Received: 1 October 2020 | Revised: 2 June 2021 | Accepted: 27 July 2021

DOI: 10.1111/geb.13383

METHOD

Global Ecology and Biogeography | A Journal of Macroecology | WILEY

Novel methods to correct for observer and sampling bias in presence-only species distribution models

Yohann Chauvier[1] | Niklaus E. Zimmermann[1] | Giovanni Poggiato[2] | Daria Bystrova[2] | Philipp Brun[1] | Wilfried Thuiller[2]

[1]Swiss Federal Research Institute (WSL), Birmensdorf, Switzerland

[2]Laboratoire d'Ecologie Alpine, LECA, CNRS, Université Grenoble Alpes, Université Savoie Mont Blanc, Grenoble, France

Correspondence
Yohann Chauvier, Swiss Federal Research Institute (WSL), 8903 Birmensdorf, Switzerland.
Email: yohann.chauvier@wsl.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 310030L_170059; Agence Nationale de la Recherche, Grant/Award Number: ANR-10-LAB-56, ANR-15-IDEX-02 and ANR-16-CE93-004

Handling Editor: Huijie Qiao

Abstract

Aim: While species distribution models (SDMs) are standard tools to predict species distributions, they can suffer from observation and sampling biases, particularly presence-only SDMs, which often rely on species observations from non-standardized sampling efforts. To address this issue, sampling background points with a target-group strategy is commonly used, although more robust strategies and refinements could be implemented. Here, we exploited a dataset of plant species from the European Alps to propose and demonstrate efficient ways to correct for observer and sampling bias in presence-only models.

Innovation: Recent methods correct for observer bias by including covariates related to accessibility in model calibrations (classic bias covariate correction, Classic-BCC). However, depending on how species are sampled, accessibility covariates may not sufficiently capture observer bias. Here, we introduced BCCs more directly related to sampling effort, as well as a novel corrective method based on stratified resampling of the observational dataset before model calibration (environmental bias correction, EBC). We compared, individually and jointly, the effect of EBC and different BCC strategies, when modelling the distributions of 1,900 plant species. We evaluated model performance with spatial block split-sampling and independent test data, and assessed the accuracy of plant diversity predictions across the European Alps.

Main conclusions: Implementing EBC with BCC showed best results for every evaluation method. Particularly, adding the observation density of a target group as a bias covariate (Target-BCC) gave the most realistic modelled species distributions, with a clear positive correlation ($r \simeq .5$) found between predicted and expert-based species richness. Although EBC must be carefully implemented in a species-specific manner, such limitations may be addressed via automated diagnostics included in a provided R function. Implementing EBC and bias covariate correction together may allow future studies to address efficiently observer bias in presence-only models, and overcome the standard need of an independent test dataset for model evaluation.

KEYWORDS

background data, cluster, covariate correction, environmental stratification, independent dataset, plant species, point process model, random stratified sampling, survey effort, target group

1 | INTRODUCTION

Analysing diversity patterns and assessing how species are distributed in space, time and along environmental gradients are central topics in ecology (Barthlott et al., 1996; MacArthur, 1965; Von Humboldt & Bonpland, 2010). Predicting current and future species distributions is frequently done using species distribution models (SDMs; Araújo et al., 2019; Pearson, 2010; Thuiller et al., 2009), which are statistical methods that relate species records to co-occurring environmental or anthropogenic conditions (Brun et al., 2020; Graham et al., 2011; Guisan & Zimmerman, 2000). Over the last two decades, SDMs have gained increasing popularity among researchers, with studies investigating and comparing SDM methods and their specific assets, as well as evaluating the reliability of their predictions compared to field data (Elith & Graham, 2009; Elith et al., 2006; Guisan & Thuiller, 2005; Guisan et al., 2017). Currently, SDMs based on presence-only data enjoy high popularity, due to the rapid growth of online databases of species observations from scientific, naturalist and citizen-science initiatives (Samy et al., 2013; Wüest et al., 2020). There are several types of presence-only SDMs, and one of the most well-known methods is Maxent (Phillips et al., 2006), a special case of point-process model (PPM; Warton & Shepherd, 2010). PPMs are common tools to model presence-only data in other fields (e.g. seismology, epidemiology, neurology and economics), and they have been recently introduced in ecology as a type of presence-only SDM. Being proportional to Maxent, but with many additional advantages (Renner et al., 2015; Renner & Warton, 2013), this method is becoming one of the tools of choice for presence-only models.

Presence-only SDMs are an appealing tool, but their implementation necessitates following good modelling practice (Araújo et al., 2019), with correcting for observer bias often being regarded as the most important part (Graham et al., 2004; Phillips et al., 2009). Observer or sampling bias of species records results from excessive effort in distinct geographic regions of the study area; for example along roads, coasts, rivers, towards low elevations, near towns or biological field stations (Chauvier et al., 2021; Fithian et al., 2015; Graham et al., 2004; Reddy & Dávalos, 2003). Moreover, bias of observational datasets in geographic space generally results in bias in environmental space (Bystriakova et al., 2012; Graham et al., 2004; Phillips et al., 2009). This may lead to non-representative ecological preferences of species, and, in turn, distorted SDM fits and predictive outputs, particularly if a part of the species environmental niche is greatly over-sampled (Hastie & Fithian, 2013). However, various methods of bias correction have been suggested and combined in the literature. For example, spatial thinning of species observations according to the study area's resolution (Aiello-Lammens et al., 2015; Kiedrzyński et al., 2017; Steen et al., 2020), model-based corrections (Komori et al., 2020; Stolar & Nielsen, 2015), or combining species observations with large survey data (Fithian et al., 2015; Fletcher et al., 2016) have been recommended. Yet overall, sampling background data with the target-group strategy—that is, using similar sampling design/bias to sample background points as observed species presence points—remains currently the most popular approach (Botella et al., 2020; Hertzog et al., 2014; Kramer-Schadt et al., 2013; Phillips et al., 2009; Righetti et al., 2019). While this approach seems to improve model predictive performance, limitations regarding the number of background points to sample, and how to define target groups remain (Cerasoli et al., 2017; Warton et al., 2013), and might lead to lack of observer bias correction (Hanberry et al., 2012; Hertzog et al., 2014; Iturbide et al., 2015).

In response to these limitations, new corrective strategies based on bias covariate correction (BCC) have recently been implemented for presence-only models (Merow et al., 2016; Warton et al., 2013). During the model calibration and on top of the environmental variables, one or several bias covariates are added to account for imperfect sampling. These covariates are then kept constant (mean or zero) when predictions are made. By this, the parameters estimated during the model inference are corrected for the observer bias. Distance to/density of roads and cities (here called Classic-BCC; Bonnet-Lebrun et al., 2020; El-Gabbas & Dormann, 2018a, 2018b; Warton et al., 2013) have been shown to be particularly useful bias covariates that can improve predictive outputs. These improvements, however, strongly depend on whether the chosen bias covariates manage to capture the geographic bias existing in the data. Efficient bias covariate correction is therefore strongly dependent on preliminary spatial diagnostics to clearly uncover the cause of geographic bias (Albert et al., 2010; Amano & Sutherland, 2013; Warton et al., 2013). This need could be overcome if the selected bias covariates were to directly summarize the patterns of effort bias in the data, allowing for automated BCC implementation. Here, as a first step, we attempt to integrate the whole density of observations in the study area (i.e. target-group observation density; Figure 1a) as bias covariate within PPMs, alone (referred to as Target-BCC), and in combination with Classic-BCC (referred to as Mixed-BCC).

