# How to best threshold and validate stacked species assemblages? Community optimisation might hold the answer

Daniel Scherrer[a], Manuela D'Amen[a], Rui F. Fernandes[a], Rubén G. Mateo[a,b] and Antoine Guisan[a,c]

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

[b] ETSI de Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040, Madrid, Spain.

[c] Institute of Earth Surface Dynamics, University of Lausanne, Géopolis, CH-1015 Lausanne, Switzerland

Corresponding author:
Daniel Scherrer
Department of Ecology and Evolution
University of Lausanne
CH-1015 Lausanne, Switzerland

Running title: Thresholding stacked assemblages

## ABSTRACT

1. The popularity of species distribution models (SDMs) and the associated stacked species distribution models (S-SDMs), as tools for community ecologists, largely increased in recent years. However, while some consensus was reached about the best methods to threshold and evaluate individual SDMs, little agreement exists on how to best assemble individual SDMs into communities, i.e. how to build and assess S-SDM predictions.
2. Here, we used published data of insects and plants collected within the same study region to test (1) if the most established thresholding methods to optimize single species prediction are also the best choice for predicting species assemblage composition, or if community-based thresholding can be a better alternative, and (2) whether the optimal thresholding method depends on taxa, prevalence distribution and/or species richness. Based on a comparison of different evaluation approaches we provide guidelines for a robust community cross-validation framework, to use if spatial or temporal independent data are unavailable.
3. Our results showed that the selection of the "optimal" assembly strategy mostly depends on the evaluation approach rather than taxa, prevalence distribution, regional species pool or species richness. If evaluated with independent data or reliable cross-validation, community-based thresholding seems superior compared to single species optimisation. However, many published studies did not evaluate community projections with independent data, often leading to overoptimistic community evaluation metrics based on single species optimisation.
4. The fact that most of the reviewed S-SDM studies reported over-fitted community evaluation metrics highlights the importance of developing clear evaluation guidelines

for community models. Here, we move a first step in this direction, providing a framework for cross-validation at the community level.

## INTRODUCTION

Past and future environmental changes may not only lead to shifts in species distributions (e.g., Parmesan & Yohe 2003; Thuiller *et al.* 2005; Dullinger *et al.* 2012), but also to changes in species assemblages and interactions (e.g., Van der Putten, Macel & Visser 2010; Nogues-Bravo & Rahbek 2011; Blois *et al.* 2013; Alexander *et al.* 2016). Information about communities, here defined as a taxonomic assemblage of distinct populations of species that co-occur in a given space at a given time (Begon, Harper & Townsend 1996), is therefore essential to make informed decisions for conservation prioritisation (D'Amen *et al.* 2011; Guisan *et al.* 2013; Mateo *et al.* 2013) and to create biodiversity indices (e.g., Essential Biodiversity Variables; Pereira *et al.* 2013) for policy decisions (Fleishman, Noss & Noon 2006; Granger *et al.* 2015).

Different approaches to model communities are available, using either correlative (e.g., Ferrier & Guisan 2006; Guisan & Rahbek 2011) or mechanistic techniques (e.g., Kearney & Porter 2009; Mokany & Ferrier 2011), with some predicting only macro-ecological properties such as species richness (e.g., Currie *et al.* 2004; Gotelli *et al.* 2009; Dubuis *et al.* 2011) and others also predicting community composition (see D'Amen *et al.* 2017b for a review). In this study, we focused on correlative approaches based on individual species distribution models (SDMs), as they are the most common technique applied to conservation strategies (Guisan *et al.* 2013), and to predict future patterns of biodiversity in the face of global change (Nogues-Bravo & Rahbek 2011; D'Amen *et al.* 2017b). Niche-based SDMs quantify the relationship between available species occurrences and different environmental factors to analyse and predict distributional patterns (Guisan & Thuiller 2005; Elith & Leathwick 2009; Guisan, Thuiller & Zimmermann 2017). By additionally stacking individual SDMs (S-SDMs), one can produce spatiotemporal projections of species richness and composition (Ferrier & Guisan 2006; Guisan & Rahbek 2011).

While there is a vast and now long-standing literature on advances and limitations of single species predictions (e.g., Guisan & Thuiller 2005; Guisan *et al.* 2006; Maggini *et al.* 2006; Elith & Leathwick 2009; Meier *et al.* 2010; Zimmermann *et al.* 2010; Merow *et al.* 2014), studies exploring how to improve community predictions based on aggregated information from individual SDMs emerged more recently (e.g., Mateo *et al.* 2012; Benito, Cayuela & Albuquerque 2013; Cord *et al.* 2014; Mod *et al.* 2015; but see Ferrier et al. 2002). A fundamental difference among the proposed solutions is whether to maintain the information on species composition in the final predictions. For instance, the simple sum of probabilities of individual SDM predictions usually gives better estimates of species richness, but the information on species identity is lost (Dubuis *et al.* 2011; Calabrese *et al.* 2014). Therefore, predictions of community composition have mainly been achieved so far by thresholding the individual continuous SDM predictions (e.g., probability or suitability index) to obtain binary maps (Liu, White & Newell 2013) and then stacking the latter at the assemblage level (e.g., Pottier *et al.* 2013; D'Amen *et al.* 2015; D'Amen, Pradervand & Guisan 2015).

There are several examples in the literature of optimizing thresholding methods for single species predictions (e.g., Liu *et al.* 2005; Jimenez-Valverde & Lobo 2007; Freeman & Moisen 2008; Liu, White & Newell 2013). These led to a mounting consensus about the most appropriate methods, with the majority of SDM studies published nowadays using either an approach maximising the true skills statistics (Max.TSS) or based on the curve in a receiver operating characteristic plot (Opt.ROC, related to AUC) (see Guisan, Thuiller & Zimmermann 2017; Table S1). However, the threshold selection can strongly influence the reliability of the predicted richness and composition of S-SDMs assemblages (Pineda & Lobo 2009; Benito, Cayuela & Albuquerque 2013). It is thus relevant to explore which thresholding approach provides the best performance in assemblage estimates, and if alternatives exist that can improve the assemblage prediction from individual SDMs.

Studies focussing on S-SDMs tend to over-predict species richness when based on (thresholded) binary predictions (e.g., Pineda & Lobo 2009; Dubuis *et al.* 2011; Mateo *et al.* 2012; Pottier *et al.* 2013; Pouteau *et al.* 2015), with some exceptions (e.g., D'Amen, Pradervand & Guisan 2015; Distler *et al.* 2015). Different factors have been proposed to explain this over-prediction: (1) a statistical bias in thresholding site-level occurrence probabilities for each species (Calabrese *et al.* 2014); (2) the implicit assumption of unsaturated communities not assuming an ecological limit for species numbers in assemblages (environmental carrying capacity; Guisan & Rahbek 2011); (3) the lack of considering different constraints on community composition (i.e., ecological, evolutionary, historical, or biological biodiversity drivers; see Mateo, Mokany & Guisan 2017).

The commonly used approach to get binary maps from continuous SDM predictions is to use a species-specific threshold, i.e. each species has a single threshold across all sites ("species threshold", Calabrese *et al.* 2014). Recently, another community-based approach, called probability ranking rule (PRR), was proposed to predict assemblage composition from individual SDMs (D'Amen *et al.* 2015). This method does not require a species-specific threshold, therefore preventing over-prediction, but site-by-site ecological constraints (e.g., macro-ecological models) are applied to assemblages to predict species richness ("site-threshold").

Surprisingly, studies aiming to test and improve S-SDM have used very different approaches to evaluate the predicted assemblages (Cord *et al.* 2014; Hespanhol *et al.* 2015; Pouteau *et al.* 2015; Thuiller *et al.* 2015; Zurell *et al.* 2016) and this evaluation aspect of the community modelling procedure has not yet received all the attention it deserves. In most studies, assemblage predictions are not adequately evaluated because the data used for the evaluation were already used for individual model fitting, not allowing anymore a correct cross-validation at the community level. Ideally, the best evaluation method should use spatial or

120 temporal independent data (Elith *et al.* 2006; Guisan, Thuiller & Zimmermann 2017), but if
121 not available, an appropriate cross-validation approach should at least be set up.

122 Here, we used published high-resolution data of insects (butterflies and grasshoppers) and
123 plants (forests and grasslands sites), collected within the same study region to (1) test if the
124 most established thresholding methods for optimal single species prediction (i.e., Max.TSS
125 and Opt.ROC) are also the best choice for species assemblages, (2) investigate if the optimal
126 thresholding method depends on taxa, prevalence distribution (Allouche, Tsoar & Kadmon
127 2006), and/or species richness and (3) provide guidelines for a correct community cross-
128 validation framework, to be used if spatially- or temporally- independent data are unavailable.

## MATERIALS AND METHODS

### Community data and environmental variables

*Study area*

The data on all taxa were collected within the same study area located in the western Swiss Alps of the canton Vaud (Fig. 1; 46°10´ to 46°30´ N; 6°50’ to 7°10’ E), covering an area of ca. 700 $km^2$, with elevation ranging from 375 to 3210 m a.s.l. and forested areas up to 1900 m a.s.l. For centuries, agriculture (farming and pasturing) has maintained grasslands among forests and altered the position of the treeline. The highly variable topography and diverse land use of the study area, in combination with our high-resolution environmental data (25 x 25 m cell size), provide a huge range of complex species-environment relationships to test our modelling framework.

*Plant data*

The forest data were part of a forest inventory of the canton Vaud conducted between 1988 and 2002 (mostly 1990 to 1994) and consisted of 3076 sites. The forest sites were distributed on a 400 m grid all across the forested area of the canton and had a circular area of 314 $m^2$ (Fig. 1; for details see Hartmann, Fouvy & Horisberger 2009). In total, 703 plant species were recorded, but only 312 (44%) had enough occurrence data (> 20 occurrences) across the dataset for modelling purposes (see Table 1 for more detailed statistics on the datasets).

The grassland dataset was collected between 2002 and 2009 following an equal random-stratified sampling of non-forested areas in the study area. In total, 911 vegetation sites of 4 $m^2$ were sampled (Fig. 1; for more information see Dubuis *et al.* 2011). A total of 905 plant species were recorded but only the 212 most frequent (>20 occurrences) were selected for modelling (Table 1).

To predict the distribution of the plant species we used five environmental variables: growing

153 degree-day (above 0 °C), moisture index over the growing season (difference between
154 precipitation and potential evapotranspiration), the sum of potential solar radiation over the
155 year, slope (in degrees), and topographic position (unit-less, indicating the ridges and valleys).
156 All these variables were at a 25 m resolution and have been shown to be useful predictors for
157 plant species in mountain environments (see Dubuis *et al.* 2011; D'Amen *et al.* 2015; Scherrer
158 *et al.* 2017 for details on predictors).

159 *Insect data*

160 Data on butterflies and grasshoppers were respectively collected in 192 and 202 squares of 50
161 m x 50 m across all the elevational range of the study area (Fig. 1; see Pellissier *et al.* 2012;
162 Pradervand *et al.* 2013, for more information). In total, 131 butterfly and 41 grasshopper
163 species were observed, but due to model limitations only the most common 67 butterfly and
164 20 grasshopper species (>=20 occurrences) were considered for modelling (Table 1).

165 For our SDMs we used the same predictors as D'Amen, Pradervand and Guisan (2015): four
166 bioclimatic variables (solar radiation, summer temperature, annual degree-days and annual
167 average number of frost days during the growing season), an index of vegetation productivity,
168 i.e. normalized difference vegetation index (as proxies for trophic resources), and the distance
169 to forest. These variables were selected as they are not highly correlated (<0.7; Dormann *et al.*
170 2013) and considered ecologically important for insects (e.g., Turner, Gatehouse & Corey
171 1987; Hawkins & Porter 2003).

172 **The modelling framework**

173 Our modelling framework used three different S-SDM based community modelling pathways
174 ("single species cross-validation", "independent data" and "community cross-validation)
175 representing the most commonly reported practices in the literature (see Fig. 2 and
176 "Evaluating community predictions" section).

