How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach

Rui F. Fernandes[a], Daniel Scherrer[a] and Antoine Guisan[a,b]

[a] Department of Ecology and Evolution, University of Lausanne, Biophore, 1015 Lausanne, Switzerland

[b] Institute of Earth Surface Dynamics, University of Lausanne, Geopolis, 1015 Lausanne, Switzerland

ABSTRACT

Correlative species distribution models (SDMs) are widely used to predict species distributions and assemblages, with many fundamental and applied uses. Different factors were shown to affect SDM prediction accuracy. However, real data cannot give unambiguous answers on these issues, and for this reason, artificial data have been increasingly used in recent years. Here, we move one step further by assessing how different factors can affect the prediction accuracy of virtual assemblages obtained by stacking individual SDM predictions (stacked SDMs, S-SDM). We modelled 100 virtual species in a real study area, testing five different factors: sample size (200-800-3200), sampling method (nested, non-nested), sampling prevalence (25%, 50%, 75% and species true prevalence), modelling technique (GAM, GLM, BRT and RF) and thresholding method (ROC, MaxTSS, and MaxKappa). We showed that the accuracy of S-SDM predictions is mostly affected by modelling technique followed by sample size. Models fitted by GAM/GLM had a higher accuracy and lower variance than BRT/RF. Model accuracy increased with sample size and a sampling strategy reflecting the true prevalence of the species was most successful. However, even with sample sizes as high as >3000 sites, residual uncertainty remained in the predictions, potentially reflecting a bias introduced by creating and/or resampling the virtual species. Therefore, when evaluating the accuracy of predictions from S-SDMs fitted with real field data, one can hardly expect reaching perfect accuracy, and reasonably high values of similarity or predictive success can already be seen as valuable predictions. We recommend the use of a ‘plot-like’ sampling method (best approximation of the species’ true prevalence) and not simply increasing the number of presences-absences of species. As presented here, virtual simulations might be used more systematically in future studies to inform about the best accuracy level that one could expect given the characteristics of the data and the methods used to fit and stack SDMs.

KEYWORDS: Virtual community ecologist; stacked species distribution models; nested design; factors importance; relative effects; sampling effect

INTRODUCTION

Important species co-existence questions have been raised in the field of community ecology over the past years (e.g. Gotzenberger *et al.*, 2012; Munkemuller *et al.*, 2014; Mittelbach & Schemske, 2015), with particular focus given to understanding what drives the distribution of assemblages (i.e. communities, sometimes used interchangeably here) and why and how their composition and richness can change in space and time. Additionally, with the increasing impacts caused by global changes (e.g. habitat fragmentation, biological invasions, climate and land-use change), it becomes critical to develop methods and tools that allow predicting the spatial distribution of species assemblages (D'Amen *et al.*, 2015b).

Species distribution models (SDMs; also called habitat suitability or ecological niche models; see Guisan et al., 2017), which statistically relate species observations, usually obtained through field observations or databases with environmental data (Guisan & Zimmermann, 2000), are useful tools in this regard as they can be stacked to predict the distribution and composition of species assemblages (e.g. Ferrier & Guisan, 2006; Dubuis *et al.*, 2011; D'Amen *et al.*, 2015b). When dealing with species richness (SR), the simplest and most common method consists in modelling the distribution of all individual species in a pool and then summing their predictions to obtain assemblages (stacked-SDM, S-SDM; Ferrier & Guisan, 2006; Dubuis *et al.*, 2011). However, this method has some limitations, such as over-predicting species richness per site (Guisan & Rahbek, 2011) or being sensitive to methodological biases (Calabrese *et al.*, 2014; Scherrer *et al.*, 2018a). Additionally, while single species models are useful, numerous factors (e.g. sample size, sampling prevalence, sampling design, modelling techniques, imperfect detection of species or the choice of environmental variables) can lead to an increase in the uncertainty of their predictions (e.g. Kadmon *et al.*, 2003; Barry & Elith, 2006; Guisan *et al.*, 2007b; Beale & Lennon, 2012), potentially propagating into the predictions of species assemblages. Until now the majority of studies used real species data to assess the effects of different factors on SDM performance at the individual species level. A recent study (Thibaud *et al.*, 2014) proposed the use of virtual or simulated data (Hirzel *et al.*, 2001; Zurell *et al.*, 2010) to assess how a set of factors affect the predictive performance of single species models. With the use of virtual data instead of real species, the “true” distribution of the species is completely known (Hirzel *et al.*, 2001) as well as the predictors that influence that distribution. Contrary, when using real species data, biological assembly rules or dispersal limitations might prevent species from coexisting even when adequate conditions exist. Other sources of uncertainty might also occur (e.g. missing environmental variables or stochasticity), making real data more difficult to use to test the relative importance of various factors. Using virtual species, whose distributions are solely determined by a set of environmental factors, ensures that the suitability of all species in each site is strictly determined by those factors with no additional biotic (e.g. competition) or dispersal restrictions. By simulating virtual sampling of these distributions with various effects (see above) and then refitting the models, one can compare the initial "true distribution" with the predicted distributions with and without 'effects' and in this way determine which factors affect models the most (Zurell *et al.*, 2010). This is still rarely done at the assemblage level, likely because there is yet no unanimous method on how to correctly predict species assemblages, or because the data and computational requirements to predict assemblages remained relatively intensive.

Here, we aimed at filling this gap by assessing how different factors affect the prediction accuracy of virtual species assemblages (obtained through S-SDMs). Specifically, we wanted to analyse the effects of five different factors - sample size, sampling method, sampling prevalence (i.e. the proportion of samples in which one found the species; not to be confused with species prevalence, the number of places occupied by a species out of the total number of places available), thresholding method and modelling technique -both separately and nested within each other- to determine: (i) what overall accuracy can be expected when sampling the known distributions in various ways (i.e. nested/non-nested, different prevalences), and (ii) which factors most affect S-SDMs performance. As a direct corollary, this should allow us to estimate the best achievable accuracy in a given modelling context where multiple factors affect the models, an aspect rarely if ever assessed in S-SDM studies (assemblage models).

METHODS

Study area

To apply our approach, we used a real landscape located in the Western Alps of Switzerland (http://rechalp.unil.ch), covering an area of approximately 700 km^2. Our analyses were conducted on 762133 sites corresponding to open, non-woody vegetation (i.e. grassland, meadow, rock, and scree). This is an intensively sampled region where many high locational accuracy biological data and high-resolution environmental data is available, providing a realistic set of species observations and predictors at very high resolution (25 meters).

Analytical framework

We followed six main steps of a virtual species simulation framework (see Figure 1), to assess how multiple factors affect the prediction accuracy of species assemblages, obtained with binary stacked species distribution models (bS-SDM). To ensure ecological realism, we defined the distributions of our virtual species based on predictions of models fitted on real data in the same study area (see section 2.2.1 below; as done in Thibaud *et al.*, 2014).

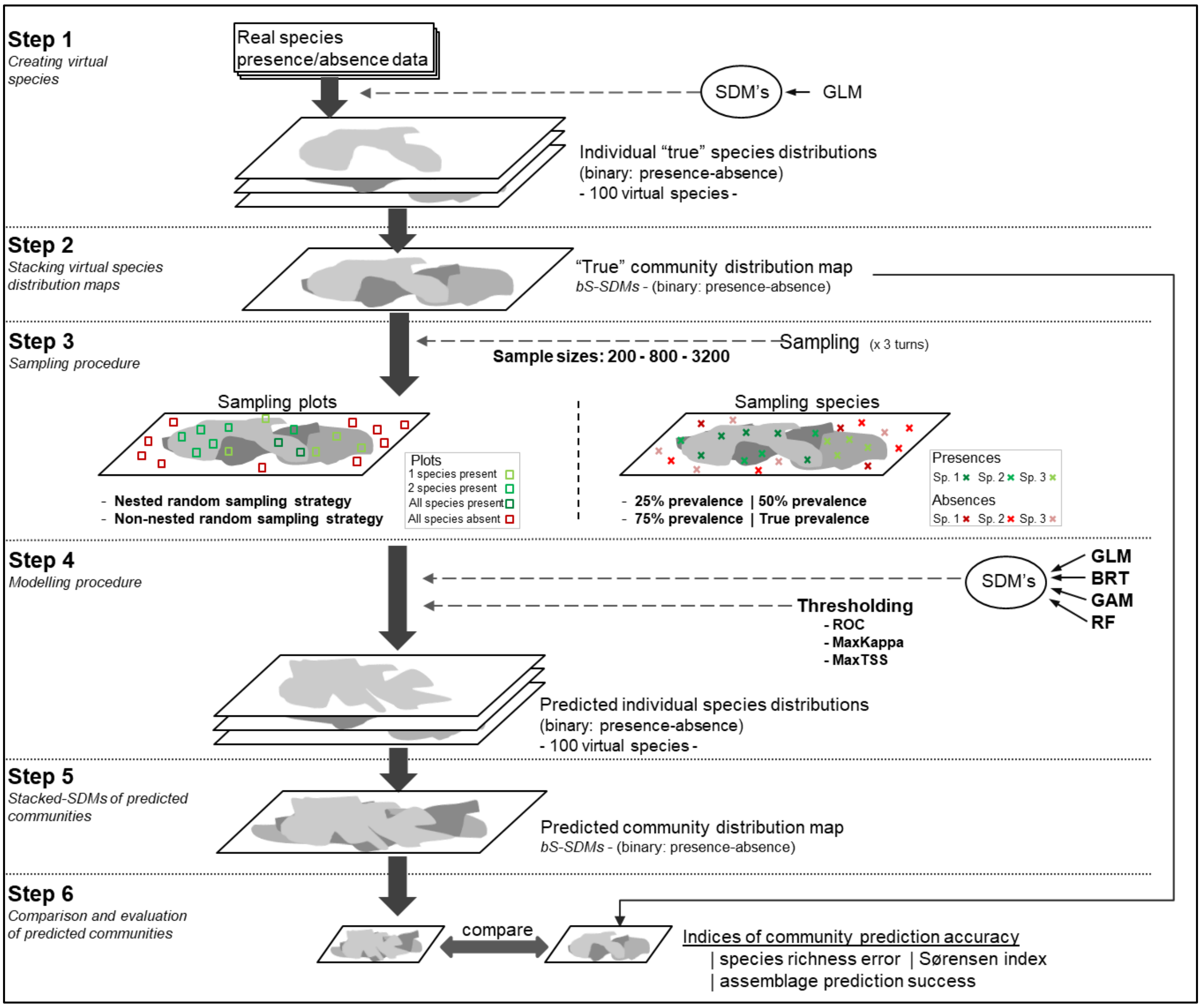
Step 1
Creating virtual species
Real species presence/absence data
SDM's
GLM
Individual "true" species distributions
(binary: presence-absence)
- 100 virtual species -
Step 2
Stacking virtual species distribution maps
"True" community distribution map
bS-SDMs - (binary: presence-absence)
Step 3
Sampling procedure
Sampling (x 3 turns)
Sample sizes: 200 - 800 - 3200
Sampling plots
Plots
1 species present □
2 species present □
All species present □
All species absent □
- Nested random sampling strategy
- Non-nested random sampling strategy
Sampling species
Presences
Sp. 1 × Sp. 2 × Sp. 3 ×
Absences
Sp. 1 × Sp. 2 × Sp. 3 ×
- 25% prevalence | 50% prevalence
- 75% prevalence | True prevalence
Step 4
Modelling procedure
SDM's
GLM
BRT
GAM
RF
Thresholding
- ROC
- MaxKappa
- MaxTSS
Predicted individual species distributions
(binary: presence-absence)
- 100 virtual species -
Step 5
Stacked-SDMs of predicted communities
Predicted community distribution map
bS-SDMs - (binary: presence-absence)
Step 6
Comparison and evaluation of predicted communities
compare
Indices of community prediction accuracy
| species richness error | Sørensen index
| assemblage prediction success

