USE it: uniformly sampling pseudo-absences within the environmental space for applications in habitat suitability models

Daniele Da Re[1,*†], Enrico Tordoni[2†], Jonathan Lenoir[3],

Jonas J. Lembrechts[4], Sophie O. Vanwambeke[1,],

Duccio Rocchini[5,6], and Manuele Bazzichetto[7†]

[1] Georges Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Place Louis Pasteur 3, 1348 Louvain-la-Neuve, Belgium.
[2] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia
[3] UMR CNRS 7058 «Ecologie et Dynamique des Systèmes Anthropisés» (EDYSAN), Université de Picardie Jules Verne, 1 rue des Louvels, 80000 Amiens, France
[4] Research Group Plants and Ecosystems, University of Antwerp, Belgium
[5] BIOME Lab., Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Via Irnerio 42, 40126 Bologna, Italy
[6] Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcka 129, 16500 Praha, Czech Republic
[7] Faculty of Environmental Sciences, Department of Spatial Sciences, Czech University of Life Sciences Prague, Kamýcka 129, 16500, Praha-Suchdol, Czech Republic

†DDR, ET and MB equally contributed to this work

Corresponding author: Daniele Da Re, daniele.dare@uclouvain.be

1

Abstract

1. Habitat suitability models infer the geographical distribution of species using occurrence data and environmental variables. While data on species presence are increasingly accessible, the difficulty to confirm real absences in the field often forces researchers to generate them *in silico*. To this aim, pseudo-absences are commonly randomly sampled across the study area (i.e., the geographical space). However, this introduces sample location bias (i.e., the sampling is unbalanced towards the most frequent habitats occurring within the geographical space) and favours class overlap (i.e., overlap between environmental conditions associated with species presences and pseudo-absences) in the training dataset.

2. To mitigate this, we propose an alternative methodology (i.e., the uniform approach) that systematically samples pseudo-absences within a portion of the environmental space delimited by a kernel-based filter, which seeks to minimise the number of false-absences included in the training set.

3. We simulated 50 virtual species and modelled their distribution using training datasets assembled with the presence points of the virtual species and pseudo-absences collected using the uniform approach and other approaches that randomly sample pseudo-absences within the geographical space. We compared the predictive performance of habitat suitability models and evaluated the extent of sample location bias and class overlap associated with the different sampling strategies.

4. Results indicated that the uniform approach: (i) effectively reduces sample location bias and class overlap; (ii) provides comparable predictive performance to sampling strategies carried out in the geographic space; and (iii) ensures gathering pseudo-

2

absences adequately representing the environmental conditions available across the study area. We developed a set of R functions in an accompanying R package called `USE` to disseminate the uniform approach.

Keywords: background points, ecological niche models, presence-only models, sample location bias, class overlap, species distribution models, reproducibility.

1 Introduction

Habitat suitability models (hereafter, HSMs) are a class of statistical models used to describe the relationship between species attributes (e.g., presence-absence, abundance) and a set of spatially-explicit variables chiefly representing abiotic, biotic and human-related factors (e.g., climate, soil, demographic parameters, land-use). These models are rooted in the niche theory (i.e., *Hutchinsonian* niche, see Guisan et al., 2017) and rely on both theoretical and practical assumptions: (i) species are assumed to be at (quasi)equilibrium with their environment (Hattab et al., 2017); (ii) the set of predictors used to fit HSMs includes all necessary information to capture the ecological niche of the species; and (iii) species distribution attributes, used as the response variable, need to be appropriate for the intended model purpose (e.g., biodiversity conservation, forecasting biological invasions, assessing the effects of global change) (Tessarolo et al., 2021; but see also Guisan et al., 2017 for a thorough review on the theoretical assumptions underpinning HSMs). Some of these assumptions are hardly, if ever, met in nature since species are seldom at equilibrium with their environment (Svenning and Skov, 2004), posing several limitations to the use and interpretation of HSMs' outputs. Acknowledging and, when possible, addressing these limitations still makes HSMs a powerful toolbox for understanding the drivers of the species' realized and potential distributions (*sensu* Jackson and Overpeck, 2000). For this reason, HSMs are still widely applied in several research fields, including biogeography (Wasof et al., 2015; Duffy et al., 2017), climate change ecology (Jarvie and Svenning, 2018), conservation biology (Newbold, 2018; Santini et al., 2021), invasion ecology (Hattab et al., 2017; Da Re et al. 2020; Bazzichetto et al. 2021), and pathogen risk assessment (Batista et. al., 2023).

One of the most critical assumptions underpinning HSMs is the appropriateness of

4

biological data for modelling the ecological niche of the species, which means that species distribution attributes, being either presence-absence or abundance data, should allow effectively describing the true species-environment relationship (Guisan et al., 2017; Baker et al., 2022). However, while information on species occurrence (i.e., presence) is usually readily accessible through field-collected observations or museum/herbaria records, trustworthy absence data are by far more difficult to gather or to confirm in the field (Jiménez-Valverde et al., 2008), as their sampling requires labour-intensive and costly field campaigns (Hattab et al., 2017). The usual lack of true absence data has led to the development of HSMs approaches that either rely solely on presence data (so-called 'presence-only models', such as the BIOCLIM model; Booth et al. 2014) or combine presence data with pseudo-absences or background points for modelling species distributions (e.g., the MaxEnt algorithm; Phillips et al., 2017).

Pseudo-absences and background points are terms often used interchangeably in scientific literature (Sillero and Barbosa, 2020), but they represent different conditions. Pseudo-absences are sampled from locations considered unsuitable for the species (Barbet-Massin et al., 2012). In contrast, background points encompass the full range of environmental conditions, including potential suitable locations for the species (presence locations; Phillips et al., 2009; Hallgren et al., 2019). The choice between pseudo-absences and background points indicates the user's uncertainty about the ecological preferences of the species, with background points used when there is no prior knowledge of unsuitable environmental conditions. Despite recognizing the distinction, we will use the term pseudo-absences for simplicity and alignment with our study.

The most common approaches for sampling pseudo-absences involve (i) randomly surveying a large number of points across the study area (e.g., 10,000; Barbet-Massin et al., 2012; Iturbide et al., 2015; Støa et al., 2019, Hysen et al., 2022) or (ii) sampling them

5

within or (iii) outside buffers created around presence locations (VanDerWal et al., 2009; Bedia et al., 2013). These approaches share the characteristic of deploying pseudo-absences randomly across the geographic space, which often leads to oversampling of the most common habitat conditions that are widespread in the study area (Tessarolo et al., 2014, 2021; Ronquillo et al., 2020). This sample location bias negatively impacts HSMs in multiple ways. Firstly, it can introduce a bias in the sampling of environmental conditions experienced by a species, potentially affecting the accurate estimation of the species response curve, particularly in heterogeneous areas (Austin 2007, Hortal et al., 2008; Albert et al., 2010; Beck et al., 2014). Secondly, it influences the predictive performance of HSMs, as reflected in the evaluation metrics used (Jiménez-Valverde et al., 2013; Sillero and Barbosa, 2020).

To overcome this issue, previous studies (Varela et al. 2014; Hattab et al., 2017) proposed to sample species presence and (true) absence data throughout a systematic sampling of the environmental conditions available across the study area, thus limiting the artificial constraint imposed by the random sampling towards the most widespread environments. More specifically, Varela et al. (2014), Hattab et al. (2017) and Perret and Sax (2022) suggested collecting species' presence and/or absence within 2- or 3-dimensional environmental spaces obtained using ordination techniques. Such approaches significantly contributed to the improvement and standardisation of the way species observations, including pseudo-absences, can be collected to calibrate HSMs reducing sample location bias. Yet, they do not explicitly consider class overlap, another relevant methodological issue encountered when collecting pseudo-absences through random sampling across the geographical space. Class overlap refers to the overlap between environmental conditions associated with species presence and absence, thus hindering the concept of pseudo-absences itself. It has negative effects on the predictive performance

6

of HSMs and it is particularly critical for machine learning techniques, while regression techniques such as Generalised Linear Models seem to be less affected (Barbet-Massin et al., 2012; Grimmett, Whitsed and Horta, 2020; Valavi et al., 2021). So far, class overlap has been addressed using resampling techniques more oriented to adjusting an unbalanced number of classes in the response variable (i.e., the 'up-' or 'down-sampling' approach; Valavi et al., 2021), irrespective of the technique to obtain pseudo-absences.

As far as we know, there are no approaches for sampling pseudo-absences that seek to mitigate both sample location bias and class overlap. Here, we present an alternative sampling strategy, which we called the 'uniform' approach, that builds upon existing strategies for systematically sampling the environmental space to select pseudo-absences. The novel aspect of the uniform approach is that, beyond reducing sample location bias, it also minimises class overlap by implementing a kernel-based filter that is used to delineate the portion of the environmental space where to collect pseudo-absences. To test our approach, we simulated 50 virtual species and compared the predictive performance of HSMs trained on pseudo-absences sampled using the uniform approach as well as other sampling strategies traditionally carried out within the geographical space: random (i.e., pseudo-absences randomly sampled within the geographical space) and buffer-out (i.e., pseudo-absences randomly collected outside buffers built around presence locations). To foster reproducibility, we provide an accompanying R package called `USE` (Uniform Sampling of the Environmental space), which bundles the R functions needed to implement the uniform approach. The package is available at https://github.com/danddr/USE. Finally, we provide a tutorial to explain how to apply the uniform approach to real case studies, using the European beech *Fagus sylvatica* L. as a target species.

7

2 Methods

2.1 Simulation of virtual species

We used virtual species (hereafter VS), a simulation tool that provides the great advantage of knowing the true generative process underlying the species geographical distribution (Meynard et al., 2019). We created the realised environmental space (*sensu* Jackson and Overpeck 2000) of 50 different virtual species using the bioclimatic variables gathered from the WorldClim database (www.worldclim.org; spatial resolution ~18.6 km at the Equator; Fick and Hijmans, 2017). We restricted the distribution of the simulated VS (and those of the bioclimatic variables) to the geographical extent spanning from -12° W to 25° E and from 36° to 60° N (approximately Western and Southern Europe) to significantly reduce the computational effort to process the entire workflow. Each VS was generated using a random set of five bioclimatic variables (out of the 19) through the function `generateRandomSp` from the R package `virtualspecies` (Leroy et al., 2016), which randomly assigns relationships between the VS and the bioclimatic variables (e.g., linear, quadratic relationships). This way, we obtained a raster layer reporting the habitat suitability index of each VS (HSI, Fig. 1a), which we then converted to a binary (i.e., presence-absence) map using the function `convertToPA`. Further details about parameters setting can be found in the R code available at https://github.com/danddr/USE_paper.

2.2 Sampling of the pseudo-absences

Regardless of the sampling approach and modelling technique used to calibrate the HSMs, the ratio between the number of presences and pseudo-absences in the training datasets

8

(i.e., sample prevalence) was kept equal to 1, which means that an equal number of presences and pseudo-absences were collected. In practice, each of the VS-specific training dataset included 300 presences, which were randomly sampled within the geographical extent using the function `sampleOccurrences` from the `virtualspecies` R package. Consequently, we collected an equal number of pseudo-absences according to the three sampling strategies presented below.

2.2.1 Uniform approach: pseudo-absences sampled within the environmental space

For each VS (i.e., iteration), we built a 2-dimensional environmental space by keeping the first two axes of a principal component analysis (PCA) performed on the correlation matrix of the five randomly selected bioclimatic variables used to generate the realised environment (Fig. 1b). Each time, we checked that the first two principal component axes accounted for at least 70% of the total bioclimatic variability. Then, we uniformly sampled pseudo-absences in the environmental space using the `uniformSampling` function. In short, each pseudo-absence is associated with a geographical location (i.e., a pixel of the environmental layers), which is in turn characterised by the set of environmental conditions encountered at that location. Such a combination of environmental conditions determines the position of the pseudo-absence within the environmental space. A pseudo-absence can thus be defined as the projection of a geographical location onto the environmental space generated through the PCA (i.e., a PC-score). Below, we present a step-by-step description of the uniform sampling performed by the function `paSampling`, which internally calls `uniformSampling` (both functions are included in the `USE` R package):

1. First, kernel density estimation (a statistical technique used to estimate the underlying probability distribution of a set of data points by smoothing them with a kernel

9

function; Scott, 1992) is used to calculate the probability density function of the presence data within the 2-dimensional environmental space. Similar uses of kernel density estimation have become popular in recent years, especially due to their increasing use in trait-based ecology to compute probabilistic hypervolumes and trait probability densities (Mammola and Cardoso, 2020 and reference therein). The PC-scores associated with a probability threshold equal to or greater than 0.75 (i.e., the default threshold value used in the `paSampling` function) are likely to bear environmental conditions associated with presence locations. Thus, we selected these presence locations and we generated the convex hull delimiting the portion of the environmental space mostly associated with this set of presence points within the environmental space (Fig. 1c). The kernel bandwidth (i.e., the width of the kernel density function that defines its shape) can be either defined by the user or automatically estimated by the function `paSampling`. In the latter case, the function uses a bandwidth selector by internally calling the function `Hpi` of the R package `ks` (Duong, 2021).

