Methods in Ecology and Evolution

DR DANIEL SCHERRER (Orcid ID : 0000-0001-9983-7510)

DR HEIDI MOD (Orcid ID : 0000-0001-7800-2688)

Article type : Research Article

Handling editor: Professor Robert B. O'Hara

How to evaluate community predictions without thresholding?

Daniel Scherrer[a,b], Heidi K. Mod[a,c] & Antoine Guisan[a,d]

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

[b] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Forest Dynamics, CH-8903 Birmensdorf, Switzerland

[c] Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, 00014 Helsinki, Finnland

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

Corresponding author:
Daniel Scherrer
Swiss Federal Institute for Forest, Snow and Landscape Research WSL
CH-8903 Birmensdorf, Switzerland

Running title: Evaluation of community predictions

Tweetable Abstract: Comparable and objective methods to evaluate probabilistic community predictions without the need of thresholding

Abstract

1. Stacked species distribution models (S-SDM) provide a tool to make spatial predictions about communities by first modelling individual species and then stacking the modelled predictions to form assemblages. The evaluation of the predictive performance is usually based on a comparison of the observed and predicted community properties (e.g., species richness, composition). However, the most available and widely used evaluation metrics require the thresholding of single species' predicted probabilities of occurrence to obtain binary outcomes (i.e., presence/absence). This binarisation can introduce unnecessary bias and error.
2. Herein, we present and demonstrate the use of several groups of new or rarely used evaluation approaches and metrics for both species richness and community composition that do not require thresholding but instead directly compare the predicted probabilities of occurrences of species to the presence/absence observations in the assemblages.
3. *Community AUC*, which is based on traditional AUC, measures the ability of a model to differentiate between species presences or absences at a given site according to their predicted probabilities of occurrence. Summing the probabilities gives the expected species richness and allows the estimation of the *probability that the observed species richness is not different from the expected species richness based on the species' probabilities of occurrence*. The traditional Sørensen and Jaccard similarity indices (which are based on presences/absences) were adapted to *maxSørensen* and *maxJaccard* and to *probSørensen* and *probJaccard* (which use probabilities directly). A further approach (*improvement over null models*) compared the predictions based on S-SDMs with the expectations from the null models to estimate the improvement in both species richness and composition predictions. Additionally, all metrics can be described against the environmental conditions of sites (e.g., elevation) to highlight the abilities of models to detect the variation in the strength of the community assembly processes in different environments.
4. These metrics offer an unbiased view of the performance of community predictions compared to metrics that requiring thresholding. As such, they allow more straightforward comparisons of model performance among studies (i.e., they are not influenced by any subjective thresholding decisions).

Keywords: community modelling, stacked species distribution models, validation, environmental gradient, insects, plants, Sørensen index, Jaccard index, null model

Introduction

In recent years, the focus of spatial ecology has shifted from analysing the distributions of species individually to examining them as part of communities or networks (e.g., D'Amen *et al.* 2017; Ovaskainen *et al.* 2017; Staniczenko *et al.* 2017), and the field of species distribution modelling has evolved from predicting the distributions of individual species towards predicting those of species assemblages (e.g., Guisan & Rahbek 2011; Wisz *et al.* 2013; Harris 2015; D'Amen *et al.* 2017). Among the different methods (see D'Amen *et al.* 2017 for a review), stacked species distribution models (S-SDMs; Dubuis *et al.* 2011; Guisan & Rahbek 2011) have become the most prevalent examples of species assemblage modelling approaches in recent literature. S-SDMs first predict the probability of occurrence of individual species using niche-based species distribution models (SDMs) based on quantification of the relationship between environmental factors (usually climatic, topographic and land-cover/use variables) and species occurrences (usually presences = 1 and absences = 0; Guisan & Thuiller 2005; Elith & Leathwick 2009; Guisan, Thuiller & Zimmermann 2017). These individual predictions are then assembled (i.e., stacked; e.g., Ferrier & Guisan 2006; Guisan & Rahbek 2011; Mateo, Mokany & Guisan 2017).

One of the major discussions around S-SDMs is whether to threshold individual species distribution predictions (i.e., to binarise the predicted probabilities of occurrence "back" to presences and absences) or to keep them as probabilities (i.e., continuous values from 0 to 1; e.g., Gastón & García-Viñas 2013; Calabrese *et al.* 2014; Scherrer *et al.* 2018). It has been argued that predictions should be kept as probabilities because thresholding is a transformation of the original model outcome, and as with all classifications, it bears some subjectivity (e.g., which threshold method to use? Fernandes, Scherrer & Guisan 2018). While there is a consensus that some community properties, such as species richness (SR), can be obtained directly based on the probabilities of all of the species predicted at a site (i.e., summing up the probabilities by assuming Poisson-binomial distribution; Dubuis *et al.* 2011; Calabrese *et al.* 2014), there is more debate on the question of how to obtain or evaluate composition information using the raw probabilities (Scherrer *et al.* 2018). The vast majority of S-SDM studies use community evaluation metrics (e.g., Sørensen or Jaccard (dis)similarity indices) that compare binary predictions with presence/absence observations and therefore require the process of thresholding the probabilistic predictions into a 0/1 outcome. The most commonly applied approach is to threshold species individually by optimising a combination of sensitivity (i.e., proportion of correctly predicted presences) and/or specificity (i.e., proportion of correctly predicted absences) and then to use a single threshold per species across all sites ("species threshold"; Calabrese *et al.* 2014). An alternative approach to thresholding individual species predictions does not require a species-specific threshold but instead uses site-specific ecological constraints (e.g., macro-ecological models or co-occurrence matrices) in combination with probability ranking rules (PRR) to threshold the species in each site ("site-threshold"; Scherrer *et al.* 2018). These methods select a