FIGURE 1 Framework of the analytical steps followed in the study (in bold the tested effects: sample size, sampling prevalence, modelling technique and thresholding method). *Step 1 – creating virtual species*: from a set of real species presence-absence data, 100 virtual species were created by fitting GLMs. *Step 2 – stacking virtual species distribution maps*: the created distribution maps were stacked together, and the resulting map was considered as the "true" assemblage distribution map. *Step 3 – sampling procedure*: the occurrence of all the virtual species present in the true assemblages were sampled using different sample sizes (200, 800 and 3200) and two different occurrence sampling methods: i) sampling plots (random nested or non-nested sampling); ii) sampling species (four different prevalence classes: 25%; 50%; 75% and a value based on the true prevalence of each species). *Step 4 – modelling procedure*: all single species models (using the occurrence data sampled in the previous step) were fitted using generalized linear models (GLM), generalized additive models (GAM), boosted regression trees (BRT) or random forests (RF). The projected species distributions were converted into binary presence/absence data using three threshold methods: ROC, MaxKappa and MaxTSS. *Step 5 – stacked-SDMs of predicted assemblages*: the binary projections created in the previous step were stacked together to create a predicted assemblage distribution map for each sample size, sampling strategy, modelling technique and thresholding method. *Step 6 – comparison and evaluation of predicted assemblages*: all the outputs resulting from the different predicted assemblages were compared with the observed true assemblage and several indices of accuracy of assemblage predictions were calculated.

Creation of virtual species

We generated presence-absence maps for 100 virtual species from a set of real species presence-absence data previously sampled in the study area. The species were randomly chosen from a set containing more than 1088 species (Dubuis *et al.*, 2011; Pottier *et al.*, 2013), but only selecting from those with more than 30 occurrences (around 627; see Appendix A for more information on the original dataset). We did this by fitting generalized linear models (GLMs; McCullagh & Nelder, 1989) to the species data as a function of five environmental predictors: summer mean monthly temperature (2 to 19° C), sum of winter precipitation (65 to 282 mm), annual sum of potential solar radiation (KJ), slope (°), and topographic position (unit-less, indicating ridges and valleys). We then used these models – in binary form – to predict the 100 species distributions across the study area. The binary presence-absence information was obtained by transforming the predicted probabilities using the threshold corresponding to the point on the receiver operating characteristic (ROC; Hanley & McNeil, 1982; Swets, 1988) plot - sensitivity against 1-specificity across successive thresholds - with the shortest distance to the top-left corner of the plot (Cantor *et al.*, 1999). We considered these binary predictions as the "true distributions". All the initial environmental and species data used in this study were available at 25 m resolution. Diverging results have been observed when different resolutions or extents were used (e.g. Thuiller *et al.*, 2004; Guisan *et al.*, 2007a), demonstrating that scale parameters can have an important influence on model predictions. This can be a result of the scale dependency of the environmental predictors (Vicente *et al.*, 2014) or the spatial stochastic effects at smaller spatial scales (Steinmann *et al.*, 2011; Scherrer *et al.*, 2018b).

Furthermore, dispersal and biotic factors can also play an important role interacting with scale (Soberon & Nakamura, 2009), with the distribution of real species not being fully explained by abiotic predictors alone. We avoid all these issues in our work by using virtual species and the same predictors to create the species and fit their distribution models, ensuring that the initial species distributions are fully explained by the chosen predictors at the study scale (extent and resolution, and using the same technique). With this approach, we tended to guarantee that the virtual species presented realistic response curves for our landscape, resulting in realistic species assemblages. The resolution should thus not matter in our study and should not affect our findings. However, we acknowledge that in real ecosystems the explanatory power of abiotic environmental factors (as used in this study) on single species distributions and assemblages (i.e. assembly rules) might strongly depend on the spatial resolution of the study, and other factors such as dispersal limitations and biotic interactions might modify the abiotic responses and interact with scale. All models were run in the R software version 3.3.3 (R Core Team, 2017), using biomod2 (Thuiller *et al.*, 2009) default settings for sake of simplicity and comparability.

Stacking virtual species distribution maps

The binary predictions for each species were then stacked to create species assemblages (i.e. binary stacked SDMs, bS-SDMs), providing both species richness and composition for each pixel in the study area. In this simplified theoretic approach, the assemblages resulting from this stacking of binary SDMs are then considered our "true" assemblages (i.e. S-SDM), meaning that all species stacked into these "true" assemblages can coexist based on the abiotic factors only, without biotic or dispersal restrictions further excluding some of them from an assemblage. This is the power and appeal of simulations, since one can restrict the niche to known factors only (in our case abiotic environment), and this way facilitate the assessment of factors affecting S-SDMs through the sampling of virtual observation sites.

Sampling procedure

Presences and absences were sampled for all species using increasing sample sizes (n=200, 800 and 3200) and according to two schemes representing the dominating types of data available to fit SDMs: 1) simulating the sampling of "vegetation plots", to reproduce real datasets obtained in field surveys where all species were sampled in the same plots (i.e. in a 'plot-like' fashion), using a nested (i.e. plots sampled in the smaller sample sizes are included in the larger sample sizes) or a non-nested random sampling strategy; here, the species prevalence cannot be controlled: and 2) simulating occurrence data as typically available in biodiversity databases where species are sampled individually from each other, but here with absences also available. In this case, we used four different sampling prevalence values: 25%; 50%; 75% and the true prevalence of each species. The complete sampling procedure was repeated three times for each of the 100 virtual/simulated species. This case can also be considered as a simulation of a situation where presence-only data is available, with pseudo-absences weighted to 25, 50 or 75%.

Modelling procedure

To test the effects of different modelling techniques, single species models were fitted with four techniques (see Guisan *et al.*, 2017 for an overview): generalized linear models (GLM; McCullagh & Nelder, 1989) , generalized additive models (GAMs; Hastie & Tibshirani, 1990), boosted regression trees (BRTs; Friedman *et al.*, 2000) and random forests (RFs; Breiman, 2001), all commonly used in SDMs (e.g. Guisan *et al.*, 2002; Prasad *et al.*, 2006; Elith *et al.*, 2008). We used the same five environmental variables to calibrate the models as previously used to create the virtual species, all at a 25-meter resolution: summer mean monthly temperature, the sum of winter precipitation, the annual sum of potential solar radiation, slope, and topographic position. Each individual model was calibrated using 80% of the available data and evaluated on the remaining 20%. This cross-validation procedure was repeated 20 times and averaged using an ensemble approach (i.e. mean probabilities across predictions). The models were evaluated on the evaluation dataset using ROC (receiver operating characteristic; Hanley & McNeil, 1982; Swets, 1988), MaxKappa (Guisan *et al.*, 1998; Huntley *et al.*, 2004) and MaxTSS (i.e. equivalent to the sensitivity-specificity sum maximization described in Liu *et al.*, 2005) (see Guisan *et al.*, 2017 for details on maximization approaches). Finally, the projected species distributions were converted into binary presence/absence using the same approach as described in section 2.2.1 (ROC plot). In parallel, we also used two other thresholding methods to transform probability distributions into presence-absence data: selecting the thresholds maximizing Kappa (MaxKappa) or maximizing TSS (MaxTSS) as presented above. The whole approach was implemented in version 3.3.3 of the open-source software R (R Core Team, 2017).

Stacked-SDMs of predicted communities

Species binary predictions were stacked together to predict assemblages for each sample size, sampling method, threshold method, and modelling technique. With these predicted assemblage maps, we simultaneously obtained information on species richness and composition for each modelled site across the whole study area. We also made three repetitions of the sampling procedure (i.e. turns; T1, T2 and T3) of different sample sizes (200, 800 and 3200), different sampling prevalences (nested or non-nested sampling of random plots and four different sampling prevalence types), three thresholding methods (ROC, MaxKappa and MaxTSS) and four different modelling techniques (GLM, GAM, BRT and RF). Consequently, we ended up with a final set of more than 64 000 models.

