (i.e., temperature and precipitation). Red lines, in panels c and d, represent the true relationship between *D. sperandii* (panel c) or *D. tundrae* (panel d) and the predictors. Comparisons between estimated and true response curves for all sampling strategies and efforts are reported in Appendix S5 (Figure S5.3) for *D. sperandii* and in Appendix S6 (Figure S6.3) for *D. tundrae*.

3.2. Results for missing covariates and incomplete sampling of the environmental space

Concerning *D. sperandii* subsp *thermophilus*, excluding northness from the fitted GLMs did not particularly affect the proximity-to-road sampling, which exhibited the worst performance in estimating all parameters except precipitation (Figure 4a), similarly to what was observed for *D. sperandii*. Conversely, simulating a missing covariate brought the performance of the uniform sampling partially closer to that of the other sampling strategies than observed for *D. sperandii* (Figure 4a). Specifically, when estimating the intercept and the parameter for precipitation, the uniform approach performed worse for *D. sperandii* subsp *thermophilus* than for *D. sperandii*. Indeed, in contrast to what was observed for *D. sperandii*, intercept and precipitation were best estimated by the topographic approach when northness was excluded (Figure 4a). However, results for the linear and the quadratic terms of temperature were in line with those obtained for *D. sperandii*, with the uniform approach scoring best across all sampling effort. As for *D. tundrae* subsp *thermophilus*, results were similar to those observed for *D. sperandii* subsp *thermophilus*. The topographic sampling scored best for intercept and precipitation, whereas the uniform sampling showed the best performance for the linear and quadratic term for precipitation, although RMSE values were not very far from those of the topographic sampling (Figure 4b).

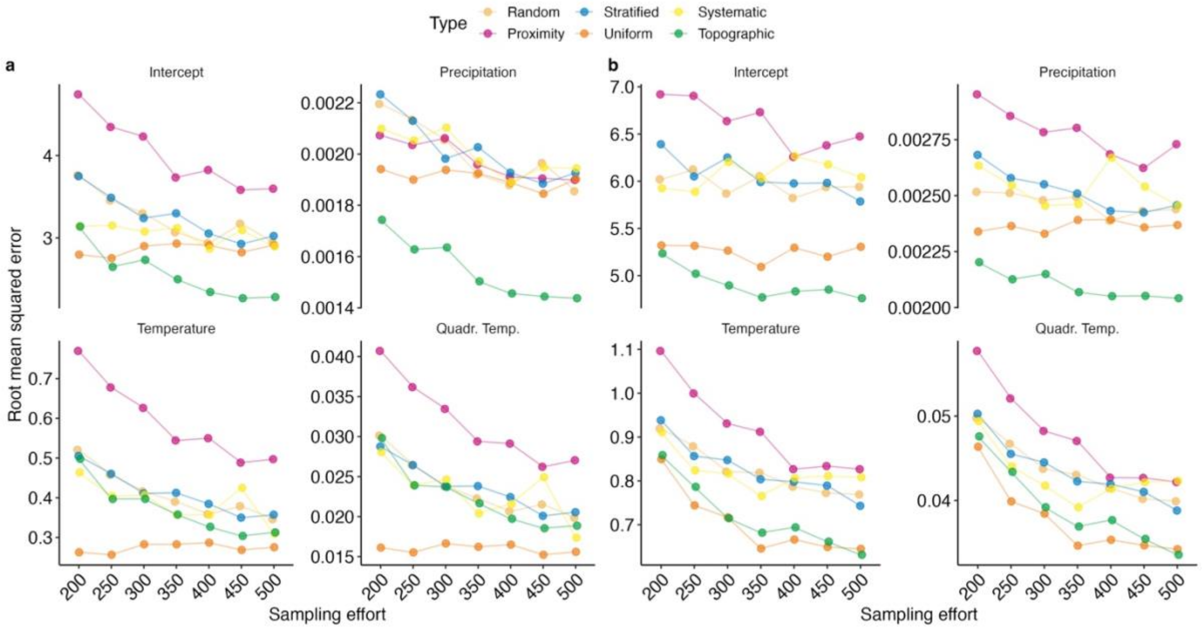


Figure 4. Root mean squared error of regression coefficients for *D. sperandii* subsp *thermophilus* (a) and *D. tundrae* subsp *thermophilus* (b) derived from GLMs fitted excluding northness.

The incomplete sampling of the environmental space overall reduced the performance of the uniform approach, except when using sampling units located above the mean temperature to estimate the parameter for precipitation. Restricting the sampling to units located below the mean temperature (left side of the environmental space; Figure S2.4, Appendix S2) reduced the performance of the uniform approach for estimating all parameters. In this case, performances were comparable to those displayed by the worst performing sampling approaches presented in *3.1.* for *D. sperandii* (Figure 3a). The performance of the uniform approach was halved when modelling the response curve for precipitation (Figure 5). On the contrary, restricting the sampling to units located above the mean temperature (right side of the environmental space; Figure S2.4, Appendix S2) did not affect the performance of the uniform approach for estimating precipitation but strongly decreased its performance for estimating the intercept and the linear and quadratic terms of the response curve for temperature (Figure 5).

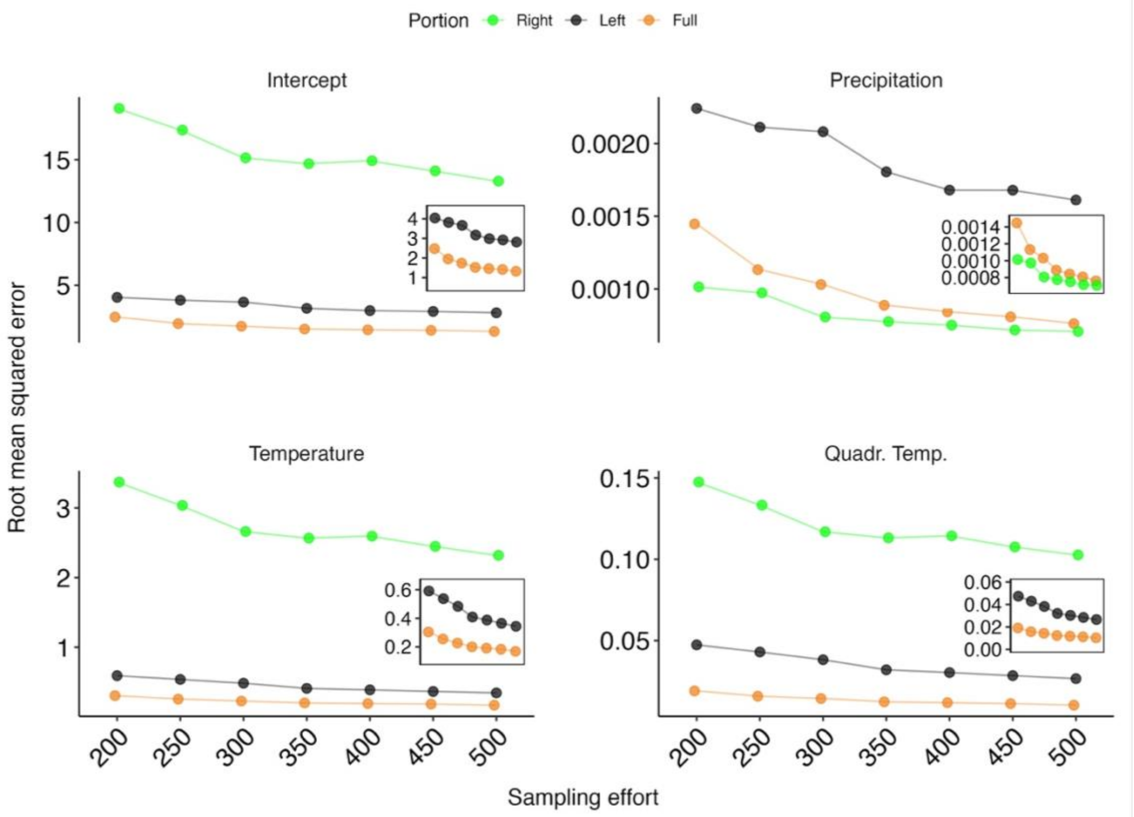


Figure 5. Root mean squared error of regression coefficients for *D. sperandii* and the uniform approach implemented in portions (left: sampling units below mean temperature; right: sampling units above mean temperature) of the whole environmental space (i.e., full). See Figure S2.4 (Appendix S2). Insets enhance the visibility of the comparison among root mean squared error trends for the portions of the environmental space with similar performances.

