# Assessing the effect of sample bias correction in species distribution models

Nicolas Dubos, Clémentine Préau, Maxime Lenormand, Guillaume Papuga,
Sophie Monsarrat, Pierre Denelle, Marine Le Louarn, Stien Heremans, Roel
May, Philip Roche, et al.

**HAL Id: hal-03833433**
**https://hal.inrae.fr/hal-03833433v1**

Submitted on 8 Dec 2022

# Assessing the effect of sample bias correction in species distribution models

Nicolas Dubos [a,*], Clémentine Préau [a], Maxime Lenormand [a,*], Guillaume Papuga [b],
Sophie Monsarrat [c,d], Pierre Denelle [a,e], Marine Le Louarn [a], Stien Heremans [f], Roel May [g],
Philip Roche [h], Sandra Luque [a]

[a] INRAE, National Research Institute on Agriculture, Food & the Environment, TETIS Unit, Montpellier, France
[b] AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France
[c] Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Biology, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark
[d] Section for Ecoinformatics and Biodiversity, Department of Biology, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark
[e] Biodiversity, Macroecology & Biogeography, University of Göettingen, Göttingen, Germany
[f] Research Institute for Nature and Forest (INBO), Brussels, Belgium
[g] Norwegian Institute for Nature Research (NINA), P.O. Box 5685 Torgarden, NO-7485 Trondheim, Norway
[h] INRAE, Aix Marseille Univ, RECOVER, Aix-en-Provence, France

## ARTICLE INFO

## ABSTRACT

**1.** Open-source biodiversity databases contain a large number of species occurrence records but are often spatially biased; which affects the reliability of species distribution models based on these records. Sample bias correction techniques require data filtering which comes at the cost of record numbers, or require considerable additional sampling effort. Since independent data is rarely available, assessment of the correction technique often relies solely on performance metrics computed using subsets of the available – biased – data, which may prove misleading.

**2.** Here, we assess the extent to which an acknowledged sample bias correction technique is likely to improve models' ability to predict species distributions in the absence of independent data. We assessed variation in model predictions induced by the aforementioned correction and model stochasticity; the variability between model replicates related to a random component (pseudo-absences sets and cross-validation subsets). We present, then, an index of the effect of correction relative to model stochasticity; the Relative Overlap Index (ROI). We investigated whether the ROI better represented the effect of correction than classic performance metrics (Boyce index, cAUC, AUC and TSS) and absolute overlap metrics (Schoener's D, Pearson's and Spearman's correlation coefficients) when considering data related to 64 vertebrate species and 21 virtual species with a generated sample bias.

**3.** When based on absolute overlaps and cross-validation performance metrics, we found that correction produced no significant effects. When considering its effect relative to model stochasticity, the effect of correction was strong for most species at one of the three sites. The use of virtual species enabled us to verify that the correction technique improved both distribution predictions and the biological relevance of the selected variables at the specific site, when these were not correlated with sample bias patterns.

**4.** In the absence of additional independent data, the assessment of sample bias correction based on subsample data may be misleading. We propose to investigate both the biological relevance of environmental variables selected, and, the effect of sample bias correction based on its effect relative to model stochasticity.

## 1. Introduction

While there is a growing demand for species data for the production of robust statistical models and evidence-based conservation actions, the availability of standardised data remains limited. In recent years, the extensive development of biodiversity databases has predominantly been supported by opportunistic, presence-only data collected by citizen science programs and naturalists associations. Despite its limitations, opportunistic, non-standardised data still constitute a promising avenue for improvement of biodiversity assessments (McKinley et al., 2017).

Such data are often limited by the heterogeneity of their sources and spatial biases as a result of uneven sampling efforts (Otegui et al., 2013; Beck et al., 2014; Bird et al., 2014; Johnston et al., 2020; Botella et al., 2021). More specifically, sampling efforts may be biased by field accessibility such as the number of observations influenced by the proximity to urban areas and roads, often leading to spatial autocorrelation among observations (Phillips et al., 2009; Stolar and Nielsen, 2015). This may incur an environmental bias and models tend to overestimate/underestimate environmental suitability in zones with higher/lower density of occurrence data. This could prove to be problematic for studies that aim to provide guidelines for management (Yackulic et al., 2013). By accounting for spatial bias in opportunistic data derived from heterogeneous sources, (e.g. citizen science, naturalist and expert associations) not only would increase the prospect of their potential use in ecological studies, it would also enable the inclusion of a broader range of species in Species Distribution Models (SDMs). SDMs are one of the most commonly used tools for testing ecological hypotheses (Anderson et al., 2009), assessing alien species invasion risks (Bellard et al., 2013; Runquist et al., 2019; Lanner et al., 2022), forecasting the potential effect of environmental change (Araújo et al., 2005), and supporting conservation management efforts (Schwartz, 2012; Leroy et al., 2014; Mikolajczak et al., 2015; Dubos et al., 2021). Although presence-only biodiversity databases are frequently used in SDMs, the adequacy of sample bias correction methods remains ambiguous (Meynard et al., 2019; Johnston et al., 2020). Spatial sampling bias is a major factor affecting the predictive performance of SDMs (Araújo and Guisan, 2006; Barbet–Massin et al., 2012; Kramer–Schadt et al., 2013; Meynard et al., 2019). A number of procedures have been developed to account for sampling bias, which include spatial filtering of presence points (Edrén et al., 2010; Boria et al., 2014; Matutini et al., 2021), environmental filtering (Varela et al., 2014; Gábor et al., 2020), the combination of presence-only and standardised presence-absence data (Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017) and the production of a similar sampling bias in non-presence background data/pseudo-absences (Phillips et al., 2009). However, presence points and environmental filtering consist in the removal of occurrence data, thereby inducing a loss of information and statistical power. This is particularly problematic when dealing with rare or poorly detected species (Lobo and Tognelli, 2011; Kramer–Schadt et al., 2013; Robinson et al., 2018; Vollering et al., 2019; Inman et al., 2021).The most widely applicable method may therefore be the production of pseudo-absences that share the same bias as the presence data. In presence/background or presence/pseudo-absence models, a range of pseudo-absence selection techniques were recently developed which reduce the effect of sampling bias, improving model performance without removing occurrence points (e.g. Senay et al. (2013), Fourcade et al. (2014), Hertzog et al. (2014), Iturbide et al. (2015)). For instance, pseudo-absence selection based on sampling bias reference maps has been acknowledged as an efficient method to account for spatially biased occurrence data (Phillips et al., 2009; Hertzog et al., 2014). Reference maps, such as the target-group (TG) approach and accessibility maps, can be used to represent a sampling bias map that is specific to a given study area. This presents a promising approach to improvement of the predictive performance of SDMs (Ranc et al., 2017; Monsarrat et al., 2019). The TG approach relies on the hypothesis that the study species share the same sampling pattern as the target group, whereas accessibility maps rely on the hypothesis that constraining features are identified (e.g. geographical barriers, social conflicts, long distances). Although neither mapping approach provides explicit information on sampling efforts they may prove appropriate when species richness patterns are heterogeneous (Ranc et al., 2017) or when treating data from heterogeneous sources with different sampling patterns (Monsarrat et al., 2019). Here we focus on a single sample bias correction technique (i.e. accessibility maps) to test a range of different methods. Accessibility maps do not require to subset the occurrence data and are therefore more appropriate for rare species. They are also more widely applicable than TG approach since TG

requires information on sampling effort throughout an entire taxon.