177 *Single species modelling, thresholding and evaluation*

178 Individual species models were run by generalised linear models (GLM; McCullagh & Nelder
179 1989), generalised additive models (GAM; Hastie & Tibshirani 1990), random forest (RF;
180 Breiman 2001) and boosted regression trees (BRT; Elith, Leathwick & Hastie 2008). Models
181 for species with more than 50 occurrences were fitted by simple SDMs using all five selected
182 predictors, followed by a weighted (AUC) ensemble forecast (Marmion *et al.* 2009). Species
183 having only between 20 and 50 occurrence records were fitted by an ensemble bivariate
184 approach optimised for rare or under-sampled species (Lomba *et al.* 2010; Breiner *et al.*
185 2015): individual models were calibrated on bivariate combinations of the selected predictors
186 with all four modelling techniques, followed by a consensus forecast from all the resulting
187 "small models" weighted by their AUC scores. We used a repeated split-sample procedure
188 (N=25) for model evaluation, followed by a weighted (AUC) ensemble forecast (across
189 techniques and split-sample runs).

190 The projected probability outputs of the ensemble models were binarised using two
191 thresholding schemes: (1) *species-specific-thresholds* (a single threshold calculated for each
192 species) and (2) *site-specific-thresholds* (differing for each site on the basis of additional
193 community information, i.e. species richness predictions). We selected seven different
194 species-specific-thresholding techniques, which can be classified in four major groups: single-
195 index based, sensitivity and specificity combined, model-building data-only-based, and
196 predicted probability-based (see Table S1; Liu *et al.* 2005; Nenzen & Araujo 2011 for details
197 on classification). As the thresholding techniques showed minimal within-group variance (see
198 Figure S1 and S2), we decided to only present the results for one thresholding technique per
199 group in the main manuscript. The chosen techniques were: Cohen's Kappa maximization
200 approach (*Max.Kappa*; single-index based), TSS maximization approach (*Max.TSS*,
201 sensitivity and specificity combined), observed prevalence (*Obs.Preval*; model-building data-

202 only-based approach), and average probability approach (*AvgProb*; predicted probability-
203 based approach; for details on techniques see Table S1). In addition, we applied two site-
204 thresholds (community-based approaches) using species richness (SR) predictions in
205 combination with a probability ranking rule (PRR). These methods selected a number of
206 species equal to the predicted SR on the basis of decreasing probabilities of presence
207 calculated by the SDMs (D'Amen *et al.* 2015; D'Amen, Pradervand & Guisan 2015).
208 Therefore, the species with the highest probabilities in a site are selected (considered present)
209 in decreasing order until the SR predicted for the site is reached. The SR predictions were
210 derived by either summing the per site probabilities of individual SDMs, obtaining a
211 prediction of richness for each site (pS-SDM; Dubuis *et al.* 2011) or by a macro-ecological
212 model (MEM; see D'Amen, Pradervand & Guisan 2015 for details), directly modelling the
213 richness of the sites. As results from the two site-thresholds were concordant, we only show
214 here the former (*pS-SDM+PRR*).

215 To evaluate the threshold independent performance of our individual species models, the area
216 under the curve of a Receiver-Operating Characteristic (ROC) plot (AUC; Fielding & Bell
217 1997) was calculated based on a repeated split sampling cross-validation (Thuiller, Georges &
218 Engler 2013). Additionally, based on our independent/cross-validation data we calculated five
219 threshold dependent metrics for each thresholding technique: the overall accuracy (PCC; i.e.
220 proportion of correctly classified presence and absences; Fielding & Bell 1997), sensitivity
221 (proportion of correctly predicted presences), specificity (proportion of correctly predicted
222 absences), the true skill statistic (i.e. [(sensitivity + specificity) -1]; TSS; Allouche, Tsoar &
223 Kadmon 2006) and Cohen's Kappa (Kappa; i.e., overall accuracy but corrected for chance
224 performance; Cohen 1968).

225 *Evaluating community predictions*

All the community predictions were built by stacking binary SDMs of individual species (S-SDMs; Dubuis *et al.* 2011; Guisan & Rahbek 2011). The three modelling pathways (Fig. 2) were identical regarding the modelling procedure for single species, thresholding and community assemblage and only varied in the selection of the data for community calibration and evaluation.

- The "single species cross-validation" (SSCV) approach (Fig. 2) has not fully "unused/independent" data for community evaluation (i.e. sites not used for the calibration of any single species). Here, in the process of the cross-validation of all individual SDMs (i.e. across all species), different sites are selected at each resampling iteration and for each species, so that all sites are most likely used in at least one split-sampling run and their information incorporated in the final ensemble model. This approach cannot thus be considered based on fully independent data. The SSCV approach has been to date the most common way to model and evaluate communities predictions based on S-SDMs (Fig. 2; e.g., Dubuis *et al.* 2011; Calabrese *et al.* 2014; D'Amen, Pradervand & Guisan 2015; Distler *et al.* 2015). As no independent data is set aside for community evaluation, this approach usually gets evaluated with all the sites used for calibration. However, to avoid bias in the results due to different numbers of evaluation sites, we evaluated the SSCV approach only on 30% of the available sites (identical to the ID and CCV approach below).
- The (spatial or temporal) "independent data" (ID) approach (Fig. 2) starts with two completely independent datasets. One is used for the calibration of the SDMs (i.e. 70% of the sites) and the other set is used (only) to evaluate the performance of the community predictions (i.e. 30% of the sites; Fig 2; e.g., Benito, Cayuela & Albuquerque 2013; Pottier *et al.* 2013; Cord *et al.* 2014; D'Amen *et al.* 2015; Zurell *et al.* 2016).

- The "community cross-validation" (CCV) approach (Fig. 2) uses a repeated split sampling of sites (100 repetitions) dividing the available sites into calibration (70%) and evaluation sets (30%) to perform all the modelling procedure from the single species prediction to the community assembly (Fig. 2). In contrast to the previous ID pathway (above), which only uses one (spatial or temporal) fixed independent evaluation dataset, in the CCV approach all SDMs are fitted at each split-sample iteration using the same training and test sets for all species, thus minimizing the risk of bias in the evaluation data (i.e. if the training and test sets differ across species, as in the ID approach). This repeated cross-validation also allows the estimation/simulation of confidence intervals for community predictions instead of just a single value per community. To our knowledge, no study used this community cross-validation method so far.

To compare the community model performance among thresholding techniques and modelling pathways, we calculated eight different community agreement metrics: 1) the deviation of the predicted from the observed species richness (SR.deviation), 2) the proportion of species correctly predicted as present (community sensitivity), 3) the proportion of species correctly predicted as absent (community specificity), 4) community accuracy (PCC; i.e. the percent correctly classified species, present or absent), 5) the 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) , 6) the community kappa (same as for TSS, for a site across species; Pottier *et al.* 2013), and 7) the Sørensen similarity (Sørensen 1948).

*Correlation of single species and community evaluation metrics*

For each combination of dataset, modelling pathway and thresholding method (4 x 3 x 9 = 108) we calculated the average evaluation metric for all five single species metrics and all seven community metrics. We then calculated the Spearman correlation of all possible

combinations of our five single species and seven community evaluation metrics. The resulting correlation matrix tells us if methods (modelling pathways or thresholding methods) that yield the highest scores in a certain single species metric also yield the highest score in the corresponding community evaluation metric.

## RESULTS

### Performance of individual SDMs

As expected the evaluation scores of the individual SDMs were similar to earlier studies published with the same data (D'Amen *et al.* 2015; D'Amen, Pradervand & Guisan 2015; Scherrer *et al.* 2017) and their performance was not affected by the chosen community evaluation approach (Table 1, Table S3). Despite their differences in site SR, prevalence distribution and species pool the average performance of individual SDMs was similar across all taxa (Table 1, Table S3). Additionally, the often reported effect of species prevalence on model performance was only marginal in our study, with rare and common species having similar average model performance within a given taxonomic group (Fig. S3).

### Correlation of single species and community evaluation metrics

The correlation between the single species and corresponding community metrics was highest (cor > 0.93; Table 2) for some combinations of metrics based on partial information from the contingency table comparing predictions to observations (i.e. PCC, specificity and sensitivity) and considerably lower for the metrics accounting for all dimensions of the contingency table, such as TSS and Cohen's Kappa (cor = 0.73; Table 2). Correlations between non-corresponding single species and community metrics (i.e. Sørensen and SR deviation) tended to be even lower, with the exception of Kappa versus Sørensen (Table 2).

### Species richness and compositional similarity

The deviation in species richness between observed and predicted communities was strongly dependent on the chosen thresholding method (Fig. 3). The thresholding approach that uses the average predicted probability (*AvgProb*) showed the highest amount of over-prediction followed by the combined sensitivity and specificity approach (*Max.TSS*). The other three thresholding methods (*Obs.Preval*, *Max.Kappa* and *pS-SDM+PRR*) performed very similar and showed overall no tendency to over-predict species richness. There were no significant differences between the three modelling pathways for any of the studied taxa (Fig. 3). The absolute number of over-predicted species was strongly related to the average number of species per plot (SR) and therefore differed among the taxa (Fig. 3). However, when corrected for the differences in SR the over-prediction did not significantly vary anymore across taxa.

The compositional similarity (Sørensen similarity index) varied significantly both among thresholding techniques and modelling pathways (Fig. 4). The compositional similarity was expectedly always much higher with the "single species cross-validation" (SSCV) pathway compared to the "independent data" (ID) or the "community cross-validation" (CCV) pathways, which both performed similarly. There was also a strong interaction between modelling pathway and thresholding technique. Using the SSCV pathway, thresholding by *Obs.Preval* and by *Max.Kappa* performed better (Fig. 4). However, if independent sites were available for the community evaluations (ID and CCV pathways), the community based approaches (*pS-SDM+PRR*) performed better than the *Obs.Preval* and *Max.Kappa* thresholds (Fig. 4). The similarity between predicted and observed communities was higher in the two insect datasets than in the two plant datasets (Fig. 4), which is likely due to the lower number of insect species compared to plant species modelled. Surprisingly, the most established thresholding methods for single species SDMs based on sensitivity and specificity (i.e. *Max.TSS, Opt.ROC* and *SenSpec*; Fig. 4 and Fig. S1 and S2) never ranked highest, as one or

more of the other thresholding method always ranked above them, both for community composition and for species richness.

## DISCUSSION

### Do the most established thresholds for single species work as well for community predictions?

In this paper, we asked if the most established methods for single species thresholding are also the optimal choice for making predictions at the community level and if there is a direct link between the individual species predictions and the corresponding community metrics. Our results confirm the existence of such a link for single-index based metrics such as sensitivity, specificity and accuracy. However, these results should be interpreted with caution as maximising sensitivity or specificity can simply be achieved by predicting the species as present or absent (respectively) everywhere. In our study system, most of the modelled species have a low prevalence (i.e. are absent at most sites), thus accuracy (PCC) can often be improved by predicting the species as "absent" nearly everywhere.

The two most commonly used community evaluation metrics, Sørensen similarity index and deviation in species richness, were only weakly correlated with most evaluation metrics used for individual species. The most established thresholding methods for individual species predictions (i.e., *Max.TSS, Opt.ROC, SenSpec*) did show lower performance when applied to community-level predictions. This is likely due to the fact that both TSS and ROC try to find the best trade-off between sensitivity and specificity (Guisan, Thuiller & Zimmermann 2017). As most of the species have a prevalence far below 50% (i.e., are absent in many more sites than present), adding a few more presences might have a big effect on the sensitivity (by increasing the chance of finding the few real presences) but only marginally affects the specificity. By definition, increasing sensitivity also increases TSS, but with the drawback of a slight over-prediction. While this might not matter much on a single species basis, for community-level predictions the over-prediction will accumulate when summing binarised maps across all species, leading to the often observed over-estimation of species richness in S-

SDMs (e.g., Pineda & Lobo 2009; Dubuis *et al.* 2011; Mateo *et al.* 2012; Pottier *et al.* 2013; Pouteau *et al.* 2015; Zurell *et al.* 2016). It is important to remark, that in the rare case of an ecosystem mostly comprising of widespread species (i.e., prevalence >50 %) this will turn into the opposite as TSS and ROC will optimise absences leading to an underestimation of species richness. The strength of the over/under prediction bias is therefore linked to the prevalence distribution of the modelled species assemblages. However, in the vast majority of natural systems, both the site SR and the regional species pool are driven by a large number of rare (low prevalence species) compared to a few widespread species (Preston 1948; Magurran & Henderson 2003).