As already mentioned, correcting observer bias in presence-only models is increasingly done with BCC. Still, this approach does not address the potential environmental bias that may occur in the sampling design of observational datasets. Before data collection, appropriate sampling design should be environmentally stratified (Albert et al., 2010; Austin & Heyligers, 1989; Hirzel & Guisan, 2002; Mohler, 1983). Indeed, sampling frequencies in environmental space may still remain skewed if species observations are not initially sampled according to an environmental stratification. Moreover, BCC effectiveness is generally known to decay proportionally to the correlation between observer bias and environmental predictors (Warton et al., 2013). For example, if a bias covariate is strongly correlated with temperature, correcting predictions for the calibrated observer bias would also correct predictions for the calibrated temperature, thus skewing the species response. As a second step, we thus address this limitation within a novel corrective method based on random stratified sampling (referred to as environmental bias correction or EBC). Correcting sampling bias by filtering, before

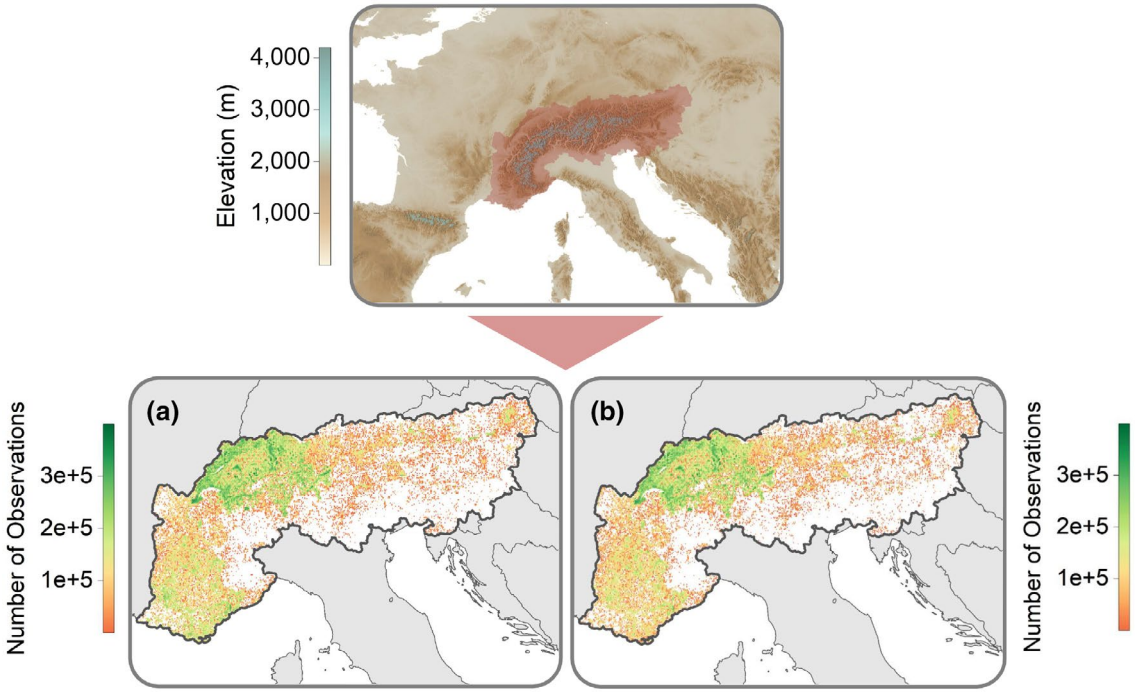


FIGURE 1 Distribution of species observation densities across the extended European Alps. (a) shows the whole target group observation densities (i.e. *Target group v1.0* layer). It includes 6,523,980 observations for 3,560 species. (b) shows observation densities used to model species distributions. It represents 5,369,442 observations for 1,900 species. Distribution of species densities was aggregated at 3-km resolution for better visual representation and log transformed

model calibrations, single-species observations based on environmental conditions has previously been addressed, with contrasting results (Fourcade et al., 2014; Varela et al., 2014). Our method extends such developments by correcting potential environmental bias by applying an artificial environmental stratification to the whole observational dataset. More specifically, environmental clusters that represent the main environmental conditions are generated across the study area. Within each cluster, the original observations of species are then concurrently subsampled without altering their ecological preferences. Depending on the research question or system studied, the type of random stratified sampling selected may be equal (sampling size equal across clusters) or proportional (sampling size proportionate to each cluster's area) across the study region (Boschetti et al., 2018; Hirzel & Guisan, 2002; Williams & Brown, 2019).

In this study, we tested the potential of the two above-mentioned EBC strategies as well as the Target-BCC approach to correct for observer and sampling bias in presence-only SDMs. More precisely, we investigated the effectiveness of 'EBC equal-stratified' (EBCe) and 'EBC proportional-stratified' (EBCp) strategies, along with different BCC strategies considering Classic-, Target- and Mixed-BCC, to model the distribution of 1,900 plant species across the European Alps. To provide a valuable comparative analysis, we evaluated the SDMs with both spatial block split-sampling (BSS) and an independent test dataset (Flora Alpina; Aeschimann et al., 2004). From the model projections, we generated diversity maps for each corrective strategy and compared them to a map of expert-based plant diversity of the European Alps.

2 | MATERIALS AND METHODS

2.1 | Study area

The study area covered the European Alps, as defined by an enlarged version of the official Alpine Convention perimeter (Permanent Secretariat of the Alpine Convention, 2009). The enlargement consisted of adding Switzerland entirely, as well as two French departments, Ain and Bouches-du-Rhône, for which we had extremely well-documented species observations. The size of the total study area is 294,994 km^2, and includes a wide range of climatic and topographic conditions.

2.2 | Species observations

The observational dataset used in this study was compiled from more than 210 individual sources, with the largest contributions from the National Data and Information Centre on the Swiss Flora (InfoFlora; *c.* 48%), the French National Alpine Botanical Conservatory (CBNA; *c.* 19%), the French National Mediterranean Botanical Conservatory (CBNMED; *c.* 5%), and the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/; *c.* 2%). All datasets were merged after unifying the species taxonomy, and after severely filtering inaccurate GBIF geo-referenced observations (see Chauvier et al., 2021 for more information on methods).

In total, the original observational dataset included 3,560 species (Figure 1a). This set was further filtered according to the prevalence

of each species (or proportion of 1-km pixels occupied); that is, species occurring in fewer than 100 pixels across the study area were removed. In total, the refined observational dataset included 1,900 species (Figure 1b) for model calibration (see Supporting Information Appendix S1: Table S1 for information on data distribution ranges). It is important to note that for species with >10,000 observations, we sampled randomly without replacement a subset of 10,000 observations for better computation efficiency (following Thuiller et al., 2018).

Additionally, an independent and unbiased test dataset, reporting the empirical distributional range of our 1,900 plant species over the European Alps, was constructed from expert-based information available in the Flora Alpina (FA) for a total of 54 political units (Aeschimann et al., 2004). For each species, we generated binary FA presence/absence rasters at a 1-km resolution by intersecting the polygons of its expert-based native political units with its expert-based elevation preference (see Supporting Information Appendix S1: Table S2, Figure S1 for more information). The 1,900 species binary layers were used for independent model evaluation, and to construct a map of expert-based plant diversity of the European Alps.

2.3 | Environmental data

We extracted all 19 bioclimatic variables from the Climatologies at High resolution for the Earth's Land Surface Areas (CHELSA) portal (Karger et al., 2017, http://chelsa-climate.org/), available for the time period 1979–2013. All predictors were kept at a 1-km spatial resolution and projected to the standard Lambert azimuthal equal area projection for Europe (EPSG:3035). A principal component analysis (PCA) was computed among all 19 variables to obtain PCA axes summarizing the main climatic patterns. We kept the first five PCA axes that cumulatively explained >90% of the total variance.

2.4 | Model calibration

We used point-process models (PPMs), where model output represents the intensity of the expected number of species occurrences per unit area, which is modelled as a log-linear function of the environmental covariates (Renner et al., 2015). PPMs were calibrated as a 'down-weighted Poisson regression' (DWPR, following Renner et al., 2015), using a generalized linear model (McCullagh, 1984) with second-order polynomials for all covariates (see Supporting Information Appendix S1: Text S1 for more details). Quadrature points (commonly referred to as background data) were here used to estimate the model log likelihood (see Supporting Information Text S1), and sampled randomly without replacement across the study area over a 1-km regular mesh. We estimated for each PPM the appropriate number of quadrature points by running 10 repeated series of DWPR while gradually increasing the number of randomly sampled points from 500 to 800,000, as explained in Renner et al. (2015) (in Figure 2b). The number of quadrature points at which the log likelihood converged was then automatically kept (see Supporting Information Appendix S1: Figure S2 and Text S1).