The efficiency of a given sample bias correction technique is often measured by comparing the performance metrics of corrected and uncorrected models. In both corrected and uncorrected groups, performance metrics quantify the degree to which models built with a subset of the original data (i.e. training/calibration dataset) accurately predict the remaining data (test/evaluation dataset). This process is commonly referred to as "cross-validation" (e.g. Senay et al. (2013), Boria et al. (2014)). The most common model performance metrics include the Area Under the operating Curve (AUC), the True Skill Statistic (TSS), the Boyce index, and Similarity indices. These metrics provide quantitative measurements of discrimination ability between models that are built with training data and those that utilise the full dataset (Fourcade et al., 2018). The vast majority of studies test the efficiency of sample bias correction by performing internal cross-validation, (see point 3B of the Standard for SDMs on data sharing the same bias in Araújo et al. (2019)) an approach that has shown strong limitations when SDM is used to extrapolate predictions to a different time/region (Araújo et al., 2005; Araújo et al., 2019; Beck et al., 2014; Hertzog et al., 2014; Fourcade et al., 2018). Ideally, the improvement conferred by a correction technique should be evaluated with an independent, unbiased dataset (Phillips et al., 2009; Hertzog et al., 2014; Norberg et al., 2019; Johnston et al., 2020). Field validation (evaluations relative to independent standardised datasets) represents the best standard practice to assess models' ability to predict species distribution (Araújo et al., 2019). Nevertheless, field validation is labour-intensive and sometimes unfeasible (e.g. taxonomically or geographically extensive scale study sites) and independent standardised datasets are rarely available (Hao et al., 2019). When relying on partitioned datasets, one possible method for assessing sample bias correction techniques is to select data subsets that are subject to different types of bias (e.g. Bean et al. (2012), Matutini et al. (2021)). However, this method may also prove highly demanding in respect to spatio-temporal coverage and may not be feasible for noumerous species (Johnston et al., 2020). A cost-effective method proposed by Hijmans (2012) to assess the potential efficiency of sample bias correction techniques uses an AUC calibrated with a null geographic model. The efficiency of this method may vary through space and between species (Hijmans, 2012), calling for the characterisation of those sites and species. Virtual species can be used to assess bias correction techniques (e.g. Phillips et al. (2009), Fourcade et al. (2014), Varela et al. (2014), Ranc et al. (2017)), by simulating a sampling bias and producing performance metrics that are relative to a perfectly known distribution. The projection of a range of virtual species on multiple real regions may represent a cost-effective approach for assessing whether a correction technique is likely to improve the accuracy of SDMs, provided that virtual and real species are sufficiently comparable. Species distribution models can be calibrated with a range of model parameters generated with a random component: (e.g. pseudo-absence selection, cross-validation subsets) inducing a stochasticity among models that are otherwise identical. These model parameters can be sources of uncertainty in model projections (Buisson et al., 2010; Thibaud et al., 2014). Sample bias correction should induce variation in the predicted values and subsequently in species range projections, but may prove negligible if the variation is of the same magnitude as the sources of uncertainty. Therefore, the effect of sample bias correction may be assessed on the basis of its effect between corrected and uncorrected modalities relative to intra-modality variation.

We present the Relative Overlap Index, which informs the extent of spatial similarity between corrected and uncorrected models relative to the variability between model replicates. We hypothesised that sample bias correction improves model predictions if its relative effect is stronger than that of the remaining model input parameters. We tested this assumption using virtual species with a generated sample bias, modelled at the same sites and with the same range of model parameters as the real species. The aim of this study was (1) to assess the effect of a sampling bias correction technique on distribution projections over a

range of terrestrial vertebrate species (n = 64) in three contrasting regions and in the absence of independent data. We measured the effect of correction by computing the degree of overlap between uncorrected and corrected models. We predicted that the effect of correction on projections differs between sites and species. We further evaluated (2) whether the effect of correction could be assessed with a range of validation metrics. We also tested whether this effect could be represented by an index of overlap between correction modalities relative to intramodality variation. We tested (3) whether sample bias correction actually improved model predictions by using virtual species with a simulated sample bias. Finally, we provided recommendations for the assessment of sample bias correction when independent data are unavailable.

## 2. Materials and methods

### 2.1. Study sites and accessibility

We focused on three regions located in the vicinity of Thau (southern France), Grote Nete (northern Belgium) and Trondheim (central Norway). These sites were characterised by contrasting distributions of roads and cities. While there was an altogether homogeneous distribution of roads and towns in Thau and Grote Nete, there was a strong gradient in road and town density in Trondheim. We produced accessibility maps for each site (Fig. 1) by computing an Accessibility Index (AI). Accessibility Indices represent the degree to which a geographic point is easily reached by an observer and is context specific (Monsarrat et al., 2019). In this instance, we have used occurrence data from heterogeneous sources. For that reason, we assumed that accessibility was – for the most part – dependent upon distance from cities and roads (e.g. Sicacha–Parada et al. (2020)). The Accessibility Index was computed as follows:

$$AI_i = \frac{1}{2}\left(e^{-\frac{1}{2}\left(\frac{dist_C}{\sigma_C}\right)^2} + e^{-\frac{1}{2}\left(\frac{dist_R}{\sigma_R}\right)^2}\right) \tag{1}$$

where $AI_i$ is the accessibility index at pixel $i$, $dist_C$ is the Euclidean distance from the closest city centre ($\geqslant 200$ inhabitants), $dist_R$ is the Euclidean distance from the closest primary and secondary road. $\sigma_C$ and $\sigma_R$ are the standard deviations of the distances distributions to the nearest city and road, respectively.

### 2.2. Environmental data

We used land use variables retrieved from Corine Land Cover habitat classes, a European biophysical dataset derived from remote sensing. At the local scale, land use variables are more relevant to species distribution models than climatic predictors at the local scale (Soberón and Nakamura, 2009; Ficetola et al., 2014). For each $200 \times 200m^2$ pixel, we measured the distance from the nearest habitat features using 8 habitat classes: artificial surface, forest edge, intensive farmland, non-intensive farmland, scrubland/herbaceous areas, coastal areas, water courses, water bodies. We also computed the proportion of a given habitat type within a range of buffer zones around occurrence points (200, 500 and 1,000 meters). This corresponded to species habitat use at the landscape scale in accordance to previous studies (e.g. in reptiles, amphibians and bats; Jeliazkov et al. (2014), Azam et al. (2016)). In birds, the landscape may be influential at larger scales (e.g. 5,000 meters; Dubos et al. (2018)). Given the scale of our study sites, the use of larger buffer zones would result in a lack of variability in environmental conditions at occurrence points, so we limited the extent of our buffer zones to 1,000 meters.

### 2.3. Occurrence records

At each of the three sites, we used occurrence records obtained from biodiversity databases (*Ligue pour la Protection des Oiseaux*[1], the *Artsobservasjoner*[2] and *Natuurpunt Studie Association*[3]) for 79 terrestrial vertebrate species (58 birds, 10 mammals, 6 amphibians and 5 reptiles). We selected one occurrence point per $200 \times 200m^2$ pixel (i.e. data thinning/resampling at the resolution of our environmental variables). Whilst this process is used as a rule of thumb to limit sampling bias driven by multiple observations within the pixel, but it does not account for sampling bias driven by aggregated observations in the surrounding pixels and at larger scales. After data thinning species with fewer than 10 occurrence points per site were discarded (n = 15). The level of sample bias can be estimated using the Boyce index, which is usually used as an evaluation metric for presence-only data to assess the extent to which a spatial layer correctly predicts presence points. More details about the species name, sample size (after thinning/resampling) and sample bias are available in Table S1 in Supporting Information.

In our case, the spatial layer pertains to the accessibility map (Fig. 1). As observed in Fig. 2, the spatial bias due to accessibility in occurrence data was negative in Grote Nete (average Boyce index = −0.75), slightly positive in Thau (average Boyce index = 0.16) and positive in Trondheim (average Boyce index = 0.59).

### 2.4. Distribution modelling

We built species distribution models using the *biomod2* R package (Thuiller et al., 2009) and an ensemble of eight modelling techniques: generalised linear modelling (GLM), generalised boosting modelling (GBM), classification tree analysis (CTA), artificial neural network (ANN), surface range envelop (SRE), flexible discriminant analysis (FDA), general additive modelling (GAM) and random forest (RF). The modelling procedure included (1) a method for pseudo-absence selection, (2) an environmental variable selection process, (3) a final model calibration, and (4) a model evaluation process as summarised in Fig. 3.

**Pseudo-absence selection.** For each group, (uncorrected and corrected groups as defined below) we included 10 different sets of generated pseudo-absences equal in number to that of presence points (Meynard et al., 2019). For models which did not account for field accessibility, (and subsequent sampling bias) hereafter referred as the "uncorrected group", we randomly and evenly selected a number of pseudo-absences in the study area equal to the number of occurrence data within the background (Barbet–Massin et al., 2012; Liu et al., 2019). For models which accounted for sampling bias, we randomly selected pseudo-absences with a sampling probability weighted by AIs after excluding presence pixels. This enabled pseudo-absences to share the same bias as presence points in accordance with the original concept proposed in Phillips et al. (2009). Thus for species that were negatively biased by accessibility, i.e. more commonly found in inaccessible areas, we weighted sampling probability using negative AIs.