The community-based thresholding methods based on the selection of the most probable species (through a probability ranking) up to the predicted site richness (*MEM+PRR*, *pS-SDM+PRR*) can overcome this problem, because they are able to constrain species predictions based on a different value of species richness in each site (i.e. making them site-specific thresholding methods). Therefore, these methods prevent over-prediction while still allowing the analyses of species composition. Our results thus support the conclusion that, when the final goal is to optimize community composition, community-thresholding methods are the best option. Yet, as discussed in the next section, two single-species thresholding methods – *maximized Kappa* and *observed prevalence* – also showed good results for predicting communities (close to the community-based approaches). However, as community-based thresholds combine the optimisation of species richness prediction and a probability ranking rule (PRR), they would always select the species with the highest predicted probabilities in each site (D'Amen, Pradervand & Guisan 2015). This could seem logic and straightforward, but there might be a bias when the species in the community have varying prevalence (D'Amen *et al.* 2017a). In fact, the maximum predicted probability is depending on the prevalence of the species, thus the common species will tend to always have greater

maximum predicted probabilities than rare species and, as a result, will be considered present an over-proportionate number of time in the final community compositions. This bias will produce high similarity scores (Sørensen index) in the prediction evaluation, as the most common species are correctly predicted in most sites. However, the drawback is that the rarest species will be often omitted in the community predictions, which can be for instance problematic if the final goal of the modelling exercise has conservation implications.

**Is there a "best" threshold for community S-SDMs?**

We also tested if different methods for binarising community S-SDMs could be superior depending on the taxonomic group, prevalence distribution or species richness. While we observed significant differences between the different groups (i.e. taxa), there is no simple statistical way to assess if these differences are attributable to the biology of the taxa themselves or simply to the differences in site species richness and prevalence distributions. Nevertheless, when we standardized the deviation in species richness by the total number of modelled species (regional species pool), no significant difference was any more visible among the different taxonomic groups. The differences in species richness deviation seem therefore a direct cause of the regional species pool. The same also seems correct for the Sørensen similarity index, as datasets with higher species richness and species pool have lower similarity scores. This likely results from the fact that the more species need to be predicted correctly, the more difficult it becomes to predict the whole communities.

A similar ranking of thresholding methods was overall observed across taxonomic group within a given modelling pathway, while among the pathways there were clear shifts in the ranking of thresholding methods: with no independent community evaluation data (SSCV), the *Obs.Preval* and *Max.Kappa* threshold showed superior results, while the pathways using independent community evaluation data (ID and CCV) indicated the community-based thresholding to be superior (*pS-SDM+PRR*). This observation is in line with published

literature, where studies not using independent community data usually report a good performance of single species optimisations methods (e.g. D'Amen, Pradervand & Guisan 2015; Distler *et al.* 2015; Thuiller *et al.* 2015), while studies using independent data usually have better results using community constraints (e.g. D'Amen *et al.* 2015). Yet, it is remarkable to notice that, although previously much criticized in the literature (e.g., McPherson, Jetz & Rogers 2004; Allouche, Tsoar & Kadmon 2006), maximized Kappa (together here with the observed prevalence) did indeed perform well as a thresholding method for predicting both single species and communities, being nearly always superior to the sensitivity-specificity thresholding methods supporting earlier findings of Manel, Williams and Ormerod (2001).

It is important to notice that the shift in ranking between modelling pathways was likely due to a lower degree of overfitting and therefore a lower decrease in performance when predicting to independent data.

**Summing up: How to evaluate community predictions correctly?**

Our results show that the "single species cross-validation" approach (SSCV), the most commonly used in the literature to evaluate community predictions (e.g., Dubuis *et al.* 2011; Calabrese *et al.* 2014; Distler *et al.* 2015), yields overoptimistic and thus not fully realistic measures of predictive power. While this approach is usually able to provide satisfying evaluation for single species, as revealed by the cross-validation of individual species runs, it shows a clear degradation of predictions when measured at the level of communities. This occurs likely because "all" sites are used at least once at some stage across all modelling runs of the split-sampling procedure, and thus no observation (or very few in the best cases) remains fully independent (i.e. unused) for the final evaluation at the community level. Additionally, the sets of training sites used at each run differ among the species, making the results not entirely comparable across species.

425 The second approach found in the literature builds on the first one (SSCV; thus including an
426 internal cross-validation evaluation), but uses spatially or temporally independent data (ID)
427 for the assessment (thus an external evaluation), thus (unlike SSCV) using the same set of
428 evaluation sites for all species (e.g., Benito, Cayuela & Albuquerque 2013; Pottier *et al.* 2013;
429 Cord *et al.* 2014). When such independent data are available, this method provides the best
430 possible evaluation, provided that the evaluation data are representative of the area where the
431 models apply. This approach – with both internal and external evaluation - is also the one
432 considered as optimal in James *et al.* (2013), and recently promoted in the field of SDMs by
433 Guisan, Thuiller and Zimmermann (2017).

434 The third approach (CVV), newly presented here, repeats the ID approach a large number of
435 times within a cross-validation procedure at the community-level (no example of this
436 approach known in the literature). By doing this, the risk of bias in the evaluation data,
437 inherent to the selection of a single evaluation data set, is minimized compared to the simple
438 ID approach. Additionally, the repeated cross-validation allows the assessments of uncertainty
439 and confidence intervals around the community predictions' performance metrics. However,
440 as this approach selects the same sites for all species, its application is only possible under
441 specific circumstances. First, all the species data need to be collected in the same sites (i.e.
442 true 'community data'). Second, as this approach leads to an unequal number of
443 presences/absences between different cross-validation runs for the same species, it can lead to
444 models failing for very rare (low sample size) species in some of the cross-validation runs if
445 not enough presence sites are selected in the training set.

446 According to our results and despite the potential limitations we advise the use of the
447 proposed community cross-validation approach (CCV) to evaluate community models in
448 future studies. In fact, we clearly showed that the common practice of evaluating the
449 community predictions on the same dataset used for calibration process (SSCV) leads to

overoptimistic estimations of model performance. In the commonest case of unavailability of truly spatial (i.e., different region) or temporal (i.e., different sampling period) independent data, often independent datasets are “created” by randomly splitting the initial dataset in two parts. However, we advocate against this practise and instead promote the community cross-validation approach, which minimizes the artefacts of randomly splitting the initial data and allows the estimation of uncertainty associated with the community evaluation metrics.

## Acknowledgements

This study was supported by the Swiss national Science Foundation (SESAM'ALP project, grant nr 31003A-1528661) to AG and by the European Commission, Marie Skłodowska-Curie Research Fellowship Programme (SESAM-ZOO project) to MDA and AG. R.G.M. was funded by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme (ACONITE, PIEF-GA-2013-622620). 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.

## Authors' contributions

DS and AG conceived the ideas; RF and MD analysed the plant and insect data; DS and RGM developed the modelling framework; DS led the writing and all authors contributed critically to the drafts and gave final approval for publication.

## Data Accessibility

A generalised version of the community cross-validation algorithm is available in the ecospat R package (Cola *et al.* 2016) on GitHub (ecospat.CCV; https://doi.org/10.5281/zenodo.1287805). All species and environmental data are available on Dryad: https://doi.org/10.5061/dryad.28d4k (Grassland species and environmental predictors for plants; Guisan, Dubuis & Vittoz 2011) and https://doi.org/10.5061/dryad.nf925ps (forest, insect species and environmental predictors for insects; Guisan *et al.* 2018).

## REFERENCES

Alexander, J.M., Diez, J.M., Hart, S.P. & Levine, J.M. (2016) When climate reshuffles competitors: a call for experimental macroecology. *Trends in Ecology & Evolution,* **31,** 831-841. doi: 10.1016/j.tree.2016.08.003

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,** 1223-1232. doi: 10.1111/j.1365-2664.2006.01214.x

Begon, M., Harper, J.L. & Townsend, C.R. (1996) *Ecology: individuals populations and communities,* Third edition edn. Blackwell Science Inc, Oxford.

Benito, B.M., Cayuela, L. & Albuquerque, F.S. (2013) The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversitymodels. *Methods in Ecology and Evolution,* **4,** 327-335. doi: 10.1111/2041-210x.12022

Blois, J.L., Zarnetske, P.L., Fitzpatrick, M.C. & Finnegan, S. (2013) Climate Change and the Past, Present, and Future of Biotic Interactions. *Science,* **341,** 499-504. doi: 10.1126/science.1237184

Breiman, L. (2001) Random forests. *Machine Learning,* **45,** 5-32. doi: 10.1023/A:1010933404324

Breiner, F.T., Guisan, A., Bergamini, A. & Nobis, M.P. (2015) Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution,* **6,** 1210-1218. doi: 10.1111/2041-210x.12403

Calabrese, J.M., Certain, G., Kraan, C. & Dormann, C.F. (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography,* **23,** 99-112. doi: 10.1111/Geb.12102

Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *American Psycological Association,* **70,** 213-220.

Cola, V.D., Broennimann, O., Petitpierre, B., Breiner, F.T., D'Amen, M., Randin, C., . . . Dubuis, A. (2016) ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography*.

Cord, A.F., Klein, D., Gernandt, D.S., la Rosa, J.A.P. & Dech, S. (2014) Remote sensing data can improve predictions of species richness by stacked species distribution models: a case study for Mexican pines. *Journal of Biogeography,* **41,** 736-748. doi: 10.1111/jbi.12225

Currie, D.J., Mittelbach, G.G., Cornell, H.V., Field, R., Guégan, J.F., Hawkins, B.A., . . . O'Brien, E. (2004) Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters,* **7,** 1121-1134.

D'Amen, M., Bombi, P., Pearman, P.B., Schmatz, D.R., Zimmermann, N.E. & Bologna, M.A. (2011) Will climate change reduce the efficacy of protected areas for amphibian conservation in Italy? *Biological Conservation*, **144,** 989-997. doi: 10.1016/j.biocon.2010.11.004

D'Amen, M., Dubuis, A., Fernandes, R.F., Pottier, J., Pellissier, L. & Guisan, A. (2015) Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, **42,** 1255-1266. doi: 10.1111/jbi.12485

D'Amen, M., Mateo, R.G., Pottier, J., Thuiller, W., Maiorano, L., Pellissier, L., . . . Guisan, A. (2017a) Improving spatial predictions of taxonomic, functional and phylogenetic diversity. *Journal of Ecology*.