Comparison and evaluation of predicted communities

Composition outputs and species richness resulting from the differently predicted assemblages (for each site) obtained through the previous steps were compared and evaluated to our observed assemblages (i.e. “true” assemblage map). We calculated three main indices of assemblage prediction accuracy by using the *ecospat.SSDMeval* function available in the “*ecospat*” R package (see Table A.1 for details on all the indices; Di Cola *et al.*, 2017): (i) species richness error (i.e. difference between predicted and observed species richness); (ii) the assemblage prediction success (i.e. proportion of species correctly predicted as present or absent); and (iii) a widely used metric of assemblages similarity, the Sørensen index (Sørensen, 1948). We calculated six additional indices to complement our analyses: (iv) community TSS (here measured for a site across all species, rather than for a species across all sites as in single SDM evaluation; Pottier *et al.*, 2013) and (v) community Kappa (same as for previous metric, for a site across species; Pottier *et al.*, 2013), (vi) over-prediction; (vii) under-prediction; (viii) sensitivity (i.e. the proportion of species correctly predicted as present); and (ix) specificity (i.e. the proportion of species correctly predicted as absent). Finally, some of those indices were used to assess the importance of the different studied factors following a procedure similar to the one proposed by Thibaud *et al.* (2014). However, contrary to the latter study, here we were not only interested to measure model accuracy for individual species models, but mainly to assess the predictive accuracy of species assemblages. To assess the importance of the studied factors, we analysed the variation of the previously mentioned indices via a linear mixed-effects model, adapting codes from Thibaud *et al.* (2014). To examine the relative importance of factors in the linear models, we calculated the marginal and conditional coefficients of determination (R2; Nakagawa & Schielzeth, 2013). The R package *nlme* (Pinheiro *et al.*, 2017) was used to fit these linear mixed-effects models, with each factor and all its interactions being excluded and compared using marginal R^2 (i.e. calculating the proportion of variance that is explained by fixed effects compared to that of the full model). By excluding one factor at a time, we can measure its contribution to the full model, and therefore its contribution to improve the predictive accuracy of SDMs (i.e. the lower the value of R^2 when compared with the full model, the greater the effect of the excluded factor).

RESULTS

The SDMs used to create the virtual species had scores ranging between 0.626 and 0.967 when evaluated by ROC (mean = 0.86 ± SD = 0.13) and between 0.1/0.73 and 0.2/0.9 when evaluated respectively by MaxKappa (0.35 ± 0.14) and MaxTSS (0.6 ± 0.14). The prevalence values for the virtual species' presence-absence distributions ranged between 0.02 and 0.74 (0.35 ± 0.17). The virtual species SDMs (i.e. fitted using virtual species sampled data) had very high evaluation scores (ROC: 0.999 ± 0.002; MaxKappa: 0.99 ± 0.02; MaxTSS: 0.99 ± 0.02).

Results from the SDMs based on a random sampling of 100 000 plots from the virtual species distribution maps (Fig. 2, MaxKappa; see Appendix A for the other thresholding methods results) revealed that modelling technique and sample size were the factors with the largest effect on prediction accuracy of our assemblages when all the factors are taking into account in a nested manner (Fig. 2 and 3). This can also be observed in the values of marginal and conditional R^2, calculated through a linear mixed-effects model to quantify our visual impressions from the previous mentioned figures (Table 1; i.e. the partial models when modelling technique or sample size are excluded have the lowest values of R^2 when compared with the full model, indicating the important effects of these factors). Independently of the calculated indices and taking into account the reduction in marginal R^2, modelling technique is more important than sample size (e.g. when Sørensen was used, marginal R^2 gets reduced from 0.913 in the full model to 0.366 when modelling technique is excluded and to 0.526 when sample size is excluded). However, the effects of the different factors can also be important when analysed separately. Models fitted with GAM and GLM provided the most accurate predictions on average (i.e. highest similarity of observed/predicted assemblages – Sørensen above 0.95 – and prediction success, also always above 0.95 on average; the values for the two techniques are similar, followed by BRT and RF (on average below 0.95 both for Sørensen and prediction success; Fig.2b, c). It's also noticeable that models fitted by BRT or RF presented higher variance for the different calculated metrics. Within each modelling technique, higher sample sizes decreased the difference between predicted and observed SR (i.e. species richness error; Fig. 2a). Higher sample sizes also increased prediction success (Fig. 2b) and assemblage similarity (i.e. Sørensen index; Fig. 2c). Additionally, both over- and under-prediction decreased with increasing sample size (Fig. 3a, b), while sensitivity (Fig. 3c) and specificity (Fig. 3d) increased with increasing sample size.

The level of sampling prevalence appeared to individually (i.e. taking into account one factor at a time) influence the accuracy of predicted assemblage models (Fig. 2 and 3), not being important when all factors are taking into account (see Table 1). Models calibrated with higher levels of sampling prevalence (i.e. 75%) presented higher species richness error (Fig. 2a), with a higher number of species predicted than those observed (Fig. 3a, b; over-prediction larger than under-prediction; plus 3 species on average) and a higher sensitivity than specificity (Fig. 3c, d). On the other hand, levels of sampling prevalence of 25% presented the inverse pattern, with a lower species richness error (Fig.2a) and a lower number of species predicted than observed (Fig. 3a and b; under-prediction larger than over-prediction; minus 1 species on average). When considering over- and under-prediction as well as sensitivity and specificity, similar patterns were observed with the true prevalence sampling method or when using the "plot-like" sampling methods (i.e. nested and non-nested random sampling). These three methods presented very similar results and more accurate predictions (i.e. higher values of predictions success and the same number of species predicted as observed, on average). Models calibrated with high sampling prevalence (75%) also appeared to have lower values of assemblage predictive success (Fig. 2b; 0.96 on average) and assemblage similarity (Sørensen, Fig. 2c, 0.94 on average), yet with relatively small differences when compared with other sampling prevalences (Sørensen around 0.95, depending on sample size and modelling technique). Furthermore, based on the calculated R^2 (Table 1), the nested importance, considering the other factors at once, of sampling prevalence was negligible (i.e. for each calculated index, the R^2 actually increased when compared with the full model). However, it could still be important when analysing individual cases (e.g. when species have values of sampling prevalence above 75%). The method employed to transform our probability distributions into presence-absence data (i.e. thresholding method) was also not an important factor influencing assemblages' prediction success, with the same patterns being observed in the three tested methods (see Appendix A) and with calculated R^2 not suffering any reduction when this factor was removed (i.e. indicating that it was not important in the overall model; see Table 1).

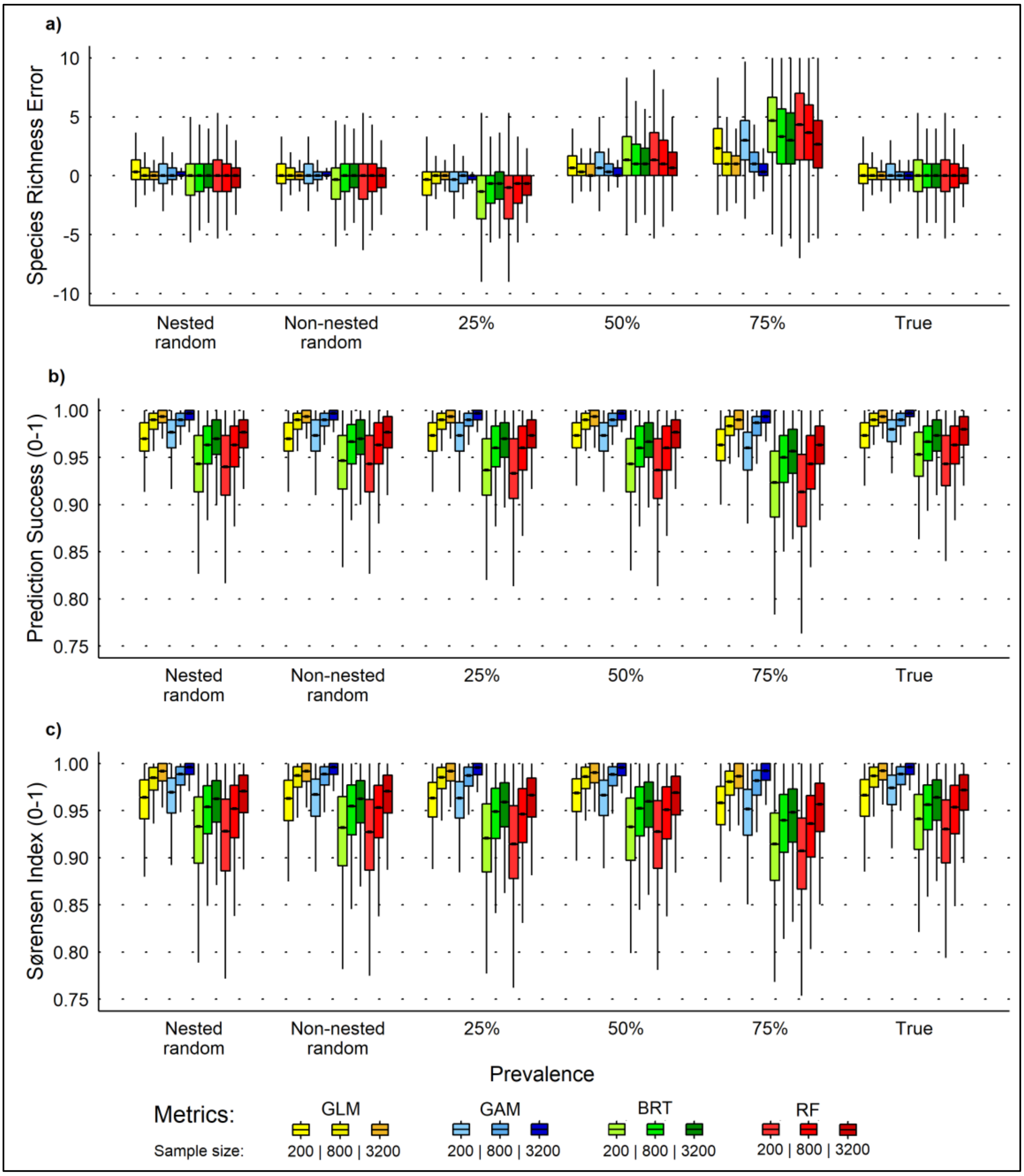

FIGURE 2 Boxplots of different indices of assemblage prediction (S-SDM) accuracy (i.e. species richness error, prediction success and Sørensen) for all the simulated species and for all the sampling strategies (based on plots (nested or not) or prevalence (25%, 50%, 75% or true) sampling; in abscissa). Each box shows the variation across all virtual species in a random subset of the study area (100 000 plots) for the binary predictions obtained using MaxKappa as thresholding technique, averaged from the three sampling turns. For each prevalence sampling, four sets of three boxplots are displayed, corresponding to models fitted using either GLMs (yellow), GAMs (blue), BRTs (green) or RFs (red), with increasing values of sample size (200, 800 and 3200).