**Variable selection.** For both correction modalities, we selected one variable per group of inter-correlated variables to avoid collinearity, treating each site separately (Pearson's r $\geqslant 0.7$). For each individual species, (Hawkins et al., 2017) we assessed the relative importance of each variable (calculated as the Pearson's coefficient between initial model predictions and model predictions made when the assessed variable is randomly permuted) with 10 permutations. The final set of variables included in the final models were those with a relative importance of $\geqslant 0.05$ across at least 50% of model runs (Bellard et al., 2016).
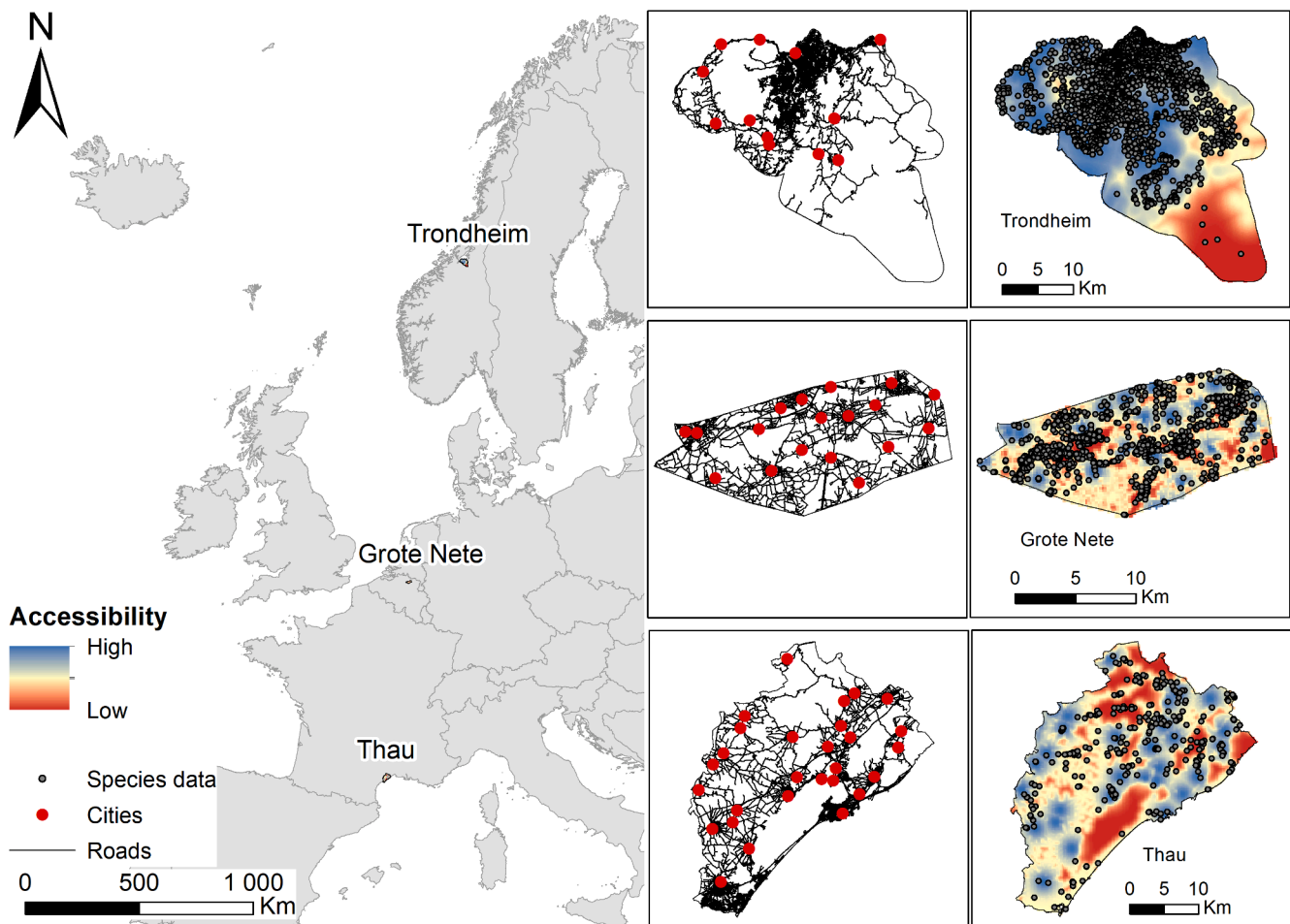
---

**Fig. 1.** Location of the three study sites in Europe and accessibility maps for each site. The accessibility index is inversely proportional to the Euclidean distance between cities and roads and represents the degree to which a geographic point is easily reached by an observer. Black dots represent occurrence records of all species pooled together (n = 46, 11 and 7 species in Trondheim, Grote Nete and Thau, respectively).

**Final models.** We used the eight aforementioned modelling techniques with 10 sets of pseudo absence, and 3 runs of calibration over 80% of the data (20% for evaluation).

### 2.5. Effect of sample bias correction

**Effect on model predictions.** We measured the "absolute" effect of sample bias correction using indices of similarity and correlation coefficients between uncorrected and corrected predictions. We computed the Schoener's D as a measure of projection overlap (computed with the *ENMTool* R package (Warren et al., 2010; Rödder and Engler, 2011)), the Pearson's correlation coefficient (Li and Guo, 2013) and the Spearman's rank coefficient (Phillips et al., 2009). Schoener's D was computed as follows:

$$D\left(p_x, p_y\right) = 1 - \frac{1}{2}\sum_i \left|p_{x_i} - p_{y_i}\right| \tag{2}$$

For each species, modelling technique, cross-validation run, and pseudo-absence run individually, $p_{x_i}$ and $p_{y_i}$ are the normalised suitability scores for uncorrected $x$ and corrected $y$ prediction in grid cell $i$. This corresponds to the comparison of 30 corrected projections and 30 uncorrected projections (10 pseudo-absence datasets and 3 cross-validation subsets). This therefore represents $8 \cdot (3 \cdot 10)^2 = 7,200$ values for each species and overlap metrics. It is important to note that some predictions failed, particularly for the GBM modelling technique, and were therefore not included in the computation of the overlap metrics. See Tables S2

and S3 in Supporting Information for more details.

For comparison, we also assessed model performance using four classical evaluation metrics based on cross-validation data subsets, namely: the Boyce index (computed with the *ecospat* R package (Cola et al., 2017)), the true skill statistic (TSS), the area under the relative operating characteristic curve (AUC), and a calibrated AUC (cAUC). The cAUC was computed following Hijmans (2012) and calibrated on the AUC of a null geographic model. The null geographic model was computed with the *geoDist* function of the *dismo* R package (Hijmans et al., 2015). The Boyce index, a reliability metric, indicates the extent to which a spatial layer correctly predicts presence points. The TSS, AUC and cAUC are discrimination metrics that indicate the ability to distinguish between occupied and unoccupied sites. We also used a one-sided Student t-test in order to evaluate whether or not the mean corrected performance was significantly greater than the mean uncorrected performance for a given performance metric, species, and modelling technique. The p-value was computed with the *rquery.t.test* function of the *ContDataQC* R package (Leppo, 2021).

**Effect relative to model stochasticity.** For each species and modelling technique, we assessed the extent to which the correction technique affected predictions relative to the sources of stochastic variation between models of the corrected group (i.e. cross-validation runs and pseudo-absence set runs for each modelling technique). Model stochasticity was quantified using the aforementioned overlap metrics (Schoener's D, Pearson's and Spearman's coefficient) between all pairwise combinations of model projections for the each of the 64 species individually (10 pseudo-absence datasets, 3 cross validation
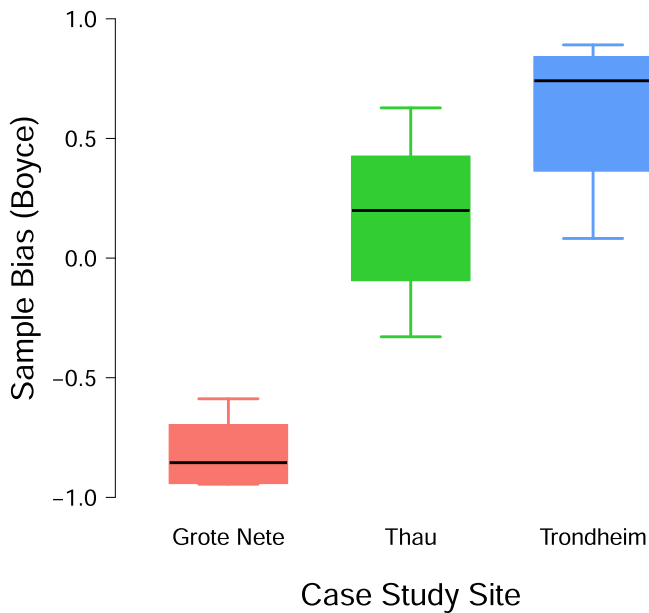
**Fig. 2.** Boxplots of measured sample bias in the spatial distribution of the 64 real species according to the case study site. The bias is assessed with Boyce indices measuring how accurately occurrence data are predicted by the Accessibility Index. Each boxplot is composed of the first decile, the first quartile, the median, the third quartile and the ninth decile.

subsets and 8 modelling techniques, resulting in $8 \cdot \frac{(3 \cdot 10)^2 - (3 \cdot 10)}{2} = 3,480$ values for each species and overlap metric). We present the Relative Overlap Index (ROI), an index of mean overlap between predictions of the uncorrected and the corrected groups, relative to the average overlap between pairwise model projections of the corrected group. The two overlap components of the ROI can be assessed either with similarity metrics or correlation coefficients. When based on Schoener's D, the ROI was computed as follows.

$$ROI = \frac{\overline{D}_0 - \overline{D}}{\overline{D}_0} \tag{3}$$

Where $\overline{D}_0$ is the mean overlap between model runs of the corrected group, $\overline{D}$ is the mean overlap between runs of the uncorrected and the corrected groups. It is important to note that the ROI is always computed for a given species and modelling technique, where $\overline{D}_0$ is based on $\frac{(3 \cdot 10)^2 - (3 \cdot 10)}{2} = 435$ overlaps between model runs, and $\overline{D}$ is based on $(3 \cdot 10)^2 = 900$ overlaps between corrected and uncorrected models runs. A value close to 0 represents a perfect match between predictions, i.e. no effect of sample bias correction. The overlaps between uncorrected and corrected groups tend to be significantly smaller than the overlaps between runs when the ROI approaches 1 (i.e. strong effect of sample bias correction). A negative value can sometimes be obtained when the sample size is small, meaning that model stochasticity is of higher magnitude than the sample bias correction. In this case, we can thereby conclude that there is no effect of sample bias correction. The formula is similar when based on Pearson's and Spearman's rank coefficient, but values were transformed in order to range between 0 and 1 by adding 1 and dividing by 2.