D'Amen, M., Pradervand, J.N. & Guisan, A. (2015) Predicting richness and composition in mountain insect communities at high resolution: a new test of the SESAM framework. *Global Ecology and Biogeography*, **24,** 1443-1453. doi: 10.1111/geb.12357

D'Amen, M., Rahbek, C., Zimmermann, N.E. & Guisan, A. (2017b) Spatial predictions at the community level: from current approaches to future frameworks. *Biological Reviews*, **92,** 169–187. 10.1111/brv.12222

Distler, T., G., S.J., Velasquez-Tibata, J. & Langham, G.M. (2015) Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change. *Journal of Biogeography*, **42,** 1-13. doi: 10.1111/jbi.12479

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., . . . Lautenbach, S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36,** 27-46. doi: 10.1111/j.1600-0587.2012.07348.x

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. *Diversity and Distributions*, **17,** 1122-1131. doi: 10.1111/j.1472-4642.2011.00792.x

Dullinger, S., Gattringer, A., Thuiller, W., Moser, D., Zimmermann, N.E., Guisan, A., . . . Hulber, K. (2012) Extinction debt of high-mountain plants under twenty-first-century climate change. *Nature Climate Change*, **2,** 619-622. doi: 10.1038/Nclimate1514

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., . . . Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29,** 129-151. doi: 10.1111/j.2006.0906-7590.04596.x

536 Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across
537 Space and Time. *Annual Review of Ecology Evolution and Systematics,* **40,** 677-697. doi:
538 10.1146/annurev.ecolsys.110308.120159
539 Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal*
540 *Ecology,* **77,** 802-813. doi: 10.1111/j.1365-2656.2008.01390.x
541 Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial
542 pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and*
543 *Conservation,* **11,** 2309-2338. doi: 10.1023/A:1021374009951
544 Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied*
545 *Ecology,* **43,** 393-404. doi: 10.1111/j.1365-2664.2006.01149.x
546 Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation
547 presence-absence models. *Environmental Conservation,* **24,** 38-49. doi: 10.1017/S0376892997000088
548 Fleishman, E., Noss, R.F. & Noon, B.R. (2006) Utility and limitations of species richness metrics for
549 conservation planning. *Ecological Indicators,* **6,** 543-553. doi: 10.1016/j.ecolind.2005.07.005
550 Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary
551 classification in terms of predicted prevalence and kappa. *Ecological Modelling,* **217,** 48-58. doi:
552 10.1016/j.ecolmodel.2008.05.015
553 Gotelli, N.J., Anderson, M.J., Arita, H.T., Chao, A., Colwell, R.K., Currie, D.J., . . . Willig, M.R. (2009) Patterns
554 and causes of species richness: a general simulation model for macroecology. *Ecology Letters,* **12,** 873–
555 886. doi: 10.1111/j.1461-0248.2009.01353.x
556 Granger, V., Bez, N., Fromentin, J.M., Meynard, C., Jadaud, A. & Merigot, B. (2015) Mapping diversity indices:
557 not a trivial issue. *Methods in Ecology and Evolution,* **6,** 688-696.
558 Guisan , A., Dubuis, A., Pellisser, L., Pradervand, J.N., Meier, S., Scherrer, D., . . . Vittoz, P. (2018) Data from:
559 How to best threshold and validate stacked species assemblages? Community optimisation might hold
560 the answer. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.nf925ps
561 Guisan, A., Dubuis, A. & Vittoz, P. (2011) Data from: Predicting spatial patterns of plant species richness:
562 acomparison of direct macroecological and species stacking modelling appraoches. *Dryad Digital*
563 *Repository*. https://doi.org/10.5061/dryad.28d4k
564 Guisan, A., Lehmann, A., Ferrier, S., Aspinall, R., Overton, R., Austin, M. & Hastie, T. (2006) Making better
565 biogeographic predictions of species distribution. *Journal of Applied Ecology,* **43,** 386-392.

Guisan, A. & Rahbek, C. (2011) SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography,* **38,** 1433-1444. doi: 10.1111/j.1365-2699.2011.02550.x

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

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017) *Habitat suitability and distribution models*. Cambridge University Press.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., . . . Buckley, Y.M. (2013) Predicting species distributions for conservation decisions. *Ecology Letters,* **16,** 1424-1435. doi: 10.1111/Ele.12189

Hartmann, P., Fouvy, P. & Horisberger, D. (2009) L'Observatoire de l'écosystème forestier du canton de Vaud: espace de recherche appliquée| The Forest Ecosystem Observatory in Canton Vaud: a field of applied research. *Schweizerische Zeitschrift fur Forstwesen,* **160,** s2-s6.

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

Hawkins, B.A. & Porter, E.E. (2003) Water–energy balance and the geographic pattern of species richness of western Palearctic butterflies. *Ecological Entomology,* **28,** 678-686. doi: 10.1111/j.1365-2311.2003.00551.x

Hespanhol, H., Cezón, K., Felicísimo, Á.M., Muñoz, J. & Mateo, R.G. (2015) How to describe species richness patterns for bryophyte conservation? *Ecology and evolution,* **5,** 5443-5455.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning*. Springer.

Jimenez-Valverde, A. & Lobo, J.M. (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica-International Journal of Ecology,* **31,** 361-369. doi: 10.1016/j.actao.2007.02.001

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters,* **12,** 334-350. doi: 10.1111/j.1461-0248.2008.01277.x

Liu, C., White, M. & Newell, G. (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography,* **40,** 778-789.

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. doi: 10.1111/j.0906-7590.2005.03957.x

595 Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J. & Guisan, A. (2010) Overcoming the
596 rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant.
597 *Biological Conservation,* **143,** 2647-2657. doi: 10.1016/j.biocon.2010.07.007
598 Maggini, R., Lehmann, A., Zimmermann, N.E. & Guisan, A. (2006) Improving generalized regression analysis
599 for the spatial prediction of forest communities. *Journal of Biogeography,* **33,** 1729-1749. doi:
600 10.1111/j.1365-2699.2006.01465.x
601 Magurran, A.E. & Henderson, P.A. (2003) Explaining the excess of rare species in natural species abundance
602 distributions. *Nature,* **422,** 714-716.
603 Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence-absence models in ecology: the need to
604 account for prevalence. *Journal of Applied Ecology,* **38,** 921-931. doi: 10.1046/j.1365-
605 2664.2001.00647.x
606 Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009) Evaluation of consensus
607 methods in predictive species distribution modelling. *Diversity and Distributions,* **15,** 59-69. doi:
608 10.1111/j.1472-4642.2008.00491.x
609 Mateo, R.G., de la Estrella, M., Felicisimo, A.M., Munoz, J. & Guisan, A. (2013) A new spin on a
610 compositionalist predictive modelling framework for conservation planning: A tropical case study in
611 Ecuador. *Biological Conservation,* **160,** 150-161. doi: 10.1016/j.biocon.2013.01.014
612 Mateo, R.G., Felicisimo, A.M., Pottier, J., Guisan, A. & Munoz, J. (2012) Do Stacked Species Distribution
613 Models Reflect Altitudinal Diversity Patterns? *PLoS ONE,* **7,** e32586. doi:
614 10.1371/journal.pone.0032586
615 Mateo, R.G., Mokany, K. & Guisan, A. (2017) Biodiversity Models: What If Unsaturation Is the Rule? *Trends in*
616 *Ecology & Evolution,* **32,** 556-566. doi: http://dx.doi.org/10.1016/j.tree.2017.05.003
617 McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models. 2nd edition*. Chapman and Hall, London.
618 McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of
619 distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology,* **41,** 811-
620 823. doi: 10.1111/j.0021-8901.2004.00943.x
621 Meier, E.S., Kienast, F., Pearman, P.B., Svenning, J.C., Thuiller, W., Araujo, M.B., . . . Zimmermann, N.E.
622 (2010) Biotic and abiotic variables show little redundancy in explaining tree species distributions.
623 *Ecography,* **33,** 1038-1048. doi: 10.1111/j.1600-0587.2010.06229.x

624 Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., . . . Elith, J. (2014) What do
625 we gain from simplicity versus complexity in species distribution models? *Ecography*, **37,** 1267-1281.
626 doi: 10.1111/ecog.00845
627 Mod, H.K., le Roux, P.C., Guisan, A. & Luoto, M. (2015) Biotic interactions boost spatial models of species
628 richness. *Ecography*, **38,** 913-921. doi: 10.1111/ecog.01129
629 Mokany, K. & Ferrier, S. (2011) Predicting impacts of climate change on biodiversity: a role for semi-
630 mechanistic community-level modelling. *Diversity and Distributions*, **17,** 374-380. doi: 10.1111/j.1472-
631 4642.2010.00735.x
632 Nenzen, H.K. & Araujo, M.B. (2011) Choice of threshold alters projections of species range shifts under climate
633 change. *Ecological Modelling*, **222,** 3346-3354. doi: 10.1016/j.ecolmodel.2011.07.011
634 Nogues-Bravo, D. & Rahbek, C. (2011) Communities Under Climate Change. *Science*, **334,** 1070-1071. doi:
635 10.1126/science.1214833
636 Parmesan, C. & Yohe, G. (2003) A globally coherent fingerprint of climate change impacts across natural
637 systems. *Nature*, **421,** 37-42. doi: 10.1038/nature01286
638 Pellissier, L., Pradervand, J.-N., Pottier, J., Dubuis, A., Maiorano, L. & Guisan, A. (2012) Climate-based
639 empirical models show biased predictions of butterfly communities along environmental gradients.
640 *Ecography*, **35,** 684-692. doi: 10.1111/j.1600-0587.2011.07047.x
641 Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., . . . Wegmann, M. (2013)
642 Essential Biodiversity Variables. *Science*, **339,** 277-278. doi: 10.1126/science.1229931
643 Pineda, E. & Lobo, J.M. (2009) Assessing the accuracy of species distribution models to predict amphibian
644 species richness patterns. *Journal of Animal Ecology*, **78,** 182-190. doi: 10.1111/j.1365-
645 2656.2008.01471.x
646 Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C.F., . . . Guisan, A. (2013) The accuracy
647 of plant assemblage prediction from species distribution models varies along environmental gradients.
648 *Global Ecology and Biogeography*, **22,** 52-63. doi: 10.1111/j.1466-8238.2012.00790.x
649 Pouteau, R., Bayle, E., Blanchard, E., Birnbaum, P., Cassan, J.J., Hequet, V., . . . Vandrot, H. (2015) Accounting
650 for the indirect area effect in stacked species distribution models to map species richness in a montane
651 biodiversity hotspot. *Diversity and Distributions*, **21,** 1329-1338. doi: 10.1111/ddi.12374

Pradervand, J.N., Dubuis, A., Reymond, A., Sonnay, V., Gelin, A. & Guisan, A. (2013) Quels facteurs influencent la richesse en orthoptères des Préalpes vaudoises? *Bulletin de la Société Vaudoises des Sciences Naturelles,* **93,** 155-173.

Preston, F.W. (1948) The Commonness, and Rarity, of Species. *Ecology,* **29,** 254-283.

Scherrer, D., Massy, S., Meier, S., Vittoz, P. & Guisan, A. (2017) Assessing and predicting shifts in mountain forest composition across 25 years of climate change. *Diversity and Distributions,* **23,** 517-528.

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.

Thuiller, W., Georges, D. & Engler, R. (2013) biomod2: Ensemble platform for species distribution modeling. *R package version,* **2,** r560.

Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the United States of America,* **102,** 8245-8250. doi: 10.1073/pnas.0409902102

Thuiller, W., Pollock, L.J., Gueguen, M. & Münkemüller, T. (2015) From species distributions to meta-communities. *Ecology Letters,* **18,** 1321-1328. doi: 10.1111/ele.12526

Turner, J.R., Gatehouse, C.M. & Corey, C.A. (1987) Does solar energy control organic diversity? Butterflies, moths and the British climate. *Oikos,* 195-205.

Van der Putten, W.H., Macel, M. & Visser, M.E. (2010) Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. *Philosophical Transactions of the Royal Society B-Biological Sciences,* **365,** 2025-2034. doi: 10.1098/rstb.2010.0037

Zimmermann, N.E., Edwards, T.C., Graham, C.H., Pearman, P.B. & Svenning, J.-C. (2010) New trends in species distribution modelling. *Ecography,* **33,** 985-989. doi: 10.1111/j.1600-0587.2010.06953.x

Zurell, D., Zimmermann, N.E., Sattler, T., Nobis, M.P. & Schröder, B. (2016) Effects of functional traits on the prediction accuracy of species richness models. *Diversity and Distributions.*

## Figure legends

**Figure 1:** Map of the study area with the forested sites (dark green triangles, N=3076), the grassland sites (light green circles and red squares, N=903) and the insect sites (red squares, butterflies N=192, grasshoppers N=202).

**Figure 2:** The modelling framework illustrating the three different community modelling approaches: “single species cross-validation” (SSCV), “independent data” (ID) and “community cross-validation” (CCV).

**Figure 3:** Deviation in site specific species richness between observations and predictions for the four different datasets (top to bottom) and the three different modelling pathways (left to right). The boxplots are sorted by the median and the colours indicate the different thresholding techniques used to binarise predictions. The line in the box indicates the median, boxes range from the $25^{th}$ to the $75^{th}$ percentile and the whiskers indicate ± 2 standard deviations. Letters above the boxplots indicate significant differences (Wilcoxon rank sum test, $p < 0.05$).