number of species equal to the expected species richness (e.g., derived from macro-ecological models) on the basis of the decreasing probabilities of occurrence calculated by the SDMs (D'Amen *et al.* 2015; D'Amen, Pradervand & Guisan 2015). Therefore, the species with the highest probabilities for a site are selected (considered present) in decreasing order until the SR predicted for the site is reached.

Nevertheless, no matter which approach is used, thresholding bears the risk of introducing bias in the modelling outcome (see Nenzen & Araujo 2011; Calabrese *et al.* 2014). To date, only very few studies have used threshold-independent community evaluation metrics (but see Gastón & García-Viñas 2013; Harris 2014). Therefore, while the process of thresholding might ultimately be necessary for some applications (e.g., to create community maps in space or time for conservation purposes), we herein advocate that this bias-prone "a priori" thresholding can be avoided in the process of model evaluation. Using an evaluation method prior to any thresholding could therefore give a more realistic, bias-free estimation of the community prediction.

In addition to the potential biases associated with thresholding, there is also the question of which information to include in the model evaluation. The most commonly used metrics (e.g., the Sørensen or Jaccard similarity indices) only consider the species that are present or that have been predicted to be present, while the correctly predicted absences bear no influence on the metrics. Metrics that also consider the species as absent or predicted to be absent from a site (i.e., assigning '0' to all species of the modelled species pool that are not (predicted to be) in the plot) might provide a more comprehensive evaluation of the performance of community predictions (Baroni-Urbani & Buser 1976; Wolda 1981).

Herein, we present five groups of valuable but rarely used approaches and propose new metrics to evaluate predictions of both species richness and community composition that do not require thresholding but instead directly compare the predicted probabilities of occurrence of species assemblages to the presence/absence observations (see Table 1 for an overview). Our aim is to show the power of these threshold-independent metrics and approaches to compare any probabilistic predictions to real presence-absence observations (note that we do not consider the special case of presence-only or presence-background data here but that we do discuss these in some places in this study). We illustrate their uses and performances by applying them to re-evaluate published predictions of both plant and insect communities with varying assemblage characteristics (D'Amen *et al.* 2015; D'Amen, Pradervand & Guisan 2015) as well as to a set of 100 virtual species in a controlled environment with a known "true" distribution and bias.

Materials and Methods

We first present the data and models used to obtain the community predictions, which were existing datasets and published community models; later, we present the existing and new metrics and approaches that can be used to evaluate them without requiring any subjective thresholding.

Community data and environmental variables

To model and predict species assemblages, we used presence-absence data on vascular plants, butterflies, grasshoppers and bumble bees that were exhaustively surveyed in a study area located in the western Swiss Alps (46°10′ to 46°30′ N; 6°50’ to 7°10’ E). Sites were placed following an equal random-stratified sampling of grasslands, covering an area of ca. 700 km^2 and spanning an elevation range from 375 to 3210 m a.s.l.

For the plants, 909 sites (4 m^2 plots) were surveyed (for more details see Dubuis *et al.* 2011), whereas the data on butterflies, grasshoppers and bumble bees were collected at 208, 202 and 202 sites, respectively, by sampling 50 x 50 m areas (for more details see Pradervand *et al.* 2011 (bumble bees); Pellissier *et al.* 2012 (butterflies); Pradervand *et al.* 2013 (grasshoppers)). Due to data and model restrictions, only 37-69% of the originally observed species (those with >10 occurrences) were considered in this study (Table 2 and Figure S1). The loss of a large proportion of the observed species is a common disadvantage of methods such as S-SDM. However, while the rare species form an essential part of the biodiversity of the area, within the datasets used here, they only marginally contributed to shaping the patterns of species richness across sites (Figure S2). Furthermore, not accounting for bias in data can affect community models and predictions (e.g., Dorazio *et al.* 2006; Kery, Gardner & Monnerat 2010; Fernandes, Scherrer & Guisan 2019). For the sake of ecological realism, we demonstrated the use of the metrics by evaluating existing model predictions based on real data with varying community characteristics (see Table 1). Additionally, we used artificial data to explore the influence of detection issues and species misidentifications (more detailed analysis in Fernandes, Scherrer & Guisan 2018; Fernandes, Scherrer & Guisan 2019).

For all species in all taxonomic groups, we used the same set of six environmental predictors calculated at a 25 m × 25 m resolution: annual mean temperature [°C], annual temperature range [K], annual precipitation sum [mm], sum of potential solar radiation over the year [KJ], slope [°] and topographic position [unit-less, indicating ridges or valleys]. These variables captured the topo-climatic conditions of the mountain environments.