We also used a one-sided Student t-test to evaluate whether or not $\overline{D}_0$
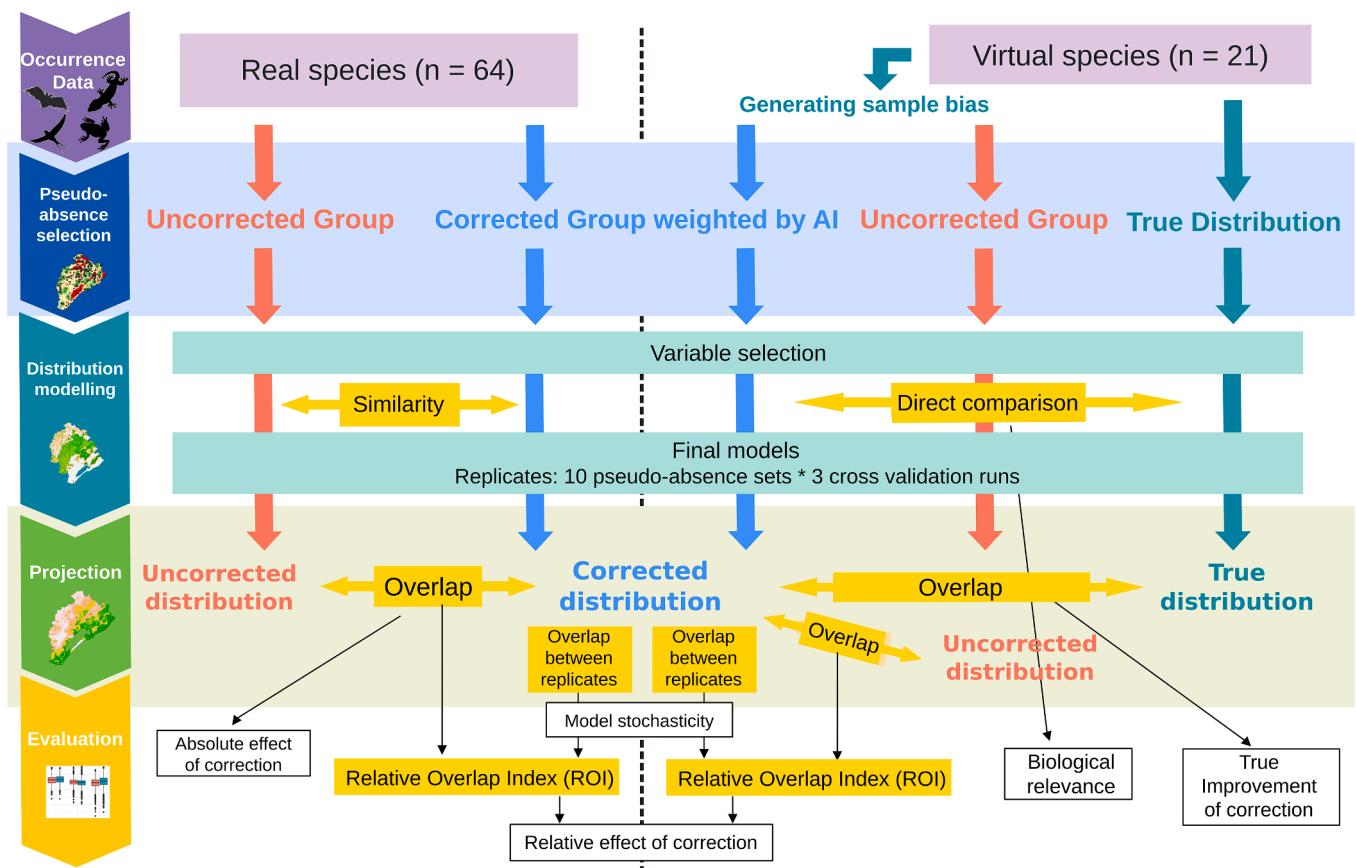


**Fig. 3.** Steps of the methodology developed to study the effect of sample bias correction applied on the species distribution modelling of 64 real and 21 virtual species.

was significantly greater than $\overline{D}$ for a given overlap metric, species, and modelling technique. The p-value was computed with the *rquery.t.test* function of the *ContDataQC* R package (Leppo, 2021).

**Effect on variable selection.** We estimated the degree of similarity in the selected variables between the uncorrected and corrected groups.

We used the Jaccard index (Jaccard, 1912) only when considering whether or not the variable was selected. We also computed the Bray-Curtis index (Bray and Curtis, 1957) when accounting for variable importance.

**Table 1**
**Variables and response functions used to generate virtual species, and variables selected during the modelling process for the uncorrected and the corrected group.**[a]

| Variable used for generation | Response function | Site | Uncorrected group Selected variable | Variable importance | Corrected group Selected variable | Variable importance |
|---|---|---|---|---|---|---|
| Distance from non-intensive agriculture | Logistic (beta = −500, alpha = −5000) | Grote Nete | Proportion of non-intensive agriculture within 500 m | 0.579 | Proportion of non-intensive agriculture within 500 m | 0.592 |
| | | Trondheim | **Distance from non-intensive agriculture** | **0.781** | **Distance from non-intensive agriculture** | **0.804** |
| | | Thau | **NDVI** | **0.478** | **NDVI** | **0.453** |
| | | | Grassland Index | 0.127 | Grassland Index | 0.143 |
| | | | Distance from forest edges | 0.149 | Distance from forest edges | 0.158 |
| | | | Distance from intensive agriculture | 0.124 | Distance from intensive agriculture | 0.141 |
| | | | | | Distance from water courses | 0.117 |
| Distance from intensive agriculture | Logistic (beta = 500, alpha = −5000) | Grote Nete | **Distance from intensive agriculture** | **0.792** | **Distance from intensive agriculture** | **0.809** |
| | | Trondheim | **Distance from intensive agriculture** | **0.751** | **Distance from intensive agriculture** | **0.754** |
| | | | Distance from forest edges | 0.097 | Distance from forest edges | 0.113 |
| | | Thau | **Distance from intensive agriculture** | **0.851** | **Distance from intensive agriculture** | **0.892** |
| | | | **NDVI** | **0.049** | | |
| Distance from water courses | Gaussian (mean = 100, SD = 100) | Grote Nete | **Distance from water courses** | **0.871** | **Distance from water courses** | **0.897** |
| | | Trondheim | **Distance from water courses** | **0.818** | **Distance from water courses** | **0.880** |
| | | | **Distance from coastal areas** | **0.066** | | |
| | | Thau | **Distance from water courses** | **0.932** | **Distance from water courses** | **0.937** |
| Distance from water courses | Gaussian (mean = 100, SD = 500) | Grote Nete | **Distance from water courses** | **0.647** | **Distance from water courses** | **0.770** |
| | | | **Distance from non-intensive agriculture** | **0.110** | | |
| | | Trondheim | **Distance from water courses** | **0.594** | **Distance from water courses** | **0.747** |
| | | | **Distance from coastal areas** | **0.066** | | |
| | | Thau | **Distance from water courses** | **0.891** | **Distance from water courses** | **0.902** |
| Grassland Index | Logistic (beta = −1000, alpha = −50) | Grote Nete | **Grassland Index** | **0.568** | **Grassland Index** | **0.575** |
| | | | **Proportion of non-intensive agriculture within 200m** | **0.074** | | |
| | | Trondheim | **Grassland Index** | **0.802** | **Grassland Index** | **0.834** |
| | | Thau | **Grassland Index** | **0.864** | **Grassland Index** | **0.879** |
| Distance from open natural areas | Logistic (beta = −500, alpha = −5000) | Grote Nete | **Proportion of open natural areas within 200 m** | **0.495** | **Proportion of open natural areas within 200 m** | **0.393** |
| | | | **Distance from open natural areas** | **0.315** | **Distance from open natural areas** | **0.42** |
| | | Trondheim | **Distance from open natural areas** | **0.367** | **Distance from open natural areas** | **0.391** |
| | | | Proportion of open natural areas within 200 m | 0.394 | Proportion of open natural areas within 200 m | 0.439 |
| | | | **Distance from forest edges** | **0.117** | | |
| | | Thau | Proportion of non-intensive agriculture within 500m | 0.297 | Proportion of non-intensive agriculture within 500m | 0.298 |
| | | | NDVI | 0.313 | NDVI | 0.326 |
| | | | **Distance from forest edges** | **0.184** | **Distance from forest edges** | **0.176** |
| Distance from open natural areas | Logistic (beta = −500, alpha = −2000) | Grote Nete | **Distance from open natural areas** | **0.786** | **Distance from open natural areas** | **0.808** |
| | | Trondheim | **Distance from open natural areas** | **0.496** | **Distance from open natural areas** | **0.609** |
| | | | **Distance from intensive agriculture** | **0.197** | | |
| | | Thau | Distance from forest edges | 0.55 | Distance from forest edges | 0.617 |
| | | | **Distance from intensive agriculture within 200m** | **0.234** | **Distance from intensive agriculture within 200m** | **0.12** |
| | | | **Distance from water courses** | **0.108** | | |