**Figure 4:** Sørensen similarity between observations and predictions for the four different datasets (top to bottom) and the three different modelling pathways (left to right). The boxplots are sorted by the median and the colours indicate the different thresholding techniques. The line in the box indicates the median, boxes range from the $25^{th}$ to the $75^{th}$ percentile and the whiskers indicate ± 2 standard deviations. Letters above the boxplots indicate significant differences (Wilcoxon rank sum test, $p < 0.05$).

**Figure 1**

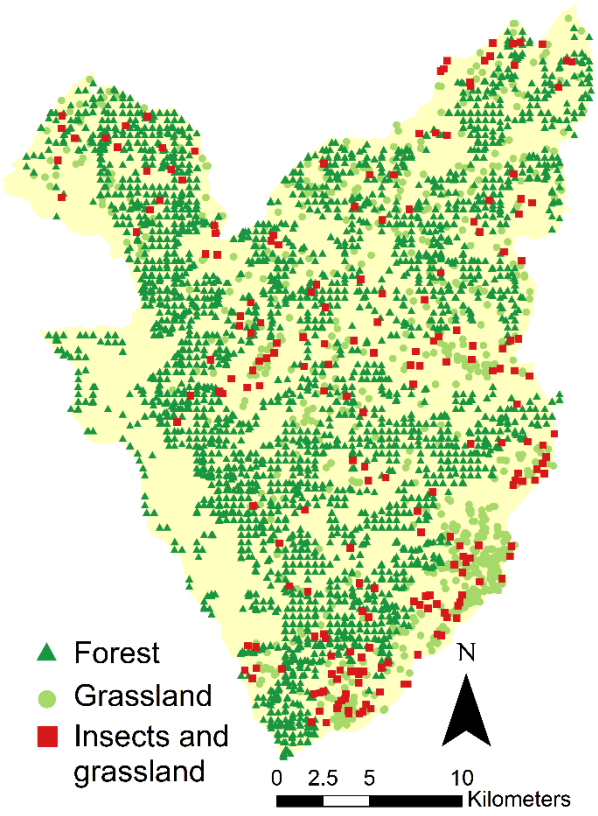


**Figure 1**: Map of the study area with the forested sites (dark green triangles, N=3076), the grassland sites (light green circles and red squares, N=903) and the insect sites (red squares, butterflies N=192, grasshoppers N=202).

**Figure 2**

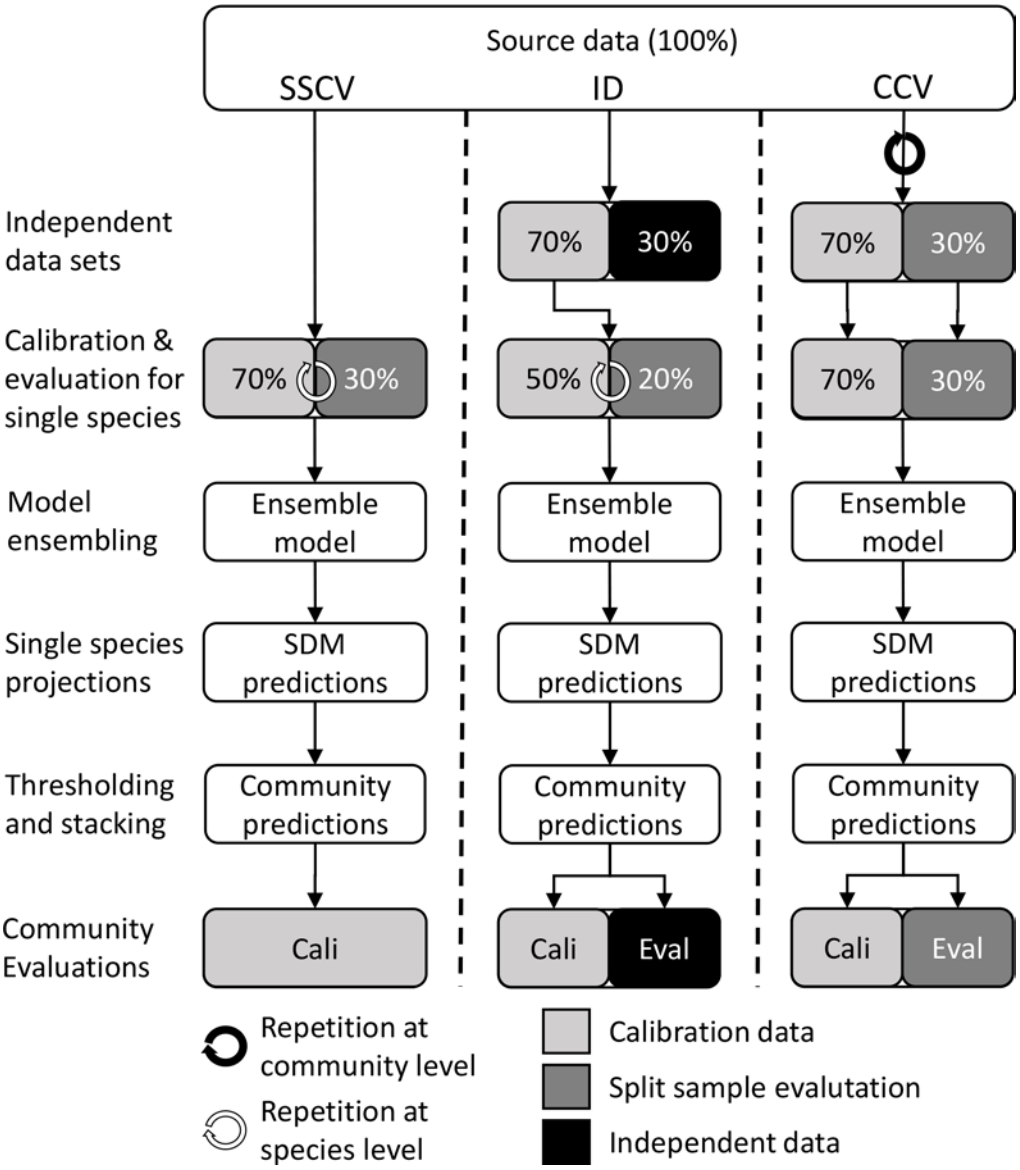


**Figure 2:** The modelling framework illustrating the three different community modelling approaches: "single species cross-validation" (SSCV), "independent data" (ID) and "community cross-validation" (CCV).

707 **Figure 3**

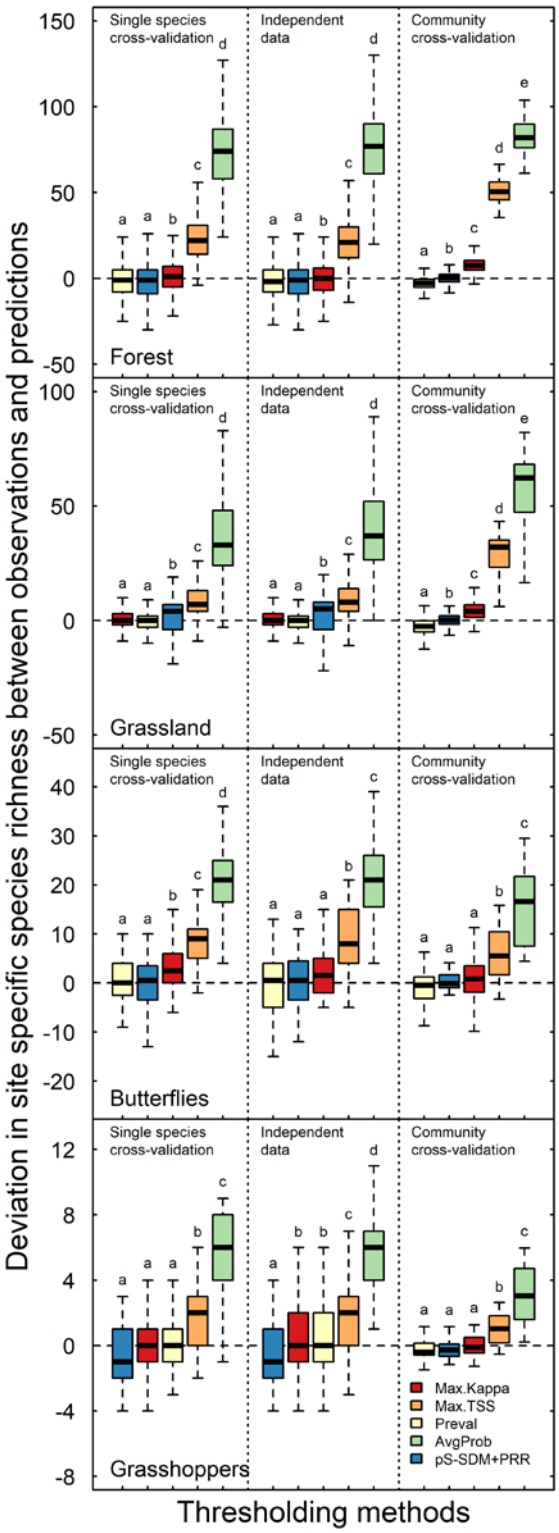


**Figure 3:** Deviation in site specific species richness between observations and predictions for the four different datasets (top to bottom) and the three different modelling pathways (left to right). The boxplots are sorted by the median and the colours indicate the different thresholding techniques used to binarise predictions. The line in the box indicates the median, boxes range from the 25$^{th}$ to the 75$^{th}$ percentile and the whiskers indicate ± 2 standard deviations. Letters above the boxplots indicate significant differences (Wilcoxon rank sum test, $p < 0.05$).

**Figure 4**

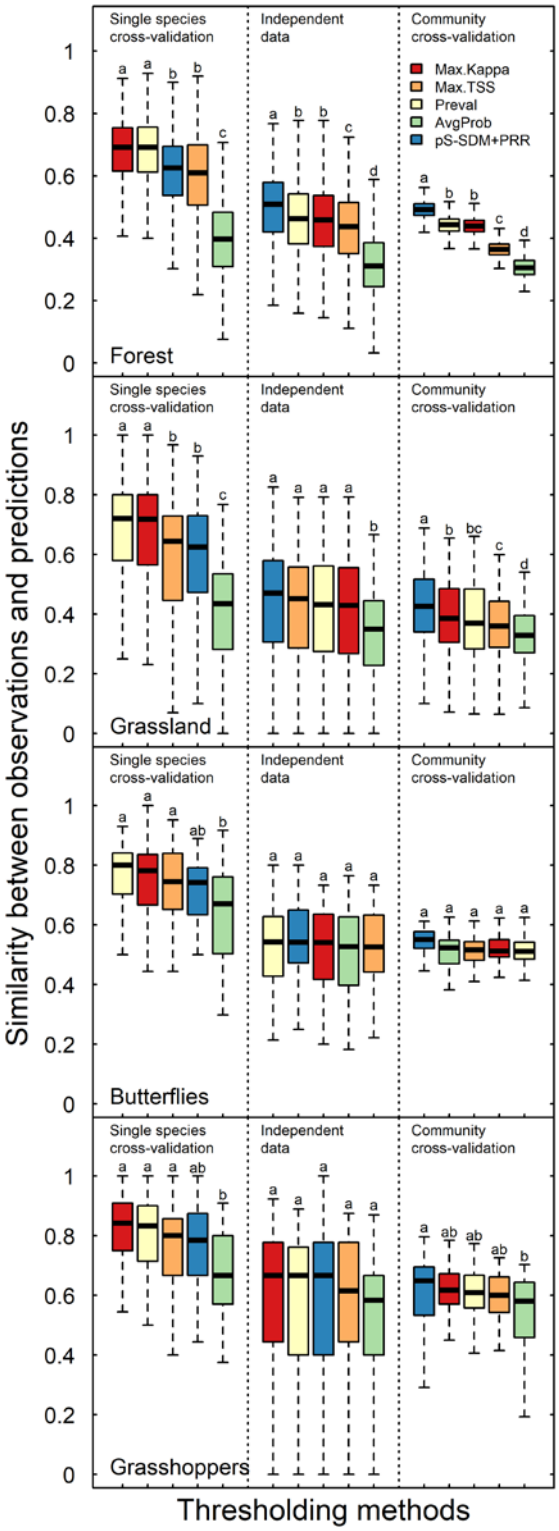


**Figure 4:** Sørensen similarity between observations and predictions for the four different datasets (top to bottom) and the three different modelling pathways (left to right). The boxplots are sorted by the median and the colours indicate the different thresholding techniques. The line in the box indicates the median, boxes range from the 25$^{th}$ to the 75$^{th}$ percentile and the whiskers indicate ± 2 standard deviations. Letters above the boxplots indicate significant differences (Wilcoxon rank sum test, $p < 0.05$).