Species distribution models

Due to the low prevalence of many modelled species, we fitted our models with an ensemble of small models approach optimised for rare or under-sampled species (ESMs; Lomba *et al.* 2010; Breiner *et al.* 2015; Breiner *et al.* 2018). Individual models were calibrated with bivariate combinations of the predictors using four modelling techniques: generalised linear models (GLM; McCullagh & Nelder 1989), boosted regression trees (BRT; Elith, Leathwick & Hastie 2008), classification tree analysis (CTA; Strobl, Malley & Tutz 2009) and artificial neural networks (ANN; Lek *et al.* 1996). All the converged bivariate models were then averaged into one ensemble model weighted by their respective AUC. To evaluate the

performance of our models both on an individual species level and on the species assemblage level, we used a community-cross-validation approach (N=10, 80%/20% training/validation; CCV; Scherrer *et al.* 2018). The CCV uses the same set of study sites for all species within an assemblage for each CV and therefore allows correct cross-validation at both the species and community levels.

The single species models were evaluated using three threshold independent evaluation metrics: the area under the curve of a receiver-operating-characteristic (ROC) plot (AUC; Hanley & McNeil 1982; Swets 1988), the maximum true skill statistics (maxTSS; Allouche, Tsoar & Kadmon 2006) and the maximum Cohen's Kappa (maxKappa; Cohen 1960; see Guisan, Thuiller & Zimmermann 2017 for details on maximisation approaches). All models were run in the R software version 3.4.2 (R Core Team 2017) using the ESM functions from the ecospat package (Broenniman, Di Cola & Guisan 2017; Di Cola *et al.* 2017) in combination with biomod2 (Thuiller *et al.* 2009; Thuiller *et al.* 2016).

Evaluation of species assemblage predictions

We used four different approaches to evaluate the species assemblages' predictions based directly on the probabilistic output of the ESMs, thus avoiding the bias introduced by thresholding (see the next section). These threshold-independent metrics were then compared to the corresponding evaluation metrics based on the binary stacked species distribution models (bS-SDMs) using traditional thresholding techniques (Scherrer *et al.* 2018) and/or examined along the elevational gradient to the identify environmental conditions where the models performed best or worst.

To create binary predictions for all species (needed for the bS-SDMs), we used three "species-specific" (i.e., for each species same threshold across all sites) and one "site-specific" thresholding method (i.e., same threshold for all species at a given site). The "species-specific" thresholds were either fixed to 0.5 or determined by maximising the true skill statistics (maxTSS) or Cohen's Kappa values (maxKappa), and the "site-specific" threshold was based on the sum of probabilities in combination with a probability ranking rule (pS-SDM+PRR; see D'Amen *et al.* 2015; D'Amen, Pradervand & Guisan 2015; Scherrer *et al.* 2018 for details on thresholding).

To estimate the sensitivity of our evaluation approaches to bias in the initial calibration data, we not only tested them on the above-mentioned "real world" datasets but also on a set of 100 virtual species. The simulations with virtual species allowed us to observe the behaviours of the evaluation approaches to different types of error (i.e., omission errors due to detectability issues and omission and commission errors due to misidentification) under controlled conditions (i.e., known truth and bias). All details about these virtual simulations can be found in Appendix 2.

All the presented approaches are intended for and tested on accurate presence-absence data. Using any of the suggested approaches on presence-only (or presence-background) data/models is technically incorrect, as these models will not give a probability of occurrence (due to the missing absence information) but rather a habitat suitability index (i.e., violating the underlying assumptions of the Poisson-binomial distribution). Nevertheless, some of the approaches (especially those with low influence of absences; Table 1) might work well in practice with presence-only models, as the same sites are used for all species within a community.

Metrics proposed for evaluating species assemblage predictions without thresholding

1. Community AUC

The community AUC (cAUC, equation 1) is mathematically identical to the standard AUC used in the evaluation of single species distribution models (SDMs), but instead of calculating the area under an ROC curve (i.e., the false positive rate vs true positive rate) of one species across all sites, the cAUC calculates the area under the ROC curve of all the species at one site (resulting from the S-SDM predictions). In this case, the cAUC can be interpreted as the abilities of the models to distinguish species presence/absence based on the ranking of the predicted probabilities of occurrence of all species at a given site (identical to Harrell's Concordance Index; Harrell Jr *et al.* 1982). The cAUC is defined as follows:

eqn 1
$$\text{cAUC} = \frac{\sum R_P - \frac{n_P(n_P+1)}{2}}{n_P n_A}$$

where $\sum R_P$ is the rank sum of all species present at a site (i.e., the ranks sorted from lowest to highest probability), n_P is the number of species present and n_A is the number of species absent (see Marrocco, Duin & Tortorella 2008 for mathematical details). Therefore, a cAUC value of 1 means that all the species that are present at a site have higher predicted probabilities of occurrence than any of the species that are absent from the site. As a result, a probability threshold can be set that perfectly distinguishes the species that are present from those that are absent. A value of cAUC lower than 1 indicates that no such perfect threshold is possible, and some commission or omission errors are unavoidable (note: 0.5 = random model, 0 = all the species present in a site have lower predicted probabilities of occurrence than all the species absent from the site; i.e., counter-prediction). While the AUC is the most commonly used evaluation metric for assessing the predictive performance of individual species SDMs (i.e., performance across sites; Fourcade, Besnard & Secondi 2018), it has rarely been used thus far to evaluate predictions at the community level (i.e., across species in a site; but see Gastón & García-Viñas 2013).