4. Discussion

By creating virtual species with different thermal tolerances and, as a result, different distribution extents across the Abruzzo region in Italy (wide for *D. sperandii* and narrow for *D. tundrae*), we tested the impact of different sampling strategies 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, Lobo, Hortal, 2011). We also found that collecting more data (increasing the sample size *N*) alleviates the impact of the sampling strategy on the variance and RMSE of the coefficients, thereby confirming results from previous studies (Albert et al., 2010; 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 and 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 S3.1 in Appendix S3; 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 (Albert et al., 2010;

Figure S5.2 in Appendix S5). 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 -0.60, topographic -0.61, others -0.66), which may partly explain the higher precision of the coefficients estimated from parametric SDMs through the uniform sampling approach. Importantly, coefficient estimators derived from the uniform approach remained the most efficient even after accounting for the (low) proportion of sampling units consistently included in the datasets used to fit the GLMs. Instead, an incomplete sampling of climatic gradients, and more specifically, systematically sampling only restricted portions of the environmental space (mimicking real-life situations of SDMs fitted on a limited amount of environmental data), had an overall negative impact on the performance of the uniform approach. While this seems to impair the efficacy of the uniform approach, similar effects on the other sampling strategies exist (Figure S2.3, Appendix S2).

In contrast to what we observed for the uniform approach, the proximity-to-road strategy exhibited the worst performance in terms of RMSE (Albert et al., 2010). 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 warn ecologists against using data gathered through the proximity-to-road strategy for fitting SDMs, except for specific circumstances under which this is the investigated factor (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 3c, S5.3), 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 (Albert et al., 2010). 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 best for all regression parameters but precipitation. The similar performance of the six sampling strategies we tested for the specialist species might be due to species with a low relative occurrence area (i.e., the proportion of area occupied by the species over the whole study 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) (Albert et al., 2010). 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 is 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).

Although not strictly related to any sampling strategy, model misspecification, which includes the problem of missing covariates, is widely acknowledged as an important (and common) issue in SDMs (Elith & Leathwick 2009). Our results indicate that excluding one key covariate, like northness which was used to generate the distribution of the two subspecies of *Dianthus*, had little impact on the ranking of the performance of the sampling approaches. This suggests that the effect of 'missing covariates' may equally and negatively impact all sampling approaches. At the same time, the impact of excluding northness might have been weakened by the low correlation between the variable and the climatic predictors observed in the area of interest. Indeed, the amount of bias associated with missing covariates is a function of the correlation between the omitted predictor and both (i) the response variable and (ii) the predictors kept in the fitted model (Gelman & Hill 2007). Interestingly, we found that excluding northness led to the topographic sampling outperforming the uniform approach for modelling the response curve along the precipitation gradient. A possible explanation could be that, being based on northness (among other topographic variables), the topographic approach was probably less impacted by the bias introduced by model misspecification. This evidences how using a sampling strategy associated with 'important' predictors excluded from SDMs can reduce the impact of missing covariates on model predictions of the species response curve. Yet, we also observed that the uniform sampling still scored best for describing the unimodal response curve of the generalist *D. sperandii* subsp *thermophilus* to temperature, pointing to its best performance for modelling generalist species.

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 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. Hence, and because no sampling strategy clearly stands out to sample specialist species, uniformly sampling the environmental space may in the end be the best bet irrespective of the degree of specialization of a given species as it will allow to at least optimize the accuracy of the response curves generated for generalist and widespread species. Importantly, real-life issues related to SDMs can strongly affect the performance of the sampling strategies. Here, we considered the effect of incompletely sampling the environmental space or excluding important predictors, but we acknowledge that other, equally important factors, e.g., location and survey error (Gábor et al., 2022) or the lack of *a priori* knowledge on the shape of the species response curve may affect the performance of the tested sampling approaches. We, therefore, envisage the use of simulation-based approaches for testing the performance of different sampling strategies under a wider set of potential modelling related issues. Also, although our study is designed with plant ecology in mind, its rationale could be extended to improve the modelling of the distribution of other biological organisms.

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.

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 review on the manuscript.

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.

Funding statement

MB acknowledges funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101066324. MGS acknowledges funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101090344. DDR is supported by the FRS-FNRS. ET acknowledges funding from the Estonian Research Council (grant MOBJD1030).

Conflict of interest disclosure

The authors declare no conflict of interest.

Ethics approval statement

Not applicable.

References

Albert, C. H., Yoccoz, N. G., Edwards Jr, T. C., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). Sampling in ecology and evolution–bridging the gap between theory and practice. *Ecography*, *33*(6), 1028-1037.

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 modelling 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.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, *17*(1), 43-57.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, *40*(1), 677-697.

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., Jetz, W., Lu, M., Rocchini, D., Cord, A., Malavasi, M., ... Moudrý, V. (2022). Positional errors in species distribution modelling are not overcome by the coarser grains of analysis. *Methods in Ecology and Evolution*, *13*(10), 2289-2302.

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.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, New York, New York, USA.

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.