**Table 1:** Basic statistics of the data sets used for the case study and the evaluation metrics (AUC) for the individual species distribution models using the three different community evaluation approaches. SSCV = Single species cross-validation, ID = Independent data, CCV = Community cross-validation

| Data set | Number of species modelled (recorded) | Prevalence (mean ± sd) | Species richness (mean ± sd) | AUC SSCV (mean ± sd) | AUC ID (mean ± sd) | AUC CCV (mean ± sd) |
|---|---|---|---|---|---|---|
| Forest | 312 (703) | 0.044 ± 0.090 | 29.5 ± 11.8 | 0.80 ± 0.09 | 0.80 ± 0.08 | 0.79 ± 0.09 |
| Grassland | 212 (905) | 0.098 ± 0.089 | 23.5 ± 13.8 | 0.82 ± 0.07 | 0.83 ± 0.06 | 0.81 ± 0.06 |
| Butterflies | 77 (131) | 0.235 ± 0.137 | 18.1 ± 9.2 | 0.76 ± 0.10 | 0.75 ± 0.12 | 0.76 ± 0.10 |
| Grasshoppers | 20 (41) | 0.256 ± 0.193 | 5.1 ± 3.3 | 0.84 ± 0.07 | 0.86 ± 0.08 | 0.84 ± 0.06 |

**Table 2:** Pearson Correlation of single species and community evaluation statistics. The asterisks indicate the significance level. Correlations of the single species evaluation metrics and their corresponding community evaluation metric are highlighted in bold.

| | **Community metrics** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Single species** | **Accuracy** | **Sensitivity** | **Specificity** | **KAPPA** | **TSS** | **Sørensen similarity** | **SR deviation** |
| **Accuracy** | **1.00 ***** | -0.37 * | 0.95 *** | 0.70 *** | 0.37 * | 0.37 * | -0.58 *** |
| **Sensitivity** | -0.36 ** | **0.93 ***** | -0.54 *** | 0.01 n.s. | 0.56 *** | 0.18 n.s. | -0.44 *** |
| **Specificity** | 0.97 *** | -0.53 *** | **0.99 ***** | 0.64 *** | 0.20 n.s. | 0.31 * | -0.63 *** |
| **KAPPA** | 0.41 ** | 0.50 * | 0.27 * | **0.79 ***** | 0.72 *** | 0.82 *** | -0.3 * |
| **TSS** | 0.06 n.s. | 0.85 *** | -0.14 n.s. | 0.35 n.s. | **0.79 ***** | 0.38 ** | -0.20 n.s. |

The asterisks indicate the significance level (n.s.= not significant, * $p<0.05$, ** $p<0.01$, *** $p<0.001$)

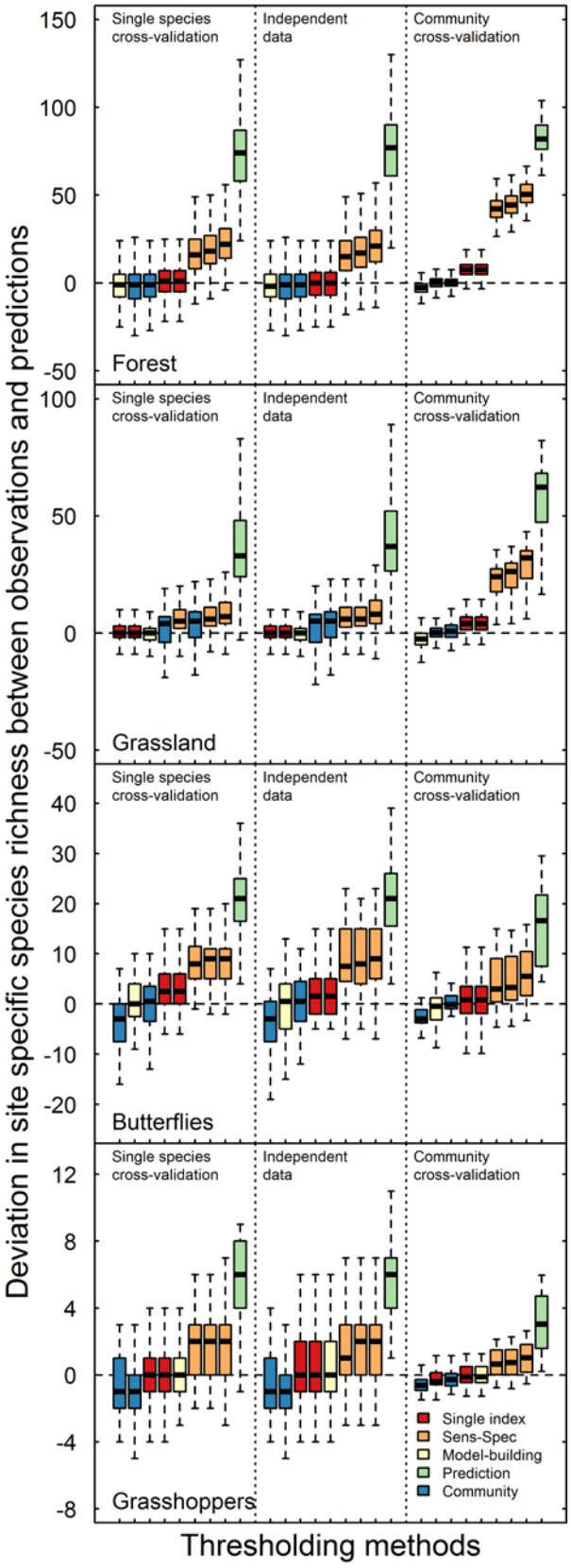


**Figure S1:** Deviation in site specific species richness between observations and predictions for the four different datasets (top to bottom) and the three different modelling pathways (left to right). The boxplots are sorted by the median and the colours indicate the different thresholding techniques. The line in the box indicates the median, boxes range from the 25$^{th}$ to the 75$^{th}$ percentile and the whiskers indicate ± 2 standard deviations. For details on the method used within each threshold group see Table S1.

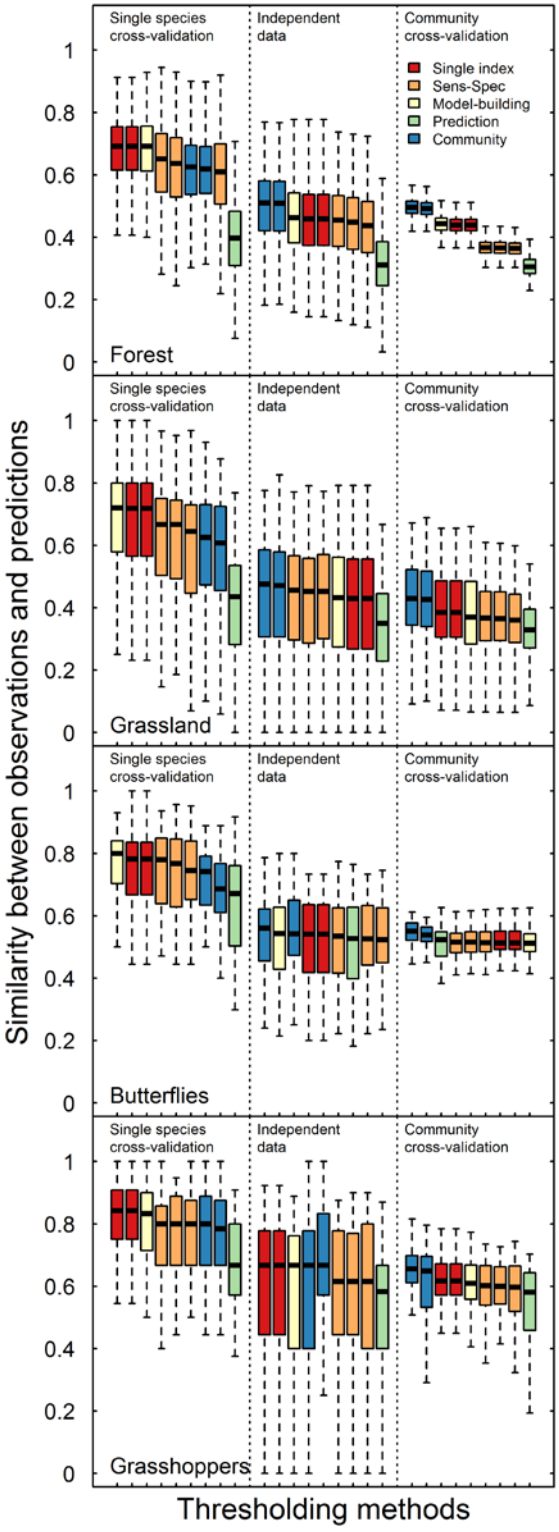


**Figure S2:** Sørensen similarity between observations and predictions for the four different datasets (top to bottom) and the three different modelling pathways (left to right). The boxplots are sorted by the median and the colours indicate the different thresholding techniques. The line in the box indicates the median, boxes range from the 25$^{th}$ to the 75$^{th}$ percentile and the whiskers indicate ± 2 standard deviations. For details on the method used within each threshold group see Table S1.

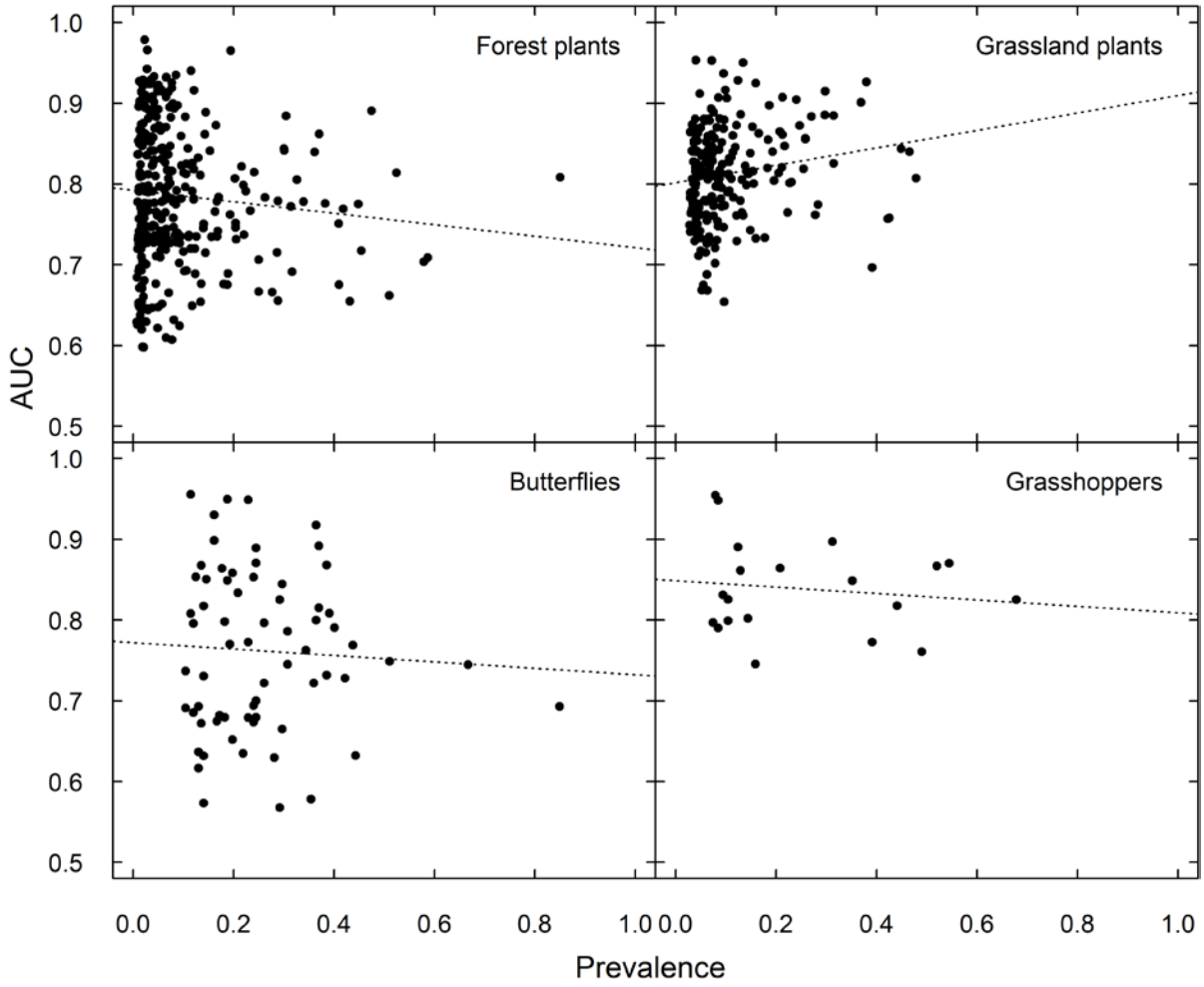


**Figure S3:** The relationship of the prevalence of a species (i.e., percentage of sites inhabited) to the performance of the SDMs (i.e., as measured by AUC) for the four studied data sets (taxa).