[a] We have provided the distribution family and the parameters of the response function. The variables used to generate the species for which the relative importance increased after correction, and variables which were not used for which the relative importance decreased are shown in bold.

## 2.6. Testing whether changes correspond to actual improvements using virtual species

For each of the three study sites, we generated a set of 7 virtual species, with different ecological niches, generated at each site (21 species in total) using the *Virtualspecies* R package (Leroy et al., 2016). The probability of their presence was generated according to a relationship with a single environmental variable (parameters and shape of the relationships are provided in Table 1). We produced presence-absence maps using a probability threshold of 0.6 and sampled 300 presence points. For species associated with the grassland index and the proportion of herbaceous areas within 500 meters, for which there were fewer than 300 pixels, we respectively selected 100 and 80 presence points. We then produced a spatially biased sample for the 21 virtual species by weighting the probability of selection using the Accessibility Index. More details about the species name, sample size, and sample bias are available in Table S4 in Supporting Information.

For the generated species, we used the same distribution modelling framework and metrics to assess the effect of sample bias correction as those used for the real species. It is again worth noting that some predictions failed (particularly for the GBM modelling technique) and were therefore excluded from the computation of the overlap metrics (see Tables S5 and S6 in Supporting Information for more details).

**Did the correction effect correspond to an improvement?** The virtual species generated with simulated sample bias enabled us to test whether the correction technique actually improved the models' ability to predict the "true" distribution (Meynard and Kaplan, 2012). Using the Schoener's D overlap, the Pearson's correlation coefficient and the root-mean-square error (RMSE) to quantify the degree to which model predictions were improved by the correction technique, we compared the predicted probability of occurrence of the corrected and uncorrected groups with the "true" probability of occurrence.

We relied on a one-sided Student t-test to evaluate whether or not the overlap with the "true" distribution was significantly greater for the corrected group than the uncorrected group for a given overlap metric, species, and modelling technique. The p-value was computed with the *rquery.t.test* function of the *ContDataQC* R package (Leppo, 2021).

We used the following procedure to assess Schoener's D and classic performance metrics' (Boyce, cAUC, AUC and TSS) ability to assess the effect of correction. We took the overlap with the "true" distribution as reference for dividing each species and modelling technique into two groups (effect/no effect). This was based upon the value of the significance threshold $\delta$ determined by the one-sided Student t-test p-value (lower than $\delta$/higher than $\delta$)., We followed the same process with the Schoener'D overlap (p-value associated with the comparison between $\overline{D}_0$ and $\overline{D}$), the Boyce, cAUC, AUC and TSS (p-values associated with the comparison of performance metrics for the corrected and uncorrected projections). We then built five tables of confusion to see how the five metrics classified the species for each modelling technique when compared with the reference for a given significance threshold $\delta$. To measure the proportion of species and modelling techniques classified into the same category (effect and no effect) we computed "accuracy" for a given significance threshold $\delta$.

**Did the correction improve the biological relevance of variable selection?** We determined whether variable selection was relevant by comparing the variables used to generate the virtual species with those selected before and after correction.

## 3. Results

### 3.1. Effect of sample bias correction

**Effect on model predictions.** The effect of correction was largely consistent between sites as shown by the overlap between projections built from uncorrected and corrected pseudo-absences, plotted in Fig. 4a. Model projections shared about 80% of information (Schoener's D) common to uncorrected and corrected models. Maps of the projection are available in Supporting Information (Figures S1 to S17). Generally, there were no significant differences between Boyce, AUC, and TSS indices between groups (Fig. 4c). Only the cAUC showed a significant effect of correction for a majority of species and modelling techniques over the three sites, as indicated by the low median p-value (Fig. 4c).

**Effect on variable selection.** At Thau and Trondheim, we found important differences in variable selection (Jaccard indices) and variable importance (Bray-Curtis indices) between corrected and uncorrected models (Fig. 4b).

### 3.2. Relative effect of correction

The effect of correction was high compared to model stochasticity in Trondheim, and to a lesser extent in Thau. This is shown by the lower overlaps between treatments (i.e. corrected and uncorrected groups) compared to overlaps between model replicates (Fig. 5). At the remaining site, the effect of correction was of similar magnitude to that of model replicates. This result remained consistent when using Pearson's and Spearman's coefficients as a measure of overlap (Fig. S28 in Supporting Information).

A direct relationship exists between the ROI and the one-sided Student t-test p-value used to evaluate whether the overlap between runs ($\overline{D}_0$) was significantly greater than the overlap between corrected and uncorrected projections ($\overline{D}$). Indeed, we can show that all the p-values associated with ROI strictly higher than 0.02 are strictly lower than 0.05 (Fig. S29 in Supporting Information). The majority of ROI values in Thau and Trondheim were higher than 0.02 (as shown in the inset of Fig. 5).

### 3.3. Did the correction actually improve predictions?

The correction remarkably improved distribution predictions in Trondheim and Grote Nete for the vast majority of species and modelling techniques (Fig. 6a). The results were less impressive in Thau with less than 42% of the species and modelling techniques showing an improvement. Maps of the projection are available in Supporting Information (Figs. S18 to S23).

In terms of "accuracy" with regards to the classification of species and modelling techniques into two groups (effect/no effect), the results obtained using Schoener's D (p-value associated with the comparison between $\overline{D}_0$ and $\overline{D}$) are more in line with the classification (considering the "true" distribution) than those obtained using the classic performance metrics (Boyce, cAUC, AUC and TSS). This may be observed in Fig. 6b.

### 3.4. Did the correction improve the biological relevance of the selected variables?

Sample bias correction improved the biological relevance for 17 virtual species (Table 1). The improvement was characterised by an increase in relative importance for the relevant variable (i.e. used to generate the species; n = 9), or by an increase in the importance of the relevant variable and a decrease the importance of irrelevant variables (n = 8). For two such instances, the improvement was mitigated by an increase in the relative importance of irrelevant variables.