2. The portion of the environmental space defined by the above-mentioned convex hull is removed from the whole environmental space. Then, a sampling grid was generated from a pre-selected resolution (e.g., 10 × 10 cells) and overlaid on the 2-dimensional environmental space (Fig. 1d). The optimal resolution of the sampling grid within the environmental space can be determined using the function `optimRes` from the `USE` package. This function operates as follows:

- Within each cell of the sampling grid, the average (squared) Euclidean distance between the pseudo-absences (PC-scores) in the cell and the centroid of their convex hull is computed;

10

- Once this metric is computed across all cells of the sampling grid, the average mean value is computed across all cells (hereafter, grid average);
- The procedure above is separately repeated on different sampling grids of increasing resolution (i.e., increasing number of cells);
- The resulting set of grid averages (one per resolution) are used as a measure of the aggregation among pseudo-absences within the cells of the sampling grids. This value is compared across resolutions and the best grid is chosen as the one providing the best trade-off between resolution and average distance among points within cells (i.e., resolution that allows uniformly sampling the environmental space without overfitting it). More specifically, the best grid is the one whose resolution is just below that which would not allow the average distance among pseudo-absences to be reduced by more than 10% (other values can be set by the user).

3. Once the optimal resolution is set, the sampling grid is sequentially scanned (i.e., cell by cell) by the `uniformSampling` function called via the `paSampling` function and, from each grid cell, a given number of pseudo-absences is randomly collected. At this stage, the pseudo-absences associated with environmental conditions too close to those of the presence locations are already excluded (see step 1). Note that the pseudo-absences are randomly selected within the area of each cell of the sampling grid, and not at the centroid nor at the nodes.

The total number of pseudo-absences sampled within each cell of the sampling grid can be set by the user (using the argument `n.tr`, default `n.tr` = 5), who can also indicate a desired sample prevalence. If the sample prevalence is not specified, fewer pseudo-absences are likely to be eventually sampled than expected (i.e., `n.tr` × number of cells). This happens because (i) no pseudo-absence points are collected in empty cells, and (ii) less pseudo-

11

absence points than `n.tr` are available within the cells at the boundary of the environmental space (see zooming window in Figure 1d). Similarly, no pseudo-absences are collected within the core area of the presences (excluded in step 1). If a sample prevalence is set by the user, the sampling grid is surveyed until the chosen sample prevalence is reached by the algorithm.

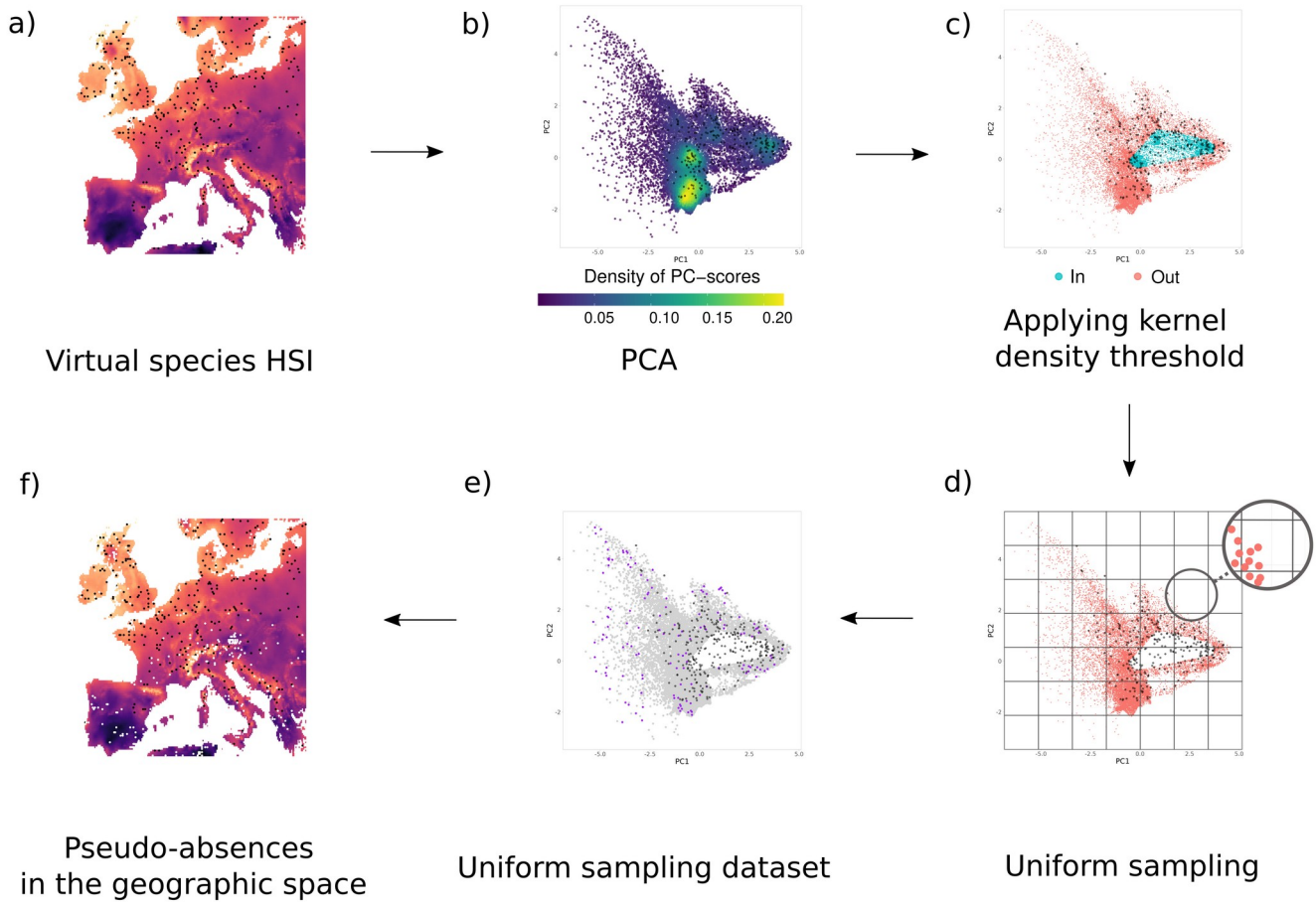


Figure 1: Flowchart representing the step-by-step procedure for implementing the uniform approach: a) habitat suitability index (HSI) of the *i-th* virtual species (VS; lighter colours indicate higher habitat suitability and black dots represent presence points in the geographical space); b) PCA performed on the environmental variables in the study region

12

(lighter colours indicate high PC-scores densities and black dots represent the presence points within the environmental space); c) application of the kernel-based filter, which splits the environmental space in two sub-spaces associated with either the environmental conditions more suitable for the species (in blue) or those associated with less/not suitable environmental conditions (in red; with black dots still depicting presence points); d) pseudo-absences are uniformly sampled across a sampling grid of a chosen resolution overlaid to the 2-dimensional environmental space. Specifically, pseudo-absences are sampled within each cell of the 2-d grid. The inset map shows an example of a grid cell at the boundary of the environmental space (i.e., a grid cell containing low density of pseudo-absences), black dots represent presence points; e) the purple dots represent the pool of randomly selected pseudo-absences after running the uniform sampling approach; f) the white dots represent the selected set of pseudo-absences after running the uniform sampling approach, but displayed in the geographical space this time, black dots still represent presence points from the focal VS.

2.2.2 Pseudo-absences sampled within the geographical extent

The sampling of pseudo-absences within the geographical extent was conducted using the random and buffer-out approaches. For the random approach (Barbet-Massin et al. 2012; Iturbide et al., 2015; Støa et al., 2019), we simply generated 300 random pseudo-absences across the studied geographical extent. For the buffer-out approach (Bedia et al., 2013), we created a buffer of 50 km radius around each presence location, and we then randomly sampled pseudo-absences outside the presence-specific buffers, but within the convex hull of the species geographical distribution (i.e., the convex hull that connects the outer presences of the species and thus delimits the range actually covered by the species in the geographical space). To test the effect of the length of the radius on the buffer-out

13

approach, we performed a sensitivity analysis on 10 VS using the following radius lengths: 50, 100 and 200 km.

2.3 Comparison among sampling strategies

2.3.1 Predictive performance comparison

For each of the 50 VS and for each of the three sampling strategies (i.e., uniform, random, buffer-out), we built a specific dataset combining the presence records with the pseudo-absences sampled within the environmental and the geographical space. First, we modelled the presence and pseudo-absences data as a function of the same five bioclimatic variables used to generate each of the 50 VS. To this aim, we randomly partitioned each dataset (specific for a sampling strategy) in 5 training (70% observations) and testing (30%) sets, which we used to calibrate and validate five modelling algorithms: (i) binomial generalised linear models with 'logit' link (GLMs); (ii) generalised additive models (GAMs); (iii) random forests (RFs); (iv) boosted regression trees (BRTs); and (v) MaxEnt. In total, we fitted 3,750 HSMs (50 VS species × 3 different sets of pseudo-absences × 5 modelling algorithms × 5 replicates of 70-30% partitions). To fit the HSMs, we used the R package `sdm` (Naimi and Araújo, 2016). Although we acknowledge the importance of fine-tuning HSMs (Fourcade, 2021), we kept model settings at their default value since it would have been unfeasible to individually parametrise each algorithm for all 50 VS and sampling strategies. A detailed representation of the workflow of the analyses is shown in Fig. 2. Furthermore, we acknowledge that our use of MaxEnt did not conform with the general recommendations for its adequate implementation (e.g., using 10,000 background points; Cobos et al., 2019; Kaas et al., 2021). Nonetheless, we included it in the comparison of models' performance due to its wide usage within the HSMs community.

After fitting HSMs for all the 50 VS, we compared the predictive performance

14

associated with each combination of sampling approaches and modelling techniques by computing the following metrics: (i) Area Under the receiver operating characteristic Curve (AUC); (ii) Continuous Boyce Index (CBI); (iii) sensitivity; (iv) specificity; (v) True Skill Statistics (TSS); and (vi) Root Mean Squared Error (RMSE). The RMSE was computed by comparing the true (i.e., simulated) habitat suitability of the focal VS against the one predicted by each combination of modelling and sampling approach. A detailed description of the above-mentioned modelling techniques and validation metrics can be found in Guisan et al. (2017). To compare the predictive performance of the HSMs fitted under different combinations of sampling strategy and modelling technique, we visually assessed the results of the 50 VS simulations using violin plots reporting the distribution of the values of the predictive performance metrics listed above. Furthermore, we tested for statistical differences among the predictive performance of the sampling strategies using Kruskall-Wallis tests, followed by Dunn's post hoc rank sum comparisons using the `dunn.test` R package (Dinno, 2017) (p-values for multiple comparisons adjusted using Holm correction).

To test the potential effect on our comparison of varying the number of bioclimatic variables used, we repeated the entire workflow on 50 VS using all the 19 bioclimatic variables for both the selection of pseudo-absences and as predictors in the five HSM algorithm. To test the potential effect on our comparison of varying sample prevalence, we repeated the entire workflow on 10 VS using two additional prevalence values, namely 0.5 and 0.1. Specifically, for each VS, we generated two additional training datasets with 300 presences, but we combined them with 600 and 3,000 pseudo-absences, to achieve sample prevalence of 0.5 and 0.1 respectively.

2.3.2 Sample location bias and class overlap

To assess the intensity of sample location bias associated with the different sampling

15

strategies, we extracted the pseudo-absences of a single VS and map their aggregation within the environmental space using bivariate density plots. The aim was to identify which, among the three sampling strategies, was more subject to oversampling particular environmental conditions within the geographical space. In principle, the sampling strategies more affected by sample location bias would exhibit a clear aggregation of pseudo-absences within the environmental space. We visually assessed the areas of the environmental space sampled by the different sampling strategies using the function `geom_density_2d` of the `ggplot2` R package (Wickham, 2016). This function performs a 2D kernel density estimation using the `kde2d` function of the `MASS` R package (Venables and Ripley, 2002) and displays the results with contours.