Appendix S1 - Study area

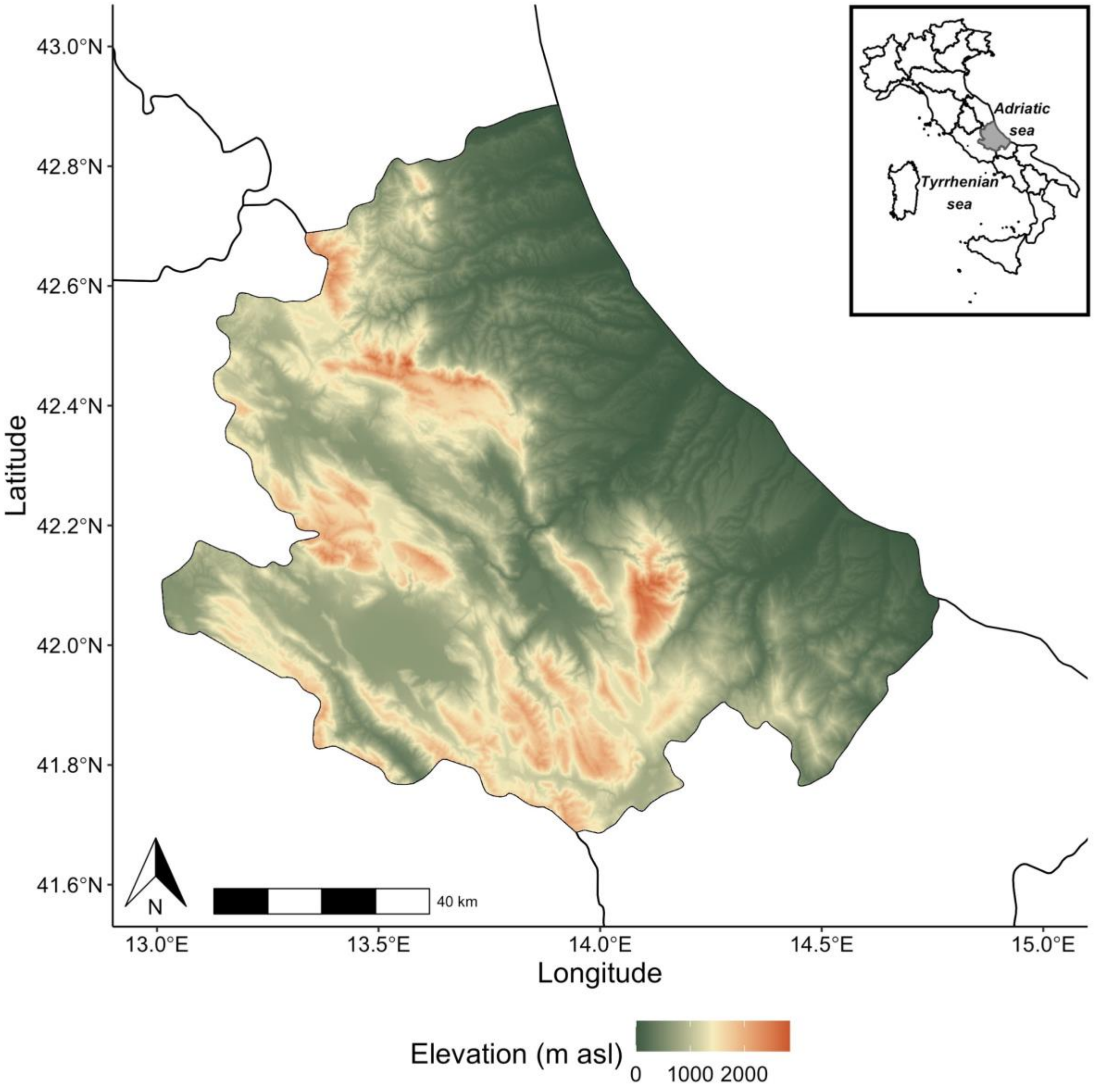


Figure S1.1. Elevation map of the Abruzzo region. The inset map shows the geographic position of Abruzzo in Italy.

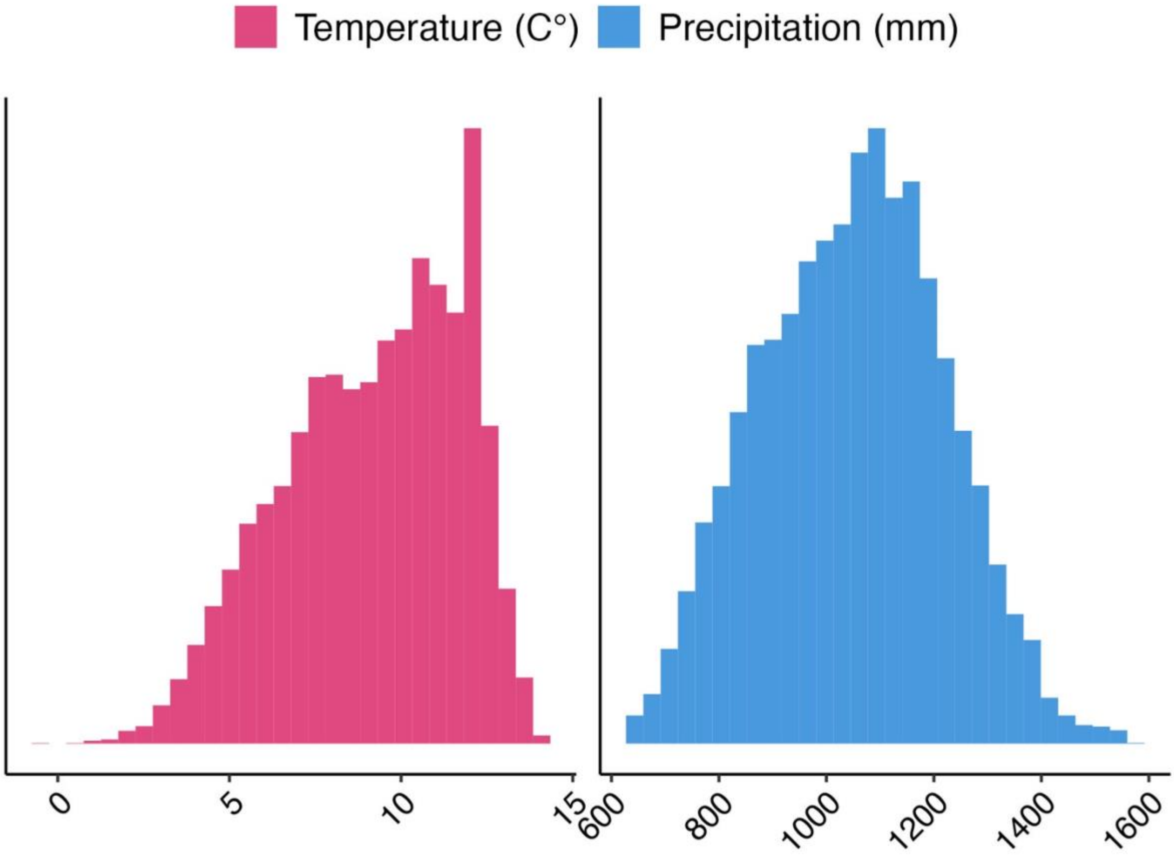


Figure S1.2. Frequency distribution of temperature and precipitation in the area of interest of Abruzzo.

Appendix S2 - Environmental space

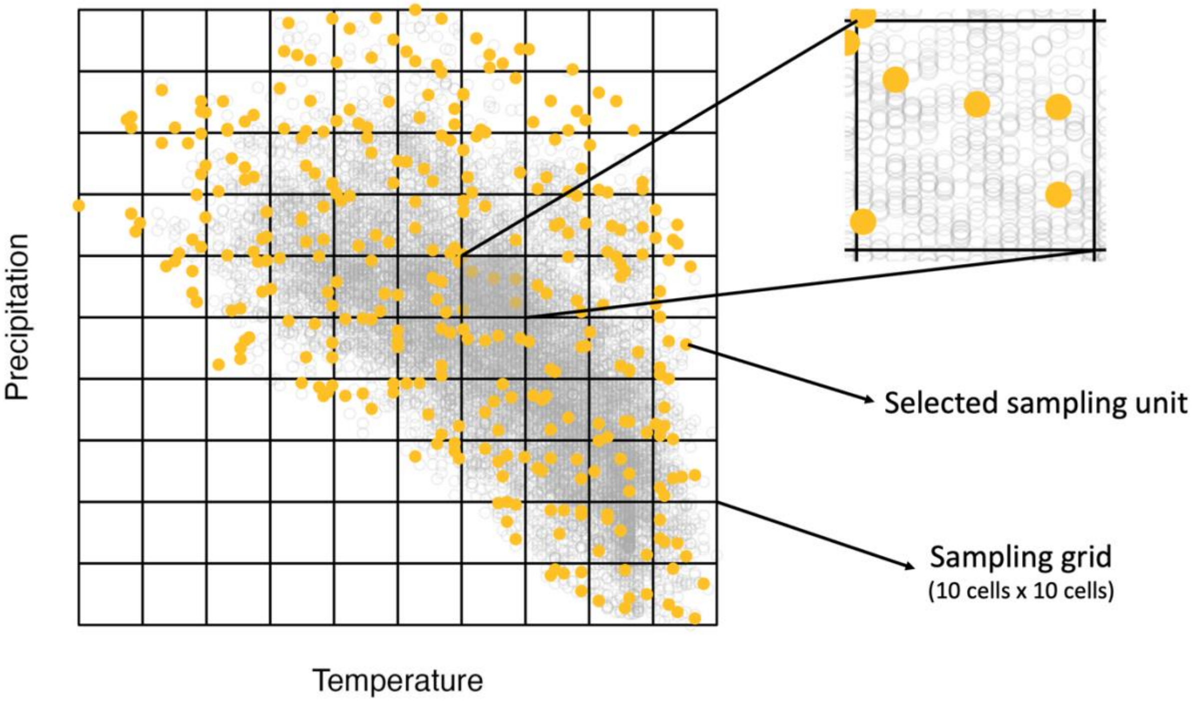


Figure S2.1. Representation of the uniform approach. The sampling grid is scanned cell by cell and, from each cell, a fixed number of sampling units (here in gold) is selected. No sampling units are collected in empty cells, while a lower number is collected in cells including less sampling units than the set number.