## 4. Discussion

The efficiency of sample bias correction based on accessibility varies between sites and species. The effect of correction depends upon the landscape complexity at each site, and on the degree of spatial bias in the occurrence data and predictor variables. In absence of independent data, the effect of the correction cannot be assessed with classic evaluation metrics (difference in performance, or absolute overlap between
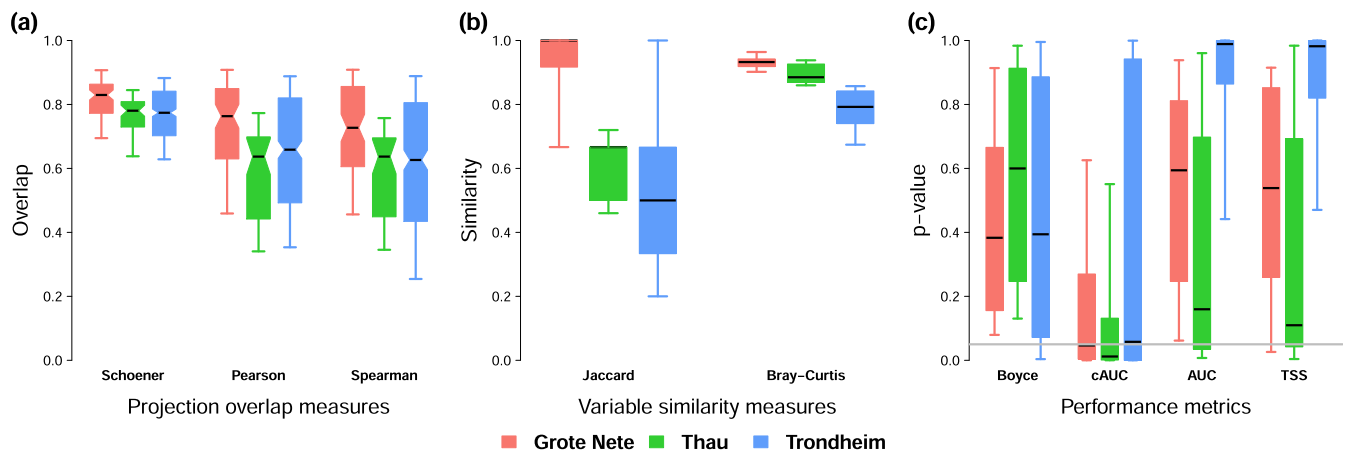
**Fig. 4. Site-specific variation in the effect of sample bias correction.** (a) Boxplots of the three overlap measures between corrected and uncorrected predictions. Each boxplot is composed of the average overlap between 30 corrected and 30 uncorrected projections obtained for each species and each modelling technique. (b) Boxplots of the variable selection similarity (one value per species). (c) Boxplots of the p-value of a one-sided Student t-test to evaluate whether the mean performance of corrected group is greater than the mean performance of corrected group. The grey horizontal line represents the 0.05 significance threshold. Each boxplot is composed of the first decile, the first quartile, the median, the third quartile and the ninth decile.
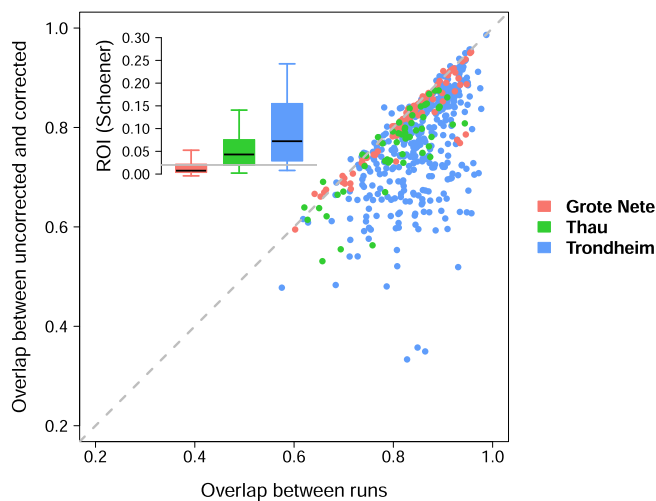


**Fig. 5. Site-specific variation in the relative effect of sample bias correction.** Relationship between the effect of sample bias correction (Schoener's D overlap between uncorrected and corrected predictions $\overline{D}$) and the effect of model stochasticity (Schoener's D overlap between model runs $\overline{D}_0$). Points located below the y = x line represent models for which the relative effect of sample bias correction exceeds that of model stochasticity. The inset shows the boxplots of the Relative Overlap Index (ROI) according to the case study site. The horizontal grey line represents the ROI threshold value 0.02 associated with the one-sided Student t-test significance threshold 0.05 (see Fig. S29 in Supporting Information for more details). Each boxplot is composed of the first decile, the lower hinge, the median, the upper hinge and the ninth decile.

corrected and uncorrected projections). The sample generated bias in virtual species enabled us to verify that the Relative Overlap Index represented better the effect of correction than alternative approaches.

### 4.1. Effect of sample bias correction

In this study, we found that the impact of sample bias correction primarily depended upon the study site and, to a lesser extent, on species. Sample bias correction primarily affected species distribution predictions in Trondheim and, to a lesser extent, in Thau. The fact that the impact of correction between sites differed may be explained by variations in the spatial scale and landscape complexity between sites.

The accessibility index was based upon site-specific characteristics and reflects a relative measure of distance. The less accessible zones were geographically closer to the more accessible ones at smaller spatial scales, and the absolute distance may prove unproblematic in terms of accessibility. This suggests that sampling bias correction, when based on accessibility, is less necessary when studying areas with homogeneous distributions of roads and towns at smaller spatial scales.

Additionally, accessibility maps may not reflect sample bias equally for all species and sites. Accessibility maps can be built on the basis of topography, land use, and property, to better represent species or site-specific sample bias. In any case, the comparisons with the "true" probability of occurrence of virtual species suggested that the correction was likely to improve model predictions at each site for a high number of species and modelling techniques (Fig. 6a).

We also investigated whether the effect of correction differed between sites and according to the sample bias, sample size and the modelling technique. Methodological details, results and related discussion are available in Supporting Information. The effect of correction on predictions was greater when sample bias was positive but remained highly heterogeneous, even among species with strong sample biases. Variation in ROI value tends to decrease with increasing sample size. Species with low sample sizes (fewer than 100 occurrences after thinning/resampling) cover a very wide range of ROI values from negative value (close to −0.1) to very high ROI value (up to 0.6). This is presumably because the model stochasticity may be higher as a result of lower accuracy (Stockwell and Peterson, 2002). More details are available in Supporting information.

### 4.2. Biological relevance of variable selection

Variable selection differed the most after correction at Trondheim, where correction was most effective and accessibility gradients were clearest. For the most part, models of the uncorrected group selected additional variables that were not biologically meaningful, presumably because they were locally correlated with the accessibility index (intensive agriculture, water courses and distance to the coast in Trondheim; Figs. S1 to S17 in Supporting Information). In Trondheim, models tended to explain the absence of occurrence data in the inaccessible area by the environmental variable which was most represented there. Virtual species indicated that the correction increased the relative importance of biologically meaningful variables, since they were often used to generate species distribution. This result is supported by the virtual species analysis, wherein the effect of correction decreased when
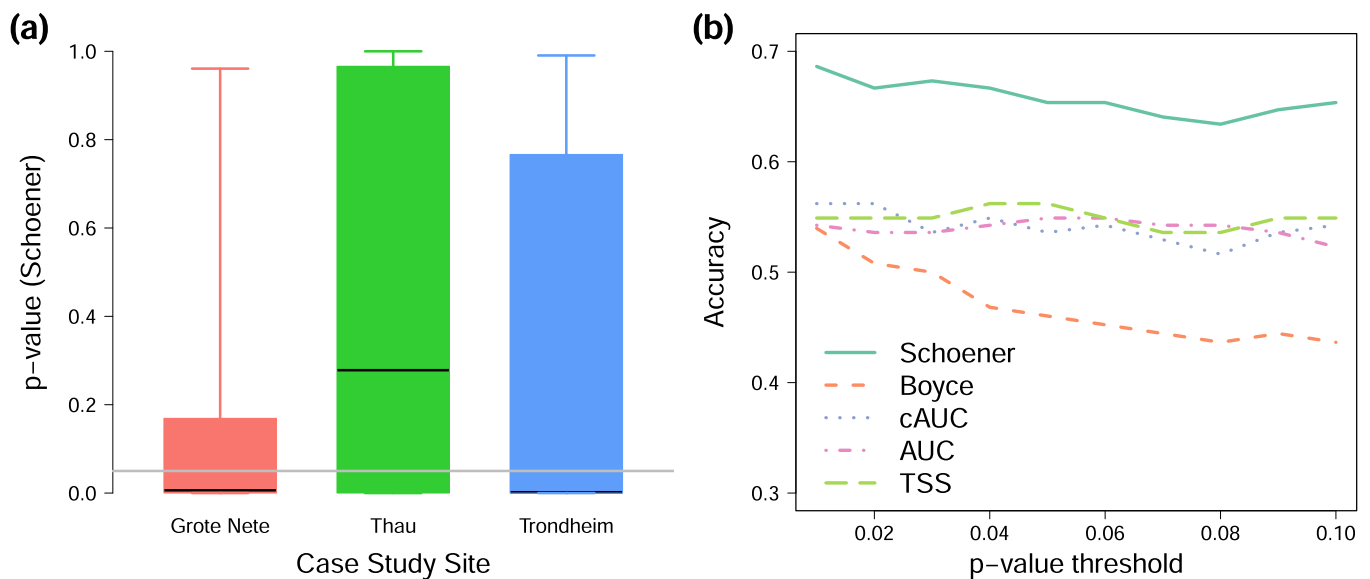
**Fig. 6. Comparison with the "true" probability of occurrence for the uncorrected and the corrected groups at three sites (virtual species).** (a) Boxplots of the p-value obtained from a one-sided Student t-test evaluating whether the overlap between Schoener's D and the "true" distribution is significantly greater for the corrected group than the uncorrected group, for a given species and modelling technique. Similar plots obtained with the Pearson's coefficient and the RMSE are available in Fig. S30 in Supporting Information. (b) Evolution of the accuracy values as a function of the significance threshold $\delta$. The five tables of confusion pertaining to accuracy values obtained within a significance threshold $\delta = 0.05$ are available in Table S7 in Supporting Information.

the species habitat variable was correlated with accessibility (e.g. distance from intensive agriculture in Trondheim correlated with the AI, Pearson's r = 0.80). This result underlines that the biological relevance of the variable selected before and after correction needs to be carefully investigated, in accordance with the recommendations given in Hijmans (2012) and Fourcade et al. (2018).