2. Deviation in species richness

Following Dubuis *et al*. (2011) and Calabrese *et al.* (2014), the expected species richness *SR* of a site *j* was calculated by summing the predicted probabilities ($p_{j,k}$) of all *K* species at site *j*. This was based on the assumption that the site-level species richness prediction SR_j follows a Poisson-binomial distribution with a probability mass function, which was calculated as follows:

eqn 2 $$\Pr\left(SR_j \middle| p_j\right) = \frac{1}{K+1}\sum_{n=0}^{K}\left(e^{\frac{-i2\pi nS_j}{K+1}}\prod_{k=1}^{K}\left[p_{j,k}e^{\frac{i2\pi n}{K+1}} + (1-p_{j,k})\right]\right)$$

where $i = \sqrt{-1}$ is the imaginary unit. Therefore, the expected mean species richness $E(SR_j)$ and its standard deviation σ(SRj) of site *j* are as follows:

eqn 3 $$E(SR_j) = \sum_{k=1}^{K} p_{j,k}$$

and

eqn 4 $$\sigma(SR_j) = \sqrt{\sum_{k=1}^{K}(1-p_{j,k})p_{j,k}}$$

which provide a formal theoretical basis for stacking predictions of probability of occurrence values from SDMs (see Calabrese *et al.* 2014 for more details). Based on the probability mass function (equation 1), we can also calculate the probability (p-value) of a site-level species richness prediction SR_j that is equal to or lower than the observation SR_{obs_j} (equation 5) or equal to or higher than the observation SR_{obs_j} (equation 6) based on the probabilities of occurrence in a site p_j, which can be written as follows:

eqn 5 $$\Pr\left(SR_j \leq SR_{obs_j}\right) = \sum_{SR_j=0}^{SR_{obs_j}} \Pr\left(SR_j | p_j\right)$$

eqn 6 $$\Pr\left(SR_j \geq SR_{obs_j}\right) = 1 - \sum_{SR_j=0}^{SR_{obs_j}-1} \Pr\left(SR_j | p_j\right)$$

Based on the null hypothesis (H0) that there is no difference between the observed and expected SR and a predefined α (the probability of making a Type I error, which is usually set at 0.05), we can then use equation 5 (if observed SR <= expected SR) or equation 6 (if observed SR >= expected SR) to decide whether to accept or reject H0 (see example scripts in Appendix 1). All these metrics are based on the assumption that the probabilities (p_j) are fixed, known quantities. In reality, the p_j contain uncertainty, and ignoring this uncertainty might lead to errors in the estimation of confidence intervals for SR_j. Therefore, in the case of known uncertainty, error propagation techniques could be used to account for uncertainty in the site-level richness predictions.

As the absolute SR error (i.e., the difference between the observed and expected SR) is strongly dependent on the modelled species pool (i.e., number of species modelled) as well as the average site SR, we

standardised the SR error by dividing it by the average site SR of each taxa (see Table 2). In this way, the results from the different taxa with differently sized species pools and the average site SR can be compared.

3. Maximisation approaches for community composition metrics

To evaluate the abilities of our models to predict the community composition, we propose two types of new similarity metrics that are closely related to the commonly used the Sørensen (equation 7; Sørensen 1948) and Jaccard (dis)similarity (equation 8; Jaccard 1901) indices. The two traditional indices were defined as follows:

eqn 7 $$Sørensen\ similarity = \frac{2TP}{2TP + FP + FN}$$

eqn 8 $$Jaccard\ similarity = \frac{TP}{TP + FP + FN}$$

where *TP* are true positives (i.e., species are both observed and predicted to be present), *FP* are false positives (i.e., species that are not observed but are predicted to be present) and *FN* false negatives (i.e., species that are observed but are predicted to be absent). All these metrics only consider the species that are either predicted or observed to be present in at least one of the samples/sites and ignore species that are absent from both samples/sites.

The first type of new similarity metrics uses a maximisation of both the Sørensen *(maxSørensen*) and Jaccard similarity indices (*maxJaccard*) under the premise that the "site-thresholds" perform as well or better than the classical species-thresholds in S-SDMs when aiming for optimised community predictions (as evaluated by the Sørensen/Jaccard indices; D'Amen *et al.* 2015; Scherrer *et al.* 2018). By considering the "site-thresholds" instead of the "species-thresholds", we can calculate the maximum possible values of the evaluation metrics (i.e., Sørensen and Jaccard similarity indices) at each site individually (i.e., optimal site specific threshold). This process is similar to the calculation of the maximum values of standard evaluation metrics (e.g., TSS or Kappa; i.e., *maxTSS* or *maxKappa*) when evaluating single species predictions (Guisan, Thuiller & Zimmermann 2017). To calculate the site-specific *maxSørensen* and *maxJaccard*, we took the maximum of the evaluation values calculated at "all" possible thresholds (0 to 1 with 0.001 increments).