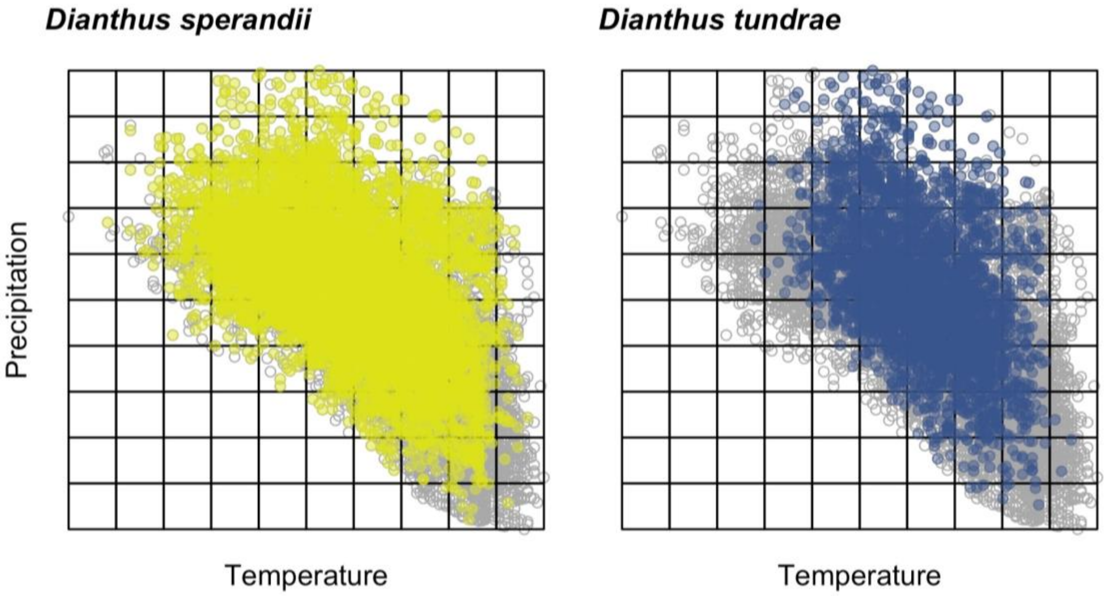


Figure S2.2. Position occupied by *D. sperandii* and *D. tundrae* within the environmental space spanned by temperature and precipitation within the area of interest. Coloured dots represent presence locations (lime for *D. sperandii*, blue for *D. tundrae*), while grey dots represent all climatic pixels, i.e., cells of the temperature and precipitation raster layers, included in the area of interest of Abruzzo.

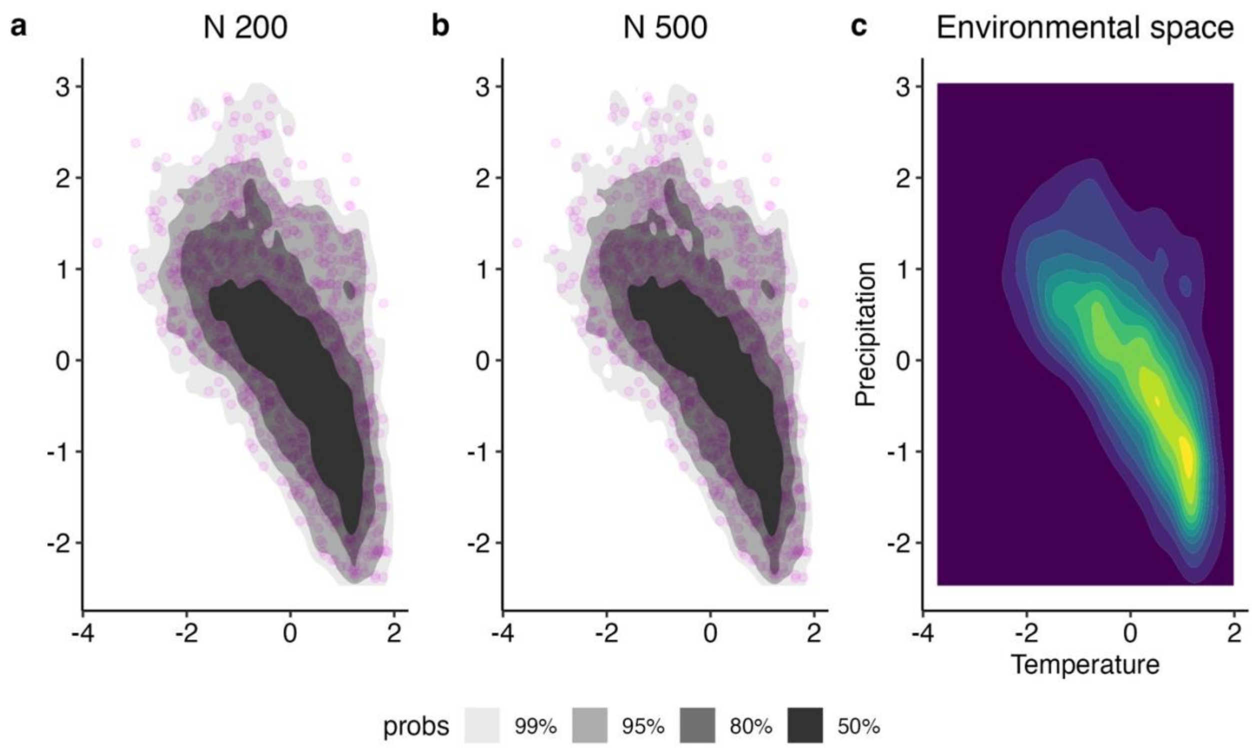


Figure S2.3. Effect of randomly sampling the environmental space. Panels (a) and (b) show the density of points randomly selected within the environmental space by extracting 500 times 200 and 500 sampling units, respectively. Purple points represent a random sample of 2,000 sampling units. Panel (c) shows the density of sampling units considering the whole environmental space (from low density areas in blue, to high density areas in yellow). Probs: probabilities.

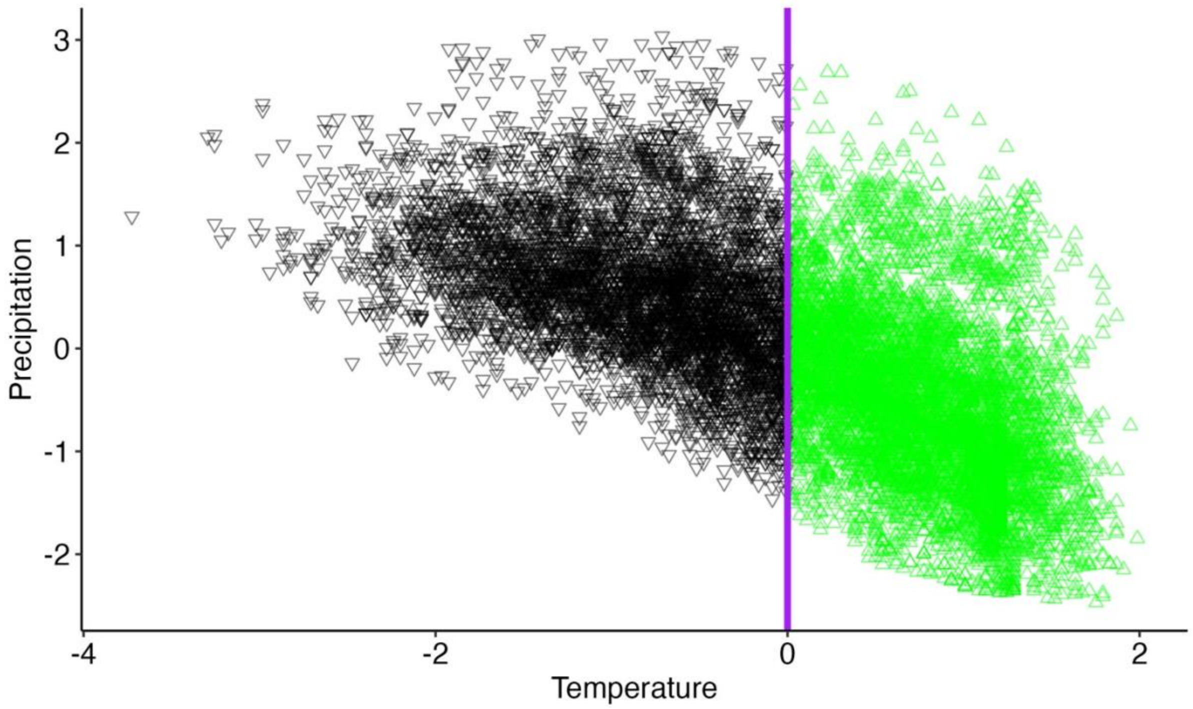


Figure S2.4. The incomplete sampling of the environmental space was carried out within two sub-spaces. These included all sampling units located either below (black triangles) or above (green triangles) mean temperature (the vertical purple line separates the environmental spaces in the two sub-spaces). Note that temperature and precipitation values are standardised.