2.5 | Bias covariate correction (BCC)

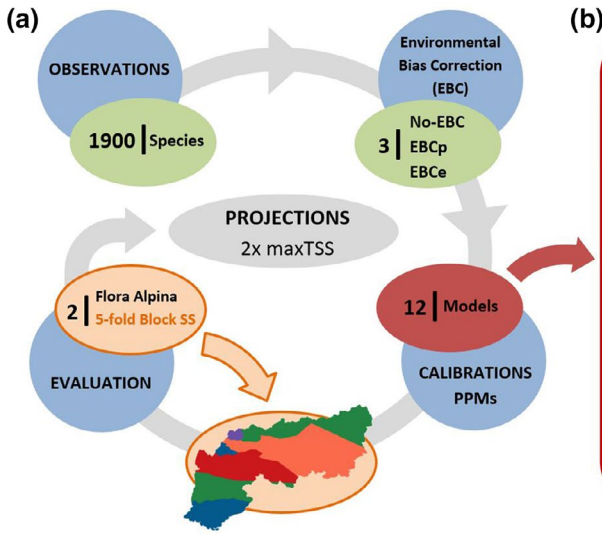
In total, we generated six bias covariates approximating survey effort. For Classic-BCC, four covariates were based on roads and cities. *Distance to roads* (a) and *to cities* (b) were generated based on OpenStreetMap (https://www.openstreetmap.org). All roads and cities of our study area were extracted from this source and converted into two binary 100-m grids. Distances to roads and cities were, respectively, calculated with the Geospatial Data Abstraction Library (GDAL; https://gdal.org/) and aggregated by sum to 1-km resolution (Bonnet-Lebrun et al., 2020; El-Gabbas & Dormann, 2018a). *Density of roads* (c) and *of cities* (d) were obtained by aggregating the binary 100-m layers by sum to 1-km resolution.

For Target-BCC, two covariates were based on the target group observation density (of the original observational dataset of 3,560 species) across the study area (Figure 1a). *Target group v1.0* (e) represented the observation density of all 3,560 species at 1-km resolution (see Figure 1a), whereas *Target group v2.0* (f) was a kernel-smoothed sampling intensity surface of the former (see Supporting Information Appendix S1: Figure S3).

All bias covariates were projected to the standard Lambert azimuthal equal area projection for Europe (EPSG:3035), after square root transformation (following Renner et al., 2015; Warton et al., 2013).

For each species, 12 PPMs were calibrated (see Figure 2a). Each PPM included as predictors, the first five climate PCA axes, and as for BCC, a unique combination of bias covariates (Figure 2b). No-BCC had no bias covariate included. Classic-BCC used *Density of roads* and *of cities* (BCCde), *Distance to roads* and *to cites* (BCCdi) or both (BCCdd) as bias covariates. Our proposal, Target-BCC, used *Target group v1.0* (BCCtg1) and its analogue *Target group v2.0* (BCCtg2) as bias covariates. Finally, we tested six combinations of Classic- and Target-BCC (Mixed-BCC) in order to test whether combined BCC strategies improve observer bias correction; that is, BCCtg1.de, BCCtg1.di, BCCtg1.dd, BCCtg2.de, BCCtg2.di and BCCtg2.dd.

Implementing BCC in PPMs allows for modelling species observation intensities both as a function of environmental predictors, and of an assumed observer bias (Renner & Warton, 2013; Warton et al., 2013, see Supporting Information Appendix S1: Text S2 for formula). Once calibrated, all models were spatially projected, with bias covariates being set to a constant value of 0 for all cells to correct for the fitted observer bias (following Warton et al., 2013). It is important to note that all bias covariates were weakly correlated with all climate PCA predictors; that is, Pearson's $|r| < .33$ (see Supporting Information Appendix S1: Figure S4 for more details). Environmental effects were therefore hardly masked by observer-bias effects during model calibration (Warton et al., 2013).



(b)

Bias Covariate Correction (BCC)	ID	ROAD/CITY	TARGET GROUP
No-BCC	No-BCC	-	-
Classic-BCC	BCCde	Density	-
	BCCdi	Distance	-
	BCCdd	Both	-
Mixed-BCC	BCCtg1.de	Density	v1.0
	BCCtg1.di	Distance	v1.0
	BCCtg1.dd	Both	v1.0
	BCCtg2.de	Density	v2.0
	BCCtg2.di	Distance	v2.0
	BCCtg2.dd	Both	v2.0
Target-BCC	BCCtg2	-	v2.0
	BCCtg1	-	v1.0

FIGURE 2 Model framework of the study. (a) shows the pipeline applied for our analysis in order to generate 684,000 projections for 1,900 species. For each species, 12 types of bias covariate correction (BCC) were applied, and each model was calibrated with three types of environmental bias correction (EBC); no (No-EBC), proportional-stratified (EBCp) and equal-stratified EBC (EBCe). Two types of model evaluation were performed: with fivefold spatial block split-sampling (5-fold Block SS) and the Flora Alpina independent dataset (FA). Each projection was converted to binary presences/absences using two different estimates of the maximum true skill statistic threshold (maxTSS); that is, for each evaluation. (b) describes the 12 BCC strategies used along five climate principal component analysis (PCA) axes to calibrate the point-process models (PPMs)

2.6 | Environmental bias correction (EBC)

The model design described so far was repeated for all 1,900 species with and without an additional, novel corrective method based on random stratified sampling (EBC). Stratified sampling design of species observations may be equal or proportional (Boschetti et al., 2018; Hirzel & Guisan, 2002; Williams & Brown, 2019). We therefore accounted within our models for three different strategies: no (No-EBC), proportional-stratified (EBCp) and equal-stratified EBC (EBCe; Figure 2a). To this end, 100 clusters of similar environmental conditions were first generated over the European Alps (Figure 3a, see Supporting Information Appendix S1: Figure S5 for non-summarized legends). Environmental conditions were defined based on the first five climate PCA axes, from which we identified clusters with the clustering large applications (CLARA) method, using the R package *cluster* (Maechler et al., 2019; R Core Team, 2020; Reynolds et al., 2006; Schubert & Rousseeuw, 2019). The number of clusters was selected to be large enough to account for the environmental complexity of the study area.

EBC corrects, before model calibrations, potential environmental bias in the design of an observational dataset, by artificially subsampling original species observations according to the environmental stratification of the study area (Figure 3b,c). The equal-stratified design, EBCe, scales the number of species observations in each cluster proportional to the total number of observations, across all species, found in the cluster with the highest observation density (cluster no. 31 in Figure 3). The correction is thus not applied to species individually, but rather across all species, which means that although the number of observations per species in a cluster is changed, its proportion relative to the other species remains constant. By this, each species' environmental bias is corrected among clusters without altering its ecological preference. The proportional-stratified design, EBCp, additionally multiplies the number of observations per species and clusters by the cluster's area.

The resulting output then indicates a new number of observations per cluster and species that may be subsampled with replacement over the original observational dataset (see Supporting Information Appendix S1: Figure S6 for observation counts along climate predictors). EBCp and EBCe are implemented in the R function *wsl.ebc* (R Core Team, 2020; see Supporting Information Appendix S2 for function, description, examples, code and parameters).

2.7 | Model evaluation

We assessed the predictive performance of each PPM under fivefold spatial block split-sampling tests (BSS; Roberts et al., 2017). This approach requires preliminarily delineating independent spatial blocks to partition observations in geographic space. Here we evenly partitioned each set of species observations and quadrature points into 10 blocks and combined them into five folds (two blocks per fold were combined so that differences in number of observations across folds were minimal). To this end, we first ran a partitioning around medoids clustering on the observation coordinates and, in a second step, *k*-nearest neighbour classification to assign quadrature points to observation blocks (see Brun et al., 2020 for more details). Model evaluation was also performed against the independent Flora Alpina dataset. For each PPM, the same number of Flora Alpina presences/absences as quadrature points were sampled, to ensure balanced repartitioning among folds. To make the