4. Probability sum-based community composition metrics

The second type of new similarity metrics *(probSørensen*, equation 9 and *probJaccard*, equation 10) are directly based on the probability outputs of the SDMs. We wanted to stay close to the original Sørensen and Jaccard similarity indices, both of which are based on the concept of shared species ($A \cap B$) versus the union of species ($A \cup B$). Therefore, we defined the shared species as the sum of the predicted probabilities for all species present in the observation and the union of species as the sum of the predicted probabilities

that are higher than the lowest predicted probability of an observed species. Identical to the Sørensen and Jaccard similarity indices based on the binary data (presence/absence), our probabilistic Sørensen (*probSørensen*) index gives double the weight to the shared species compared to the weighting strategy of the Jaccard index (*probJaccard*):

eqn 9 $$probSørensen = 2\sum_{k \in P_j} p_{j,k} / (2\sum_{k \in P_j} p_{j,k} + \sum_{p_{j,k} \geq \min(p_{j,k})_{k \in P_j} \& k \in A_j} p_{j,k})$$

eqn 10 $$probJaccard_j = \sum_{k \in P_j} p_{j,k} / \sum_{p_{j,k} \geq \min(p_{j,k})_{k \in P} p_{j,k}}$$

where $p_{j,k}$ is the predicted probability of species k in site j and P_j is the list of species present at site j and A_j is the species absent at site j.

All these new metrics vary from 0 to 1, similar to the original Sørensen and Jaccard similarity indices, where 1 means perfect agreement and 0 means no species in common between observations and predictions.

5. Improvement over null models

To test our S-SDMs against random expectations, we created two different null models containing different amounts of available information. The first null model (null.SR) had information on the modelled species pool N and the mean observed SR ($\overline{SR}$). This model therefore assumes the same probability $p_{j,k}$ for each species k in each site j, calculated as follows:

eqn 11 $$p_{j,k} = \frac{\overline{SR}}{N}$$

In this way, the sum of the probabilities at each site adds up to the mean observed species richness across all sites.

The second null model (null.Prev) had all the information on species assemblages fed to the SDMs, i.e., modelled species pool N, mean observed species richness $\overline{SR}$ and prevalence of each species *Prev*. This model therefore predicts for each species k in each site j a probability $p_{j,k}$ identical to the observed prevalence of the species and can be written as follows:

eqn 12 $$p_{j,k} = Prev_k$$

Based on our two null models, we then calculated for each site j the probability of obtaining the observed SR correctly based on the probability mass functions (equation 2). Additionally, we calculated for each site j the probability of obtaining the observed community composition C_j (presence/absence of species) correctly based on the following:

eqn 13 $$\Pr(C_j \mid p_j) = \prod_{k \in P_j} p_{j,k} \prod_{k \in A_j} (1 - p_{j,k})$$

where *P* is the list of species that are present at a site and *A* is the species that are absent at a site.

The probability of obtaining the SR or composition correct with our null models was then compared to the probability of obtaining it correctly based on our SDM predictions (i.e., the predicted probability of occurrence of species) by dividing the probability based on our SDM predictions by the probability of the null models (see Box 1 for more information). For example, a value of 2 would mean that the SDM predictions are twice as likely to produce a correct result as the null model. Therefore, if the chance of obtaining the correct SR and composition based on our SDM predictions were higher than those based on our null models, we concluded that the environmental information (i.e., predictors) in the SDMs had explanatory power beyond the assemblage characteristics (i.e., species pool modelled, mean observed SR and prevalence of species; see Table 2).

Results

The AUC values of the individual SDMs were 0.83 ± 0.05 (mean ± sd across species), 0.70 ± 0.07, 0.87 ± 0.06 and 0.80 ± 0.08 for plants, bumble bees, grasshoppers and butterflies, respectively (see Figure S3 for results on maxTSS and maxKappa). As the purpose of this work was the demonstration of new probabilistic community evaluation methods rather than model optimisation, we consider these models to be appropriate and focused on the evaluation of their community predictions.

Community AUC

The values for the community AUC (cAUC) were 0.87 ± 0.08 (mean ± sd across sites), 0.82 ± 0.12, 0.92 ± 0.09 and 0.86 ± 0.07 for plants, bumble bees, grasshoppers and butterflies, respectively (Figure S4). Additionally, there were 5.3%, 4.0%, 26.8% and 0.5% of plant, bumble bee, grasshopper and butterfly sites with a cAUC of 1, indicating that a perfect separation (i.e., one probability threshold) of species present/absent was possible. Our results showed that the variation in cAUC was not random but rather varied strongly according to elevation and had consistently higher cAUC values at the lowest and highest elevations and the worse predictions at mid elevations across all studied taxa (Figure 1).