Appendix S3 - Sampling effort

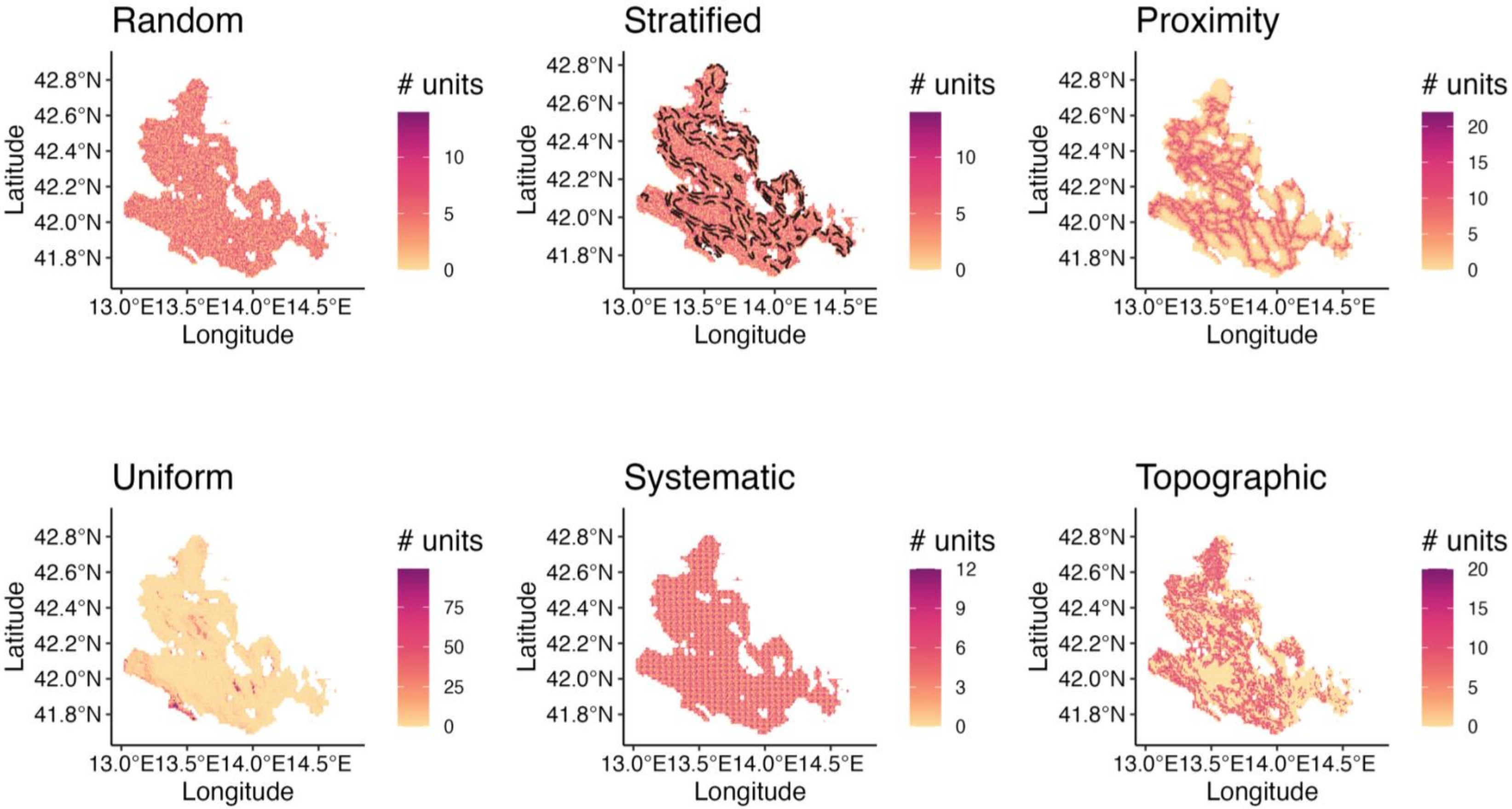


Figure S3.1. Number of sampling units selected by the different sampling approaches within the area of interest (all lands approx. between 500 and 1,800 m a.s.l.). Maps were generated replicating each sampling activity 100 times and counting the number of times each cell was selected. Proximity: proximity-to-road; Uniform: uniform sampling of the environmental space.

Appendix S4 - Effect of fixed (i.e. consistently included) sampling units on the performance of the uniform approach

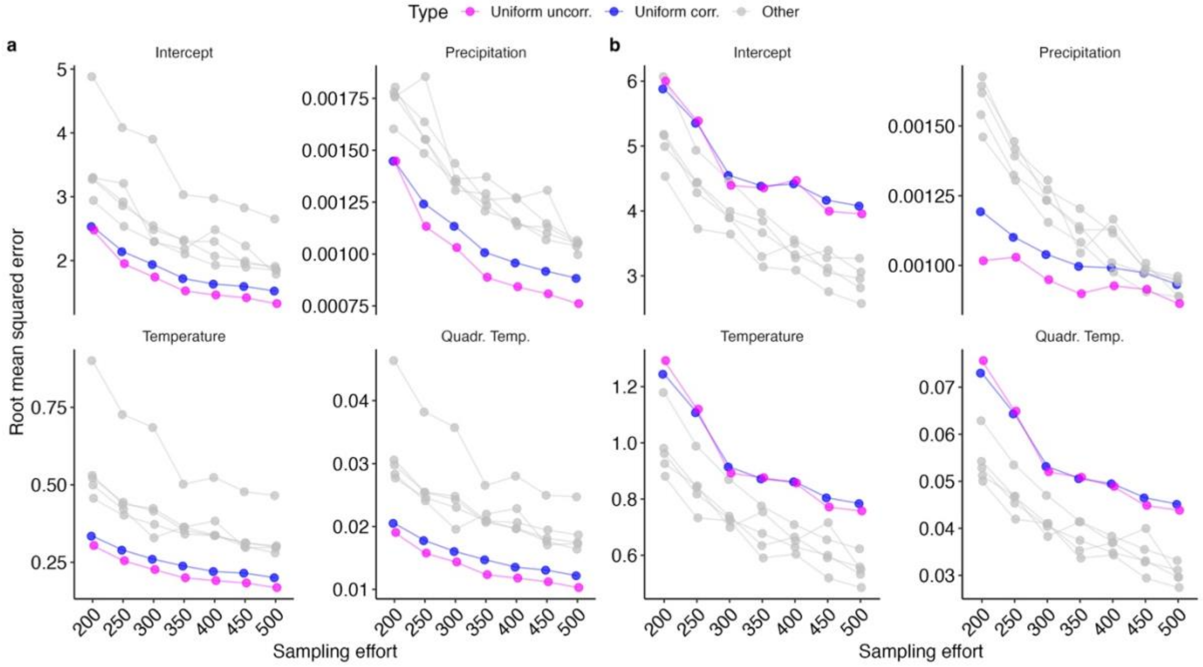


Figure S4.1. Effect of the sampling units consistently included in all datasets used to fit the generalised linear models (across simulations) on the performance of the uniform approach. Left panel (a): results for *D. sperandii*; right panel: results for *D. tundrae*. Purple line (uniform uncorrected): trend in root mean squared error computed using the 'uncorrected' variance of the estimators. Blue line (uniform corrected): trend in root mean squared error computed using the 'corrected' variance of the estimators. Grey line (other): trend in root mean squared error for the other sampling approaches.