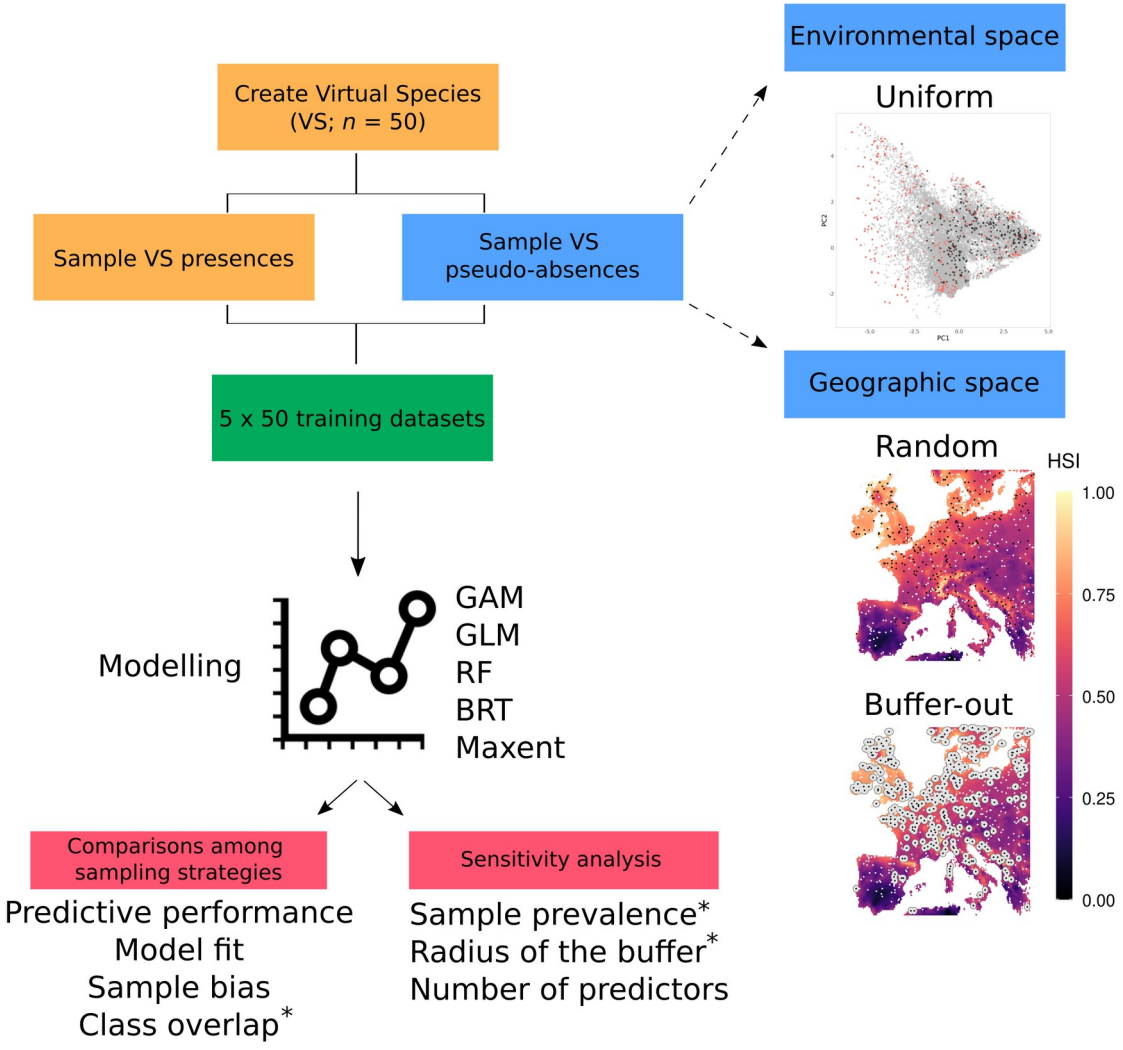
To assess the effectiveness of the uniform approach for mitigating class overlap, we simulated 10 new VS, sampled their presences and pseudo-absences using the three sampling strategies and mapped the position of the presence and pseudo-absence points within the environmental space following the procedure explained in section 2.2.1 and Figure 1a,b. Then, we computed the Gaussian hypervolume of the presences and pseudo-absences using the `hypervolumes` R package (Blonder, 2022), and calculated the overlap between them. Statistically significant differences in the degree of overlap were tested using one-way ANOVA and Tukey HSD test.

2.4 Real-case study

To illustrate how to apply the uniform approach with the `USE` R package, we modelled the realised distribution of *Fagus sylvatica* in Italy, France and Spain. We chose *F. sylvatica* as a target species because its distribution and biogeographic history is well-known across Europe (Magri et al., 2006; Poli et al., 2022). The whole analysis of *F. sylvatica* is described in S5, and the R code to replicate it can be found at: https://github.com/danddr/USE_paper.

16

352



353 **Figure 2** Overall workflow of the analysis described in the Methods section. The '*' is
354 associated with analyses (i.e., class overlap, sample prevalence, radius of the buffer)
355 performed on *n* = 10 VS.

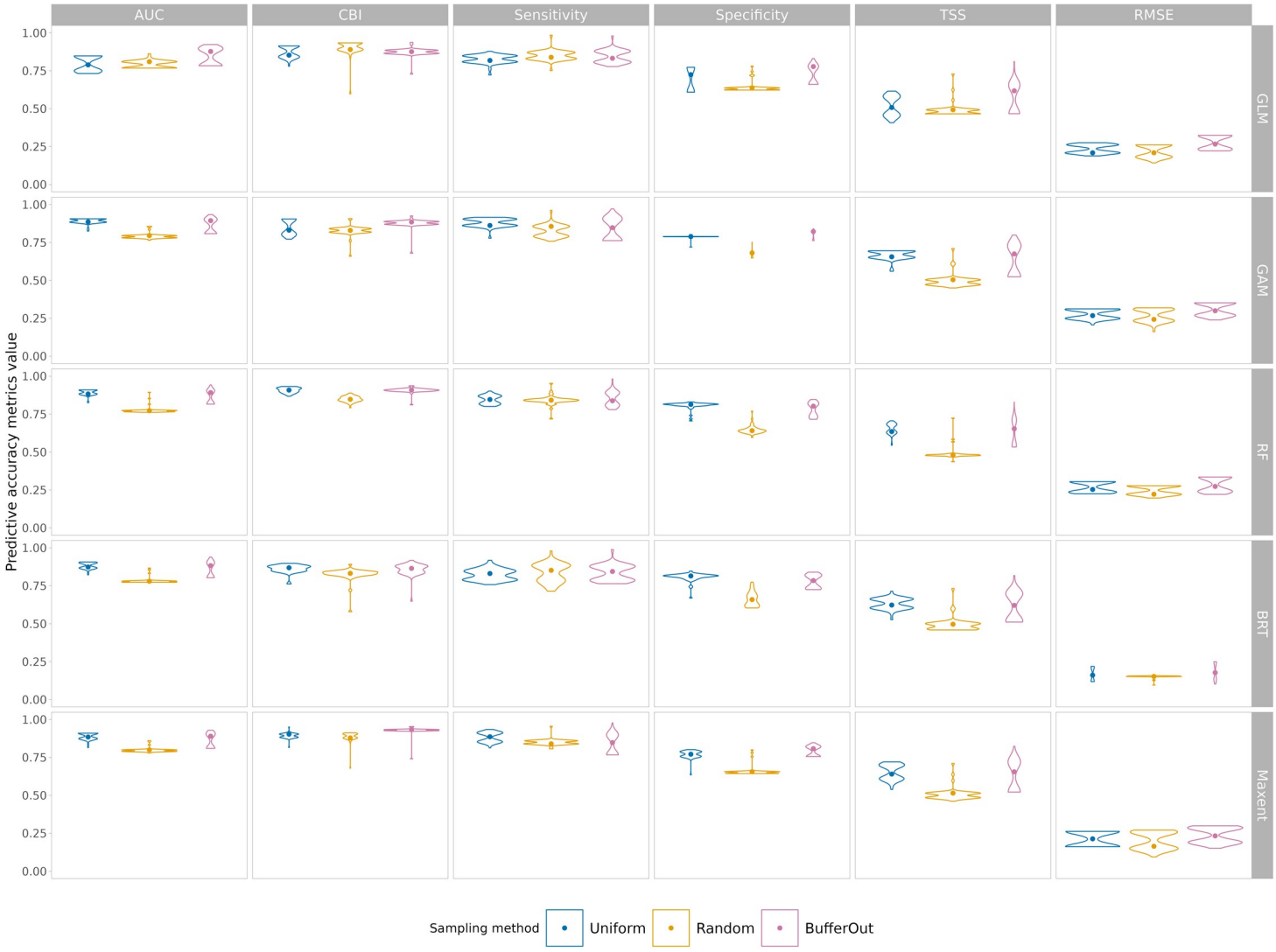
17

3 Results

3.1 Comparison of the predictive performance associated with geographical vs environmental sampling

Overall, the uniform approach performed equal to or better than the geographical approaches in terms of out-of-sample prediction (Fig. 3). Pairwise comparisons between the predictive accuracy performance of the random and buffer-out approaches against the uniform approach showed statistically significant differences in 73% and 53% of the combinations, respectively. However, these differences were algorithm- and metric-dependent and did not point to a higher predictive performance of the uniform approach (Fig. 3, Tab. S1, Fig. S1.1). The pattern of the differences among predictive performance metrics was consistent among prevalence values (Fig. S2.1-2.2) and number of bioclimatic variables used in the models (Fig. S3). Increasing the buffer radius length (Fig. S4), resulted in higher predictive performance of the buffer-out approach for some metrics (AUC, TSS, Specificity), while for CBI, Sensitivity and RMSE results remained comparable with those presented in Fig. 3.

18



371 **Figure 3**: Violin plots reporting the distribution of the values of the metrics of predictive
372 performance for the HSMs of the 50 VS, as modelled using 5 randomly selected bioclimatic
373 predictors and setting sample prevalence equal to 1 (i.e., same number of presences and pseudo-
374 absences). Dots represent median values of the metrics of predictive accuracy. Columns indicate
375 the different performance metrics, while rows are associated with the modelling techniques used to
376 fit the HSMs. Higher values in all metrics but RMSE reflect higher predictive performace. AUC =
377 Area Under the Curve; CBI = Continuous Boyce Index, TSS = True Skill Statistic; RMSE = Root
378 Mean Squared Error; GLM = Generalized Linear Model; GAM = Generalized Additive Model; RF =
379 Random Forest; BRT = Boosted Regression Trees.