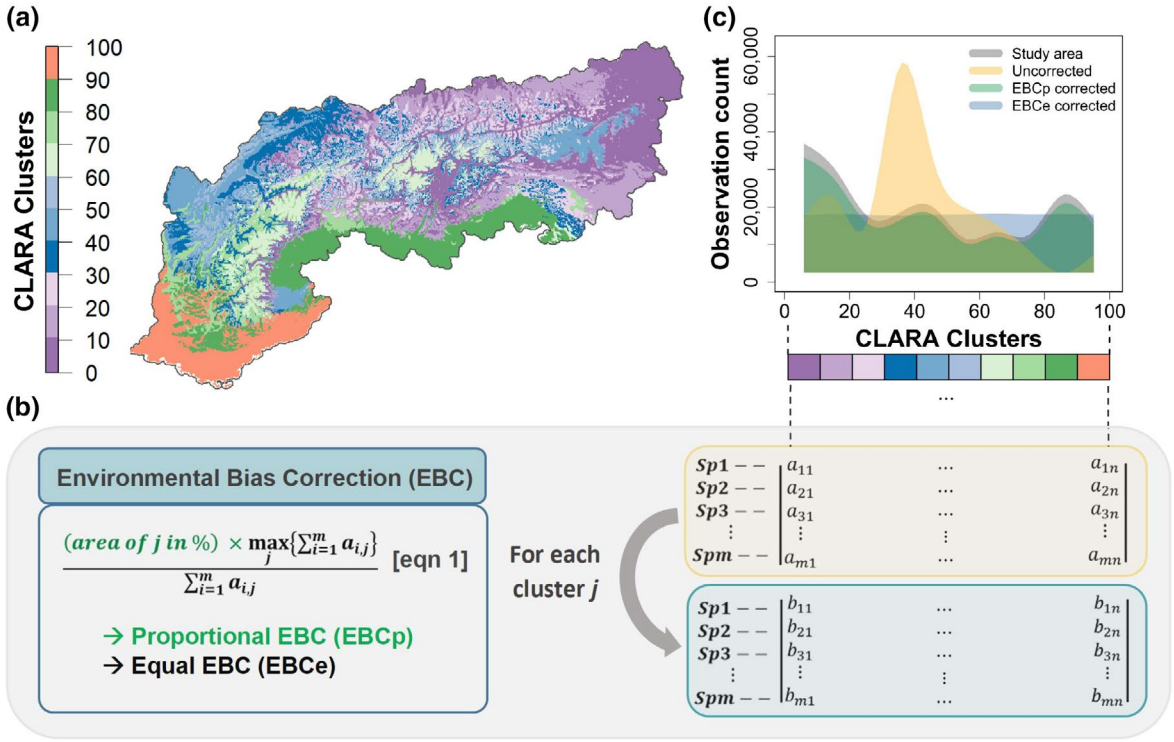


FIGURE 3 Methodology framework of the environmental bias correction (EBC). (a) summarizes with a 10-colour palette the 100 environmental clusters obtained using the clustering large applications (CLARA) method for better visualization. (b) defines the EBC methodologies that were applied in the study for 1,900 species. Each cluster *j* (matrix columns) summaries per species *m* (matrix rows) the original number of observations (yellow box). Equal-stratified EBC (EBCe) here multiplies and divides each cluster *j*, by the total number of observations across all species in cluster no. 31 and *j*, respectively (black terms in Equation 1). Proportional-stratified EBC (EBCp) additionally multiplies each cluster *j* by its area in pixels (added green term in Equation 1). Results indicate a new number of observations (EBCe or EBCp) per cluster and species (blue box), which are subsampled with replacement from the original observational dataset. (c) summarizes per cluster the observation densities in the different datasets, for uncorrected (yellow shade), EBCp (green shade) and EBCe (blue shade) corrections, relative to the cluster distribution across the whole study area (grey shade). For better visualization, 200,000 and 180,000 cluster values were sampled randomly without replacements over the study area and for the three sets of observations (uncorrected, EBCp and EBCe corrected). For more clarity, an interpolation spline between palette counts was applied

two evaluation strategies comparable, the same block structure was applied to the Flora Alpina presences/absences (see Supporting Information Appendix S1: Figure S7 for spatial block split-sampling details). PPM performance was thus concurrently calculated using presences/quadrature points, and Flora Alpina presences/absences of the left-out fold, for spatial block split-sampling and Flora Alpina evaluations, respectively.

Model predictions were evaluated with the true skill statistic (TSS; Allouche et al., 2006). For each fold, projected observation intensities were converted to a binary presence/absence raster using two different estimates of the maximum TSS threshold (maxTSS); that is, one per evaluation type. For each species, EBC-BCC strategy, and evaluation, binary rasters of each fold were summed, and presences were assigned to pixels when at least half of the folds predicted so (Araújo & New, 2007; Guisan et al., 2017; Thuiller et al., 2009). In total, we generated and stacked 1,900 binary rasters of species presence per EBC-BCC strategy ($n = 36$) and evaluation type ($n = 2$), obtaining 72 maps of modelled species diversity over the European Alps that we could compare to an expert-based map.

Finally, we used nonparametric Friedman tests to detect differences in model predictive performances (Friedman X^2) between all BCC treatments tested in our study, when EBCp/e applies or not. Post-hoc Nemenyi tests were applied to assess pairwise differences between treatment means (*p-value < .05) using the R package *PMCMR* (Pohlert, 2014; R Core Team, 2020).

3 | RESULTS

When No-EBC was considered, models including No-BCC showed higher TSS than models including Target-BCC for the split-sampling evaluation (Friedman $X^2 = 546.1$, p-value < .001, Figure 4b), but lower TSS for the independent Flora Alpina evaluation (Friedman $X^2 = 4{,}313.4$, p-value < .001, Figure 4e). Classic-BCC had overall the same performance as No-BCC models, while Mixed-BCC models showed similar performance to that of Target-BCC models. All results were confirmed with the Area Under the ROC Curve (AUC) and Boyce index tests (see Supporting Information Appendix S1: Figure S8, Table S3.

Projected species diversity for No-BCC models was more biased towards Switzerland under split-sampling than independent evaluation (Figure 4a,d). Target-BCC decreased this spatial bias for both evaluation strategies, although more strongly for the independent evaluation (BCCtg1: Figure 4c,f; BCCtg2: see Supporting Information Appendix S1: Figure S9a). Indeed, Target-BCC projected diversity displayed stronger correlations with the expert-based diversity for independent evaluation ($r_{(BCCtg1)}$ = .32, $r_{(BCCtg2)}$ = .35) rather than under split-sampling ($r_{(BCCtg1)}$ = −.12, $r_{(BCCtg2)}$ = −.11; see Supporting Information Appendix S1: Figure S10).

When EBC was considered, the quality of model predictions and correlations with expert-based diversity increased, and differences in model performances between all models decreased. For split-sampling evaluations, models including Target-BCC performed similarly to models including No-BCC (EBCp: Friedman X^2 = 372.2, p-value < .001, Figure 4h; EBCe: Friedman X^2 = 257.3, p-value < .001, see Supporting Information Appendix S1: Figure S11b). For the independent evaluation, Target-BCC performed better than No-BCC (EBCp: Friedman X^2 = 2,144.2, p-value < .001, Figure 4k; EBCe: Friedman X^2 = 2,806.5, p-value < .001, see Supporting Information Figure S11e). Classic-BCC had overall the same performances as did No-BCC models, and Mixed- as Target-BCC models. Again, our results were not sensitive to the evaluation metric used (see Supporting Information Figure S8, Table S3).

Adding EBCp/e to No-BCC led to very similar projections for both evaluation methods, with a less biased pattern towards Switzerland (Figure 4g,j; see Supporting Information Figure S11a,d). The same trend applied when adding EBCp/e to BCCtg1/2, but with higher diversity patterns at low elevations (Figure 4i,l and Supporting Information Figures S9b,c). Finally, Target-BCC projected diversity displayed strongest correlations with the expert-based diversity for both split sampling (EBCp: $r_{(BCCtg1)}$ = .37, $r_{(BCCtg2)}$ = .37; EBCe: $r_{(BCCtg1)}$ = .33, $r_{(BCCtg2)}$ = .33) and Flora Alpina evaluation (EBCp: $r_{(BCCtg1)}$ = .51, $r_{(BCCtg2)}$ = .51; EBCe: $r_{(BCCtg1)}$ = .56, $r_{(BCCtg2)}$ = .56; Figure 5 and Supporting Information Figure S10).

4 | DISCUSSION

In this paper, we compared different strategies to correct for observer and sampling bias in presence-only SDMs, and demonstrated that model predictions of plant species distributions in the European Alps are considerably improved when EBC is implemented (Figures 4 and 5). While BCC focuses on the correction of observer bias via covariate adjustment, EBC implements a complementary correction based on random stratified sampling (Austin & Heyligers, 1989; D'Antraccoli et al., 2020; Hirzel & Guisan, 2002; Mohler, 1983). Environmentally imbalanced sampling designs in observational datasets is a recurrent issue in ecological studies (Albert et al., 2010; Hirzel & Guisan, 2002). EBC corrects for this environmental bias by applying an artificial environmental stratification to the observational dataset through two design variants (EBCp, EBCe). This stratification corrects over all species observations the excessive sampling effort detected in specific environments of the study area, allowing known limitations regarding BCC to be overcome. Bias covariate correction works generally well to address geographic observer bias (Bonnet-Lebrun et al., 2020; El-Gabbas & Dormann, 2018a; Warton et al., 2013). However, this correction is only meaningful (a) if the observer bias covariate remains weakly to moderately correlated with the environmental predictors used in the model (Warton et al., 2013), and (b) if it adequately describes the origin of the bias in the data. EBC overcomes both limitations and may therefore be implemented independently to BCC.