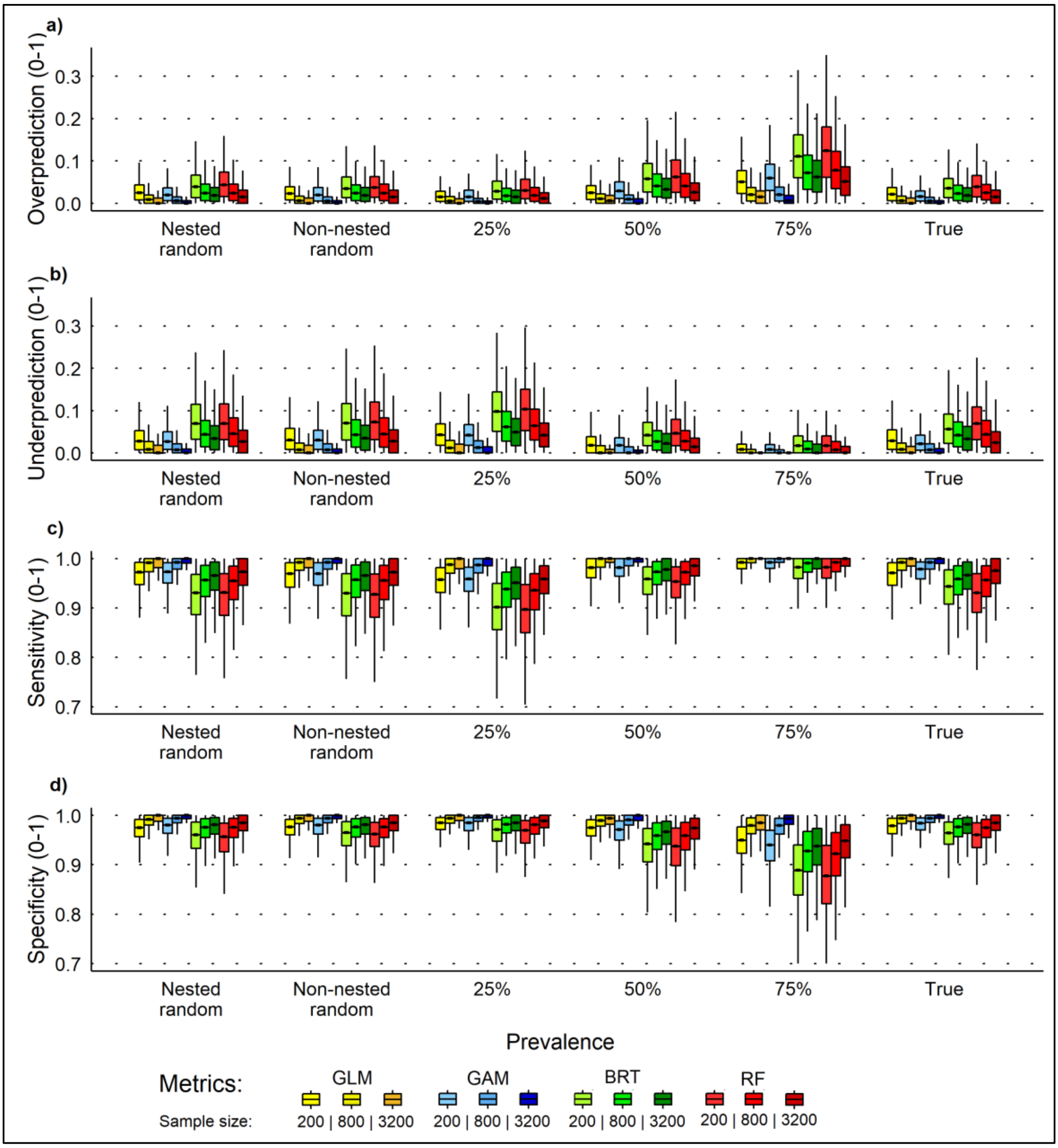

FIGURE 3 Boxplots of different indices of assemblage prediction (S-SDM) accuracy (i.e. over and underprediction, sensitivity and specificity) for all the simulated species and for all the sampling strategies (based on plots (nested or not) or prevalence (25%, 50%, 75% or true) sampling; in abscissa). Each box shows the variation across all virtual species in a random subset of the study area (100 000 plots) for the binary predictions obtained using MaxKappa as thresholding technique, averaged from the three sampling turns. For each prevalence sampling, four sets of three boxplots are displayed, corresponding to models fitted using either GLMs (yellow), GAMs (blue), BRTs (green) or RFs (red), with increasing values of sample size (200, 800 and 3200).

TABLE 1. Marginal and conditional R^2 for the five calculated indices used to evaluate our assemblage prediction (S-SDM) accuracy. Marginal and conditional R^2 are reported for each index, containing information for the full model (all factors together) and the submodels for each factor (and all its interactions with the other factors) excluded at a time. The lower the R^2 value when compared with the full model, the greater the effect of the excluded factor.

Calculated indices	Model	Marginal R^2	Conditional R^2
Sørensen	Full model	0.913	0.985
	Prevalence	0.918	0.985
	Sampling procedure	0.915	0.984
	Modelling technique	**0.366**	**0.388**
	Sample size	**0.526**	0.974
	Thresholding method	0.914	0.986
Predictive Success	Full model	0.895	0.982
	Prevalence	0.901	0.982
	Sampling procedure	0.898	0.981
	Modelling technique	**0.332**	**0.368**
	Sample size	**0.540**	0.967
	Thresholding method	0.895	0.983
Community TSS	Full model	0.942	0.988
	Prevalence	0.943	0.987
	Sampling procedure	0.943	0.987
	Modelling technique	**0.347**	**0.347**
	Sample size	**0.567**	0.973
	Thresholding method	0.943	0.989
Community Kappa	Full model	0.880	0.977
	Prevalence	0.887	0.978
	Sampling procedure	0.884	0.977
	Modelling technique	**0.324**	**0.371**
	Sample size	**0.532**	0.959
	Thresholding method	0.881	0.979

DISCUSSION

In this paper, we used a virtual ecologist approach to test the effects and importance of five factors – modelling techniques, sample size, sampling method, sampling prevalence and prediction thresholding - on the predictive accuracy of stacked binary predictions (bS-SDM) of species distribution models (SDMs). Our framework takes inspiration on the methodology first proposed by Thibaud *et al.* (2014), following a nested approach to test the relative effects of various factors on SDMs, but differing markedly from the latter by focusing here mainly on species assemblage models (S-SDMs). Furthermore, we used a much greater number of virtual species (n=100) and additionally assessed the importance of effects like sampling nestedness and prevalence on the accuracy of assemblage predictions. However, unlike Thibaud *et al.* (2014), we did not assess the effects of sampling bias and spatial autocorrelation. We took this decision because the previously mentioned paper reported that sample size and modelling technique were the factors that contributed most to the variation in prediction accuracy, while sampling bias and spatial autocorrelation had smaller and negligible effects respectively. To our knowledge, the present study is the first to use a virtual ecology framework to assess and report cumulative effects of different factors on species assemblage predictions (S-SDMs).

We found that modelling technique and sample size were the most important factors, relative to the others tested (i.e. taking into account the importance of all factors at once), affecting the accuracy of SDMs and - most importantly here - of their assemblage predictions. Additionally, we found that the overall accuracy that can be expected from the S-SDMs depended on the options made when fitting models (e.g. the choice of technique, sample size, how to sample), with inaccurate assemblage predictions being obtained after simulated sampling even when the initial distribution of the species and the environmental factors determining them are completely known.

Which methodological factors most affect the performance of S-SDM?

Modelling technique

The importance of modelling technique and the fact that different algorithms can provide different predictions and predictive performances is something widely reported (e.g. Guisan *et al.*, 2007b; Graham *et al.*, 2008; Elith & Graham, 2009; Marmion *et al.*, 2009a), but the nested framework used here allowed to further discuss these differences in the light of (i.e. relative to) other factors: sample size, sampling design, thresholding criteria and sampling prevalence. Our findings partly confirmed the results obtained by Thibaud *et al.* (2014) for single SDMs, but here applied to species assemblage predictions (S-SDMs). However, we observed that modelling technique had a larger impact than sample size in affecting assemblage prediction accuracy (see Table 1), contrary to the previously mentioned study for single SDMs. Also, while the main results reported here derived from virtual species created using a GLM, the models obtained through this technique were not better than models fitted by GAM. This was contrary to the pattern observed in the aforementioned study, where GLMs clearly presented the best results when the virtual species were also created using that technique. Additionally, models fitted here both by BRT and RF presented results with higher variance and lower predictive success. This is contrary to some SDM studies (e.g. Elith *et al.*, 2006; Guisan *et al.*, 2007a; Graham *et al.*, 2008; Williams *et al.*, 2009), which found that these techniques performed better than GLM or GAM, while other studies showed no major difference in the performance of models fitted by the different techniques (e.g. Elith & Graham, 2009; Roura-Pascual *et al.*, 2009). These differences might result from the fact that in our study the virtual species' distributions were created using a regression model (GLM) and were then resampled to fit models with various techniques, whereas in other studies the same technique was always used to both create a virtual species and then fit the models and assess the effect of the different factors on it. BRT and RF would thus be good at finding a signal in the training data, but less good at predicting to independent data (i.e., in this case, the random subset of the study area -100 000 plots - used to calculate the different indices). Another reason for the discrepancy among studies might be that, in our study, the virtual species were fully explained by the predictors used and, since they were created by GLMs, showed clearly defined unimodal response curves. However, these response curves can be much more complex when real species are used, especially due to interactions between species and the environment and among themselves. More complex techniques like GAM, RF or BRT would fit these more complex curves better, but at the cost of then predicting worse to independent data. This had been shown already for single SDMs (e.g. Randin *et al.*, 2006) and discussed elsewhere (e.g. Merow *et al.*, 2014), but had never been shown so far for S-SDMs.

Sample size

The well-known impact of sample size on single SDMs (e.g. Stockwell & Peterson, 2002; Wisz *et al.*, 2008; Mitchell *et al.*, 2017) is in large part explained by the fact that a greater number of presence-absence data provides a larger amount of information about the occupied multi-dimensional environmental space, allowing to fit more reliable species response curves along all the considered environmental gradients, improving the species' niche quantification and associated predictions. However, we showed here that even when using a large amount of sampled data (>3000 sites) we can still obtain some inaccuracy in assemblage predictions, depending on the modelling and sampling technique used. However, if these other factors – modelling and sampling - are taken into account, one can achieve relatively high or very acceptable values of prediction success and assemblage similarity (Sørensen index) even with the smallest sample size assessed here (200 sites).