Additionally, our simulations with virtual species showed that the cAUC was not strongly affected by detection issues as long as the bias levels were similar among the species of a community (Appendix 2).

Deviation in species richness

As expected, our probabilistic species richness predictions (i.e., the sum of the probabilities) resulted in a mean standardised SR error (i.e., the difference between the observed and expected SR standardised by the average site SR) across sites and species of approximately 0 (-0.01 ± 0.00, mean ± sd), while the

predictions using a "species-specific" threshold have a standardised SR error of -0.48 ± 0.23, 0.47 ± 0.46 and -0.08 ± 0.04 for the fixed threshold, maxTSS and maxKappa, respectively (Figure S4). As the "site-specific" threshold (pS-SDM+PRR) also uses the sum of probabilities to determine the SR, the results were identical to the probabilistic approach. Based on the H0 of no difference between the observed and expected SR and a predefined α of 0.05 (i.e., significance level), our results showed that, depending on the taxonomic group, 45.1–86.1% of the sites did not show a significant ($p<0.05$) difference between the observed and predicted SR (Figure 2). However, our simulation of virtual species in a controlled environment showed that this metric is strongly influenced by omission errors (Appendix 2).

Maximisation and probability-sum-based community composition metrics

Our four newly proposed threshold independent similarity metrics for community composition (i.e., *maxSørensen*, *maxJaccard* and *probSørensen*, *probJaccard*) are generally highly correlated with the corresponding similarity metrics based on thresholded presence/absences (bS-SDMs; Figure 3, Figure S6). The correlation between the binary Sørensen and Jaccard similarity indices and our threshold-independent counterparts increased with mean site SR and with the size of the modelled species pool and was consequently lowest for the grasshoppers and highest for the plants. "Species-specific" thresholding techniques generally led to lower correlations with our new metrics (Spearman correlation 0.53-0.83) than with the "site-specific" thresholding techniques (Spearman correlation 0.74-0.91).

Our simulations with virtual species showed that these metrics were not generally strongly affected by detection issues as long as the bias level was similar among the species of a community (Appendix 2).

Improvement over null models

The improvement of the SR predictions based on our SDMs compared to the null model (null.SR) was considerably higher at the sites at high elevation compared to the sites at mid or low elevation (Figure 4). This pattern seems directly linked to the SR gradient of the modelled species pool along an elevational gradient, showing that most species-rich sites were at mid-elevation and that the SR drastically decreased with elevation for all taxa above ~2000 m.

A similar pattern was observed for the improvement in composition predictions compared to the most informed null model (null.Prev) with the highest improvement at low and high elevations (Figure 4). The improvement in composition predictions compared to the null model based on average SR only (null.SR) was less clear but mostly seems to have decreased with elevation. However, the improvement over the null model based on the average SR was always much higher than that over the most informed null model independent of taxa.

Our simulations with the virtual species showed that the improvement over null models were highly affected by any bias in the initial data that was used for model calibration (Appendix 2).

Discussion

Herein, we presented a range of existing and new evaluation approaches and metrics for community predictions that do not require binary thresholding but that directly compare the probabilistic outputs of stacked species distribution models (S-SDMs) to presence-absence observations of the species occurrences that make up community composition and richness. While we illustrated the use of these metrics in the context of the S-SDM predictions, they were equally applicable for comparing observed binary species assemblages to probabilistic community predictions obtained with any type of modelling framework, such as the joint-SDMs (Warton *et al.* 2015) or the other more dynamic and/or mechanistic approaches (see D'Amen *et al.* 2017). The fact that no thresholding was found to be necessary – which is contrary to the large majority of existing community prediction evaluations– makes the comparisons of studies of different taxa or ecosystems less prone to effects other than those from the models themselves, such as the choice of a thresholding method (e.g., Gastón & García-Viñas 2013; Calabrese *et al.* 2014; Fernandes, Scherrer & Guisan 2018; Scherrer *et al.* 2018).

Community AUC

The cAUC is the most basic of the threshold independent community evaluation methods. In general, our cAUC values were high (>0.85), indicating that for a majority of the sites, a good separation of the species that are present or absent at the site was possible based on the predicted probabilities but that some error was unavoidable in most sites. Therefore, the cAUC gave us a direct indication of the minimum commission/omission error rates (false positives/false negatives) in the models. As the cAUC does not depend on the actual probabilities but rather on their ranking, cAUC was not highly affected by (uniform) detection issues (for detailed explanation see Appendix 2), and as a result might also work reasonably well for evaluating presence-only models.

By analysing the cAUC along elevation, we could see that the models perform best at the two ends of the gradient. This indicates that in the “warmest” and “coldest” environments, the models were much better at correctly predicting which species were present or absent (i.e., high or low predicted probability of occurrence, respectively) and could be interpreted in an ecological sense as the strength of the assembly processes (i.e., habitat filtering or competitive exclusion) caused by the predictors (i.e., abiotic environment). Low cAUCs might hint at the fact that the chosen predictors were not able to explain the observed species compositions. While the performance of the assembly predictions seems directly linked to

the environmental conditions, it is important to acknowledge that their performances also co-varied with the SR of the modelled species pool, although to a much lower degree.