Nevertheless, EBC should (a) only be applied if environmental bias is apparent, (b) ideally implemented for specific species, and (c) preserve the influence of the environmental cluster in which the species was originally most sampled. (a) Environmental bias in observational data is usually a consequence of spatial bias (Bystriakova et al., 2012; Phillips et al., 2009). The observational dataset must therefore be carefully explored in geographic and environmental space prior to analysis to detect potential environmental biases. Here, this bias was geographically (Figure 1) and environmentally (Figure 3) detected, with disproportionately high sampling within the administrative border of Switzerland. (b) Subsampling species observations in environmental space should not be applied to species whose environmental bias strongly differs from the overall one (i.e. if the number of original species observations per cluster is weakly correlated with that of the full dataset). In our analysis, this correlation was strong for *Abies alba* (r = .84), but weak for *Pinus cembra* (r = .13; Figure 6). Therefore, while EBC corrected the environmental bias of the former, it distorted the originally well-distributed observations of *Pinus cembra* in the latter, correcting for a non-existent environmental bias. (c) After implementing EBC, the cluster in which the species was originally most sampled no longer necessarily contains the highest number of subsampled observations. Subsampled observations may indeed be relatively more frequent in other clusters with originally stronger under-sampling. Although such corrective behaviour is pursued, one could require per species the cluster with the most original observations to be as representative as the cluster with the most corrected observations, especially if EBCp is applied (see arrows and *Abies alba*, *Datura stramonium*; Figure 6). Such optional adjustment could be particularly beneficial for SDM studies seeking to preserve the environmental influence of heavily sampled regions. In respect of (b) and (c), further descriptions and codes addressing such limitations are detailed in Supporting Information Appendix S2.

Although EBC may be implemented independently, we found that combining it with bias covariate correction yields better predictive performance and better agreement with our independent diversity map (Figures 4 and 5; as predicted in Table 1). Unless strong correlations between the bias and environmental covariates are found, BCC should be implemented together with EBC, as bias covariate correction remains strongly relevant if EBC does not initially succeed in correcting for the environmental bias (see *Pulmonaria obscura*; Figure 6). Particularly, Target-BCC improved model performance/predictions compared to Classic-BCC, with BCCtg1/2 producing very similar results (Supporting Information Figures S9–S11); that is,

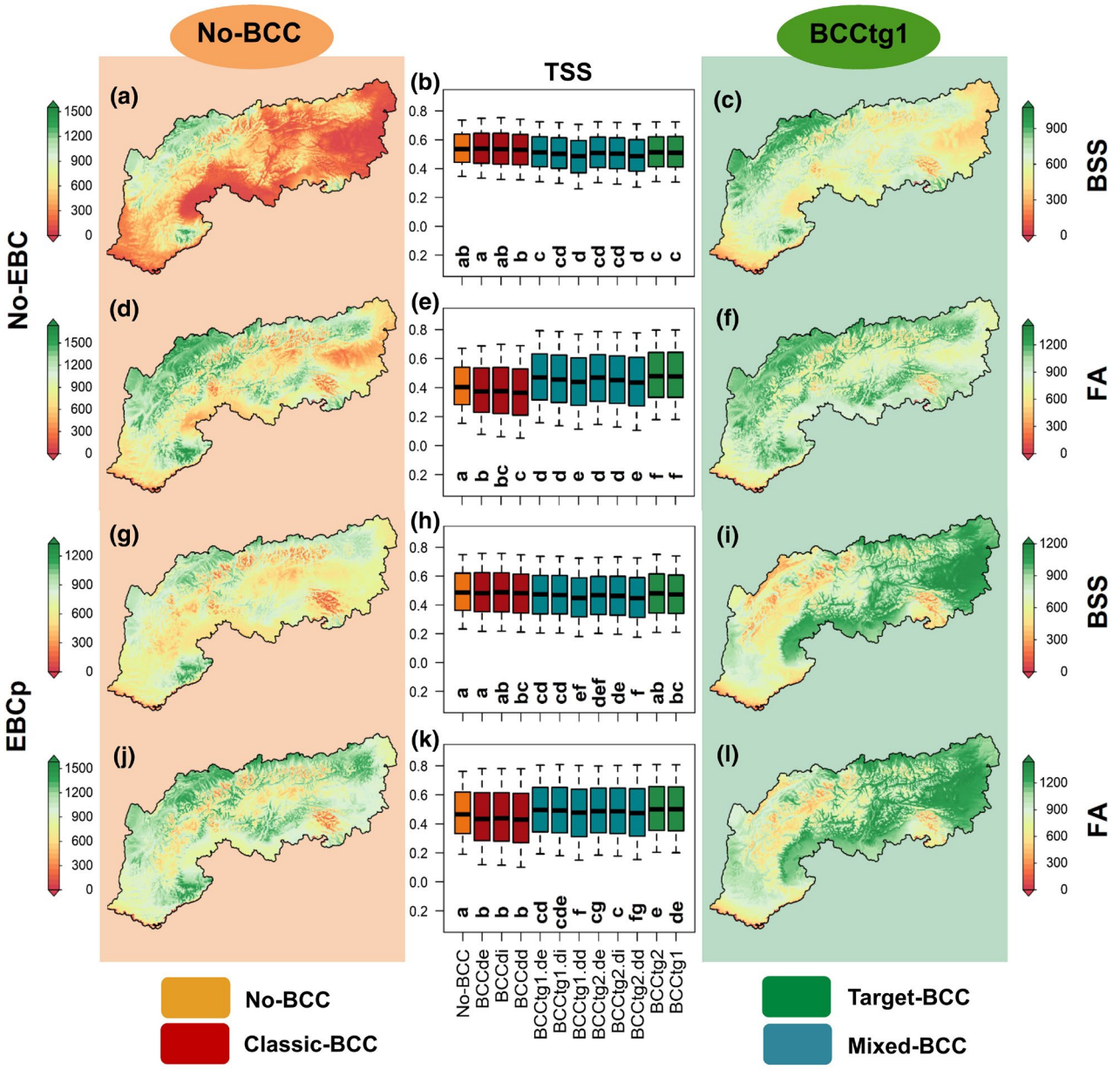


FIGURE 4 True skill statistic (TSS) of species distribution models when 12 bias covariate corrections (BCC), no environmental bias correction (No-EBC) and proportional-stratified EBC (EBCp) apply, for both block split-sampling (BSS) and Flora Alpina (FA) evaluations (b, e, h, k). Related species diversity (1,900 species) of No-BCC and BCC *Target group v1.0* (BCCtg1) are shown on the left and right panels respectively. Projected diversity is described here for No-EBC (a, c, d, f) and EBCp (g, i, j, l). The 12 BCC are categorized as follows: No-BCC (orange), Classic-BCC (red), Mixed-BCC (blue) and Target-BCC (green). Friedman tests were here applied for each of the four centre panels (***p-value < .001 for all): Friedman X^2 = 546.1, 4,313.4, 372.2 and 2,144.2 (b, e, h and k, respectively). All pairwise comparisons were run with post-hoc Nemenyi tests and displayed following a letter-based representation (p-value > .05)

an approximate density map of observations (*Target group v2.0*) is as efficient as the pixel-by-pixel target group observation density of the study area (*Target group v1.0*) when applying Target-BCC. While this correction can improve and simplify the implementation of BCC, we advise Mixed-BCC to be used by default, as Classic- and Target-BCC could perform differently depending on the observational dataset, and their combination may correct more robustly. In our study, Target-BCC likely performed best because the survey effort across our study area showed strong large-scale variations, with peaks in Switzerland and southern France. Although these regional or administrative biases are quite common across observational datasets, superimposing biases related to accessibility (e.g. roads, coasts, elevation or cities) are also frequent, and often species-specific. Including concurrently covariates targeting different types of bias should therefore make BCC more adaptive and performant.