Appendix S5 – Comparisons of sampling strategies for *Dianthus sperandii*

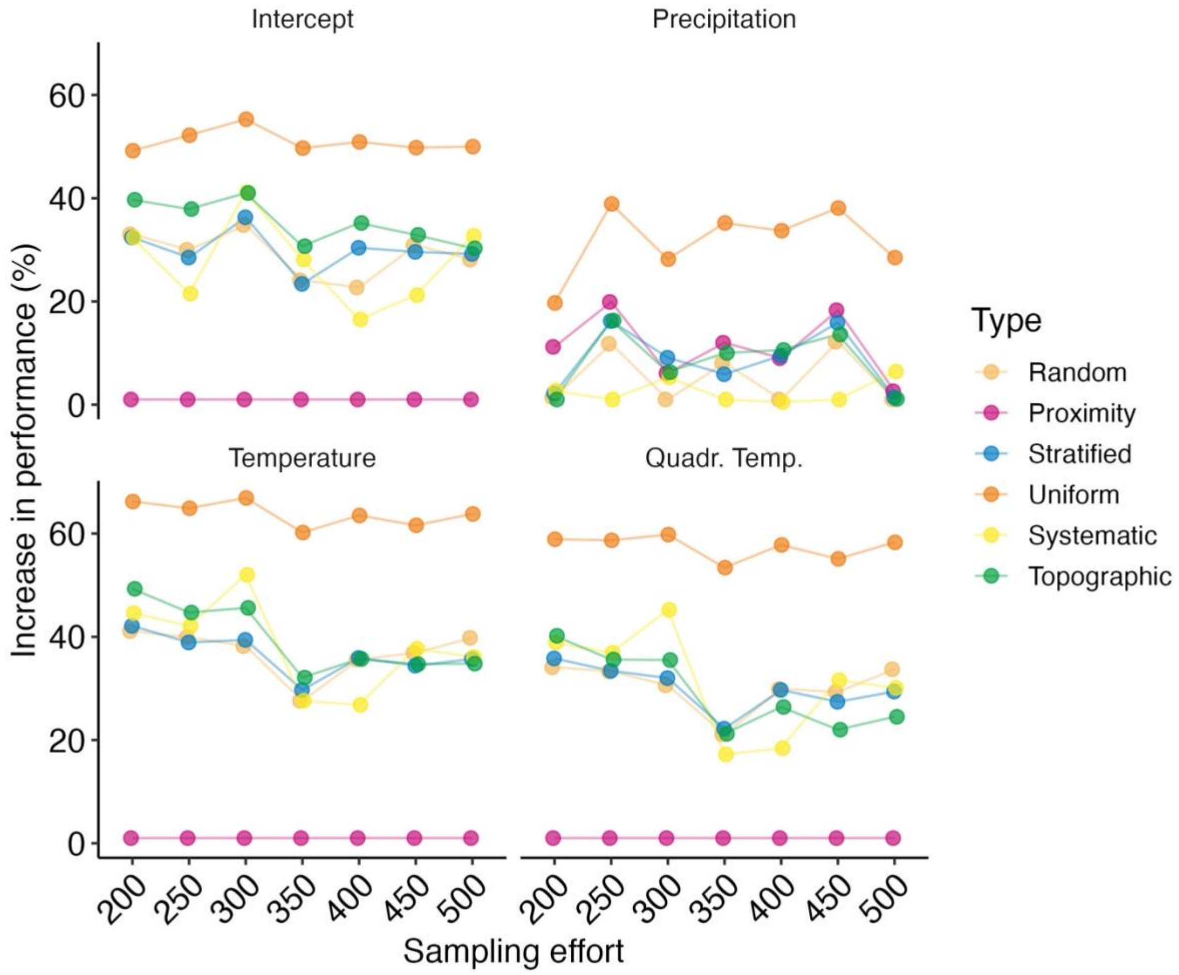


Figure S5.1. Performance (expressed by percentage decrease in RMSE values with respect to the worst performing approach) obtained for the different sampling strategies used to record the presence/absence of *D. sperandii*. Values are reported for increasing sampling effort. Quadr. Temp.: quadratic term for temperature.

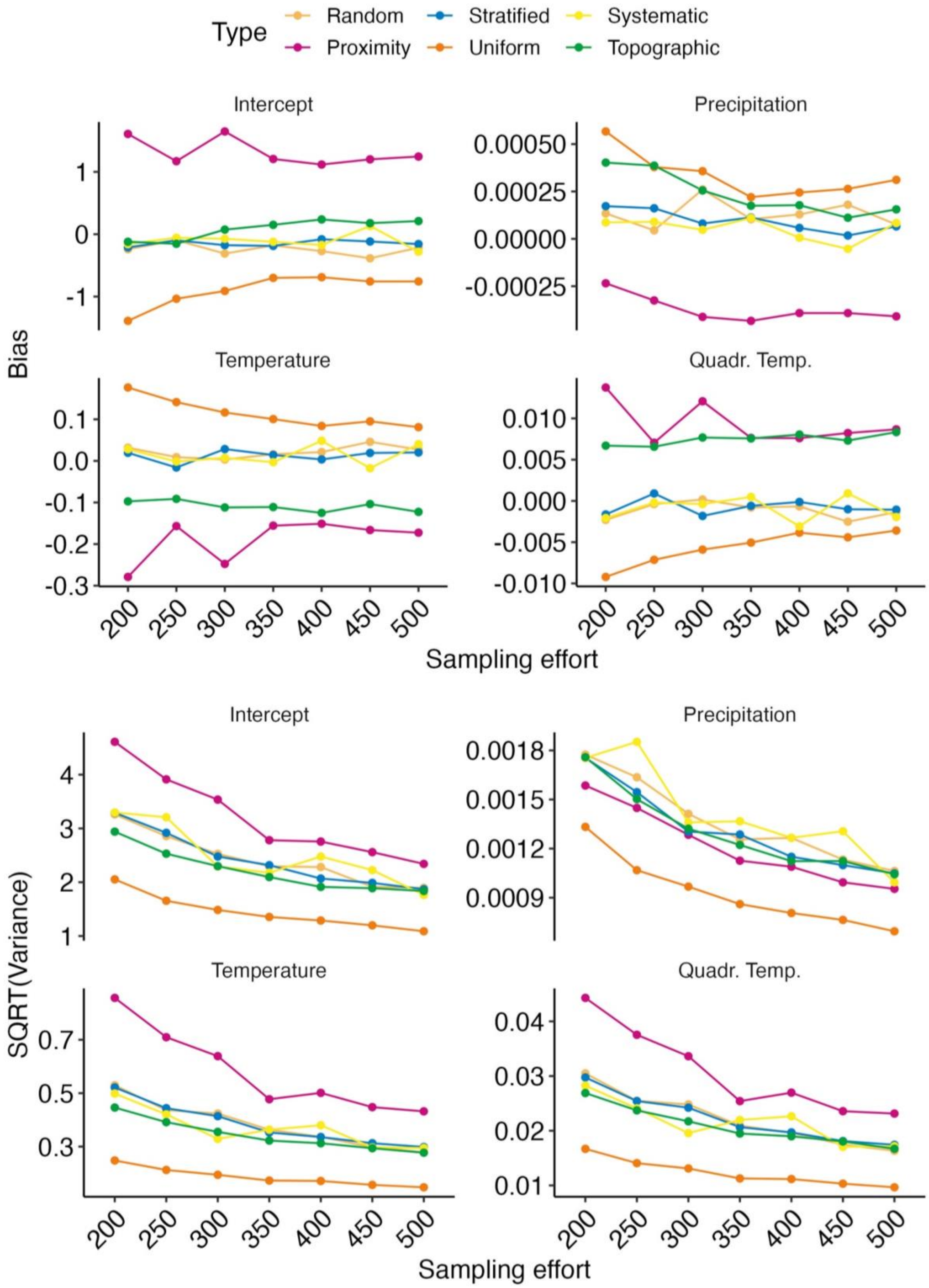
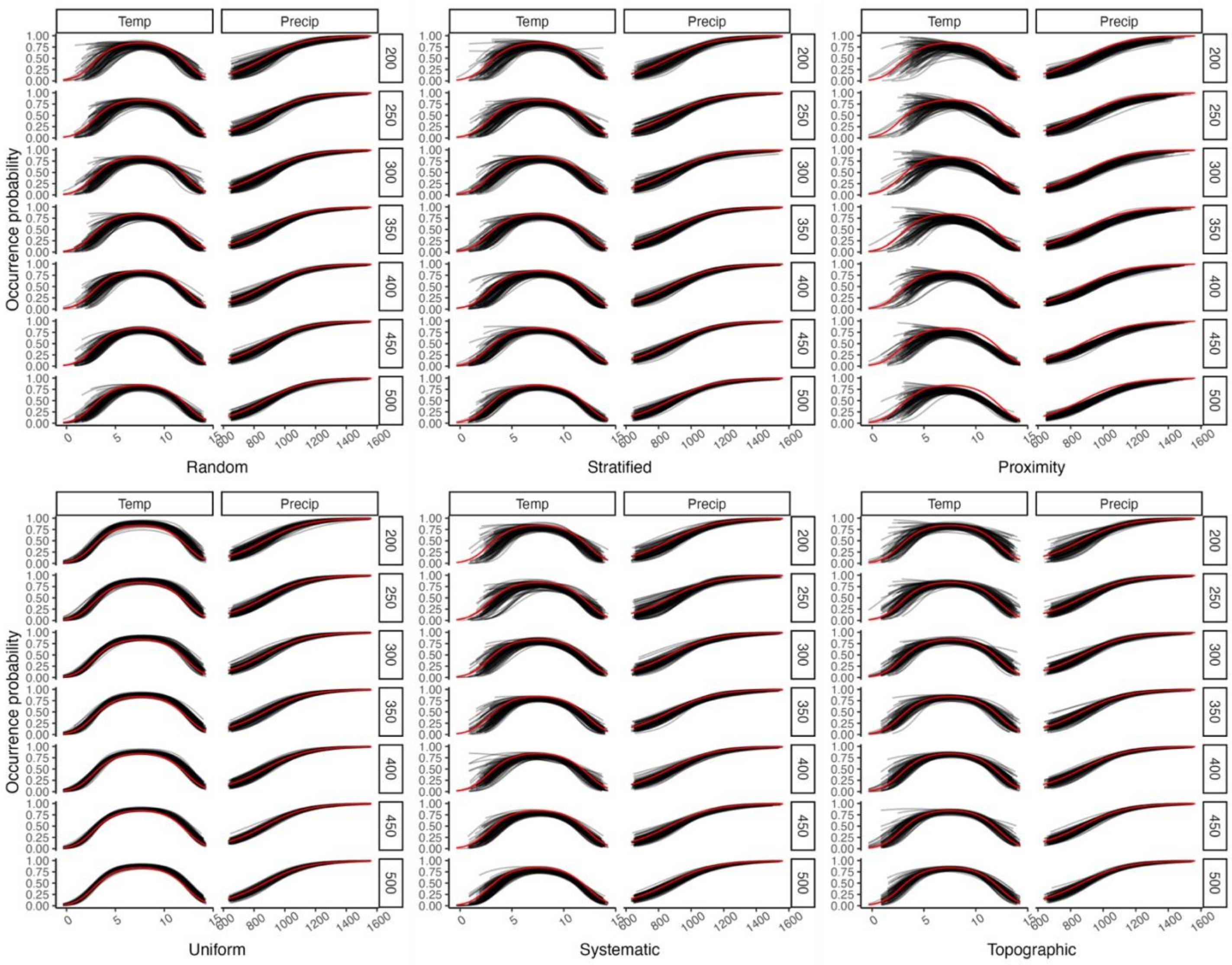