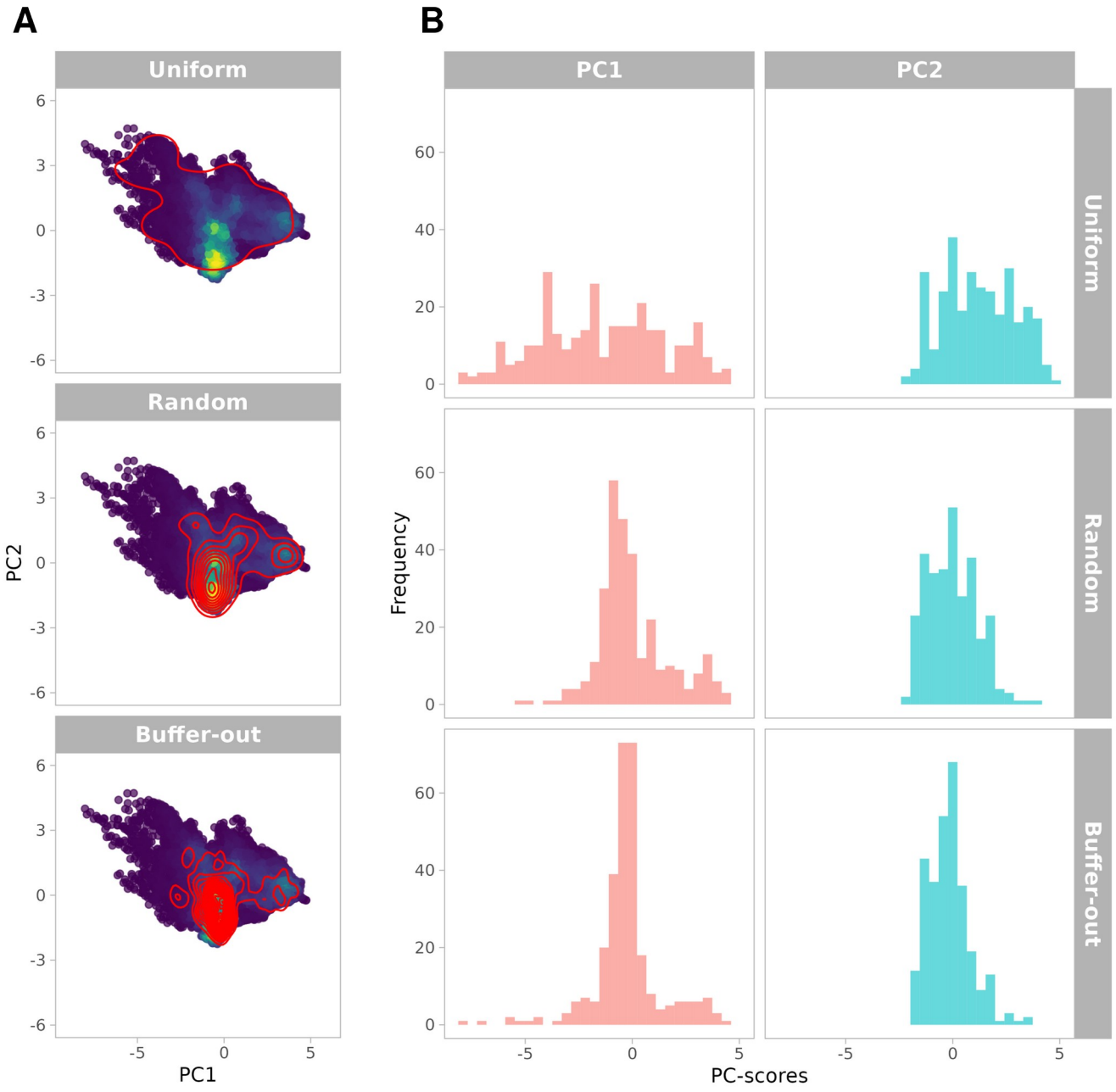
19

3.2 Effect of sample location bias and class overlap

The bivariate density plots of the pseudo-absences sampled within the environmental and geographical space highlighted that the uniform approach had the widest and most homogeneous coverage of environmental conditions throughout the environmental space (Fig. 4, see Figure S1.2 for a more detailed representation of the density of pseudo-absences sampled within the environmental space when running the uniform approach). In contrast, the random and buffer-out approaches appeared to be prone to sample location bias, with peaks of high density of pseudo-absences occurring in specific areas of the environmental space, i.e. those associated with the most frequent habitat conditions encountered within the geographical space (Fig. 4).

Regarding class overlap, we detected a statistically significant difference in the overlap between the portions of the environmental space occupied by presences and pseudo-absences (one-way ANOVA $F(2, 27) = 5.83$, p-value = 0.008). Specifically, the uniform approach exhibited the lowest overlap in comparison to the other sampling strategies (Fig. 5).

20



395 **Figure 4**: A) Bivariate plots showing the environmental space generated by a PCA carried out on 5
396 bioclimatic variables. Red lines represent the density of pseudo-absences for an individual VS, as
397 sampled by the random and buffer-out approaches within the geographical space, and by the
398 uniform approach within the environmental space. A more detailed representation of the density of
399 pseudo-absences sampled by the uniform approach is reported in Figure S1.2. B) Histograms

21

showing the frequency distribution of the first two principal components (columns) associated with the different sampling strategies (rows).

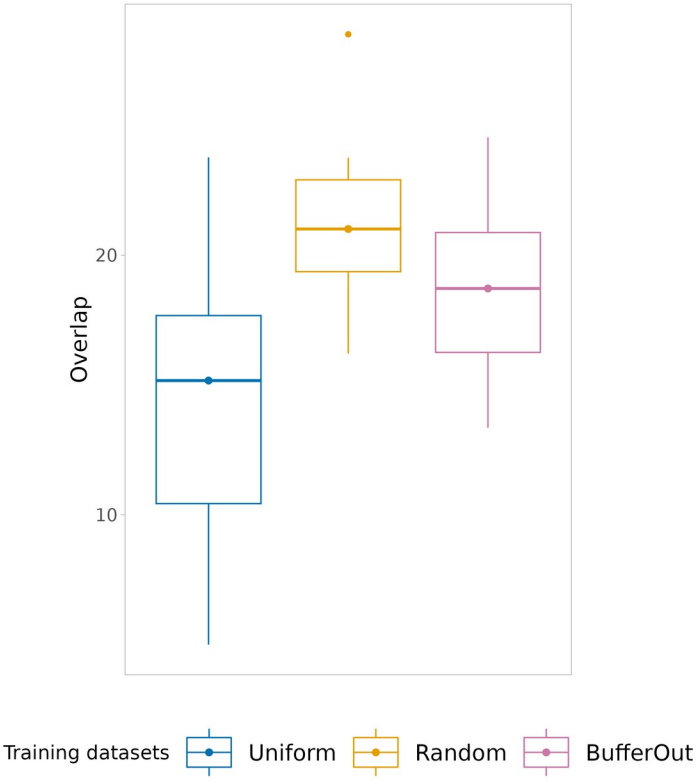


Figure 5: Box plots showing the overlap between environmental spaces generated by presences and pseudo-absences of the VS. Colours are associated with the three sampling strategies used to generate the pseudo-absences (uniform in blue, random in yellow and buffer-out in pink). Dots represent median values of overlap, as computed across 10 VS.

22

4 Discussion

In this study, we proposed the uniform approach as an alternative strategy to sample pseudo-absences within the environmental space. In contrast to existing techniques, our approach systematically samples pseudo-absences from portions of the environmental space excluding the conditions that are likely to be suitable for the species to establish. As a result, the uniform approach reduces the chance of including false-absences in the training dataset. From a more theoretical perspective, data collected after the application of the kernel-based filter are much closer to the concept of pseudo-absences than those obtained through traditional, geographical sampling approaches. Our findings show that the uniform approach represents a valid strategy for gathering pseudo-absences, resulting in out-of-sample predictive accuracy comparable to the sampling strategies implemented within the geographical space. In addition, the uniform sampling significantly reduces sample location bias and class overlap, which is critical to obtain ecologically meaningful pseudo-absences. Importantly, the uniform approach is flexible, as it allows the user to set parameters (e.g., kernel bandwidth, sample prevalence, sampling grid resolution) that control how pseudo-absences are sampled within the environmental space. Such flexibility is particularly valuable to mimic different ecological processes that are easier to capture within the environmental space than within the geographical space (e.g., source-sink dynamics). In all cases, by generating informative pseudo-absences, the uniform approach allows satisfying one of the most critical assumptions underpinning habitat suitability modelling: the need for adequate species distribution attributes (i.e., pseudo-absence data here) to model the species-environment relationship (Guisan et al., 2017).

4.1 Effect of the sampling approaches on models' predictive performances

23

Results of the VS simulations showed that the uniform approach performed well in terms of out-of-sample prediction regardless of the modelling technique, metric of predictive performance, and sample prevalence used. HSMs calibrated on pseudo-absences sampled with the uniform approach consistently showed high predictive performance, especially for the metrics related to the capacity of a model to correctly predict presences (i.e., sensitivity and CBI). Concerning the metrics associated with the models' ability to predict absences (e.g., specificity), the uniform sampling showed values comparable to the other strategies. This suggests that the uniform approach reduces omission error without necessarily increasing commission error. This is coherent with Fei and Yu (2016), who reported an increase in overall model predictive performance when pseudo-absences were systematically collected within the environmental space.

In this sense, results for the CBI, which is currently the go-to accuracy metric for validating HSMs fitted on pseudo-absences (or background points), and for the RMSE were particularly encouraging since the uniform approach scored, together with the buffer-out approach, the highest CBI values and lowest RMSE values across all modelling techniques. The high predictive performance associated with the uniform approach can be attributed to its two main underlying properties: the systematic sampling of the environmental space and the kernel-based filter on the presence data. In particular, the coherent trends of the predictive accuracy metrics (Fig. 3) and the environmental overlap among pseudo-absences collected through the different sampling approaches (Fig. 5) highlight the importance of the kernel-based filter for favouring the discrimination between the environmental features associated with presences and pseudo-absences.

Notwithstanding the positive results obtained in terms of predictive performance, we argue that a comparison of metrics of model predictive accuracy may not be the best means for

24

evaluating the adequacy of different sampling strategies carried out within the environmental rather than the geographical space. Indeed, previous studies showed that these metrics are affected by several factors, including sample prevalence (Guisan et al., 2017; Leroy et al., 2018), sample bias (Dubos et al., 2022, Rocchini et al., 2023) or the spatial extent of the study area (Lobo et al., 2008). Moreover, AUC and TSS tend to score high even in case of poor models calibrated on data exhibiting strong sample location bias (Fourcade et al., 2018, Jiménez-Valverde, 2021). Assessing HSMs predictive performance using a set of different predictive accuracy metrics might help the user to critically evaluate the outputs of the models.

4.2 Effect of the uniform sampling on sample location bias and class overlap

The uniform approach proved to significantly reduce sample location bias, since pseudo-absences were homogeneously scattered across the bivariate density plot of the two principal component axes (Fig. 4a,b, Fig. S1.2 in Supplementary Materials). On the contrary, the two sampling approaches carried out within the geographical space exhibited prominent peaks of density of pseudo-absences in correspondence with the most frequently encountered environmental conditions within the geographical space. As a consequence, the random and buffer-out approaches are likely to provide sub-optimal pseudo-absences for modelling the species-environment relationship (Thuiller et al. 2004; Austin 2007). This aspect gets increasingly relevant as environmental conditions are more heterogeneously distributed across the geographical space (e.g., in mountain regions with high topographic heterogeneity). Therefore, HSMs calibrated on training datasets adequately representing environmental variability rather than wide geographical coverage represent a crucial step to better capture and discriminate species niche breadth (Tessarolo et al., 2014, 2021; Varela

25

et al., 2014; Bazzichetto et al., 2022; Perret and Sax 2022).

The uniform approach proved to also significantly reduce class overlap. The `thres` argument passed to the `paSampling` function controls the portion of the environmental space associated with the species presence, thus inherently limiting class overlap by the exclusion of environmental conditions suitable to the species (see Fig. 1c, Fig. 5 and Fig. S1.3). This results in a set of pseudo-absences theoretically much closer to the species' true absences. Given that presence points are unevenly distributed within the environmental space, different kernel thresholds might also be used to handle the sampling of pseudo-absences under particular scenarios. As an example, setting a low kernel threshold would allow excluding accidental presences from unsuitable locations (e.g., 'sink populations') from the training dataset, while potentially including observations from these areas as pseudo-absences. Unfortunately, there is no *a priori* choice about the value of the threshold without having preliminary information on species' ecology, the study area and the goal of the research. For this reason, we provided the `thresh.inspect` function, which produces plots depicting the entire environmental space alongside the portion that would be excluded based on a specific kernel density threshold.

4.4 Limitations and usage notes

4.4.1 Limitations

The first limitation of the uniform approach, which is anyway a general limitation in HSMs (e.g., Cayuela et al., 2012), is that its effectiveness depends on the amount (sample size) and quality (e.g., geographically unbiased data *sensu* Fourcade 2014) of presence data. Indeed, if few presence data are available and/or presence data are geographically biased, the kernel-based filter might not accurately delimit the area associated with suitable

26

conditions for the species. As a consequence, the capacity of discriminating between suitable and unsuitable conditions of the uniform approach might be negatively affected.

A second limitation is that, although the uniform approach proved to be robust to varying sample prevalence, its effectiveness might diminish if a very large number of pseudo-absences is sampled (e.g., in case of low sample prevalence) (Fig. S2.1-2.2). Since the uniform approach samples a user-defined number of pseudo-absences within a grid overlaid to a bi-dimensional environmental space, if the number of pseudo-absences grows indefinitely, the advantage of the systematic sampling decreases. Indeed, oversampling the environmental space would generate datasets suffering from sample location bias as much as those based on the random sampling carried out within the geographical space.

From a more practical perspective, the uniform approach can currently operate only across 2-dimensional environmental spaces, but 3-dimensional spaces might be supported in the future.

Finally, although the idea behind `USE` and the uniform sampling approach is to provide users with an easy-to-use tool to generate more ecologically meaningful pseudo-absences, we acknowledge the existence of other techniques designed to avoid generating pseudo-absences altogether. Notable examples are point-process analyses (e.g., Isaac et al., 2020), which model the density of presence-only points per unit area, rather than the probability of presences and (pseudo-)absences. More recently, machine-learning methods based on isolation forests were also proposed, with the R package `ITSDM` specifically dedicated to HSMs (Song and Estes, 2023). We believe, however, that our approach provides a simpler and more intuitive way to deal with the issue of presence-only data, and thus has a lower threshold for end-users to implement in their workflow.