When the geographic sampling bias translates into a bias in the environmental predictors, the benefit conferred by sample bias correction with accessibility maps may be limited – as is consistent with the target-group approach (Ranc et al., 2017). In other terms, when the distribution of environmental variables matches that of accessibility, issues related to sample biases cannot be corrected with accessibility maps. In our study regions, this may be the case for species that are most impacted by urbanisation (Geslin et al., 2013) and intensive agriculture (Jeliazkov et al., 2016; Olivier et al., 2020).

### 4.3. Correction effect relative to model stochasticity

In the absence of independent, standardised data, the performance of SDMs and correction methods cannot be properly assessed. Here, we propose to measure the effect of correction relative to the within-model stochasticity (between runs of varying input parameters, individually, for each modelling technique) to inform the potential benefit conferred by correction. We show that the Relative Overlap Index was in better agreement than the classic cross-validation performance metrics when concerning changes between corrected and uncorrected predictions of virtual species. The ROI also yielded better results with regards to changes in variable selection and the relative importance across sites and species. The use of this index may be generalisable to species for which habitat is not restricted to the same section of the accessibility gradient. This metric can be used to indicate whether species distribution models are likely to be improved by sample bias correction.

### 4.4. Cross-validation metrics

The performance metrics based on cross-validation failed to detect an improvement in species range predictions, even for cAUC. This differs from the findings of Hertzog et al. (2014), which assessed the performance of a variety of bias correction techniques based on the model's

ability to predict the range of a dung beetle (Coprophagous Scarabaeidae). This study found a striking difference between the evaluation of partitioned datasets and field validation. In their study, cAUC was in agreement with field validation. In our case, the cAUC showed an effect of correction on real species but not virtual ones with perfectly known distribution and generated bias. This calls into question the reliability of cAUC to properly identify a change after correction. As specified by Hijmans (2012), a null geographic model may not be relevant when a species occurs in a single continuous range. Perhaps this may explain why no improvement of correction was detected in some cases. Another recently developed evaluation method considers the accumulation curve of occurrences within the area predicted as suitable as well as the amount of uninformative niche space predicted (Jiménez and Soberón, 2020). This method is appropriate when absence data are unavailable and is similar to our approach in that the evaluation is relative to a random component. It is recommended that multiple metrics describing various aspects of model performance be reported, so as to improve the understanding and transparency of SDMs (Araújo et al., 2019). Our results suggest that the performance of sample bias correction should be assessed using the Relative Overlap Index, and that alternative metrics may be misleading. The ROI is to be interpreted as the degree of change in spatial predictions whilst accounting for variation between model replicates.

### 4.5. Generalisation and limitations

Our analyses included a wide range of species (n = 64) of various taxa (four classes of vertebrates) with varying responses to environmental predictors. We generated a set of virtual species with contrasting ecological preferences (e.g. dependence on water courses or open, natural areas) and varying degrees of specialisation commonly found in amphibians, bats, and birds (Godet et al., 2015; Jeliazkov et al., 2014; Dubos et al., 2021). We also attempted to represent the diversity of species sensitivity to anthropogenic disturbance by including species with a negative response to intensive agriculture (Chiron et al., 2014). However, our study does not encompass the entirety of species' distribution, nor does it account for climatic predictors, which may reduce the scope of our conclusions. The correction method best improved our predictions at the largest site which suggests, that our correction

method, based on accessibility, may prove efficient at larger geographic scales where terrain accessibility is a clear issue. Aside from terrain accessibility, the efficiency of a correction technique may also differ depending upon the type of bias. Further studies should be designed at broader taxonomic and geographic scales and should assess potential differences between various correction techniques.

### 4.6. Increasing the potential use of biodiversity databases

The limited availability of high-quality data is still a major hindrance to effective decision-making, despite the increasing availability of open-source, publicly available data and EU policies to promote open access and data sharing. Accounting for sample bias is challenging, especially for rare species and those for which the distribution range is small with subsequently fewer occurrence points. Pseudo-absence selection, weighted by accessibility maps, enables us to account for sample bias without the use of filtering techniques that reduce the amount of data available. Sample bias correction is also viable for when concerning broadly distributed species in the event that occurrence data are spatially biased – provided that the effect of correction is assessed (Hertzog et al., 2014). When standardised data are unavailable, effectively assessing the efficiency of sample bias correction techniques remains challenging. However, the relative measure of its effect and the use of virtual species could prove to be critical for the increased inclusion of large, heterogeneous, biased datasets in species distribution models and biodiversity assessments.

### Author contributions

ND, ML and SL designed the analyses. ND and ML performed the analyses. CP built the environmental variables and the map. GP, SM and PD aided in interpreting the results and worked on the manuscript. MLL and FR, SH and RM helped to organize data collected in France, Belgium and Norway, respectively. ND led the writing of the manuscript. All authors read, commented and validated the final version of the manuscript.

### Data accessibility statement

All relevant data and code are available online (https://gitlab.com/maximelenormand/sample-bias-correction-sdms).

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecolind.2022.109487.

### References

Anderson, B.J., Arroyo, B.E., Collingham, Y.C., Etheridge, B., Fernandez-De-Simon, J., Gillings, S., Gregory, R.D., Leckie, F.M., Sim, I.M.W., Thomas, C.D., Travis, J.,
Redpath, S.M., 2009. Using distribution models to test alternative hypotheses about a species' environmental limits and recovery prospects. Biol. Conserv. 142, 488–499.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.

Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species–climate impact models under climate change. Glob. Change Biol. 11, 1504–1513.

Araújo, M.B., Anderson, R.P., Barbosa, A.M., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Sci. Adv. 5, eaat4858.

Azam, C., Le Viol, I., Julien, J.-F., Bas, Y., Kerbiriou, C., 2016. Disentangling the relative effect of light pollution, impervious surfaces and intensive agriculture on bat activity with a national-scale monitoring program. Landscape Ecol. 31, 2471–2483.

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3, 327–338.

Bean, W.T., Stafford, R., Brashares, J.S., 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35, 250–258.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15.

Bellard, C., Thuiller, W., Leroy, B., Genovesi, P., Bakkenes, M., Courchamp, F., 2013. Will climate change promote future invasions? Glob. Change Biol. 19, 3740–3748.

Bellard, C., Leroy, B., Thuiller, W., Rysman, J.-F., Courchamp, F., 2016. Major drivers of invasion risks throughout the world. Ecosphere 7, e01241.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N., Frusher, S., 2014. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv. 173, 144–154.

Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77.

Botella, C., Joly, A., Bonnet, P., Munoz, F., Monestiez, P., 2021. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. In: Methods Ecol. Evol..

Bray, J.R., Curtis, J.T., 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. 27, 325–349.

Buisson, L., Thuiller, W., Casajus, N., Lek, S., Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Change Biol. 16, 1145–1157.

Chiron, F., Chargé, R., Julliard, R., Jiguet, F., Muratet, A., 2014. Pesticide doses, landscape structure and their relative effects on farmland birds. Agricult., Ecosyst. Environ. 185, 153–160.

Cola, V.D., Broennimann, O., Petitpierre, B., Breiner, F.T., D'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R.G., Hordijk, W., Salamin, N., Guisan, A., 2017. ecospat: an R package to support spatial analyses and modeling of species niches and distributions. Ecography 40, 774–787.

Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Glob. Ecol. Biogeogr. 23, 1472–1484.

Dubos, N., Viol, I.L., Robert, A., Teplitsky, C., Ghislain, M., Dehorter, O., Julliard, R., Henry, P.Y., 2018. Disentangling the effects of spring anomalies in climate and net primary production on body size of temperate songbirds. Ecography 41, 1319–1330.

Dubos, N., Kerbiriou, C., Julien, J.-F., Barbaro, L., Barré, K., Claireau, F., Froidevaux, J., Le Viol, I., Lorrillière, R., Roemer, C., Verfaillie, F., Bas, Y., 2021. Going beyond species richness and abundance: robustness of community specialisation measures in short acoustic surveys. Biodivers. Conserv. 30, 343–363.

Dubos, N., Montfort, F., Grinand, C., Nourtier, M., Deso, G., Probst, J.-M., Razafimanahaka, J.H., Andriantsimanarilafy, R.R., Rakotondrasoa, E.F., Razafindraibe, P., Jenkins, R., Crottini, A., 2021. Are narrow-ranging species doomed to extinction? Projected dramatic decline in future climate suitability of two highly threatened species. In: Perspectives in Ecology and Conservation.