Employing semi-dependent (spatial split block sampling) and independent (Flora Alpina dataset) evaluation strategies, as well as corresponding estimates of the maxTSS, revealed an 'evaluation

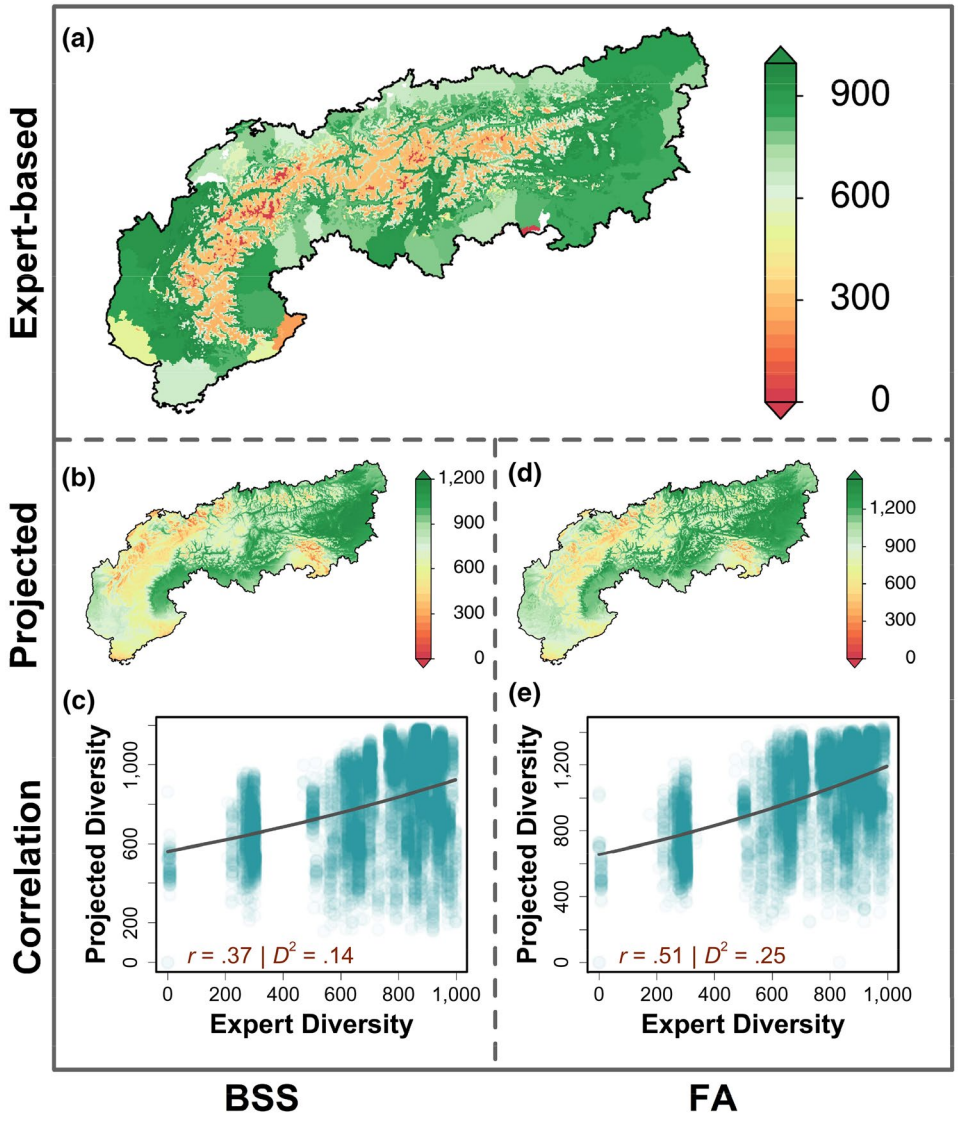


FIGURE 5 Comparison between expert-based species diversity (Flora Alpina) and Target-BCC (BCC *Target group v1.0*; BCCtg1) projected and binarized (maximum true skill statistic threshold, maxTSS) diversity, when proportional-stratified EBC (EBCp) is applied for split sampling and Flora Alpina evaluation. (a) represents the expert-based plant diversity constructed from the independent test dataset. (b) and (d) represent projected diversity maps found in Figure 4i and 4l, respectively. (c) and (e) show the statistical relationship between expert-based diversity and projected diversity, for block split-sampling (BSS) and Flora Alpina (FA) evaluation. For this analysis, 10,000 values were spatially sampled randomly without replacement per diversity layer. Both panels display Spearman's rank correlation coefficient *r*, and the explained deviance D^2 from a fitted Poisson-based generalized linear model (GLM) with first-order polynomials

paradox', which is conditional on whether the test data are truly independent. In agreement with other studies, we confirm that non-corrected model predictions seem to perform better than corrected ones, if the observations used for testing and calibrating originate from the same spatially biased observational dataset (Smith et al., 2021; Stolar & Nielsen, 2015; Warton et al., 2013). This paradox was particularly apparent under split-sampling evaluation, and when comparing No-BCC with Target-BCC models (Figure 4a–f). This paradox may be quite detrimental if SDMs are selected based on TSS, or other evaluation metrics, calculated from test data lacking independence. This may lead to inappropriate model selection, and consequently to erroneous spatial predictions of species distributions. Interestingly, when EBC is applied, the amplitude of the paradox strongly diminishes, and similar evaluation performances are obtained for split sampling and independent tests when predicting plant diversities (Figure 4g–l). This means that implementing EBC decreases the risk of inappropriate evaluations and therefore predictions, regardless of the evaluation procedure.

However, the correlations we found between predicted and expert-based diversities (Figure 5) remained moderate, and this may have different reasons: (a) the expert-based diversity is only based on filtered coarse political units, which may not represent

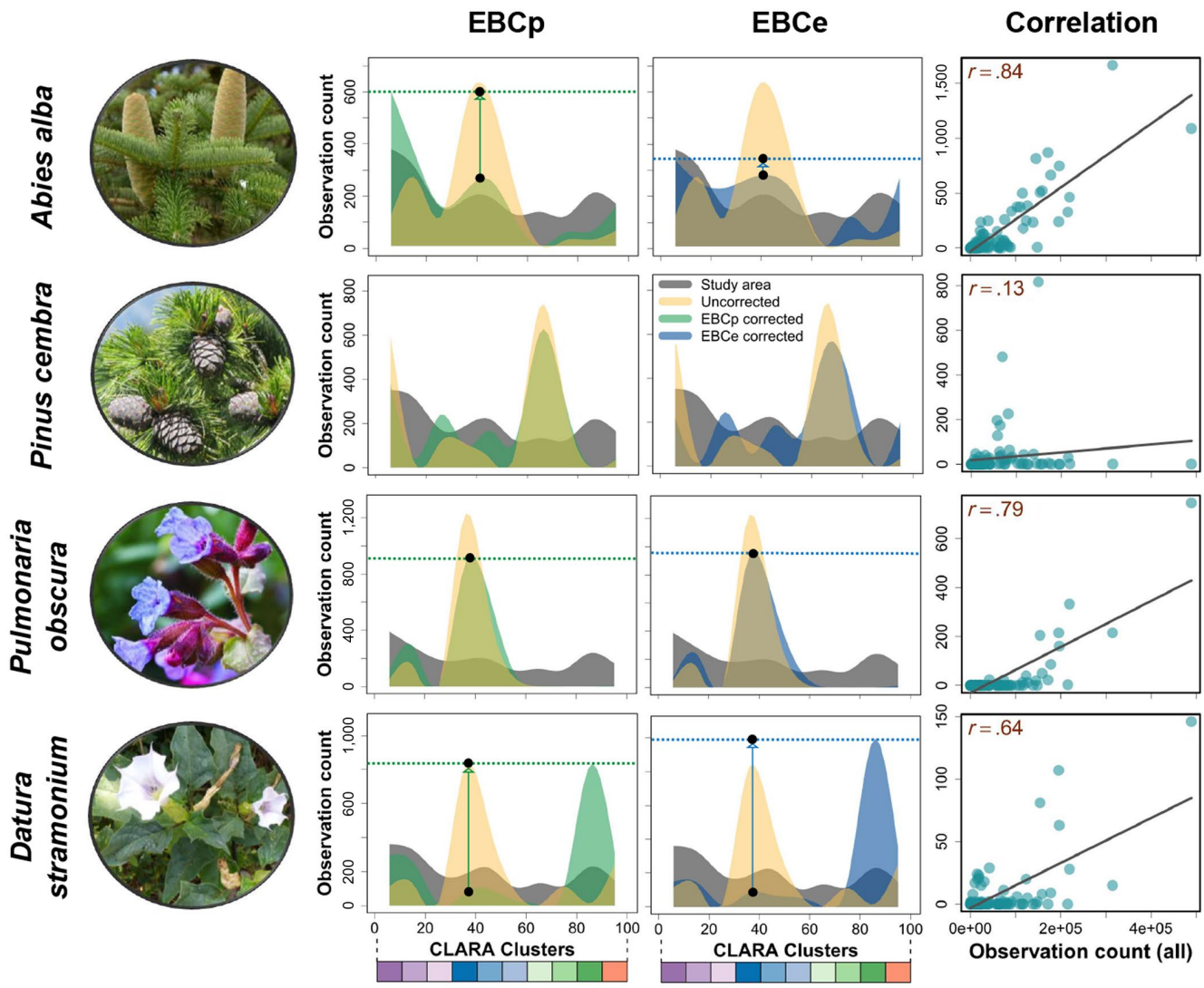


FIGURE 6 Summary of observation densities per cluster for four selected species, using uncorrected (yellow shade), proportional (EBCp; green shade; first column) and equal-stratified EBC (EBCe; blue shade; second column) corrected data. The data are shown relative to the cluster distribution across the study area (grey shade). Two thousand cluster values were sampled randomly without replacements within the study area and for the three sets of observations (uncorrected, EBCp and EBCe corrected). For more clarity, an interpolation spline between palette counts was applied. Arrows indicate the difference between EBC corrected observations of the densest original cluster (with the highest number of original observations) and the densest corrected cluster (with the highest number of corrected observations). The third column displays Pearson's correlations between the number of original observations per cluster for *Abies alba*, *Pinus cembra*, *Pulmonaria obscura* and *Datura stramonium* (first, second, third and fourth row, respectively) and that of the full dataset (all; see Figure 3c). CLARA = clustering large applications