Deviation in species richness

As shown in earlier studies, the sum of all the probabilities in a site gives a good threshold-independent estimation of the species richness of the modelled species pool (Lehmann, Leathwick & Overton 2002; Gelfand *et al.* 2005; Dubuis *et al.* 2011; Calabrese *et al.* 2014). As expected, the average species richness was correct if based on the sum of the probabilities, but the most species-poor and most species-rich sites were over- and under-estimated, respectively (Dubuis *et al.* 2011; D'Amen *et al.* 2015). However, the approach presented here allowed an additional estimation of the probability (p-value) that a value equal to or lower/higher than the observed SR was the result of the predicted probabilities of occurrence. While the absolute differences between predicted and observed species richness were highly dependent on the study system (i.e., modelled species pool, average site species richness), the *p*-values should be more objectively comparable among studies and allow the identification of the sites (e.g., environmental conditions, site species richness) where the differences between observations and predictions are most significant. However, as expected, this metric was quite sensitive to detectability issues (omission of species; Appendix 2), as these lead to an underestimation of the expected species richness. However, this was not problematic as long as the calibration and evaluation datasets shared the same omission rates (e.g., when using cross-validation) but might lead to greater errors when evaluated with independent data (i.e., possibly with a different bias).

Maximisation and probability-sum-based community composition metrics

Our four newly proposed similarity metrics for community composition (i.e., *maxSørensen*, *maxJaccard* and *probSørensen*, *probJaccard*) were highly correlated with the classical binary Sørensen and Jaccard similarity indices. As expected, the maximisation approaches usually showed higher values of the evaluation metrics, illustrating that all the other thresholding techniques (maxTSS, maxKappa and pS-SDM+PRR) rarely found the optimal threshold possible at a site (i.e., per community). Therefore, we think this simple maximisation approach is an efficient way to make studies predicting and analysing communities more comparable as it eliminates the problem of different thresholding choices that make posterior comparisons difficult among different study groups and systems. As a very similar maximisation approach is becoming the standard in single species modelling (e.g., maxTSS or maxKappa; see Guisan, Thuiller & Zimmermann 2017), the community prediction evaluation version proposed here should be "easy-to-apply", well received in community modelling, and adaptable to other similarity metrics (as listed e.g., in Cheetham & Hazel 1969; Legendre & Legendre 1998). As these maximisation approaches mostly depend on the ranking of probabilities rather than the actual values, these metrics are not strongly affected

by systematic omission errors (detectability issues) and might also work well with presence-only data (see Appendix 2 for detailed explanations).

Our new community composition evaluation approach (*probSørensen* and *probJaccard*) based on the sums of predicted probabilities shows promising potential as a truly threshold-independent metric. Furthermore, this metric mostly depends on the relative differences between the sums of probabilities rather than their actual values and is therefore reasonably resilient to a systematic bias in the initial data (see Appendix 2 for detailed explanations).

Improvement over null models

The third approach evaluated the performance improvement of S-SDM assemblage predictions over null models. This approach allowed us to identify if and where (in environmental space) the SDMs outperformed the null models. As the null models were fed with the same species data as the SDMs, all the improvements in the predictions could be directly related to the predictors. This enabled us to determine where the environmental variables had the strongest role in defining both the species richness and species composition. The improvement over null models was the highest at the two ends of the elevational gradient. This finding is in line with both ecological and mathematical expectations. As mentioned before, from an ecological perspective, the biological constraints are expected to be the strongest at the extreme ends of the environmental gradients (Michalet *et al.* 2006; Sexton *et al.* 2009; Louthan, Doak & Angert 2015). From a mathematical perspective, it is obvious that predictions that are close to the mean species richness and most common compositions cannot be improved much more by S-SDMs. Thus, it is important to state that improvement over the null model is not necessarily linked to the model performance. If a site experiences exactly average conditions, both the null models and S-SDMs might predict the species assemblage perfectly, and therefore, no improvement would occur. The average improvement over null models therefore mostly allows us to determine how (un-)uniform the assemblages are, while the analysis of the improvement along gradients allows us to identify the sites where the species richness and composition are most affected by environmental constraints.

In summary, although our results derived from real data showed the value of testing the improvement over null models to obtain more insight into community predictions, our virtual simulations additionally suggested that these metrics were very sensitive to differences in the bias between calibration and evaluation data (see Appendix 2), and care should thus be taken when two distinct datasets with different properties (e.g., different sampling designs) are used. It also confirms the value of complementing any tests of new metrics or approaches on real data (including the many uncertainties) with virtual simulations where the truth is known.