Figure S5.2. Bias and variance (square rooted) of regression coefficients for *D. sperandii*.



Random

Stratified

Proximity

Uniform

Systematic

Topographic

Figure S5.3. Comparisons between modelled (black) and true (red) response curves for *D. sperandii*. Modelled response curves were derived by fitting, for each sampling strategy and sample size, 100 binomial generalised linear models (link logit).

Appendix S6 – Comparisons of sampling strategies for *Dianthus tundrae*

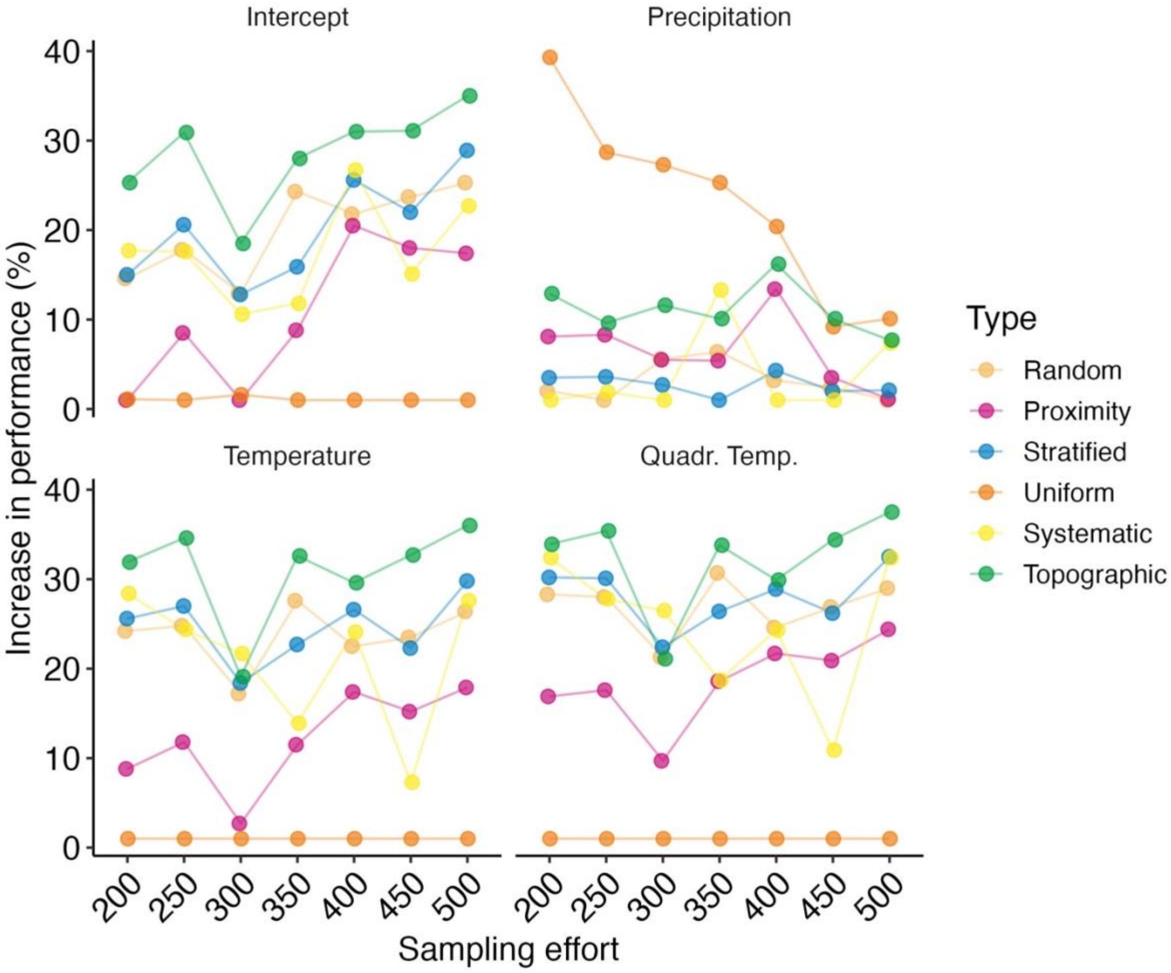


Figure S6.1. Performance (expressed by percentage decrease in RMSE values with respect to the worst performing approach) obtained for the different sampling strategies used to record the presence/absence of *D. tundrae*. Values are reported for increasing sampling effort. Quadr. Temp.: quadratic term for temperature.