Edrén, S.M.C., Wisz, M.S., Teilmann, J., Dietz, R., Söderkvist, J., 2010. Modelling spatial patterns in harbour porpoise satellite telemetry data using maximum entropy. Ecography 33, 698–708.

Ficetola, G.F., Bonardi, A., Mücher, C.A., Gilissen, N.L.M., Padoa-Schioppa, E., 2014. How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height. Int. J. Geograph. Inform. Sci. 28, 1723–1739.

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol. Evol. 6, 424–438.

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, e97122.

Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27, 245–256.

Geslin, B., Gauzens, B., Thébault, E., Dajoz, I., 2013. Plant Pollinator Networks along a Gradient of Urbanisation. PLOS ONE 8, e63421.

Godet, L., Gaüzere, P., Jiguet, F., Devictor, V., 2015. Dissociating several forms of commonness in birds sheds new light on biotic homogenization. Glob. Ecol. Biogeogr. 24, 416–426.

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? Int. J. Geograph. Inform. Sci. 1567–1584.

Hao, T., Elith, J., Guillera-Arroita, G., Lahoz-Monfort, J.J., 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Divers. Distrib. 25, 839–852.

Hawkins, B.A., Leroy, B., Rodríguez, M., Singer, A., Vilela, B., Villalobos, F., Wang, X., Zelený, D., 2017. Structural bias in aggregated species-level variables driven by repeated species co-occurrences: a pervasive problem in community and assemblage data. J. Biogeogr. 44, 1199–1211.

Hertzog, L.R., Besnard, A., Jay-Robert, P., 2014. Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. Divers. Distrib. 20, 1403–1413.

Hijmans, R., Phillips, S., Leathwick, J., Elith, J., 2015. Dismo: Species distribution modeling. R Package Version 1.0–12.

Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93, 679–688.

Inman, R., Franklin, J., Esque, T., Nussear, K., 2021. Comparing sample bias correction methods for species distribution modeling using virtual species. Ecosphere 12, e03422.

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., Gutiérrez, J.M., 2015. A framework for species distribution modelling with improved pseudo-absence generation. Ecol. Model. 312, 166–174.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. New Phytol. 11, 37–50.

Jeliazkov, A., Chiron, F., Garnier, J., Besnard, A., Silvestre, M., Jiguet, F., 2014. Level-dependence of the relationships between amphibian biodiversity and environment in pond systems within an intensive agricultural landscape. Hydrobiologia 723, 7–23.

Jeliazkov, A., Mimet, A., Chargé, R., Jiguet, F., Devictor, V., Chiron, F., 2016. Impacts of agricultural intensification on bird communities: New insights from a multi-level and multi-facet approach of biodiversity. Agricul. Ecosyst. Environ. 216, 9–22.

Jiménez, L., Soberón, J., 2020. Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. Methods Ecol. Evol. 11, 1571–1586.

Johnston, A., Moran, N., Musgrove, A., Fink, D., Baillie, S.R., 2020. Estimating species distributions from spatially biased citizen science data. Ecol. Model. 422, 108927.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R.M., White, M., Stone, L., 2017. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. Methods Ecol. Evol. 8, 420–430.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19, 1366–1379.

Lanner, J., Dubos, N., Geslin, B., Leroy, B., Hernández-Castellano, C., Dubaić, J.B., Bortolotti, L., Calafat, J.D., Ćetković, A., Flaminio, S., Le Féon, V., Margalef-Marrase, J., Orr, M., Pachinger, B., Ruzzier, E., Smagghe, G., Tuerlings, T., Vereecken, N., Meimberg, H., 2022. On the road: Anthropogenic factors drive the invasion risk of a wild solitary bee species. Sci. Total Environ. 827, 154246 https://doi.org/10.1016/j.scitotenv.2022.154246.

Leppo, E., 2021. ContDataQC. https://github.com/leppott/ContDataQC.

Leroy, B., Bellard, C., Dubos, N., Colliot, A., Vasseur, M., Courtial, C., Bakkenes, M., Canard, A., Ysnel, F., 2014. Forecasted climate and land use changes, and protected areas: the contrasting case of spiders. Divers. Distrib. 20, 686–697.

Leroy, B., Meynard, C.N., Bellard, C., Courchamp, F., 2016. Virtualspecies, an R package to generate virtual species distributions. Ecography 39, 599–607.

Li, W., Guo, Q., 2013. How to assess the prediction accuracy of species presence–absence models without absence data? Ecography 36, 788–799.

Liu, C., Newell, G., White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. Ecography 42, 535–548.

Lobo, J.M., Tognelli, M.F., 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. J. Nature Conserv. 19, 1–7.

Matutini, F., Baudry, J., Pain, G., Sineau, M., Pithon, J., 2021. How citizen science could improve species distribution models and their independent assessment. Ecol. Evolut. 11, 3028–3039.

McKinley, D.C., Miller-Rushing, A.J., Ballard, H.L., Bonney, R., Brown, H., Cook-Patton, S.C., Evans, D.M., French, R.A., Parrish, J.K., Phillips, T.B., Ryan, S.F., Shanley, L.A., Shirk, J.L., Stepenuck, K.F., Weltzin, J.F., Wiggins, A., Boyle, O.D., Briggs, R.D., Chapin, S.F., Hewitt, D.A., Preuss, P.W., Soukup, M.A., 2017. Citizen

science can improve conservation science, natural resource management, and environmental protection. Biol. Conserv. 208, 15–28.

Meynard, C.N., Kaplan, D.M., 2012. The effect of a gradual response to the environment on species distribution modeling performance. Ecography 35, 499–509.

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

Mikolajczak, A., Maréchal, D., Sanz, T., Isenmann, M., Thierion, V., Luque, S., 2015. Modelling spatial distributions of alpine vegetation: A graph theory approach to delineate ecologically-consistent species assemblages. Ecolog. Inform. 30, 196–202.

Monsarrat, S., Boshoff, A.F., Kerley, G.I.H., 2019. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. Ecography 42, 125–136.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S.D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N.A., Holt, R.D., Hui, F.K.C., Husby, M., Kålås, J.A., Lehikoinen, A., Luoto, M., Mod, H.K., Newell, G., Renner, I., Roslin, T., Soininen, J., Thuiller, W., Vanhatalo, J., Warton, D., White, M., Zimmermann, N.E., Gravel, D., Ovaskainen, O., 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecol. Monogr. 89, e01370.

Olivier, T., Thébault, E., Elias, M., Fontaine, B., Fontaine, C., 2020. Urbanization and agricultural intensification destabilize animal communities differently than diversity loss. Nature Commun. 11, 2686.

Otegui, J., Ariño, A.H., Encinas, M.A., Pando, F., 2013. Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). PLOS ONE 8, e55144.

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

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., Maiorano, L., 2017. Performance tradeoffs in target-group bias correction for species distribution models. Ecography 40, 1076–1087.

Robinson, O.J., Ruiz-Gutierrez, V., Fink, D., 2018. Correcting for bias in distribution modelling for rare species using citizen science data. Divers. Distrib. 24, 460–472.

Briscoe Runquist, R.D., Lake, T., Tiffin, P., Moeller, D.A., 2019. Species distribution models throughout the invasion history of Palmer amaranth predict regions at risk of future invasion and reveal challenges with modeling rapidly shifting geographic ranges. Sci. Rep. 9, 2426.

Rödder, D., Engler, J.O., 2011. Quantitative metrics of overlaps in Grinnellian niches: advances and possible drawbacks. Glob. Ecol. Biogeogr. 20, 915–927.

Schwartz, M.W., 2012. Using niche models with climate projections to inform conservation management decisions. Biol. Conserv. 155, 149–156.

Senay, S.D., Worner, S.P., Ikeda, T., 2013. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. PLOS ONE 8, e71218.

Sicacha-Parada, J., Steinsland, I., Cretois, B., Borgelt, J., 2020. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. Spatial Stat. 100446.

Soberón, J., Nakamura, M., 2009. Niches and distributional areas: Concepts, methods, and assumptions. Proc. Nat. Acad. Sci. 106, 19644–19650.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Model. 148, 1–13.

Stolar, J., Nielsen, S.E., 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. Divers. Distrib. 21, 595–608.

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A.C., Guisan, A., 2014. Measuring the relative effect of factors affecting species distribution model predictions. In: Methods Ecol. Evol., p. 5.

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD – a platform for ensemble forecasting of species distributions. Ecography 32, 369–373.

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, 1084–1091.

Vollering, J., Halvorsen, R., Auestad, I., Rydgren, K., 2019. Bunching up the background betters bias in species distribution models. Ecography 42, 1717–1727.

Warren, D.L., Glor, R.E., Turelli, M., 2010. ENMTools: a toolbox for comparative studies of environmental niche models. Ecography 33, 607–611.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Grant, E.H.C., Veran, S., 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods Ecol. Evol. 4, 236–243.