Sampling prevalence

Species prevalence is often a key factor affecting model performance (e.g. Manel *et al.*, 2001; Jiménez-Valverde *et al.*, 2009; Santika, 2011; Lawson *et al.*, 2014). Here, we showed that sampled prevalence also has an effect in some components of the assemblage evaluation, with higher sampling prevalence (here 75%) causing species richness over-prediction and favouring sensitivity, whereas lower sampling prevalence (here 25%) causes species richness under-prediction and favors specificity. Yet, it did not affect greatly our assemblage predictions (prediction success and Sørensen index) in our nested analysis taking into account simultaneously for the other methodological factors (see Table 1). However, if we consider the patterns observed in the different sample prevalence groups, assemblage prediction accuracy increased slightly (i.e. close to 1 at large sample sizes, around 0.95 at smaller sizes) when sampled prevalence reflected the true prevalence of the species (i.e. species prevalence) in the study area. The same occurred when low sampling prevalences were used or when the sampling was done in a 'plot-like' fashion – nested or non-nested random sampling – (Fig. 2 and 3). We observed high prediction success when using the species true prevalence mainly because, when using those values, the information given to the SDMs (presence-absence) allows unbiased estimates of species richness (Calabrese *et al.*, 2014). It can thus be expected that, if one obtains individual SDMs reflecting true prevalence, one should also get more accurate S-SDM predictions of species richness. By sampling prevalences of 25, 50 or 75%, one wrongly defines the initial level of SR in the model. As the virtual species' true prevalence was between 25 and 50% (35% on average), this then explains why we under-predict models when sampling at the 25% prevalence level and over-predict at the 50% and 75% prevalence levels.

Previous studies like the one of Jiménez-Valverde *et al.* (2009) on single SDMs further showed that species prevalence strongly interacts with sample size, with the distribution of species being over-predicted when sample sizes are small and species prevalences high. These patterns were also observed in our data but at the assemblage level (Fig. 2 and 3). The authors of the previously mentioned study additionally showed that when the sampled presence-absences cover the entire environmental gradient, high or low species prevalences have less or no effect on model accuracy. In our study, this representative sample of the entire environmental gradient was best reflected by the true prevalence sampling method (if perfect conditions were possible), which simulated the most correct distribution of the species (considering also that the sites were randomly sampled), reducing the probability of sampling all the presences (or absences) in a reduced part of the environmental gradients. Accordingly, one should obtain minimal or no error in predicted species richness, as we observed. However, in real-world conditions where sampling the true species prevalence is impossible, the most appropriate method appears to be to randomly (or random-stratified) sample in a 'plot-like' fashion (i.e. similar results to true species prevalence; Fig.2 and 3), an approach used in many studies with real data (e.g. Dubuis *et al.*, 2011; Pottier *et al.*, 2013; D'Amen *et al.*, 2015a).

Sampling strategies

We also observed that the different sampling strategies (i.e. between sampling information in a 'plot-like' fashion - inventorying all species in each plot as sampling unit - and sampling species individually and independently of each other) is not one of the most important factors affecting assemblage predictions (i.e. considering all factors at once - i.e. relative effects; Table 1). Nevertheless, when analysing specific cases (like the individual sampling prevalence groups), we can say that sample prevalence is important to take into account, thus sampling plots is preferable to single occurrences as it is the best approximation to true species prevalence sampling (see above).

Thresholding methods

While the effects of different thresholding methods on species distribution predictions were widely studied for single SDMs (see e.g. Liu *et al.*, 2005; Lobo *et al.*, 2008; Lawson *et al.*, 2014; Vale *et al.*, 2014), our results for S-SDMs showed that this factor had negligible importance on the prediction success of our assemblages, with different methods presenting the same patterns (see Appendix A). This was surprising because, as showed by Nenzén and Araújo (2011), the choice of the threshold explained 25% of the variability in their results (with modelling technique explaining 35% and their interaction 19%). The fact that the thresholding method had no effect on the predictive success of our modelled species and assemblages might be associated with the values of the selected thresholds. In our case, these were always below 0.5 (and often <0.4), independent of the thresholding method (thus not showing a great variation in threshold values).

What overall accuracy can be expected when perfectly known species distributions are sampled?

We showed clearly here that, even with complete initial knowledge of the distribution of the species and assemblages and of the environmental factors determining those, fitting the models on samples of the data (of varying size) quickly brings some error in assemblage predictions, even with quite large samples (>3000 sites) and even if models with high evaluation values (e.g. ROC or MaxKappa > 0.9) were still obtained. This means that even if our individual species models present very high evaluation scores (i.e. on average close to 1 for the three metrics), we are unable to fully recover the initial assemblages, based on the stacking of all virtual species' distributions. This could be caused by the fact that even when one is able to obtain accurate individual species (i.e. SDMs), small errors (i.e. falsely predicted presences or absences) can occur in each of them. These errors can then accumulate and prevent us from getting accurate assemblage predictions (i.e. S-SDMs). If this is the case with virtual species, one can expect an equivalent or likely higher error accumulation with real species. Therefore, obtaining inaccurate community predictions using real species data might also occur due to methodological problems (e.g. Calabrese *et al.,* 2014) and not only because of missing dispersal or biotic constraints (e.g. Guisan & Rahbek, 2011). Similar tests could also be performed when using more mechanistic or process-based methods to determine if the same patterns are observed. Nevertheless, despite not being able to completely predict species assemblages, depending on the factors used to model these distributions (i.e. modelling technique, sample size or sampling method), one can still yield valuable accurate predictions (e.g. assemblage prediction success around or above 0.95). More particularly, we showed that one can obtain very good assemblage predictive success (i.e. predicted assemblages very similar to the observed ones across the whole area; Fig. 1) particularly when large sample sizes are available, when GLMs or GAMs are used (rather than more complex techniques like RF or BRT) to fit the models and when the sampling prevalence reflects either the true prevalence (possible with artificial data) or a 'plot-like' sampling of the species (realistic method that samples an approximation of the species true prevalence). Large sample sizes as the one used here (3200 plots) might be difficult to obtain for the majority of species and taxonomic groups, but we showed that even with smaller samples (200) we are still able to obtain very good assemblage predictions (e.g. prediction success and Sørensen similarity index above 0.95) when GLM or GAMs were used. So, more effort should be put into getting a representative sample of species distributions in a certain area using a 'plot-like' sampling method, and not simply increasing the number of presences or absences for some species. This will guarantee that the sampled data would potentially cover all the spatial and environmental gradients in a certain area while also reflecting the species true prevalence (or a very close approximation).

A critical assessment of our framework and how to go forward

Our framework, combined with the use of realistic virtual species can be used in future studies to test an even larger number of factors that might affect the accuracy of models, with the choice of those factors depending on the availability of computational power that is necessary in order to perform all the required steps. A more complex nested approach could include factors like the extent of the study area, different spatial scales, a larger set of available modelling techniques, the use of different environmental variables to create virtual species and fit the models (i.e. missing covariates), sampling bias (e.g. random, clustered, close to roads, stratified), spatial autocorrelation, the effect of using presence-absence data or only presences or the effects of different methods to create pseudo-absences.

We also recognize that the method used to create our virtual species might be considered simplistic and that other methods (threshold vs probabilistic approach; see Meynard & Kaplan, 2013) and different packages (e.g. SDMvspecies, NicheLim or virtualspecies - Duan *et al.*, 2015; Huang *et al.*, 2016; Leroy *et al.*, 2016) could be used to create the virtual species. This might be another opportunity to test if the same conclusions can be obtained when virtual species are created by different methods and further contribute to refine a specific methodological choice or approach. Even larger sample sizes might also be investigated to assess how large samples would need to be to reach near perfect predictions.

The majority of experimental studies are by essence oversimplifications of the real world, and alike our objective was to use a simplified and controlled artificial reality instead of simulations on real community data. Our goal was to assess the effects of a set of methodological factors on species assemblage modelling (S-SDMs when assembly rules are purely determined by abiotic constraints, thus ruling out effects from biotic interactions and dispersal limitations. While this is of course ecologically not fully realistic, it allowed us to test a scenario that was simple and with (nearly perfectly) known abiotic assembly rules. Considering that this study was performed to determine the influence of certain methodological factors on assemblage predictions, one can reasonably think that if the methods used were not good enough to correctly predict assemblages given such simplistic environmental drivers, then adding other factors (not considered here; e.g. dispersal limitations, biotic interactions, sampling bias or wrongly parameterized techniques) can only reduce the chances of obtaining accurate predictions.

We showed one illustration of how the use of virtual species can be helpful in spatial modelling of species assemblages, but it can also prove useful in numerous other situations in ecology. For SDMs, virtual species were already used in several instances (see Miller, 2014 for examples of recent applications) to validate proposed methods (e.g. Zurell *et al.*, 2016; Guisande *et al.*, 2017; Hattab *et al.*, 2017), but also to test the effects of observation errors on model performance and the efficiency of currently used evaluation metrics (Fernandes *et al.*, in press), to test different approaches to sample species data (Hirzel & Guisan, 2002), to test different downscaling methods (Bombi & D'Amen, 2012) or to assess the effectiveness of different hierarchical modelling frameworks when compared to more traditional methods (Fernandes *et al.*, unplub.). Other potential uses worth exploring might include the testing of methods or software used for spatial conservation planning (e.g. ConsNet, Zonation or Marxan - Ciarleglio *et al.*, 2009; Watts *et al.*, 2009; Moilanen *et al.*, 2014), to determine their real effectiveness in defining prioritization areas and identify strengths and weaknesses of different alternative approaches.

CONCLUSION AND MAIN MESSAGES

With this paper, we wished (i) to contribute to the ongoing discussion on the usefulness and validity of stacking species distribution models (SDMs) to predict species assemblages (S-SDMs), and (ii) to propose a way (taken from single SDMs; Thibaud *et al*., 2014) to analyse the relative effects and behaviour of different methodological factors potentially affecting S-SDM predictive success. We also discussed different features and potentialities that can help virtual species and simulations become a more useful tool in ecological or evolutionary research, e.g. to test the efficiency of alternative modelling frameworks in a fully controlled abiotic environment.