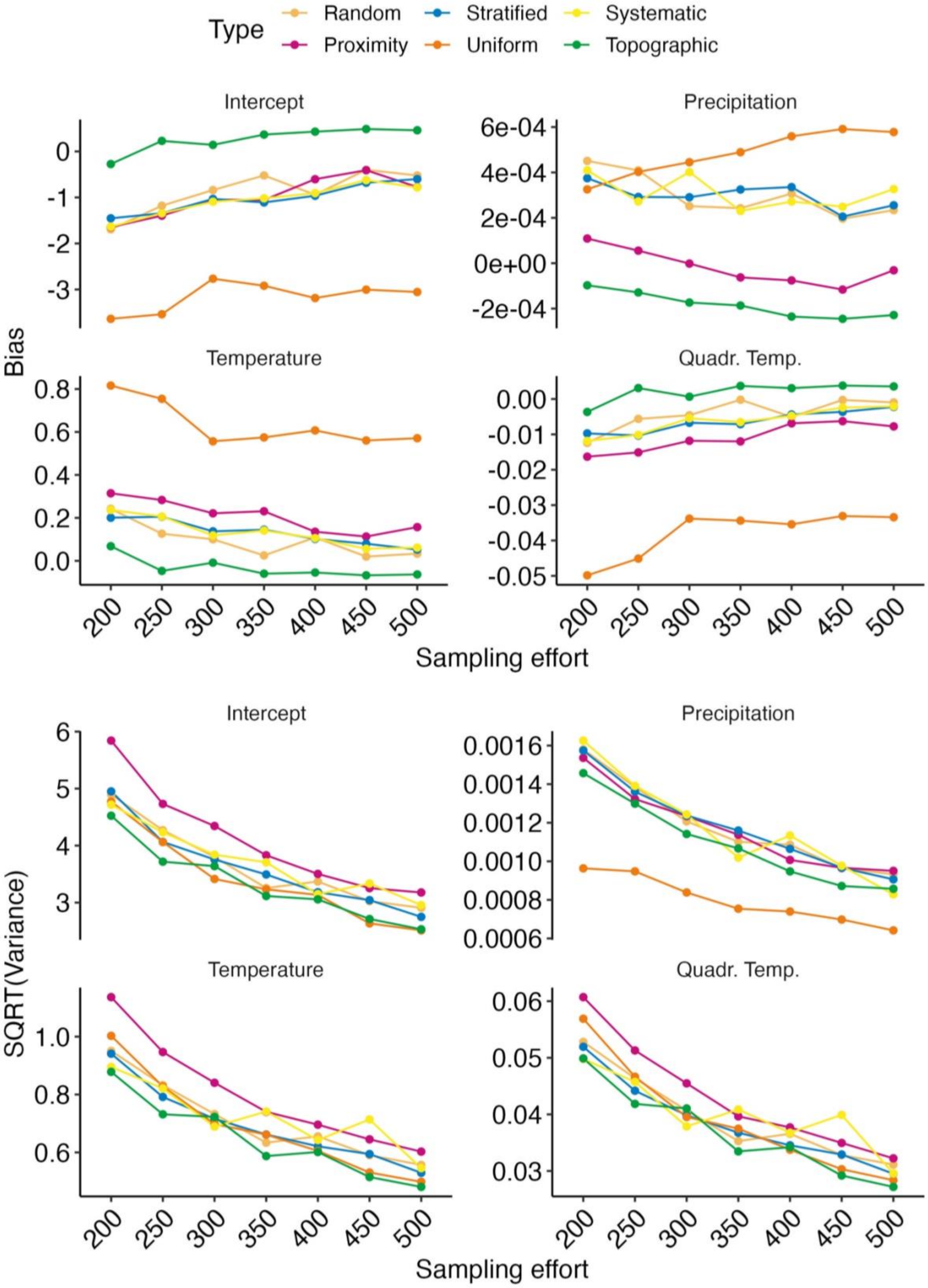


Figure S6.2. Bias and variance (square rooted) of regression coefficients for *D. tundrae*.

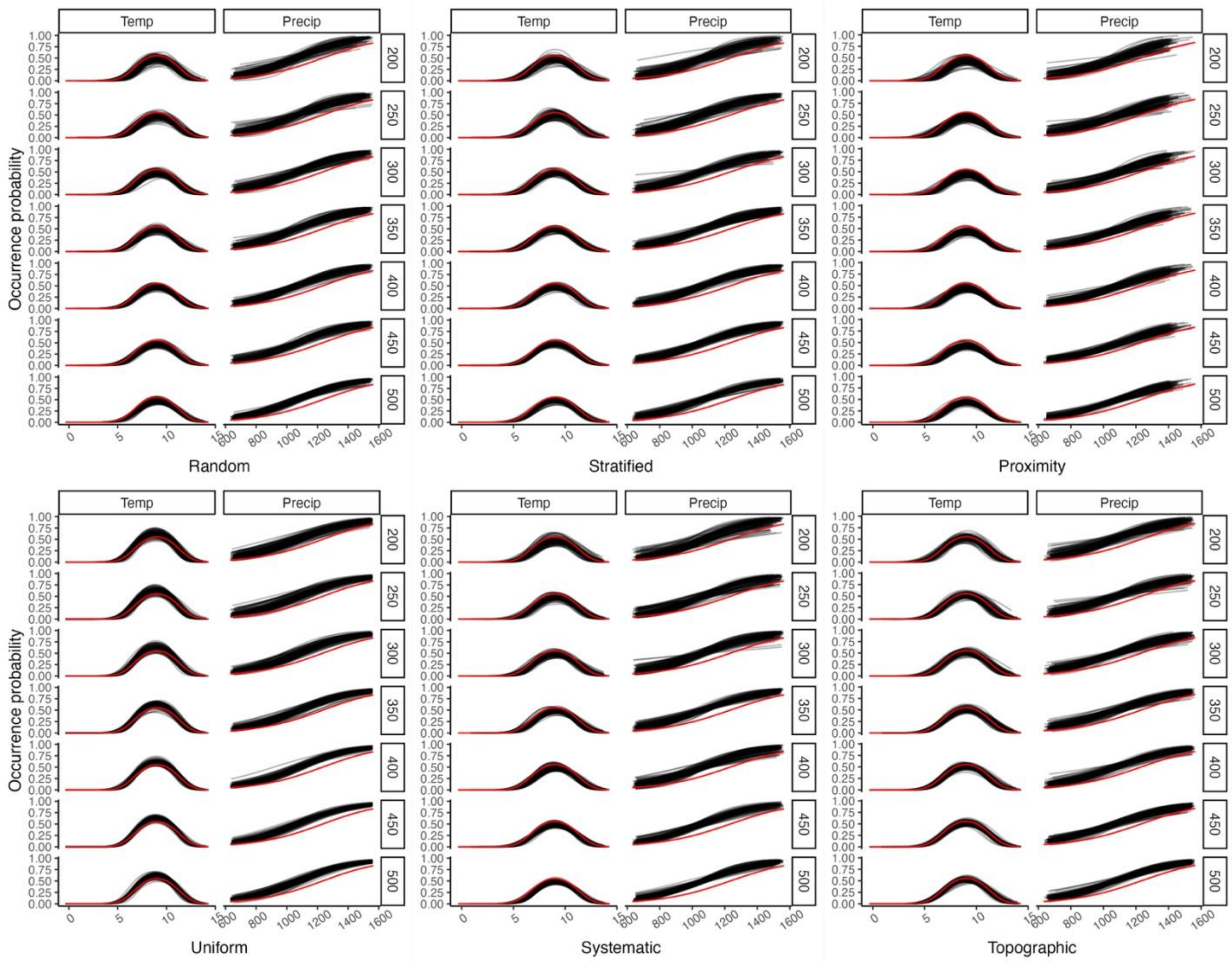


Figure S6.3. Comparisons between modelled (black) and true (red) response curves for *D. tundrae*. Modelled response curves were derived by fitting, for each sampling strategy and sample size, 100 binomial generalised linear models (link logit).