**Table S1:** Description of the ten thresholding methods based on Liu *et al.* (2005) and Nenzen and Araujo (2011).

| Approach | Accronym | Definition | Reference |
|---|---|---|---|
| **Single index-based approaches** | | | |
| 1. Kappa maximization approach | Max.Kappa | Kappa statistic is maximized | (Huntley *et al.* 1995; Guisan, Theurillat & Kienast 1998) |
| 2. Maximum commission error | MCE05 | Allowed a maximum commission error of 5% | (Mateo et al. 2012) |
| **Sensitivity and specificity-combined approaches** | | | |
| 3. TSS maximization approach | Max.TSS | TSS statistic is maximized | (Allouche, Tsoar & Kadmon 2006) |
| 4. Sensitivity-specificity equality approach | SensSpec | Difference of sens-spec is minimized | (Cantor *et al.* 1999) |
| 5. ROC plot-based approach | Opt.ROC | ROC statistic is maximized | (Cantor *et al.* 1999) |
| **Model-building data-only-based approach** | | | |
| 6. Prevalence approach | Preval | Prevalence of the calibration data | (Cramer 2003) |
| **Predicted probability-based approaches** | | | |
| 7. Average probability approach | AvgProb | Taking the average predicted probability of the model-building data as threshold | (Cramer 2003) |
| **Community based approaches** | | | |
| 8. pS-SDM+PRR | pS-SDM+PRR | Probability stacked SDM | (Dubuis *et al.* 2013) |
| 9. MEM+PRR | MEM+PRR | Macroecological model for SR | (Guisan & Rahbek 2011) |

**Table S2 :** Community evaluation metrics used in this study.

| Metric | Definition |
|---|---|
| **Species richness** | |
| Deviation in species richness | $Dev.SPR = n_{pred} - n_{obs}$ |
| **Prediction success** | |
| Sensitivity | $Sens = \frac{TP}{TP + FA}$ |
| Specificity | $Spec = \frac{TA}{TA + FP}$ |
| Community accuracy | $Acc = \frac{TP + TA}{N}$ |
| Community TSS | $TSS = Sens + Spec - 1$ |
| Community Kappa | $K = \frac{Acc - p_e}{1 - p_e}$ |
| **Community composition** | |
| Sørensen | $S = \frac{2 * TP}{2 * TP + FP + FA}$ |

$n_{pred}$ =Number of species predicted
$n_{obs}$ = Number of species observed
N = Number of events
TP = Correctly predicted present species
TA = Correctly predicted absent species
FP = Falsely predicted present species
FA = Falsely predicted absent species
$p_e = \frac{(TP+FA)(TP+FP)+(TA+FP)(TA+FA)}{N^2}$

**Table S3 :** Evaluation scores of individual SDMs by TSS (A), Kappa (B), PCC (C), Sensitivity (D) and Specificity (E) for the three community evaluation approaches and four datasets. SSCV = Single species cross-validation, ID = Independent data, CCV = Community cross-validation, FO = Forest plants, GL = Grassland plants, BF = Butterflies, GH = Grasshoppers.

(A) TSS

| Thresholding | SSCV | | | | ID | | | | CCV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Approach | FO | GL | BF | GH | FO | GL | BF | GH | FO | GL | BF | GH |
| Max.Kappa | 0.2 ± 0.14 | 0.27 ± 0.2 | 0.31 ± 0.23 | 0.42 ± 0.17 | 0.21 ± 0.17 | 0.25 ± 0.21 | 0.3 ± 0.22 | 0.43 ± 0.26 | 0.23 ± 0.13 | 0.28 ± 0.14 | 0.31 ± 0.18 | 0.37 ± 0.15 |
| MCE05 | 0.3 ± 0.17 | 0.32 ± 0.17 | 0.27 ± 0.18 | 0.43 ± 0.16 | 0.28 ± 0.17 | 0.34 ± 0.17 | 0.27 ± 0.23 | 0.45 ± 0.21 | 0.25 ± 0.17 | 0.34 ± 0.12 | 0.31 ± 0.17 | 0.42 ± 0.12 |
| Max.TSS | 0.35 ± 0.2 | 0.38 ± 0.14 | 0.34 ± 0.24 | 0.47 ± 0.12 | 0.33 ± 0.22 | 0.38 ± 0.21 | 0.34 ± 0.23 | 0.5 ± 0.26 | 0.35 ± 0.14 | 0.39 ± 0.11 | 0.35 ± 0.18 | 0.44 ± 0.12 |
| SensSpec | 0.32 ± 0.16 | 0.36 ± 0.14 | 0.33 ± 0.18 | 0.51 ± 0.18 | 0.31 ± 0.2 | 0.37 ± 0.21 | 0.34 ± 0.24 | 0.51 ± 0.26 | 0.35 ± 0.13 | 0.38 ± 0.11 | 0.35 ± 0.18 | 0.45 ± 0.11 |
| Opt.ROC | 0.34 ± 0.21 | 0.36 ± 0.19 | 0.33 ± 0.17 | 0.44 ± 0.26 | 0.32 ± 0.21 | 0.37 ± 0.21 | 0.34 ± 0.23 | 0.49 ± 0.25 | 0.35 ± 0.13 | 0.38 ± 0.11 | 0.35 ± 0.18 | 0.44 ± 0.12 |
| Preval | 0.18 ± 0.15 | 0.27 ± 0.15 | 0.3 ± 0.19 | 0.41 ± 0.23 | 0.2 ± 0.16 | 0.26 ± 0.2 | 0.31 ± 0.22 | 0.4 ± 0.23 | 0.18 ± 0.14 | 0.23 ± 0.15 | 0.3 ± 0.17 | 0.37 ± 0.17 |
| AvgProb | 0.43 ± 0.16 | 0.5 ± 0.12 | 0.41 ± 0.21 | 0.55 ± 0.14 | 0.47 ± 0.16 | 0.53 ± 0.13 | 0.38 ± 0.23 | 0.56 ± 0.15 | 0.44 ± 0.16 | 0.49 ± 0.11 | 0.4 ± 0.18 | 0.54 ± 0.14 |
| pS-SDM+PRR | 0.12 ± 0.19 | 0.17 ± 0.24 | 0.28 ± 0.24 | 0.28 ± 0.22 | 0.14 ± 0.18 | 0.2 ± 0.24 | 0.24 ± 0.24 | 0.29 ± 0.28 | 0.14 ± 0.17 | 0.19 ± 0.21 | 0.27 ± 0.23 | 0.27 ± 0.22 |
| MEM+PRR | 0.16 ± 0.17 | 0.2 ± 0.23 | 0.25 ± 0.24 | 0.3 ± 0.24 | 0.14 ± 0.18 | 0.2 ± 0.24 | 0.22 ± 0.24 | 0.32 ± 0.28 | 0.14 ± 0.17 | 0.2 ± 0.22 | 0.25 ± 0.22 | 0.3 ± 0.22 |

(B) KAPPA

| Thresholding | SSCV | | | | ID | | | | CCV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Approach | FO | GL | BF | GH | FO | GL | BF | GH | FO | GL | BF | GH |
| Max.Kappa | 0.2 ± 0.14 | 0.24 ± 0.18 | 0.28 ± 0.22 | 0.35 ± 0.19 | 0.21 ± 0.15 | 0.24 ± 0.19 | 0.29 ± 0.22 | 0.42 ± 0.25 | 0.2 ± 0.14 | 0.24 ± 0.15 | 0.29 ± 0.18 | 0.36 ± 0.15 |
| MCE05 | 0.21 ± 0.13 | 0.23 ± 0.15 | 0.24 ± 0.22 | 0.31 ± 0.16 | 0.21 ± 0.13 | 0.28 ± 0.15 | 0.27 ± 0.22 | 0.41 ± 0.18 | 0.11 ± 0.12 | 0.16 ± 0.14 | 0.24 ± 0.17 | 0.32 ± 0.16 |
| Max.TSS | 0.19 ± 0.12 | 0.27 ± 0.17 | 0.3 ± 0.19 | 0.4 ± 0.14 | 0.21 ± 0.14 | 0.27 ± 0.16 | 0.3 ± 0.21 | 0.41 ± 0.23 | 0.17 ± 0.13 | 0.22 ± 0.15 | 0.3 ± 0.17 | 0.36 ± 0.15 |
| SensSpec | 0.21 ± 0.14 | 0.21 ± 0.17 | 0.29 ± 0.19 | 0.41 ± 0.17 | 0.22 ± 0.14 | 0.28 ± 0.17 | 0.3 ± 0.21 | 0.42 ± 0.23 | 0.17 ± 0.13 | 0.23 ± 0.15 | 0.3 ± 0.17 | 0.37 ± 0.14 |
| Opt.ROC | 0.15 ± 0.15 | 0.22 ± 0.14 | 0.32 ± 0.18 | 0.43 ± 0.17 | 0.22 ± 0.14 | 0.27 ± 0.16 | 0.3 ± 0.21 | 0.42 ± 0.23 | 0.17 ± 0.13 | 0.23 ± 0.15 | 0.3 ± 0.18 | 0.37 ± 0.14 |
| Preval | 0.2 ± 0.13 | 0.22 ± 0.18 | 0.3 ± 0.2 | 0.34 ± 0.17 | 0.21 ± 0.15 | 0.25 ± 0.18 | 0.3 ± 0.22 | 0.39 ± 0.22 | 0.19 ± 0.14 | 0.23 ± 0.15 | 0.3 ± 0.17 | 0.36 ± 0.16 |
| AvgProb | 0.17 ± 0.12 | 0.21 ± 0.14 | 0.26 ± 0.17 | 0.38 ± 0.15 | 0.17 ± 0.13 | 0.22 ± 0.15 | 0.26 ± 0.19 | 0.37 ± 0.18 | 0.16 ± 0.13 | 0.22 ± 0.15 | 0.29 ± 0.16 | 0.37 ± 0.16 |
| pS-SDM+PRR | 0.15 ± 0.16 | 0.17 ± 0.18 | 0.26 ± 0.23 | 0.28 ± 0.25 | 0.14 ± 0.17 | 0.18 ± 0.2 | 0.22 ± 0.23 | 0.29 ± 0.29 | 0.14 ± 0.16 | 0.17 ± 0.19 | 0.26 ± 0.21 | 0.28 ± 0.22 |
| MEM+PRR | 0.14 ± 0.16 | 0.17 ± 0.21 | 0.24 ± 0.22 | 0.32 ± 0.26 | 0.14 ± 0.17 | 0.18 ± 0.2 | 0.22 ± 0.23 | 0.34 ± 0.29 | 0.15 ± 0.16 | 0.19 ± 0.19 | 0.25 ± 0.21 | 0.32 ± 0.22 |