The main conclusions for factors affecting S-SDMs, based on our findings and given our study settings, are:

1) even when starting with the full knowledge of the species (i.e. all abiotic factors influencing its distribution being known) and sampling a large number of sites (>3000), "perfect" predictions of assemblage are difficult to attain, but very good predictions are reachable;
2) modelling technique and sample size were the most important factors, relative to the others tested (i.e. accounting for the importance of all factors at once);
3) contrary to previous studies on single SDMs, we showed that the choice of the modelling technique used to fit the models had a larger impact than sample size on S-SDM prediction success;
4) accuracy increases with sample size, but depending on the modelling technique (GLM or GAM) and sampling method (sampling 'plot-like' methods), accurate predictions could already be obtained with relatively small sample sizes (200 sites);
5) sampling species data using a 'plot-like' method is more desirable than sampling species individually, as it proves a better approximation of the true species prevalence and provides more accurate assemblage predictions.

ACKNOWLEDGEMENTS

This work was supported by the SNF project “SESAM'ALP - Challenges in simulating alpine species assemblages under global change”. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. We thank Damaris Zurell and Catherine Graham for their valuable comments on an early stage of this work.

REFERENCES

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423.

Beale, C.M., Lennon, J.J., 2012. Incorporating uncertainty in predictive species distribution modelling. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 367, 247–258.

Bombi, P., D'Amen, M., 2012. Scaling down distribution maps from atlas data: a test of different approaches with virtual species. J. Biogeogr. 39, 640–651.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Calabrese, J.M., Certain, G., Kraan, C., Dormann, C.F., 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. Glob. Ecol. Biogeogr. 23, 99–112.

Cantor, S.B., Sun, C.C., Tortolero-Luna, G., Richards-Kortum, R., Follen, M., 1999. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. J. Clin. Epidemiol. 52, 885–892.

Ciarleglio, M., Barnes, J.W., Sarkar, S., 2009. ConsNet: new software for the selection of conservation area networks with spatial and multi-criteria analyses. Ecography 32, 205–209.

D'Amen, M., Dubuis, A., Fernandes, R.F., Pottier, J., Pellissier, L., Guisan, A., 2015a. Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. J. Biogeogr. 42, 1255–1266.

D'Amen, M., Pradervand, J.-N., Guisan, A., 2015b. Predicting richness and composition in mountain insect communities at high resolution: a new test of the SESAM framework. Global Ecol. Biogeogr. 24, 1443–1453.

Di Cola, V., 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.

Duan, R.Y., Kong, X.Q., Huang, M.Y., Wu, G.L., Wang, Z.G., 2015. SDMvspecies: a software for creating virtual species for species distribution modelling. Ecography 38, 108–110.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.P., Guisan, A., 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. Divers. Distrib. 17, 1122–1131.

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.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-

Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.

Fernandes, R.F., Scherrer, D., Guisan, A., in review. Effects of simulated observation errors on the performance of species distribution models. Divers. Distrib.

Fernandes, R.F., Scherrer, D., Guisan, A., unplub. Predicting current and future virtual species assemblages: are hierarchical models useful?

Ferrier, S., Guisan, A., 2006. Spatial modelling of biodiversity at the community level. J. Appl. Ecol. 43, 393–404.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407.

Gotzenberger, L., de Bello, F., Brathen, K.A., Davison, J., Dubuis, A., Guisan, A., Leps, J., Lindborg, R., Moora, M., Partel, M., Pellissier, L., Pottier, J., Vittoz, P., Zobel, K., Zobel, M., 2012. Ecological assembly rules in plant communities–approaches, patterns and prospects. Biol. Rev. Camb. Philos. Soc. 87, 111–127.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B.A., Gro, N.P.S.W., 2008. The influence of spatial errors in species occurrence data used in distribution models. J. Appl. Ecol. 45, 239–247.

Guisan, A., Rahbek, C., 2011. SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. J. Biogeogr. 38, 1433–1444.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Guisan, A., Theurillat, J.P., Kienast, F., 1998. Predicting the potential distribution of plant species in an Alpine environment. J. Veg. Sci. 9, 65–74.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.

Guisan, A., Graham, C.H., Elith, J., Huettmann, F., 2007a. Sensitivity of predictive species distribution models to change in grain size. Divers. Distrib. 13, 332–340.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., Peterson, A.T., 2007b. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? Ecol. Monogr. 77, 615–630.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press.

Guisande, C., Garcia-Rosello, E., Heine, J., Gonzalez-Dacosta, J., Vilas, L.G., Perez, B.J.G., Lobo, J.M., 2017. SPEDInstabR: An algorithm based on a fluctuation index for selecting predictors in species distribution modeling. Ecol. Inform. 37, 18–23.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

Hattab, T., Garzon-Lopez, C.X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., Feilhauer, H., Honnay, O., Kempeneers, P., Schmidtlein, S., Somers, B., Van De Kerchove, R., Rocchini, D., Lenoir, J., 2017. A unified framework to model the potential and realized distributions of invasive species within the invaded range. Divers. Distrib. 23, 806–819. Hirzel, A., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecol. Model. 157, 331–341.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.

Huang, M.Y., Kong, X.Q., Varela, S., Duan, R.Y., 2016. The Niche Limitation Method (NicheLim), a new algorithm for generating virtual species to study biogeography. Ecol. Model. 320, 197–202.

Huntley, B., Green, R.E., Collingham, Y.C., Hill, J.K., Willis, S.G., Bartlein, P.J., Cramer, W., Hagemeijer, W.J.M., Thomas, C.J., 2004. The performance of models relating species geographical distributions to climate is independent of trophic level. Ecol. Lett. 7, 417–426.

Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Commun. Ecol. 10.

Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecol. Appl. 13, 853–867.

Lawson, C.R., Hodgson, J.A., Wilson, R.J., Richards, S.A., 2014. Prevalence, thresholds and the performance of presence-absence models. Methods Ecol. Evol. 5, 54–64.

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

Liu, C.R., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393.

Lobo, J.M., Jimenez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.

Marmion, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. Ecol. Model. 220, 3512–3520.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd edition. Chapman and Hall, London.

Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E., Elith, J., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281.

Meynard, C.N., Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. J. Biogeogr. 40, 1–8.

Miller, J.A., 2014. Virtual species distribution models: Using simulated data to evaluate aspects of model performance. Prog. Phys. Geogr. 38, 117–128.

Mitchell, P.J., Monk, J., Laurenson, L., 2017. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. Methods Ecol. Evol. 8, 12–21.

Mittelbach, G.G., Schemske, D.W., 2015. Ecological and evolutionary perspectives on community assembly. Trends Ecol. Evol. 30, 241–247.

Moilanen, A., Pouzols, F.M., Meller, L., Veach, V., Arponen, A., Leppänen, J., Kujala, H., R.F. Fernandes et al. Ecological Informatics 48 (2018) 125–134 133 2014. ZONATION: Spatial Conservation Planning Framework and Software. User Manual. Atte Moilanen/Metapopulation Research Group, University of Helsinki, Finland.

Munkemuller, T., Gallien, L., Lavergne, S., Renaud, J., Roquet, C., Abdulhak, S., Dullinger, S., Garraud, L., Guisan, A., Lenoir, J., Svenning, J.C., Van Es, J., Vittoz, P., Willner, W., Wohlgemuth, T., Zimmermann, N.E., Thuiller, W., 2014. Scale decisions can reverse conclusions on community assembly processes. Global Ecol. Biogeogr. 23, 620–632.

Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed effects models. Methods Ecol. Evol. 4, 133–142.

Nenzén, H.K., Araújo, M.B., 2011. Choice of threshold alters projections of species range shifts under climate change. Ecol. Model. 222, 3346–3354.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., Maintainer, R., 2017. Package ‘nlme’. Linear and nonlinear mixed effects models. pp. 3–13.

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C.F., Vittoz, P., Guisan, A., Field, R., 2013. The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. Glob. Ecol. Biogeogr. 22, 52–63.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., 2006. Are niche-based species distribution models transferable in space? J. Biogeogr. 33, 1689–1704.

Roura-Pascual, N., Brotons, L., Peterson, A.T., Thuiller, W., 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biol. Invasions 11, 1017–1031.

Santika, T., 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Glob. Ecol. Biogeogr. 20, 181–192.

Scherrer, D., D'Amen, M., Mateo, M.R.G., Fernandes, R.F., Guisan, A., 2018a. How to best threshold and validate stacked species assemblages? Community optimisation might hold the answer. Methods Ecol. Evol. 00, 1–12.

Scherrer, D., Mod, H.K., Pottier, J., Litsios-Dubuis, A., Pellissier, L., Vittoz, P., Götzenberger, L., Zobel, M., Guisan, A., 2018b. Disentangling the processes driving plant assemblages in mountain grasslands across spatial scales and environmental gradients. J. Ecol. 00, 1–14.

Soberon, J., Nakamura, M., 2009. Niches and distributional areas: concepts, methods, and assumptions. Proc. Natl. Acad. Sci. U. S. A. 106 (Suppl. 2), 19644–19650.

Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, 1–34.

Steinmann, K., Eggenberg, S., Wohlgemuth, T., Linder, H.P., Zimmermann, N.E., 2011. Niches and noise—Disentangling habitat diversity and area effect on species diversity. Ecol. Complex. 8, 313–319.

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

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A.C., Guisan, A., 2014. Measuring the relative effect of factors affecting species distribution model predictions. Methods Ecol. Evol. 5, 947–955.

Thuiller, W., Brotons, L., Araujo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27, 165–172.

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

Vale, C.G., Tarroso, P., Brito, J.C., 2014. Predicting species distribution at range margins: testing the effects of study area extent, resolution and threshold selection in the Sahara-Sahel transition zone. Divers. Distrib. 20, 20–33.

Vicente, J.R., Goncalves, J., Honrado, J.P., Randin, C.F., Pottier, J., Broennimann, O., Lomba, A., Guisan, A., 2014. A framework for assessing the scale of influence of environmental factors on ecological patterns. Ecol. Complex. 20, 151–156. Watts, M.E., Ball, I.R., Stewart, R.S., Klein, C.J., Wilson, K., Steinback, C., Lourival, R., Kircher, L., Possingham, H.P., 2009. Marxan with Zones: Software for optimal conservation based land- and sea-use zoning. Environ. Model. Softw. 24, 1513–1521.