4.4.2 Usage notes

We here used the uniform approach to sample bioclimatic spaces, although we stress the

27

importance of not only using bioclimatic variables, but also information on soil, land-use as well as other relevant variables when modelling species distributions. Also, we invite potential users of the uniform sampling approach to always check that the first two axes of the principal component analysis used to generate the environmental space explains a large portion of the variance observed in the data (e.g., ≥ 70%). Equally important is the choice of the boundaries of the geographical extent for which the 2-dimensional space has to be generated. Indeed, to avoid the "there are no elephants in the Antarctic" paradox (Lobo et al., 2010), the spatial extent of the study area should be delineated so that it excludes geographical locations, and in turn environmental conditions, less suitable for the species (e.g., collecting pseudo-absences from Mediterranean coastal dunes when modelling the distribution of an alpine plant species). In short, the uniform approach can provide exhaustive information on where the species is likely to not occur, but it remains a responsibility of the end user to carefully verify if such information is ecologically meaningful.

5 Conclusion

In this study, we compared the predictive performance of two strategies for sampling pseudo-absences carried out within the geographical space with that of the uniform approach, which operated within the environmental space. Also, we compared geographical and environmental sampling approaches in terms of their vulnerability to sample location bias and class overlap. The uniform approach proved to have good predictive performances and to reduce sample location bias and class overlap, thereby representing a valid alternative to generate pseudo-absences for HSMs. We made the uniform approach openly available to the modellers community at https://github.com/danddr/USE.

28

6 Declaration

- Ethics approval and consent to participate: Not applicable.
- Competing interests: No conflict of interest has been declared by the authors.
- Funding: DDR was supported by a FRS-FNRS Belgian grant, ET is supported by the Estonian Research Council grant (MOBJD1030), MB acknowledges funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101066324.
- Authors' contribution: MB conceived the idea of the uniform approach and wrote the related R functions, while ET and DDR integrated the kernel density-based estimation of presences and the prevalence-related settings. DDR, ET and MB performed the simulations, analysed the data and assembled the `USE` R package. JL, JJL, SOV, and DR critically commented on the results of the analyses and their interpretation; DDR, ET and MB led the writing of the manuscript and produced a first draft, which was further improved by all other authors.
- Acknowledgments: The authors are grateful to Joaquin Hortal, who provided constructive feedback and commented on a previous version of this manuscript. Simulations were carried out using the facilities of the High-Performance Computing Center of the University of Tartu.

29

7 Code and Data availability

The scripts for replicating the analyses presented in this paper are available at https://github.com/danddr/USE_paper, as well as all the raw outputs of the simulations and statistical analyses (which are available as an .RDS file).

We provide a general tutorial to explain how to apply the USE package at https://danddr.github.io/USE/articles/USE_vignette.html. In addition, we provide a tutorial on how to apply the uniform approach based on a real species (the European beech, *Fagus sylvatica* L.) in S5. The R script related to the tutorial is available at https://github.com/danddr/USE_paper.

30

References

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

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

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, *200*(1-2), 1-19.

Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3(2):327–338.

Baker, D. J., Maclean, I. M. D., Goodall, M., & Gaston, K. J. (2022). Correlations between spatial sampling biases and environmental niches affect species distribution models. Global Ecology and Biogeography, 00, 1– 13.

Batista, E., Lopes, A., Miranda, P., & Alves, A. (2023). Can species distribution models be used for risk assessment analyses of fungal plant pathogens? A case study with three Botryosphaeriaceae species. *European Journal of Plant Pathology*, 165(1), 41-56.

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., Barták, V. & Sperandii, M. G. (2022). Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. EcoRxiv https://doi.org/10.32942/osf.io/rhys3

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., and Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19:10–15.

Bedia, J., Herrera, S., and Gutiérrez, J. M. (2013). Dangers of using global bioclimatic datasets for ecological niche modeling. limitations for future climate projections. *Global and Planetary Change*, 107:1–12.

Blonder B, Morrow wcfCB, Harris DJ, Brown S, Butruille G, Laini A, Chen D (2022). _hypervolume: High Dimensional Geometry, Set Operations, Projection, and Inference Using Kernel Density Estimation, Support Vector Machines, and Convex Hulls_. R package version 3.0.4, <https://CRAN.R-project.org/package=hypervolume>.

Booth, T. H., Nix, H. A., Busby, J. R., and Hutchinson, M. F. (2014). Bioclim: the first species distribution modelling package, its early applications and relevance to most

current maxent studies. *Diversity and Distributions*, 20(1):1–9.

Cayuela, L., Golicher, D. J., Newton, A. C., Kolb, M., De Alburquerque, F. S., Arets, E. J. M. M., Alkemade, J. R. M. & Pérez, A. M. (2009). Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science*, 2(3), 319-352.

Cobos, M. E., Peterson, A. T., Barve, N., & Osorio-Olvera, L. (2019). kuenm: an R package for detailed development of ecological niche models using Maxent. *PeerJ*, 7, e6281.

Da Re, D., Tordoni, E., De Pascalis, F., Negrín-Pérez, Z., Fernández-Palacios, J. M., Arévalo, J. R., ... & Bacaro, G. (2020). Invasive fountain grass (*Pennisetum setaceum* (Forssk.) Chiov.) increases its potential area of distribution in Tenerife island under future climatic scenarios. *Plant Ecology*, *221*(10), 867-882.

Dinno, A. (2017). *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*. R package version 1.3.5,https://CRAN.R-project.org/package=dunn.test.

Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., ... & Luque, S. (2022). Assessing the effect of sample bias correction in species distribution models. Ecological Indicators, 145, 109487.

Duffy, G. A., Coetzee, B. W., Latombe, G., Akerman, A. H., McGeoch, M. A., & Chown, S. L. (2017). Barriers to globally invasive species are weakening across the Antarctic. *Diversity and Distributions*, *23*(9), 982-996.

Duong, T. (2021). *ks: Kernel Smoothing*. R package version 1.13.3.

Fei, S. and Yu, F. (2016). Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. *Landscape Ecology*, 31(1):31–42.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian Obama nungara. *Ecological Modelling*, *457*, 109686.

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

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS ONE*, 9(5), e97122.

Grimmett, L., Whitsed, R., & Horta, A. (2020). Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling*, *431*, 109194.

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

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., and Mackey, B. (2019). Species

distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, 408:108719.

Hattab, T., Garzón-López, C. X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., et al. (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.

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., and Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117(6):847–858.

Hysen, L., Nayeri, D., Cushman, S., & Wan, H. Y. (2022). Background sampling for multi-scale ensemble habitat selection modeling: Does the number of points matter?. *Ecological Informatics*, 72, 101914.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., ... & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in ecology & evolution*, *35*(1), 56-67.

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., and Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312:166–174.

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

Jarvie, S., & Svenning, J. C. (2018). Using species distribution modelling to determine opportunities for trophic rewilding under future scenarios of climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1761), 20170446.

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

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, *14*(6), 885-890.

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

Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., Soley-Guardia, M., & Anderson, R. P. (2021). ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution*, 12(9), 1602-1608.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., and Bellard, C. (2018). Without quality presence–absence data, discrimination metrics such as tss can be misleading measures of model performance. *Journal of Biogeography*,

45(9):1994–2002.

Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2016). virtualspecies, an r package to generate virtual species distributions. *Ecography*, 39(6):599–607.

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., Jiménez-Valverde, A., and Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145– 151.

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., ... & De Beaulieu, J. L. (2006). A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist*, 171(1), 199-221.

Mammola, S. and Cardoso, P. (2020). Functional diversity metrics using kernel density n-dimensional hypervolumes. *Methods in Ecology and Evolution*, 11(8):986–995.

Meynard, C. N., Leroy, B., and Kaplan, D. M. (2019). Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? *Ecography*, 42(12):2021–2036.

Naimi, B. and Araújo, M. B. (2016). sdm: a reproducible and extensible r platform for species distribution modelling. *Ecography*, 39(4):368–375.

Newbold, T. (2018). Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proceedings of the Royal Society B*, 285(1881):20180792.

Perret, D. L. and Sax, D. F. (2022). Evaluating alternative study designs for optimal sampling of species' climatic niches. *Ecography*.

Poli et al. (2022) Coupling fossil records and traditional discrimination metrics to test how genetic information improves species distribution models of the European beech Fagus sylvatica. *European Journal of Forest Research*, 141: 253–265

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., and Blair, M. E. (2017). Opening the black box: An open-source release of maxent. *Ecography*, 40(7):887–893.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rocchini, D., Tordoni, E., Marchetto, E. *et al.* A quixotic view of spatial bias in modelling the distribution of species and their diversity. *npj biodivers* **2**, 10 (2023). https://doi.org/10.1038/s44185-023-00014-6

Ronquillo, C., Alves-Martins, F., Mazimpaka, V., Sobral-Souza, T., Vilela-Silva, B., Medina,

34

N. G., and Hortal, J. (2020). Assessing spatial and temporal biases and gaps in the publicly available distributional information of iberian mosses. *Biodiversity Data Journal*, 8.

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., and Huijbregts, M. A. (2021). Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions*, 27(6):1035–1050.

Scott, D.W. (1992) Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons

Sillero, N. and Barbosa, A. M. (2020). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, pages 1–14.

Song, L., & Estes, L. (2023). ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*, 14(3), 831-840.

Støa, B., Halvorsen, R., Stokland, J. N., and Gusarov, V. I. (2019). How much is enough? influence of number of presence observations on the performance of species distribution models. *Sommerfeltia*, 39(1):1–28.

Svenning, J.-C. and Skov, F. (2004). Limited filling of the potential range in European tree species. *Ecology Letters*, 7(7):565–573.

Tessarolo, G., Lobo, J. M., Rangel, T. F., and 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., and Hortal, J. (2014). Uncertainty associated with survey design in species distribution models. *Diversity and Distributions*, 20(11):1258–1269.

Thuiller, W., Brotons, L., Araújo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, 27(2), 165– 172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742.

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know?. *Ecological Modelling*, 220(4), 589-594.

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

Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Wasof et al. (2015) Disjunct populations of European vascular plant species keep the same

767 climatic niches. *Global Ecology and Biogeography*, 24: 1401-1412

768 Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN
769 978-3-319-24277-4, https://ggplot2.tidyverse.org.

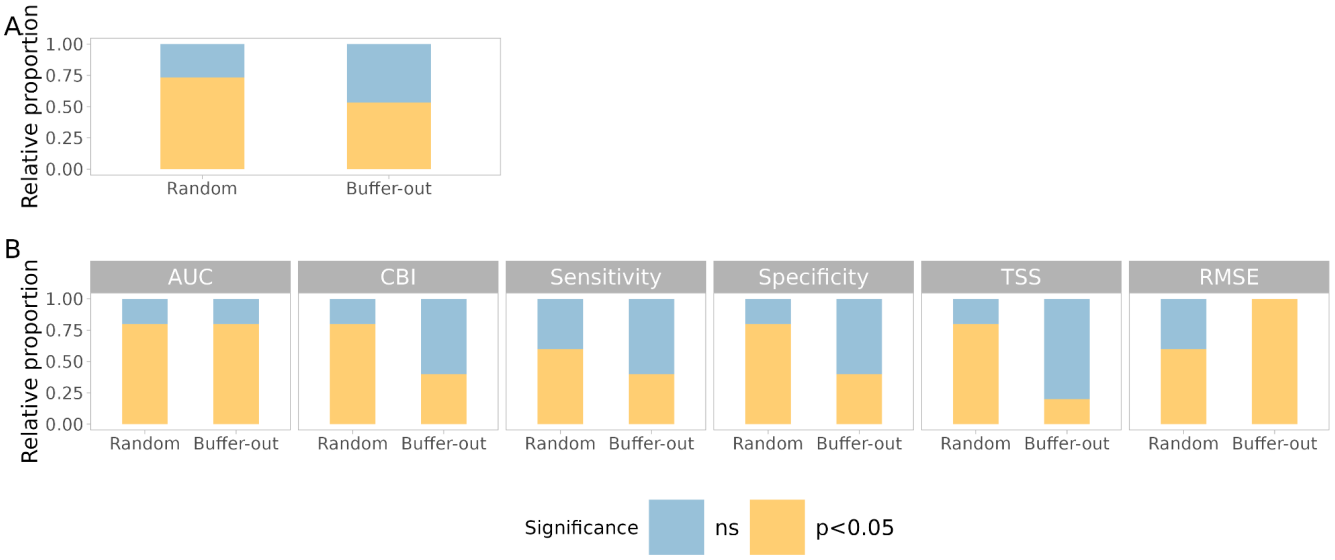
Supplementary Material 1

Tab. S1: Post-hoc multiple comparisons with Dunn's rank sum test (α = 0.05; omnibus test was always significant with P < 0.05, data not shown). All the comparisons were performed comparing the Uniform dataset with the other different sampling strategies. P-values were adjusted using Holm correction.