the true and detailed shape of plant diversity in the European Alps. To simplify model comparisons and computations, (b) our study did not remove species models with comparably low TSS, and (c) only climate predictors were used, whereas soil and land cover have been shown to be essential for predicting more accurately plant species distribution (Chauvier et al., 2021). Finally, it is essential that sampling design is based on prior knowledge of the system studied (Albert et al., 2010; Arneill et al., 2019; D'Antraccoli et al., 2020). If available, EBC can include other prior variables (than climate predictors) that are thought to have an influence on the species distribution (e.g. longitude, latitude, elevation or habitat suitability maps). No consensus exists on the best random stratified sampling design; however, not all designs should be based on the environmental space of the study area. The same problem arises when ecological studies need to choose between a proportional or equal-stratified design when inventorying or collecting species samples (Hirzel & Guisan, 2002). Accordingly, EBCp and EBCe may be both implemented via our R function (R Core Team, 2020; see Supporting Information Appendix S2).

To conclude, we propose that combining both EBC and bias covariate correction might be the best choice in presence-only SDMs when predicting current and future species distributions. Given the

TABLE 1 Expected model performances and assumed strength of the observer bias in response to the two methods of bias correction introduced in this study

		Target-BCC					
		NO		YES			
EBC	NO	Bias high	Good	Bad	Bias low	BSS	EVALUATION TYPE
		Bias low	Bad	Good	Bias low (or none*)	FA	
	YES	Bias low	Average	Optimal	Bias none	BSS	
		Bias low	Average	Optimal	Bias none	FA	
		SPECIES RANGE	MODEL PERFORMANCE		SPECIES RANGE		

Note: Abbreviations: BSS, block split-sampling test; FA, independent test dataset from Flora Alpina; *, dependent on the degree of observer bias in the dataset.

Target-group bias covariate correction (Target-BCC) uses the full observation density across the study area as bias covariate (see Figure 1a). Environmental bias correction (EBC) uses a pre-processed and balanced sampling of all species observations within the clustered environmental space of the study area.

growing interest in using citizen-science data to follow and predict invasive species, to guide the reintroduction of species, or to protect biodiversity (Grünig et al., 2020; Hunter-Ayad et al., 2020; Lehtomäki et al., 2019), using these corrections may prove to be strongly beneficial for future ecological studies, which will increasingly implement presence-only models. Interestingly, all methods presented here may be extended independently to presence–absence data not subject to the same observer bias, and might turn out to be useful for further ecological applications.

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche–Schweizerische Nationalfonds (ANR-SNF) bilateral project OriginAlps, with grant numbers 310030L_170059 (Y.C., P.B., N.E.Z.) and ANR-16-CE93-004 (W.T.). W.T. also acknowledges the ANR for the project Generating Advances in Modeling Biodiversity And ecosystem Services (GAMBAS;ANR-18-CE02-0025) and the 'Investissement d'Avenir' grants managed by the ANR (Trajectories: ANR-15-IDEX-02; Montane: OSUG@2020: ANR-10-LAB-56). Furthermore, we thank all various contributors for sharing numerous datasets of species observations, and for making possible the large scope of this study. We also thank P. Descombes, C. Botella and S. Schiess for their useful inputs in data analysis and discussion.

AUTHOR CONTRIBUTIONS

Y.C., N.E.Z. and W.T. conceived the general idea and designed the study with the help of all authors; Y.C., G.P., D.B. and W.T. developed the methodology; Y.C. performed the analysis and led the writing of the manuscript. All authors interpreted results, significantly contributed to writing and editing, and gave final approval for publication.

DATA AVAILABILITY STATEMENT

All non-copyright data and scripts supporting the findings of this study are available on the EnviDat repository (https://www.envidat.ch/dataset/correct-observer-bias-only-sdms). The R function *wsl.ebc*, used to implement the described methodology, associated examples and parameter descriptions are available in Supporting Information Appendix S2.

ORCID

Yohann Chauvier https://orcid.org/0000-0001-9399-3192
Niklaus E. Zimmermann https://orcid.org/0000-0003-3099-9604
Philipp Brun https://orcid.org/0000-0002-2750-9793
Wilfried Thuiller https://orcid.org/0000-0002-5388-5274

REFERENCES

Aeschimann, D., Lauber, K., Moser, D. M., & Theurillat, J. P. (2004). *Flora alpina: ein Atlas sämtlicher 4500 Gefässpflanzen der Alpen* (ed. by 'Haupt').

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*, 541–545. https://doi.org/10.1111/ecog.01132

Albert, C. H., Yoccoz, N. G., Edwards, T. C., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography*, *33*, 1028–1037. https://doi.org/10.1111/j.1600-0587.2010.06421.x

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Amano, T., & Sutherland, W. J. (2013). Four barriers to the global understanding of biodiversity conservation: Wealth, language, geographical location and security. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1756), 20122649. https://doi.org/10.1098/rspb.2012.2649

Araújo, M. B., Anderson, R. P., Márcia, B. A., 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. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Araujo, M., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Arneill, G. E., Perrins, C. M., Wood, M. J., Murphy, D., Pisani, L., Jessopp, M. J., & Quinn, J. L. (2019). Sampling strategies for species with high breeding-site fidelity: A case study in burrow-nesting seabirds. *PLoS ONE*, *14*, 1–17. https://doi.org/10.1371/journal.pone.0221625

Austin, M. P., & Heyligers, P. C. (1989). Vegetation survey design for conservation: Gradsect sampling of forests in North-eastern New South Wales. *Biological Conservation*, *50*(1–4), 13–32. https://doi.org/10.1016/0006-3207(89)90003-7

Barthlott W., Lauer W., & Placke, A. (1996). Global distribution of species diversity in vascular plants: Towards a world map of phytodiversity. *Erdkunde*, *50*(1). https://doi.org/10.3112/erdkunde.1996.04.03

Bonnet-Lebrun, A. S., Karamanlidis, A. A., de Gabriel Hernando, M., Renner, I., & Gimenez, O. (2020). Identifying priority conservation areas for a recovering brown bear population in Greece using citizen science data. *Animal Conservation*, *23*, 83–93. https://doi.org/10.1111/acv.12522

Boschetti, L., Stehman, S. V., & Roy, D. P. (2016). A stratified random sampling design in space and time for regional to global scale burned area product validation. *Remote Sensing of Environment*, *186*, 465–478. https://doi.org/10.1016/j.rse.2016.09.016

Botella, C., Joly, A., Monestiez, P., Bonnet, P., & Munoz, F. (2020). Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *PLoS ONE*, *15*, 1–18. https://doi.org/10.1371/journal.pone.0232078

Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z., & Zimmermann, N. E. (2020). Model complexity affects species distribution projections under climate change. *Journal of Biogeography*, *47*, 130–142. https://doi.org/10.1111/jbi.13734

Bystriakova, N., Peregrym, M., Erkens, R. H. J., Bezsmertna, O., & Schneider, H. (2012). Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and Biodiversity*, *10*(3), 305–315. https://doi.org/10.1080/14772000.2012.705357

Cerasoli, F., Iannella, M., D'Alessandro, P., & Biondi, M. (2017). Comparing pseudo-absences generation techniques in boosted regression trees models for conservation purposes: A case study on amphibians in a protected area. *PLoS ONE*, *12*, 1–23. https://doi.org/10.1371/journal.pone.0187589

Chauvier, Y., Thuiller, W., Brun, P., Lavergne, S., Descombes, P., Karger, D. N., Renaud, J., & Zimmermann, N. E. (2021). Influence of climate, soil, and land cover on plant species distribution in the European Alps. *Ecological Monographs*, *91*, 1–14. https://doi.org/10.1002/ecm.1433