Williams, J.N., Seo, C.W., Thorne, J., Nelson, J.K., Erwin, S., O'Brien, J.M., Schwartz, M.W., 2009. Using species distribution models to predict new occurrences for rare plants. Divers. Distrib. 15, 565–576.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Distribut, N.P.S., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Munkemuller, T., Nehrbass, N., Pagel, J., Reineking, B., Schroder, B., Grimm, V., 2010. The virtual ecologist approach: simulating data and observers. Oikos 119, 622–635.

Zurell, D., Thuiller, W., Pagel, J., Cabral, J.S., Munkemuller, T., Gravel, D., Dullinger, S., Normand, S., Schiffers, K.H., Moore, K.A., Zimmermann, N.E., 2016. Benchmarking novel approaches for modelling species range dynamics. Glob. Chang. Biol. 22, 2651–2664.

Supplementary material

Appendix A

Table A.1. Community evaluation metrics used in this study (see Di Cola et al., 2017 for details)

Metric	Definition	Description	References
Species Richness Error	$SRe = n_{pred} - n_{obs}$	Difference between predicted and observed species richness	(Pottier et al., 2013)
Prediction Success	$PredSuc = \frac{(TP + TA)}{N}$	Proportion of species correctly predicted as present or absent	(Pottier et al., 2013)
Sensitivity	$Sens = \frac{TP}{TP + FA}$	The proportion of species correctly predicted as present	(Pottier et al., 2013)
Specificity	$Spec = \frac{TA}{TA + FP}$	The proportion of species correctly predicted as absent	(Pottier et al., 2013)
Over-prediction	$OverPred = \frac{FP}{FP + TA}$	The proportion of species predicted as present but not observed among the species predicted as present	(Pottier et al., 2013)
Under-prediction	$UnderPred = \frac{FA}{TP + FA}$	The proportion of species predicted as absent but observed among the species observed as present	(Pottier et al., 2013)
Community TSS	$TSS = Sens + Spec - 1$	Same as TSS but measured for a site across all species, rather than for a species across all sites	(Pottier et al., 2013)
Community Kappa	$K = \frac{Acc - \ p_e}{1 - p_e}$	Same as Kappa but measured for a site across all species, rather than for a species across all sites	(Pottier et al., 2013)
Sørensen	$S = \frac{2 * TP}{2 * TP + FP + FA}$	Similarity index (compares similarity between observed and predicted assemblages)	(Sørensen, 1948)

Original dataset information

To obtain the full information about the original plant data, two datasets where combined (Dubuis et al., 2011; Pottier et al., 2013). The first one with 912 vegetation plots of 4 m^2 were selected following a random-stratified sampling design and the presence-absence of each species was recorded in each plot. Only vascular species in open and non-woody vegetation were sampled. This dataset consisted of 795 species. The second dataset with 3076 vegetation plots, from a grid of 400 m over all of the study area; therefore, a point was recorded every 400 m. If a point was falling into a forest, a field sampling was made. The field sampling was done in a circle with center at the coordinates of the point and with a radius of 10 m. All vascular plant species were recorded. This dataset consisted of 667 species. The final dataset of plants has 3967 plots composed by 1088 species (627 after removing the species with less than 30 occurrences).

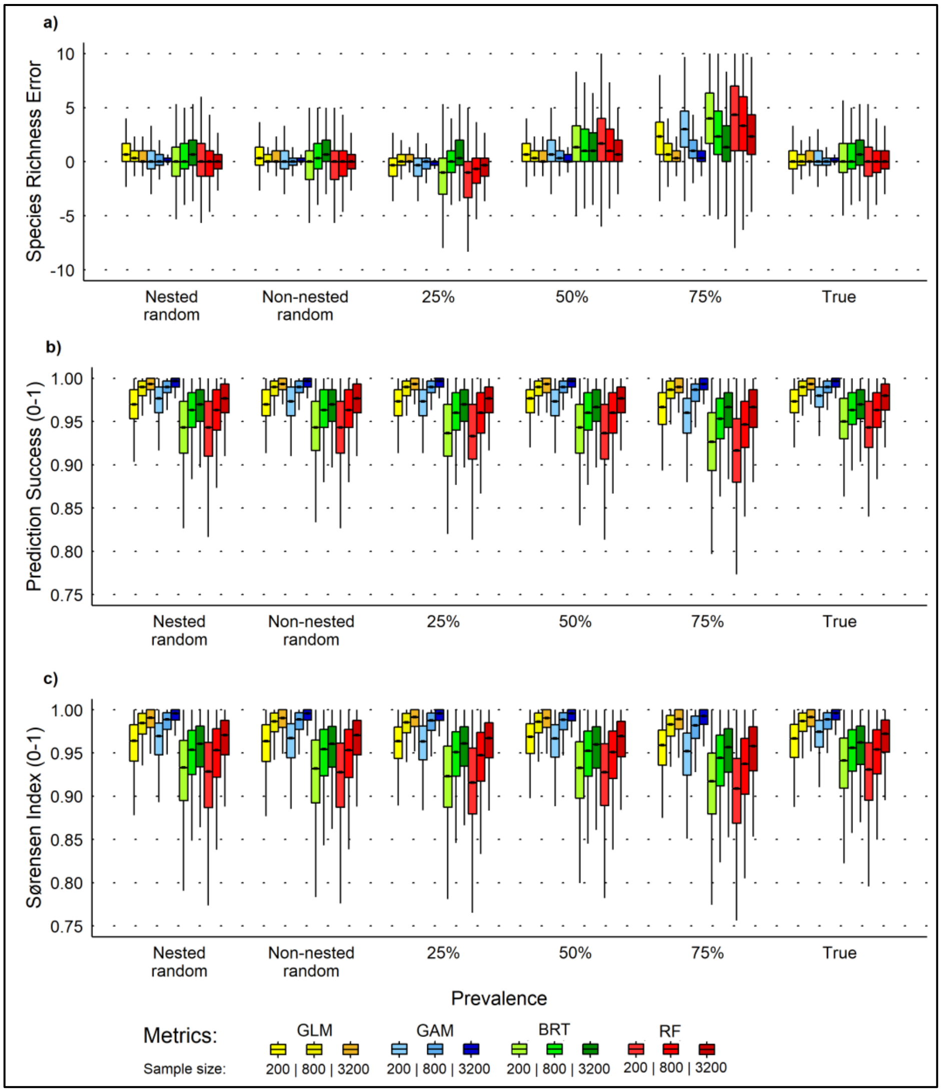

Fig. A.1 Boxplots of different indices of community prediction (S-SDM) accuracy (i.e. species richness error, prediction success and Sørensen) for all the simulated species and for all the sampling strategies (based on plots (nested or not) or prevalence (25%, 50%, 75% or true) sampling; in abscissa). Each box shows the variation across all virtual species in a random subset of the study area (100 000 plots) for the binary predictions obtained using ROC as thresholding technique, averaged from the three sampling turns. For each prevalence sampling, four sets of three boxplots are displayed, corresponding to models fitted using either GLMs (yellow), GAMs (blue), BRTs (green) or RFs (red), with increasing values of sample size (200, 800 and 3200).

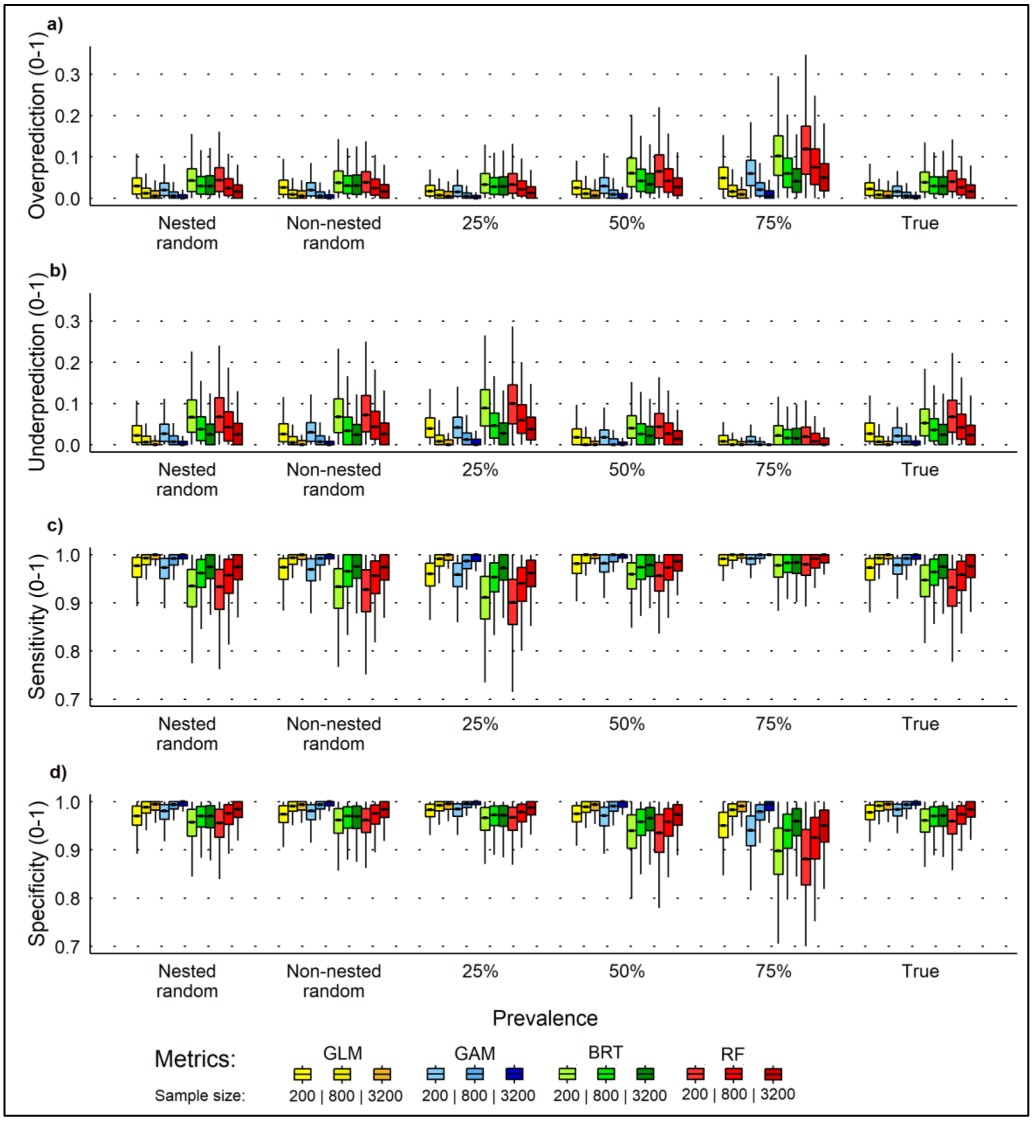

Fig. A.2 Boxplots of different indices of community prediction (S-SDM) accuracy (i.e. over and underprediction, sensitivity and specificity) for all the simulated species and for all the sampling strategies (based on plots (nested or not) or prevalence (25%, 50%, 75% or true) sampling; in abscissa). Each box shows the variation across all virtual species in a random subset of the study area (100 000 plots) for the binary predictions obtained using ROC as thresholding technique, averaged from the three sampling turns. For each prevalence sampling, four sets of three boxplots are displayed, corresponding to models fitted using either GLMs (yellow), GAMs (blue), BRTs (green) or RFs (red), with increasing values of sample size (200, 800 and 3200).