Model	Metric	Comparisons	χ2	P.val
BRT	AUC	BufferOut - Uniform	93.89	ns
BRT	AUC	Random - Uniform	93.89	p<0.05
BRT	CBI	BufferOut - Uniform	46.45	ns
BRT	CBI	Random - Uniform	46.45	p<0.05
BRT	RMSE	BufferOut - Uniform	6.08	p<0.05
BRT	RMSE	Random - Uniform	6.08	ns
BRT	Sensitivity	BufferOut - Uniform	1.97	ns
BRT	Sensitivity	Random - Uniform	1.97	ns
BRT	Specificity	BufferOut - Uniform	91.85	p<0.05
BRT	Specificity	Random - Uniform	91.85	p<0.05
BRT	TSS	BufferOut - Uniform	87.43	ns
BRT	TSS	Random - Uniform	87.43	p<0.05
GAM	AUC	BufferOut - Uniform	101.78	p<0.05
GAM	AUC	Random - Uniform	101.78	p<0.05
GAM	CBI	BufferOut - Uniform	29.26	p<0.05
GAM	CBI	Random - Uniform	29.26	p<0.05
GAM	RMSE	BufferOut - Uniform	25.43	p<0.05
GAM	RMSE	Random - Uniform	25.43	ns
GAM	Sensitivity	BufferOut - Uniform	28.82	p<0.05
GAM	Sensitivity	Random - Uniform	28.82	p<0.05
GAM	Specificity	BufferOut - Uniform	101.68	ns
GAM	Specificity	Random - Uniform	101.68	p<0.05
GAM	TSS	BufferOut - Uniform	91.55	ns

Model	Metric	Comparisons	χ2	P.val
GAM	TSS	Random - Uniform	91.55	p<0.05
GLM	AUC	BufferOut - Uniform	27.47	p<0.05
GLM	AUC	Random - Uniform	27.47	ns
GLM	CBI	BufferOut - Uniform	44.97	ns
GLM	CBI	Random - Uniform	44.97	p<0.05
GLM	RMSE	BufferOut - Uniform	35.25	p<0.05
GLM	RMSE	Random - Uniform	35.25	p<0.05
GLM	Sensitivity	BufferOut - Uniform	8.18	ns
GLM	Sensitivity	Random - Uniform	8.18	p<0.05
GLM	Specificity	BufferOut - Uniform	53.85	p<0.05
GLM	Specificity	Random - Uniform	53.85	ns
GLM	TSS	BufferOut - Uniform	25.54	p<0.05
GLM	TSS	Random - Uniform	25.54	ns
Maxent	AUC	BufferOut - Uniform	101.78	p<0.05
Maxent	AUC	Random - Uniform	101.78	p<0.05
Maxent	CBI	BufferOut - Uniform	90.22	p<0.05
Maxent	CBI	Random - Uniform	90.22	ns
Maxent	RMSE	BufferOut - Uniform	31.96	p<0.05
Maxent	RMSE	Random - Uniform	31.96	p<0.05
Maxent	Sensitivity	BufferOut - Uniform	24.26	p<0.05
Maxent	Sensitivity	Random - Uniform	24.26	p<0.05
Maxent	Specificity	BufferOut - Uniform	88.04	ns
Maxent	Specificity	Random - Uniform	88.04	p<0.05
Maxent	TSS	BufferOut - Uniform	89.73	ns
Maxent	TSS	Random - Uniform	89.73	p<0.05
RF	AUC	BufferOut - Uniform	100.83	p<0.05
RF	AUC	Random - Uniform	100.83	p<0.05
RF	CBI	BufferOut - Uniform	93.99	ns

Model	Metric	Comparisons	χ2	P.val
RF	CBI	Random - Uniform	93.99	p<0.05
RF	RMSE	BufferOut - Uniform	31.28	p<0.05
RF	RMSE	Random - Uniform	31.28	p<0.05
RF	Sensitivity	BufferOut - Uniform	0.21	ns
RF	Sensitivity	Random - Uniform	0.21	ns
RF	Specificity	BufferOut - Uniform	97.83	ns
RF	Specificity	Random - Uniform	97.83	p<0.05
RF	TSS	BufferOut - Uniform	87.93	ns
RF	TSS	Random - Uniform	87.93	p<0.05



S1.1: Post-hoc multiple comparisons with Dunn's rank sum test (α = 0.05; omnibus test was always significant with P < 0.05, data not shown). All the comparisons were performed comparing the Uniform dataset with the other different sampling strategies: A) relative proportion of the significant comparisons aggregated by sampling strategy; B) relative proportion of the significant comparisons aggregated by sampling strategy and metric.

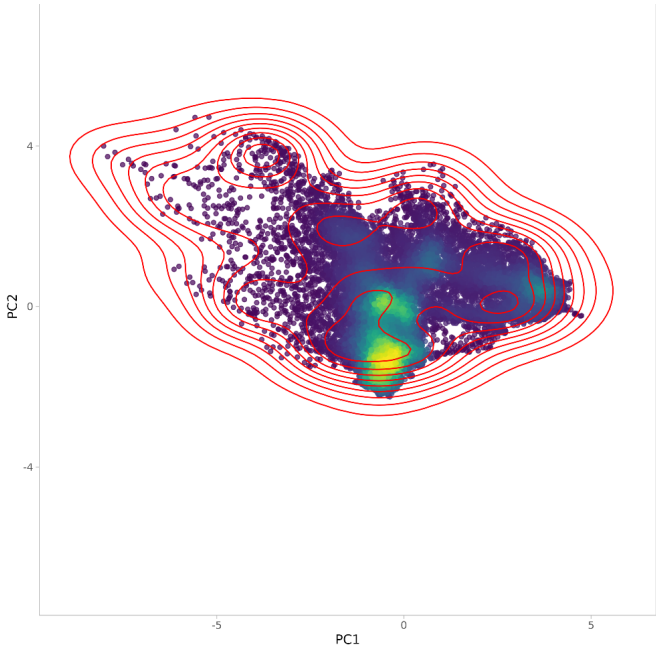


Figure S1.2: Virtual Species PC-scores bivariate plot for the Uniform approach only.

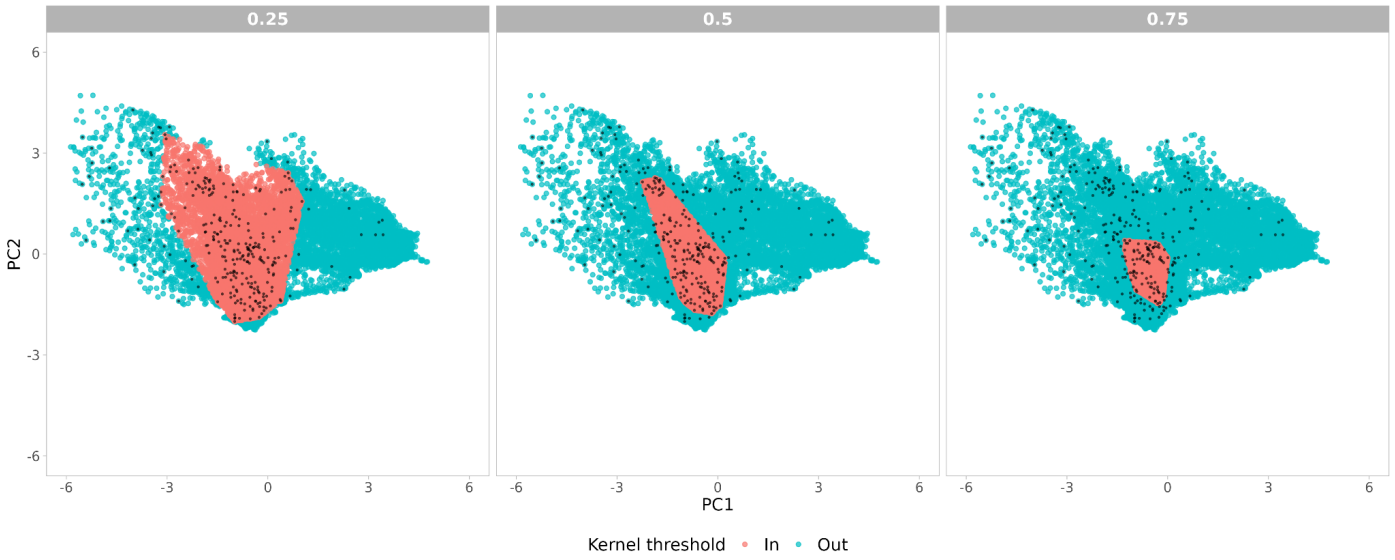


Figure S1.3: Effect of setting different kernel thresholds on the inclusion/exclusion of pseudo-absences eventually sampled using the uniform approach (black dots are the true VS presences displayed within the environmental space). Setting a low value of the kernel threshold (e.g., 0.25) increases the portion of the environmental space excluded from the uniform sampling, while, on the contrary, setting a high value of the kernel threshold increases the portion of the environmental space available for the uniform sampling.

Supplementary Material 2

To test the potential effect of different sample prevalence, we also repeated the entire workflow on 10 VS with two different prevalence values. Specifically, in both cases we kept a training dataset consisting of 300 presences but we used alternatively 600 and 3,000 pseudo-absences (sample prevalence = 0.5 and sample prevalence = 0.1, respectively).

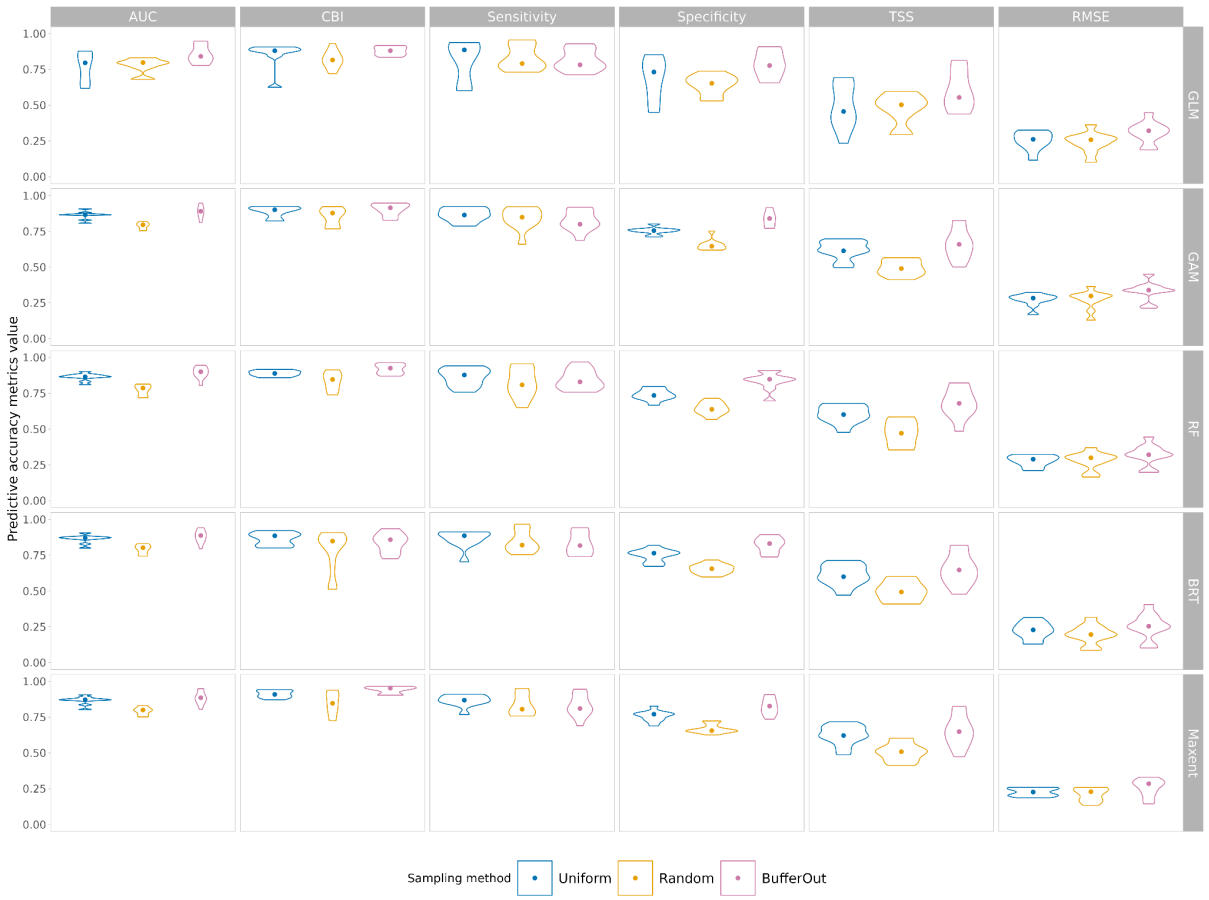


Figure S2.1: Violin plots reporting the distribution of the values of the metrics of predictive performance for the HSMs of the 10 VS (the dots are the median values), considering 5 predictors, and using a sample prevalence = 0.5. Columns indicate the different performance metrics while rows report the modelling techniques used to compute HSMs.