D'Antraccoli, M., Bacaro, G., Tordoni, E., Bedini, G., & Peruzzi, L. (2020). More species, less effort: Designing and comparing sampling strategies to draft optimised floristic inventories. *Perspectives in Plant Ecology, Evolution and Systematics*, *45*, 125547. https://doi.org/10.1016/j.ppees.2020.125547

El-Gabbas, A., & Dormann, C. F. (2018a). Improved species-occurrence predictions in data-poor regions: Using large-scale data and bias correction with down-weighted Poisson regression and Maxent. *Ecography*, *41*, 1161–1172. https://doi.org/10.1111/ecog.03149

El-Gabbas, A., & Dormann, C. F. (2018b). Wrong, but useful: Regional species distribution models may not be improved by range-wide data under biased sampling. *Ecology and Evolution*, *8*, 2196–2206. https://doi.org/10.1002/ece3.3834

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, *32*, 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

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 in Ecology and Evolution*, *6*, 424–438. https://doi.org/10.1111/2041-210X.12242

Fletcher, R. J., McCleery, R. A., Greene, D. U., & Tye, C. A. (2016). Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, *31*, 1369–1382. https://doi.org/10.1007/s10980-015-0327-9

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*, 1–13. https://doi.org/10.1371/journal.pone.0097122

Graham, C., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, *19*(9), 497–503. https://doi.org/10.1016/j.tree.2004.07.006

Graham, C. H., Loiselle, B. A., Velasquez-Tibata, J., & Cuesta, F. (2011). Species distribution modelling and the challenge of predicting future distributions. In S. K. Herzog, R. Martinez, P. M. Jørgensen, & H. Tiessen (Eds.), *Climate Change and Biodiversity in the Tropical Andes*, Chapter: 21. Inter-American Institute for Global Change Research (IAI) and Scientific Committee on Problems of the Environment (SCOPE).

Grünig, M., Mazzi, D., Calanca, P., Karger, D. N., & Pellissier, L. (2020). Crop and forest pest metawebs shift towards increased linkage and suitability overlap under climate change. *Communications Biology*, *3*(1). https://doi.org/10.1038/s42003-020-0962-9

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models*. Cambridge University Press. https://doi.org/10.1017/9781139028271

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/s0304-3800(00)00354-9

Hanberry, B. B., He, H. S., & Palik, B. J. (2012). Pseudoabsence generation strategies for species distribution models. *PLoS ONE*, *7*, 7–10. https://doi.org/10.1371/journal.pone.0044486

Hastie, T., & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, *36*, 864–867. https://doi.org/10.1111/j.1600-0587.2013.00321.x

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. *Diversity and Distributions*, *20*, 1403–1413. https://doi.org/10.1111/ddi.12249

Hirzel, A., & Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, *157*, 331–341. https://doi.org/10.1016/S0304-3800(02)00203-X

Hunter-Ayad, J., Ohlemüller, R., Recio, M. R., & Seddon, P. J. (2020). Reintroduction modelling: A guide to choosing and combining models for species reintroductions. *Journal of Applied Ecology*, *57*, 1233–1243. https://doi.org/10.1111/1365-2664.13629

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. *Ecological Modelling*, *312*, 166–174. https://doi.org/10.1016/j.ecolmodel.2015.05.018

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1). https://doi.org/10.1038/sdata.2017.122

Kiedrzyński, M., Zielińska, K. M., Rewicz, A., & Kiedrzyńska, E. (2017). Habitat and spatial thinning improve the Maxent models performed with incomplete data. *Journal of Geophysical Research: Biogeosciences*, *122*, 1359–1370. https://doi.org/10.1002/2016JG003629

Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., & Kubota, Y. (2020). Sampling bias correction in species distribution models by quasi-linear Poisson point process. *Ecological Informatics*, *55*, 101015. https://doi.org/10.1016/j.ecoinf.2019.101015

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., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*, 1366–1379. https://doi.org/10.1111/ddi.12096

Lehtomäki, J., Kusumoto, B., Shiono, T., Tanaka, T., Kubota, Y., & Moilanen, A. (2018). Spatial conservation prioritization for the East Asian islands: A balanced representation of multitaxon biogeography in a protected area network. *Diversity and Distributions*. https://doi.org/10.1111/ddi.12869

MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews*, *40*(4), 510–533. https://doi.org/10.1111/j.1469-185x.1965.tb00815.x

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source). https://CRAN.R-project.org/package=cluster

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*, 285–292. https://doi.org/10.1016/0377-2217(84)90282-0

Merow, C., Allen, J. M., Aiello-Lammens, M., Silander, J. A., & Fortin, M.-J. (2016). Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. *Global Ecology and Biogeography*, *25*, 1022–1036. https://doi.org/10.1111/geb.12453

Mohler, C. L. (1983). Effect of sampling pattern on estimation of species distributions along gradients. *Vegetatio*, *54*, 97–102. https://doi.org/10.1007/BF00035144

Pearson, R. G. (2010). Species' distribution modeling for conservation educators and practitioners. *Lessons in Conservation*, *3*, 54–89.

Permanent Secretariat of the Alpine Convention. (2009). Alpine Convention - The Alps eight countries, a single territory. Ständiges Sekretariat der Alpenkonvention.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

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

Pohlert, T. (2014). The Pairwise multiple comparison of mean ranks package (PMCMR). R Package. https://CRAN.R-project.org/package=PMCMR

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*, 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*, 366–379. https://doi.org/10.1111/2041-210X.12352

Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, *69*, 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Reynolds, A. P., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*(4), 475–504. https://doi.org/10.1007/s10852-005-9022-1

Righetti, D., Vogt, M., Gruber, N., Psomas, A., & Zimmermann, N. E. (2019). Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Science Advances*, *5*(5), eaau6253. https://doi.org/10.1126/sciadv.aau6253

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Samy, G., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., & Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodiversity Informatics*, *8*(2). https://doi.org/10.17161/bi.v8i2.4124

Schubert, E., & Rousseeuw, P. J. (2019). Faster k-Medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. In G. Amato, C. Gennaro, V. Oria, & M. Radovanović (Eds.), *Similarity search and applications. SISAP 2019. Lecture Notes in Computer Science* (Vol. *11807*, pp. 171–187). Springer.

Smith, J. N., Kelly, N., & Renner, I. W. (2021). Validation of presence-only models for conservation planning and the application to whales in a multiple-use marine park. *Ecological Applications*, *31*, 1–14. https://doi.org/10.1002/eap.2214

Steen, V. A., Tingley, M. W., Paton, P. W. C., & Elphick, C. S. (2021). Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution*, *12*(2), 216–226. https://doi.org/10.1111/2041-210x.13525

Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, *21*, 595–608. https://doi.org/10.1111/ddi.12279

Thuiller, W., Guéguen, M., Bison, M., Duparc, A., Garel, M., Loison, A., Renaud, J., & Poggiato, G. (2018). Combining point-process and landscape vegetation models to predict large herbivore distributions in space and time-A case study of Rupicapra rupicapra. *Diversity and Distributions*, *24*(3), 352–362. https://doi.org/10.1111/ddi.12684

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD - A platform for ensemble forecasting of species distributions. *Ecography*, *32*, 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

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*, no–no. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Von Humboldt, A., & Bonpland, A. (2010). *Essay on the geography of plants (1807)*. (S. T. Jackson, Ed.). University of Chicago Press.

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS ONE*, *8*, e79168. https://doi.org/10.1371/journal.pone.0079168

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, *4*(3). https://doi.org/10.1214/10-aoas331

Williams, B. K., & Brown, E. D. (2019). Sampling and analysis frameworks for inference in ecology. *Methods in Ecology and Evolution*, *10*, 1832–1842. https://doi.org/10.1111/2041-210X.13279

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2020). Macroecology in the age of Big Data—Where to go from here? *Journal of Biogeography*, *47*, 1–12. https://doi.org/10.1111/jbi.13633

BIOSKETCH

Yohann Chauvier is a macroecologist investigating ecological questions related to species, functional and phylogenetic diversity distribution, with a focus on conservation and understanding the influence of environmental drivers on these distributions. His main target organisms include vascular plants and underground taxa. Together with **Niklaus E. Zimmermann, Wilfried Thuille** and **Philipp Brun** he is involved in the 'OriginAlps' project that aims at improving the understanding of historical and contemporary processes that drive patterns of plant biodiversity in the European Alps.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Chauvier, Y., Zimmermann N. E., Poggiato G., Bystrova D., Brun P., & Thuiller W. (2021). Novel methods to correct for observer and sampling bias in presence-only species distribution models. *Global Ecology and Biogeography*, 30, 2312–2325. https://doi.org/10.1111/geb.13383