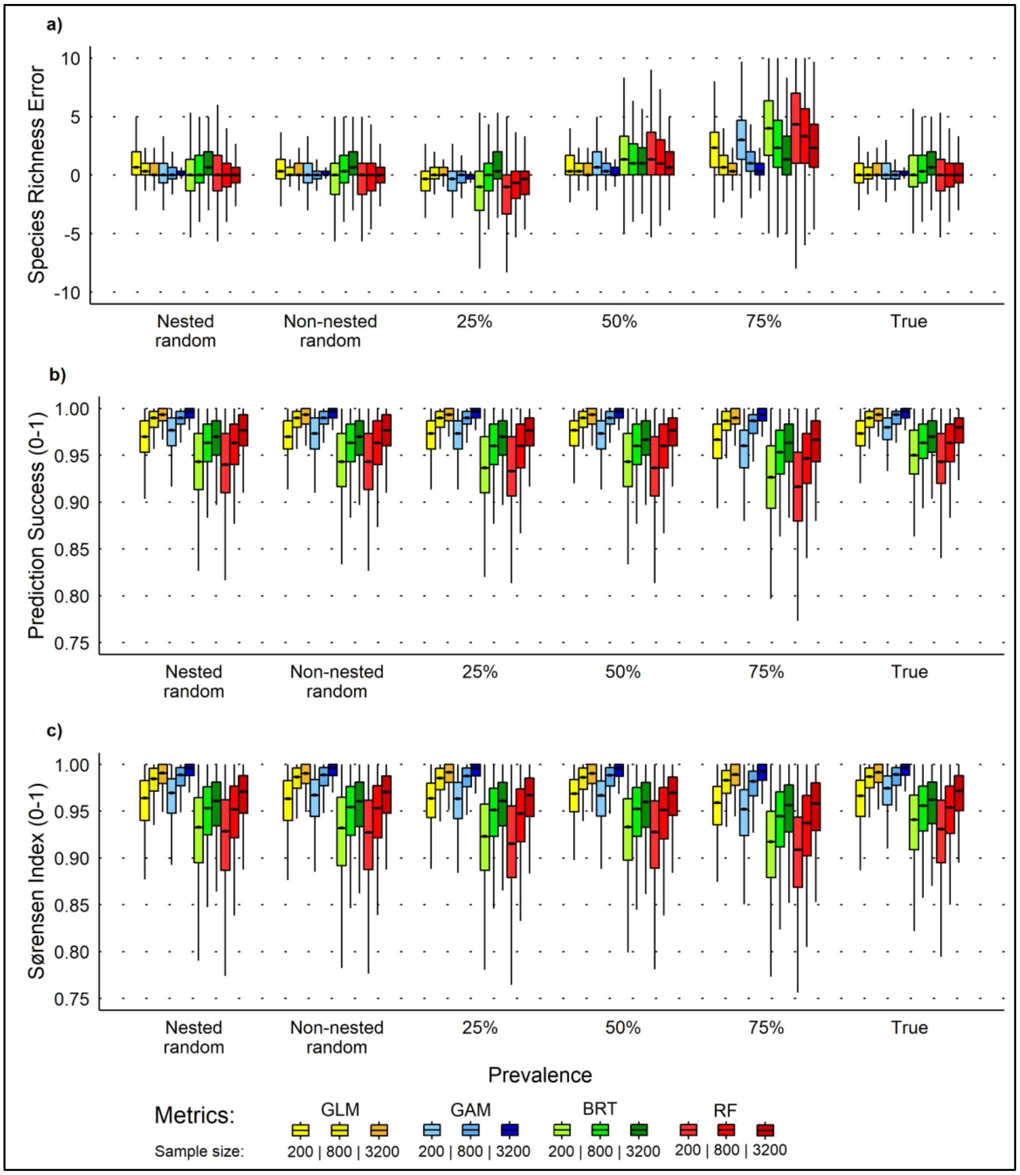

Fig. A.3 Boxplots of different indices of community prediction (S-SDM) accuracy (i.e. species richness error, prediction success and Sørensen) for all the simulated species and for all the sampling strategies (based on plots (nested or not) or prevalence (25%, 50%, 75% or true) sampling; in abscissa). Each box shows the variation across all virtual species in a random subset of the study area (100 000 plots) for the binary predictions obtained using MaxTSS as thresholding technique, averaged from the three sampling turns. For each prevalence sampling, four sets of three boxplots are displayed, corresponding to models fitted using either GLMs (yellow), GAMs (blue), BRTs (green) or RFs (red), with increasing values of sample size (200, 800 and 3200).

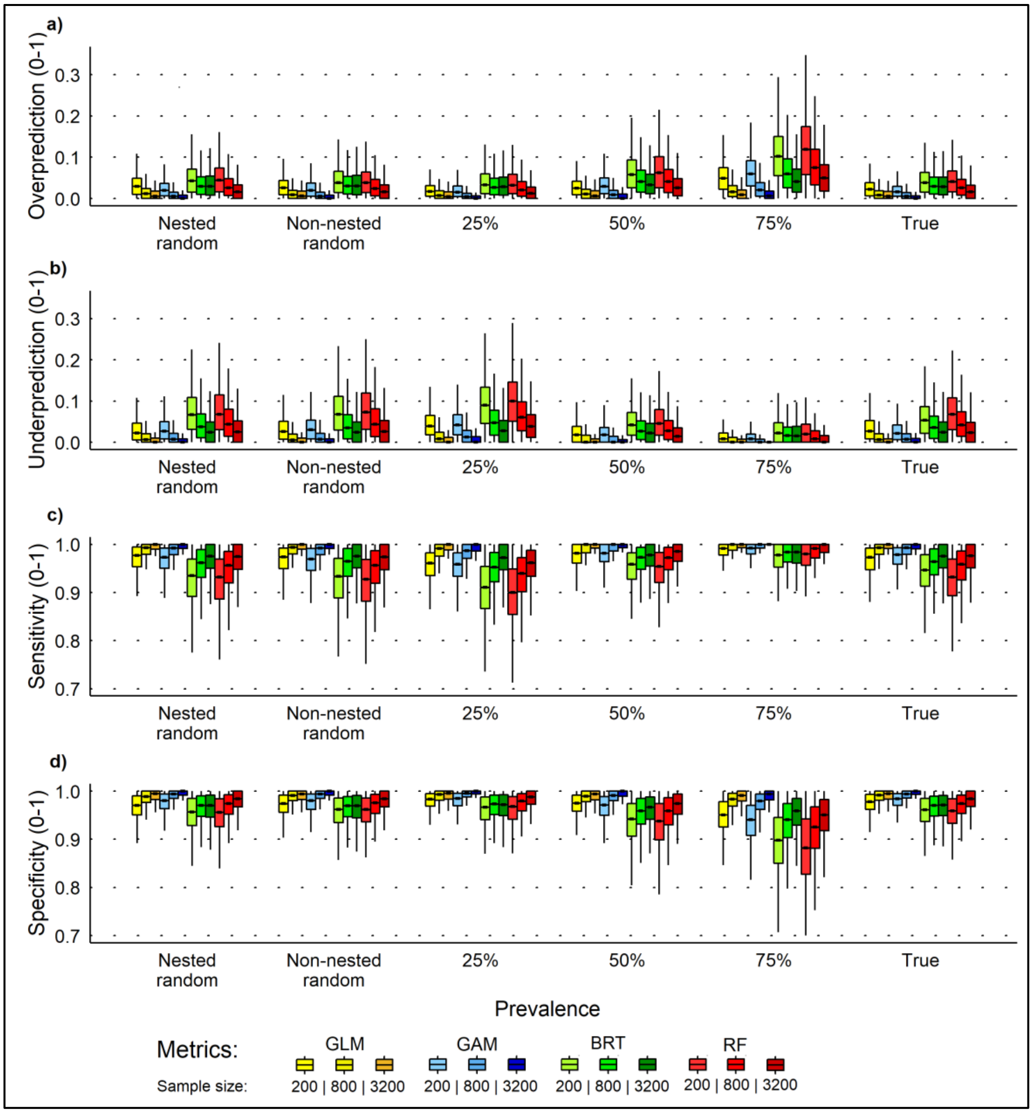

Fig. A.4 Boxplots of different indices of community prediction (S-SDM) accuracy (i.e. over and underprediction, sensitivity and specificity) for all the simulated species and for all the sampling strategies (based on plots (nested or not) or prevalence (25%, 50%, 75% or true) sampling; in abscissa). Each box shows the variation across all virtual species in a random subset of the study area (100 000 plots) for the binary predictions obtained using MaxTSS as thresholding technique, averaged from the three sampling turns. For each prevalence sampling, four sets of three boxplots are displayed, corresponding to models fitted using either GLMs (yellow), GAMs (blue), BRTs (green) or RFs (red), with increasing values of sample size (200, 800 and 3200).