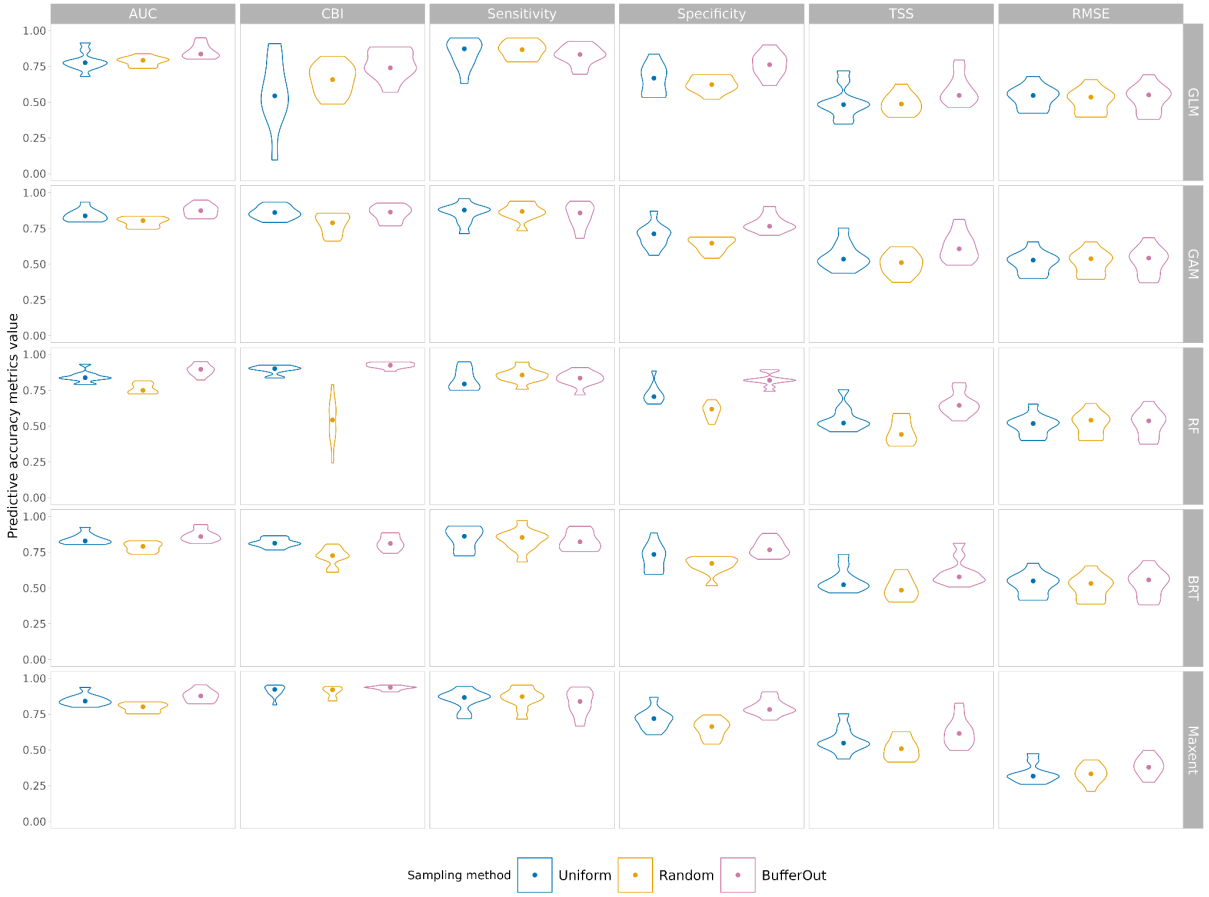


Figure S2.2: Violin plots reporting the distribution of the values of the metrics of predictive performance for the HSMs of the 10 VS (the dots are the median values), considering 5 predictors, and using a sample prevalence = 0.1. Columns indicate the different performance metrics while rows report the modelling techniques used to compute HSMs.

Supplementary Material 3

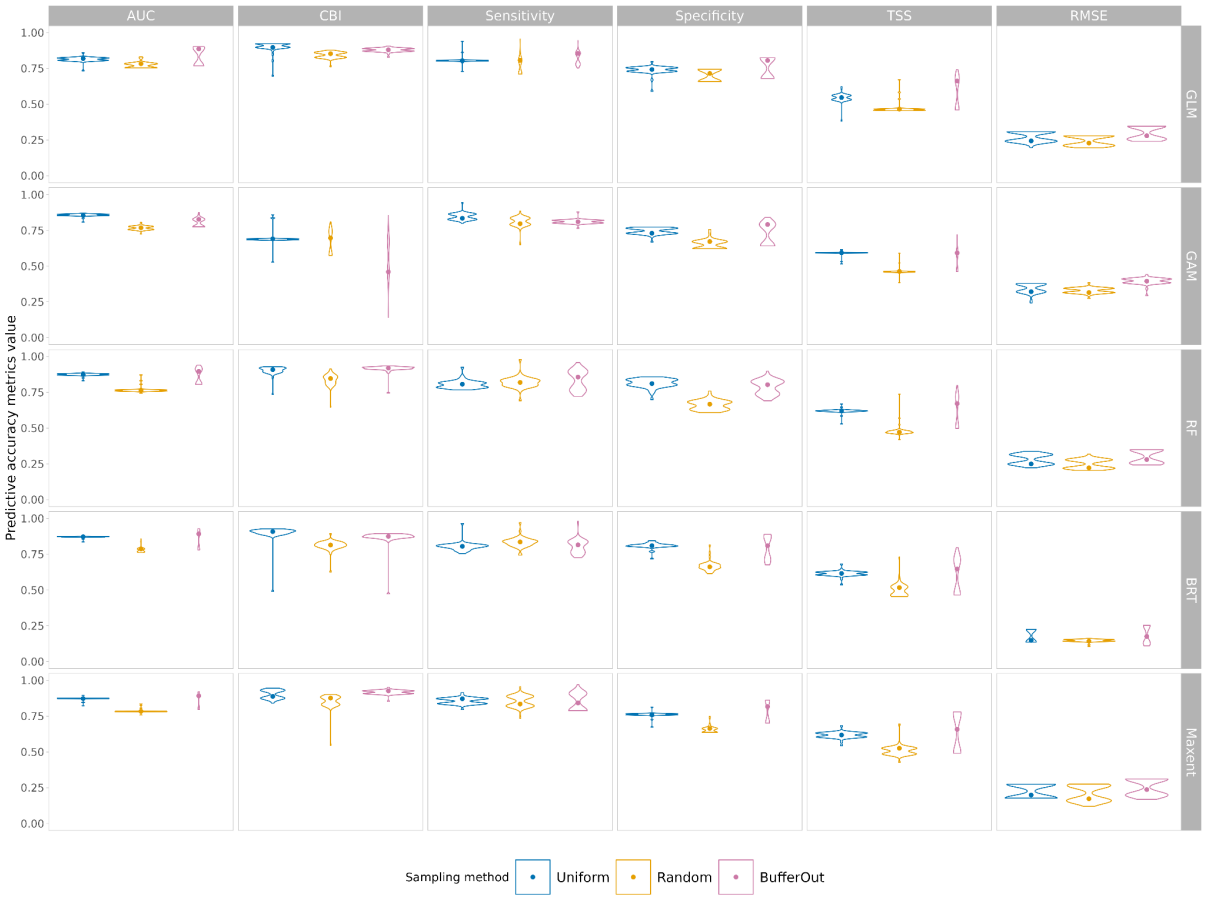


Fig. S3.1: Violin plots reporting the distribution of the values of the metrics of predictive performance for the HSMs of the 50 VS modelled as a function of the 19 bioclimatic predictors, and setting sample prevalence equal to 1 (i.e., same number of presences and pseudo-absences). Dots represent median values of the metrics of predictive accuracy. Columns indicate the different performance metrics, while rows the modelling techniques used to compute HSMs. AUC = Area Under the Curve; CBI = Continuous Boyce Index, TSS = True Skill Statistic; RMSE = Root Mean Squared Error; GLM = Generalized Linear Model; GAM = Generalized Additive Model; RF = Random Forest; BRT = Boosted Regression Trees.

Supplementary Material 4

To test the potential effect of different radius sizes on the buffer-out approach, we also repeated the entire workflow on 10 VS with three different radius sizes: 50, 100 and 200 km. Specifically, we kept the training dataset with a sample prevalence = 1, consisting of 300 presences and 300 pseudo-absences.

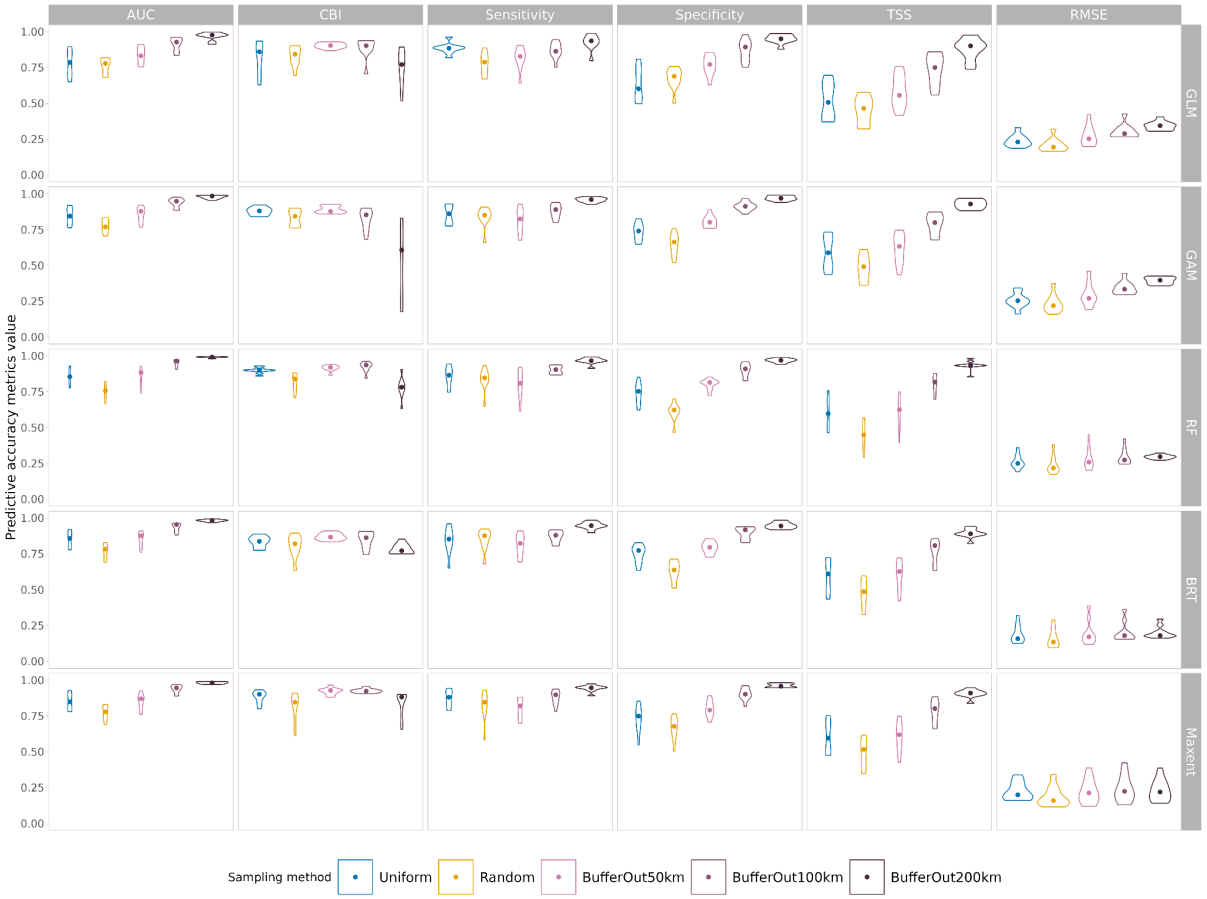


Fig. S4.1: Violin plots reporting the distribution of the values of the metrics of predictive performance for the HSMs of the 10 VS modelled as a function of 5 bioclimatic predictors, and setting sample prevalence equal to 1 (i.e., same number of presences and pseudo-absences). We varied the size of the radius for the buffer-out approach, namely: 50, 100 and 200 km. Dots represent median values of the metrics of predictive accuracy. Columns indicate the different performance metrics, while rows the modelling techniques used to compute HSMs. AUC = Area Under the Curve; CBI = Continuous Boyce Index, TSS = True Skill Statistic; RMSE = Root Mean Squared Error; GLM = Generalised Linear Model; GAM = Generalised Additive Model; RF = Random Forest; BRT = Boosted Regression Trees.