C) Percentage correct classified (PCC)

| Thresholding | SSCV | | | | ID | | | | CCV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Approach** | **FO** | **GL** | **BF** | **GH** | **FO** | **GL** | **BF** | **GH** | **FO** | **GL** | **BF** | **GH** |
| Max.Kappa | 0.91 ± 0.09 | 0.9 ± 0.07 | 0.78 ± 0.1 | 0.83 ± 0.08 | 0.9 ± 0.09 | 0.89 ± 0.07 | 0.77 ± 0.09 | 0.83 ± 0.1 | 0.88 ± 0.09 | 0.87 ± 0.06 | 0.76 ± 0.08 | 0.82 ± 0.07 |
| MCE05 | 0.85 ± 0.15 | 0.79 ± 0.07 | 0.66 ± 0.11 | 0.82 ± 0.08 | 0.88 ± 0.07 | 0.87 ± 0.05 | 0.77 ± 0.09 | 0.82 ± 0.08 | 0.59 ± 0.14 | 0.68 ± 0.09 | 0.64 ± 0.1 | 0.73 ± 0.08 |
| Max.TSS | 0.85 ± 0.08 | 0.84 ± 0.07 | 0.73 ± 0.11 | 0.83 ± 0.09 | 0.86 ± 0.1 | 0.85 ± 0.08 | 0.73 ± 0.11 | 0.81 ± 0.09 | 0.77 ± 0.07 | 0.79 ± 0.05 | 0.72 ± 0.08 | 0.79 ± 0.06 |
| SensSpec | 0.79 ± 0.1 | 0.84 ± 0.04 | 0.73 ± 0.09 | 0.81 ± 0.08 | 0.87 ± 0.09 | 0.86 ± 0.07 | 0.73 ± 0.1 | 0.81 ± 0.08 | 0.79 ± 0.07 | 0.81 ± 0.05 | 0.73 ± 0.07 | 0.8 ± 0.06 |
| Opt.ROC | 0.86 ± 0.08 | 0.86 ± 0.06 | 0.72 ± 0.11 | 0.82 ± 0.05 | 0.87 ± 0.1 | 0.86 ± 0.07 | 0.74 ± 0.1 | 0.81 ± 0.09 | 0.79 ± 0.07 | 0.81 ± 0.05 | 0.73 ± 0.08 | 0.8 ± 0.06 |
| Preval | 0.92 ± 0.08 | 0.91 ± 0.05 | 0.79 ± 0.08 | 0.83 ± 0.06 | 0.9 ± 0.08 | 0.89 ± 0.06 | 0.77 ± 0.1 | 0.83 ± 0.09 | 0.9 ± 0.08 | 0.89 ± 0.06 | 0.77 ± 0.08 | 0.82 ± 0.07 |
| AvgProb | 0.71 ± 0.07 | 0.67 ± 0.08 | 0.64 ± 0.11 | 0.74 ± 0.08 | 0.69 ± 0.08 | 0.69 ± 0.07 | 0.64 ± 0.11 | 0.73 ± 0.08 | 0.69 ± 0.07 | 0.69 ± 0.07 | 0.66 ± 0.09 | 0.73 ± 0.07 |
| pS-SDM+PRR | 0.93 ± 0.09 | 0.88 ± 0.07 | 0.76 ± 0.08 | 0.86 ± 0.09 | 0.91 ± 0.1 | 0.89 ± 0.08 | 0.77 ± 0.11 | 0.83 ± 0.1 | 0.91 ± 0.1 | 0.89 ± 0.08 | 0.78 ± 0.09 | 0.84 ± 0.1 |
| MEM+PRR | 0.92 ± 0.1 | 0.9 ± 0.09 | 0.8 ± 0.09 | 0.84 ± 0.09 | 0.91 ± 0.1 | 0.89 ± 0.09 | 0.79 ± 0.1 | 0.86 ± 0.08 | 0.91 ± 0.1 | 0.89 ± 0.08 | 0.79 ± 0.08 | 0.86 ± 0.09 |

D) Sensitivity

| Thresholding | SSCV | | | | ID | | | | CCV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Approach** | **FO** | **GL** | **BF** | **GH** | **FO** | **GL** | **BF** | **GH** | **FO** | **GL** | **BF** | **GH** |
| Max.Kappa | 0.31 ± 0.18 | 0.36 ± 0.24 | 0.47 ± 0.21 | 0.51 ± 0.26 | 0.27 ± 0.21 | 0.32 ± 0.25 | 0.46 ± 0.24 | 0.56 ± 0.26 | 0.32 ± 0.18 | 0.37 ± 0.18 | 0.5 ± 0.2 | 0.52 ± 0.22 |
| MCE05 | 0.45 ± 0.13 | 0.61 ± 0.11 | 0.66 ± 0.18 | 0.73 ± 0.12 | 0.35 ± 0.17 | 0.42 ± 0.17 | 0.42 ± 0.23 | 0.57 ± 0.19 | 0.68 ± 0.05 | 0.67 ± 0.07 | 0.7 ± 0.09 | 0.71 ± 0.11 |
| Max.TSS | 0.44 ± 0.28 | 0.55 ± 0.15 | 0.62 ± 0.14 | 0.67 ± 0.15 | 0.46 ± 0.28 | 0.52 ± 0.27 | 0.59 ± 0.22 | 0.66 ± 0.25 | 0.57 ± 0.11 | 0.59 ± 0.11 | 0.61 ± 0.14 | 0.63 ± 0.15 |
| SensSpec | 0.55 ± 0.25 | 0.49 ± 0.15 | 0.58 ± 0.18 | 0.65 ± 0.13 | 0.42 ± 0.26 | 0.49 ± 0.25 | 0.58 ± 0.22 | 0.67 ± 0.25 | 0.54 ± 0.1 | 0.55 ± 0.11 | 0.58 ± 0.14 | 0.63 ± 0.12 |
| Opt.ROC | 0.44 ± 0.2 | 0.54 ± 0.14 | 0.57 ± 0.21 | 0.67 ± 0.14 | 0.44 ± 0.26 | 0.49 ± 0.26 | 0.58 ± 0.22 | 0.65 ± 0.24 | 0.55 ± 0.1 | 0.56 ± 0.11 | 0.59 ± 0.14 | 0.62 ± 0.14 |
| Preval | 0.28 ± 0.17 | 0.28 ± 0.21 | 0.46 ± 0.21 | 0.55 ± 0.25 | 0.26 ± 0.2 | 0.32 ± 0.24 | 0.49 ± 0.22 | 0.53 ± 0.25 | 0.24 ± 0.18 | 0.29 ± 0.18 | 0.47 ± 0.19 | 0.5 ± 0.23 |
| AvgProb | 0.76 ± 0.11 | 0.82 ± 0.1 | 0.77 ± 0.12 | 0.83 ± 0.1 | 0.79 ± 0.11 | 0.85 ± 0.1 | 0.78 ± 0.17 | 0.85 ± 0.13 | 0.76 ± 0.1 | 0.81 ± 0.06 | 0.78 ± 0.1 | 0.84 ± 0.07 |
| pS-SDM+PRR | 0.21 ± 0.28 | 0.26 ± 0.32 | 0.43 ± 0.36 | 0.42 ± 0.39 | 0.21 ± 0.28 | 0.28 ± 0.33 | 0.44 ± 0.36 | 0.45 ± 0.4 | 0.21 ± 0.27 | 0.26 ± 0.3 | 0.45 ± 0.33 | 0.41 ± 0.36 |
| MEM+PRR | 0.21 ± 0.28 | 0.3 ± 0.31 | 0.38 ± 0.36 | 0.41 ± 0.38 | 0.21 ± 0.29 | 0.28 ± 0.33 | 0.38 ± 0.35 | 0.43 ± 0.38 | 0.21 ± 0.27 | 0.28 ± 0.3 | 0.39 ± 0.32 | 0.42 ± 0.34 |

E) Specificity

| Thresholding | SSCV | | | | ID | | | | CCV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Approach | FO | GL | BF | GH | FO | GL | BF | GH | FO | GL | BF | GH |
| Max.Kappa | 0.95 ± 0.09 | 0.91 ± 0.07 | 0.82 ± 0.12 | 0.87 ± 0.1 | 0.94 ± 0.07 | 0.93 ± 0.07 | 0.84 ± 0.12 | 0.88 ± 0.1 | 0.91 ± 0.09 | 0.93 ± 0.07 | 0.81 ± 0.12 | 0.85 ± 0.1 |
| MCE05 | 0.87 ± 0.15 | 0.81 ± 0.04 | 0.78 ± 0.11 | 0.79 ± 0.09 | 0.93 ± 0.03 | 0.91 ± 0.03 | 0.86 ± 0.1 | 0.88 ± 0.08 | 0.57 ± 0.15 | 0.91 ± 0.03 | 0.61 ± 0.13 | 0.7 ± 0.1 |
| Max.TSS | 0.88 ± 0.09 | 0.79 ± 0.07 | 0.73 ± 0.1 | 0.83 ± 0.09 | 0.87 ± 0.1 | 0.86 ± 0.09 | 0.75 ± 0.13 | 0.84 ± 0.09 | 0.78 ± 0.08 | 0.86 ± 0.09 | 0.74 ± 0.1 | 0.8 ± 0.08 |
| SensSpec | 0.89 ± 0.07 | 0.82 ± 0.06 | 0.77 ± 0.1 | 0.81 ± 0.09 | 0.89 ± 0.09 | 0.88 ± 0.07 | 0.75 ± 0.12 | 0.84 ± 0.09 | 0.81 ± 0.07 | 0.88 ± 0.07 | 0.76 ± 0.09 | 0.82 ± 0.07 |
| Opt.ROC | 0.84 ± 0.08 | 0.87 ± 0.08 | 0.74 ± 0.11 | 0.82 ± 0.09 | 0.88 ± 0.09 | 0.88 ± 0.08 | 0.76 ± 0.13 | 0.84 ± 0.09 | 0.8 ± 0.07 | 0.88 ± 0.08 | 0.76 ± 0.09 | 0.82 ± 0.07 |
| Preval | 0.92 ± 0.07 | 0.95 ± 0.05 | 0.85 ± 0.13 | 0.85 ± 0.09 | 0.94 ± 0.08 | 0.94 ± 0.06 | 0.82 ± 0.13 | 0.87 ± 0.11 | 0.94 ± 0.08 | 0.94 ± 0.06 | 0.84 ± 0.11 | 0.87 ± 0.09 |
| AvgProb | 0.69 ± 0.08 | 0.68 ± 0.06 | 0.63 ± 0.12 | 0.72 ± 0.09 | 0.68 ± 0.08 | 0.68 ± 0.07 | 0.6 ± 0.12 | 0.71 ± 0.09 | 0.68 ± 0.08 | 0.68 ± 0.07 | 0.62 ± 0.09 | 0.69 ± 0.08 |
| pS-SDM+PRR | 0.92 ± 0.15 | 0.91 ± 0.12 | 0.81 ± 0.19 | 0.86 ± 0.19 | 0.93 ± 0.15 | 0.92 ± 0.12 | 0.8 ± 0.23 | 0.84 ± 0.21 | 0.93 ± 0.14 | 0.92 ± 0.12 | 0.82 ± 0.19 | 0.86 ± 0.18 |
| MEM+PRR | 0.93 ± 0.14 | 0.93 ± 0.1 | 0.85 ± 0.18 | 0.91 ± 0.15 | 0.93 ± 0.15 | 0.92 ± 0.13 | 0.85 ± 0.2 | 0.89 ± 0.16 | 0.93 ± 0.14 | 0.92 ± 0.13 | 0.86 ± 0.17 | 0.88 ± 0.15 |