Supplementary Material 5: case study on the realised distribution of *Fagus sylvatica* in Western Europe

Methods

To illustrate how to apply the uniform approach with the `USE` R package, we modelled the realised distribution of *Fagus sylvatica* in Italy, France and Spain (hereafter, western Europe). We chose *F. sylvatica* as an example species because its distribution and biogeographic history is well-known across Europe (Magri et al., 2006; Poli et al., 2022). For the sake of simplicity, we restricted the area of investigation to western Europe and used two modelling algorithms. Indeed, the case study of *F. sylvatica* is only intended as a practical example to show how the `USE` package operates, while not providing a further comparison on the predictive performance of HSMs fitted on data collected through different sampling strategies (as already done using virtual species, see main manuscript). We gathered data on the presence of *F. sylvatica* from the open EU-Forest dataset (Mauri et al., 2017), which compiles observations on European tree species from national inventories and other similar sources (see Mauri et al., 2017 for further information about EU-Forest). EU-Forest data consist of presence records of tree species exhaustively collected across Europe, and then aggregated to a 1 × 1 km resolution grid. This lets us assume with a certain degree of confidence that the EU-Forest dataset provided a geographically unbiased sample of presence records for *F. sylvatica* in western Europe.

Across our study area, the EU-Forest dataset included a total of 12,444 presence records for *F. sylvatica*, which we sub-sampled within the environmental space to retrieve both a training and a testing (for internal validation) presence dataset. To this aim, we generated a 2-dimensional environmental space using all 19 bioclimatic variables available from WorldClim. Then, we used the function `USE::uniformSampling` to uniformly sample presence records within the environmental space. Note that this approach is conceptually similar to the spatial-thinning proposed by Aiello-Lammens et al. (2015), which aims at reducing the clustering of presences within the geographical space (Sillero and Barbosa, 2020), except that here we applied it within the environmental space. The obtained training and testing presence datasets were then combined to obtain the training and testing pseudo-absence datasets using the `paSampling`

function from the USE package. In particular, all presence records available for *Fagus sylvatica* were used to recover the core area of the species' bioclimatic niche within the environmental space. This allowed filtering out the pseudo-absences likely associated with suitable locations for the species (see step 1 in section 2.2.1 in the main text). The final sample size of the pseudo-absences included in the training and testing (internal validation) datasets were 1,826 and 991, respectively. Note that the sample size of the presence data included in the training and testing datasets were 1,827 and 991, respectively. Also note that prevalence was fixed to approx. 1 in both the training and testing dataset.

Finally, we derived a completely independent testing (external validation) dataset using presence and true absence data from sPlotOpen (Sabatini et al., 2021). The sPlotOpen dataset is an open-access subset of sPlot, one of the most comprehensive global databases of vegetation records (Sabatini et al., 2021). Here, we used sPlotOpen to gather *F. sylvatica* presences (*n* = 367), and to derive true absence data from those vegetation plots where *F. sylvatica* was not recorded (*n* = 4,162). As done for the EU-Forest dataset, we considered only sPlotOpen vegetation plots occurring in western Europe (i.e., Italy, France and Spain).

Then, we modelled the realised distribution of *F. sylvatica* as a function of a set of WorldClim bioclimatic variables. For simplicity, we solely focused on the climatic niche of *Fagus sylvatica*, although we acknowledge that other drivers than climate equally contribute in shaping the distribution of this species, especially so at local scales (Mellert et al., 2018). As modelling techniques, we used a 'logit' link binomial generalised linear model (binomial GLM) and random forests (RF, fitted using `ranger::ranger`; Wright and Ziegler, 2017). To reduce multicollinearity, we selected a subset of the 19 bioclimatic variables using the R function `caret::findCorrelation` function (Kuhn, 2021) (setting the pairwise-correlation threshold to 0.6). The bioclimatic variables eventually kept to fit the HSM for *F. sylvatica* were: BIO6 (minimum temperature of the coldest month); BIO7 (temperature annual range); BIO8 (mean temperature of the wettest quarter). Also, we used the latitudinal position of the presence and pseudo-absence records (hereafter, latitude) as an additional predictor to account for the effect of factors affecting the latitudinal gradient of the distribution of *F. sylvatica* that were not included in the model. An example of such factors is the

species biogeographic history of post-glacial recolonization towards northern Europe (Magri et al., 2006). To account for non-linearity in the profile of Pearson's residuals and improve the fit of the binomial GLM, we introduced second order polynomial terms for BIO6, BIO7 and latitude. The predictive performance of the fitted models was assessed on three different types of data: (i) the (internal) testing dataset derived from the EU-Forest dataset; (ii) 5 partitions of the training dataset (i.e., a 5-fold cross-validation); and (iii) the independent (external) testing dataset derived from sPlotOpen. As predictive accuracy metrics, we used the true skill statistics (TSS) and the continuous Boyce index (CBI). A TSS value greater than 0.5 is often considered to indicate good predictions. Positive values of CBI indicate that presences predicted by the model are consistent with the distribution of presences in the testing dataset. On the contrary, TSS and CBI values close to zero indicate that the model does not perform differently from a model that randomly predicts presences and absences. Finally, negative values of the CBI indicate counter predictions, i.e., predicting low suitability in areas with high density of presence records (Hirzel et al. 2006).

Beyond model predictive metrics, we computed the following measures of goodness-of-fit: Tjur's R^2 for the binomial GLM and the R^2 for the RF.

A full description of the modelling procedure (from the sub-sampling of the presences and the collection of pseudo-absences to the assessment of the model predictive performance) can be found at https://github.com/danddr/USE_paper/tree/main/Example.

Results

Both the binomial GLM and the RF for *F. sylvatica* showed high predictive performances, regardless of the dataset used for testing (Table S5.1). Concerning the binomial GLM, the TSS was always equal to or above 0.41, with the lowest value obtained for the sPlotOpen testing dataset (0.41) and the highest for the EU-Forest dataset (0.61). Similarly, the lowest CBI was scored for the sPlotOpen dataset (0.88), while the highest for the EU-Forest dataset (0.99).

We obtained comparable results for the RF, with the lowest TSS obtained when using sPlotOpen as a testing dataset (0.52), while the EU-Forest dataset and the (average across) 5-fold cross validation resulted in TSS equal to 0.79 and 0.77, respectively. With respect to the CBI, the highest value was observed

for the EU-Forest dataset (0.99), while the lowest was obtained using the sPlotOpen dataset (0.93).

Goodness-of-fit measures seemed to be affected by the modelling technique, with the R^2 of the RF being 0.66, and the Tjur's R^2 for the GLM being 0.36 (Tab. S5.1).

The pseudo-absences of *F. sylvatica* collected using the uniform approach were homogeneously distributed within the environmental space (Fig. S5.2a).

Table S5.1: Results of the HSMs for *Fagus sylvatica* (GLM and RF). Models' predictive performance was assessed through internal (5-fold CV and EU-Forest) and external (sPlotOpen) validation. TSS: true skill statistics; CBI: continuous Boyce index; R-sq: Tjur's R^2 for the GLM, and R^2 for RF. Values of TSS and CBI for the 5-fold cross-validation represent averages.

Validation dataset	**GLM**			**RF**		
	TSS	*CBI*	*Tjur's R^2*	*TSS*	*CBI*	*R^2*
5-fold CV	0.52	0.93		0.77	0.97	
EU-Forest	0.61	0.99	0.36	0.79	0.99	0.66
sPlotOpen	0.41	0.88		0.52	0.93	

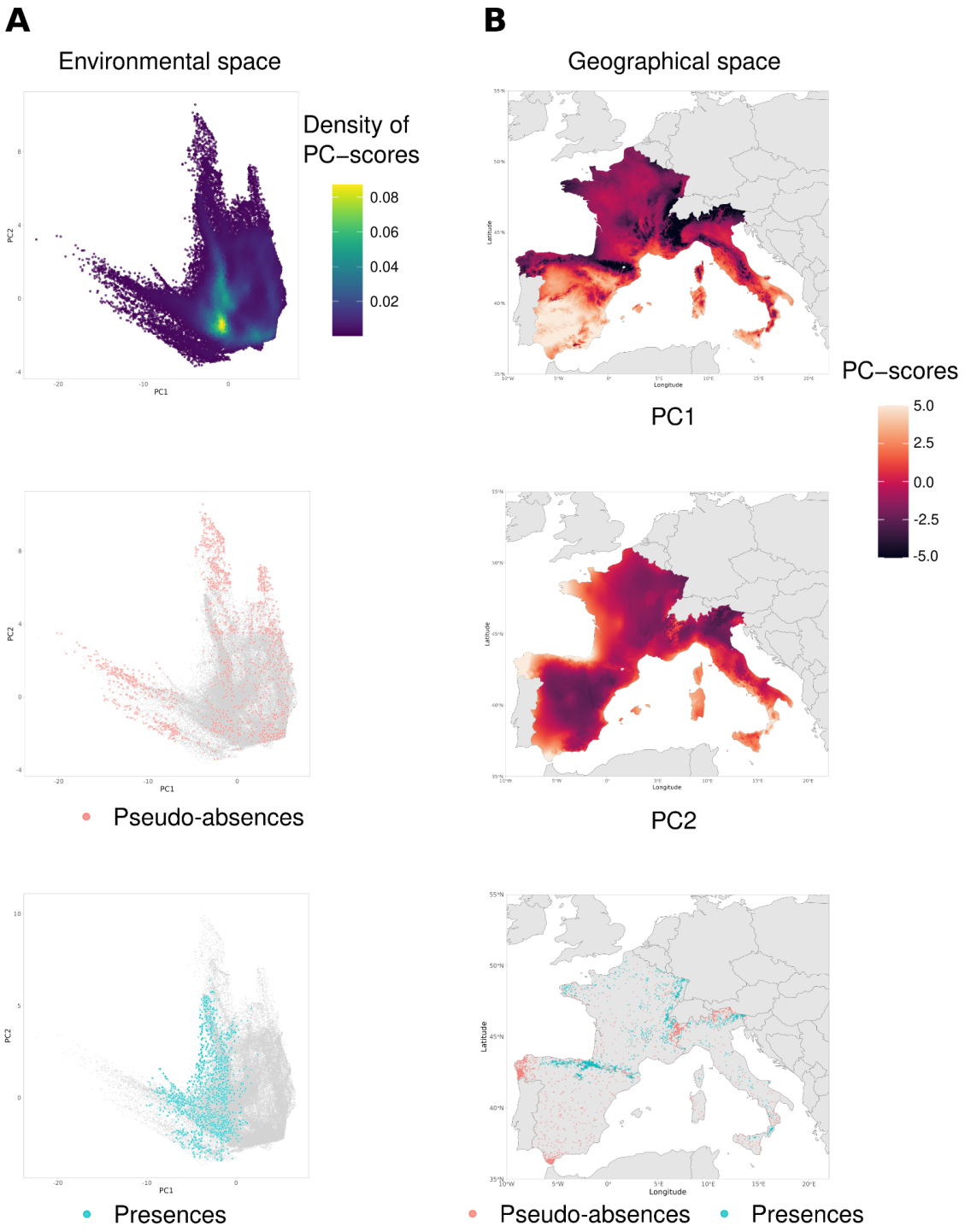


Figure S5.2: (A) environmental space available for *Fagus sylvatica* in Italy, Spain and France, and the position of presences (light blue) and pseudo-absences (red) sampled within the environmental space using the uniform approach; (B) distribution of PC-scores across the geographical space, and location (across western Europe) of presences (light blue) and pseudo-absences (red) sampled using the uniform approach.

References

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). spthin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5):541–545.

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.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-88.

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., ... & De Beaulieu, J. L. (2006). A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New phytologist*, 171(1), 199-221.

Mauri, A., Strona, G., and San-Miguel-Ayanz, J. (2017). Eu-forest, a high-resolution tree occurrence dataset for europe. *Scientific data*, 4(1):1–8.

Mellert et al. (2018) Soil water storage appears to compensate for climatic aridity at the xeric margin of European tree species distribution. *European Journal of Forest Research*, 137: 79-92.

Poli et al. (2022) Coupling fossil records and traditional discrimination metrics to test how genetic information improves species distribution models of the European beech Fagus sylvatica. *European Journal of Forest Research*, 141: 253–265

Sabatini, F. M., Lenoir, J., Hattab, T., Arnst, E. A., Chytr`y, M., Dengler, J., De Ruffray, P., Hennekens, S. M., Jandt, U., Jansen, F., et al. (2021). splotopen–an environmentally balanced, open-access, global dataset of vegetation plots. *Global Ecology and Biogeography*.

Sillero, N. and Barbosa, A. M. (2020). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, pages 1–14.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.