Conclusion

Herein, we presented five groups of approaches that allow threshold-independent evaluation of species assemblage (community) predictions. By applying these metrics to evaluate community predictions of four real species groups and virtual species, we were able to illustrate their use and identify issues that need further investigation. As with all evaluation approaches, each had its strengths and weaknesses, many of which will require further testing, yet they all offered the great advantage of providing less biased estimates of model performance than the previously used metrics that require thresholding. Furthermore, we also illustrated how the metrics could provide more robust insights into the strengths of the assembly processes that are driven by abiotic environmental predictors. Importantly, these metrics generally allowed a more straightforward comparison of the model performances among studies, as they did not depend on any thresholding choices (which are either related to the prevalence of the study species or of each study site). We herein advocate for the use of a combination of threshold-independent evaluation metrics such as cAUC, maximisation approaches for similarity indices and the improvement over null models to communicate the prediction accuracy of species assemblages rather than using a single or several community metrics and approaches based on probability thresholding. This development would be similar to the use of multiple single species evaluation metrics where threshold independent metrics such as AUC or maxTSS are currently the standard.

Acknowledgements

This study was supported by the Swiss National Science Foundation (SESAM'ALP project, grant nr 31003A-1528661 and the INTEGRALP Project, grant nr CR23I2_162754) to AG. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing at the SIB Swiss Institute of Bioinformatics.

Authors' contribution

DS and AG conceived the ideas; DS and HM developed the metrics and ran the analysis; DS led the writing and all authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

Functions to calculate all the presented evaluation metrics and the associated S-SDMs are available in the ecospat r package (v3.1; Di Cola *et al.* 2017) on GitHub (ecospat.CCV;

https://doi.org/10.5281/zenodo.3466637). All species and environmental data are available on Dryad repository: https://doi.org/10.5061/dryad.8sf7m0ch5 (Scherrer, Mod & Guisan 2019).

References

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.

Baroni-Urbani, C. & Buser, M.W. (1976) Similarity of binary data. *Systematic Zoology,* **25,** 251-259.

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.

Breiner, F.T., Nobis, M., Bergamini, A. & Guisan, A. (2018) Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution,* **9,** 802-808.

Broenniman, O., Di Cola, V. & Guisan , A. (2017) ecospat: Spatial Ecology Miscellaneous Methods.

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.

Cheetham, A.H. & Hazel, J.E. (1969) Binary (presence-absence) similarity coefficients. *Journal of Paleontology,* **43,** 1130-1136.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* **20,** 37-46.

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.

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.

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

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F.T., D'amen, M., Randin, C., Engler, R., Pottier, J., Pio, D. & Dubuis, A. (2017) ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography,* **40,** 774-787.

Dorazio, R.M., Royle, J.A., Soderstrom, B. & Glimskar, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology,* **87,** 842-854.

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.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology Evolution and Systematics,* **40,** 677-697.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology,* **77,** 802-813.

Fernandes, R.F., Scherrer, D. & Guisan, A. (2018) How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach. *Ecological Informatics,* **48,** 125-134.

Fernandes, R.F., Scherrer, D. & Guisan, A. (2019) Effects of simulated observation errors on the performance of species distribution models. *Diversity and Distributions,* **25,** 400-413.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology,* **43,** 393-404.

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

Gastón, A. & García-Viñas, J.I. (2013) Evaluating the predictive performance of stacked species distribution models applied to plant species selection in ecological restoration. *Ecological Modelling,* **263,** 103-108.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A. & Rebelo, A.G. (2005) Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society Series C-Applied Statistics,* **54,** 1-20.

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.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters,* **8,** 993-1009.

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

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.

Harrell Jr, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. & Rosati, R.A. (1982) Evaluating the yield of medical tests. *JAMA,* **247,** 2543-2546.

Harris, D.J. (2014) Building realistic assemblages with a Joint Species Distribution Model. bioRxiv.

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution,* **6,** 465-473.

Jaccard, P. (1901) *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Kery, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography,* **37,** 1851-1862.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier.

Lehmann, A., Leathwick, J.R. & Overton, J.M. (2002) Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity and Conservation,* **11,** 2217-2238.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. & Aulagner, S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling,* **90,** 39-52.

Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J. & Guisan, A. (2010) Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. *Biological Conservation,* **143,** 2647-2657.

Louthan, A.M., Doak, D.F. & Angert, A.L. (2015) Where and When do Species Interactions Set Range Limits? *Trends in Ecology & Evolution,* **30,** 780-792.

Marrocco, C., Duin, R.P. & Tortorella, F. (2008) Maximizing the area under the ROC curve by pairwise feature combination. *Pattern Recognition,* **41,** 1961-1974.

Mateo, R.G., Mokany, K. & Guisan, A. (2017) Biodiversity Models: What If Unsaturation Is the Rule? *Trends in Ecology & Evolution,* **32,** 556-566.

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

Michalet, R., Brooker, R.W., Cavieres, L.A., Kikvidze, Z., Lortie, C.J., Pugnaire, F.I., Valiente-Banuet, A. & Callaway, R.M. (2006) Do biotic interactions shape both sides of the humped-back model of species richness in plant communities? *Ecology Letters,* **9,** 767-773.

Nenzen, H.K. & Araujo, M.B. (2011) Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling,* **222,** 3346-3354.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters,* **20,** 561-576.

Pellissier, L., Pradervand, J.-N., Pottier, J., Dubuis, A., Maiorano, L. & Guisan, A. (2012) Climate-based empirical models show biased predictions of butterfly communities along environmental gradients. *Ecography,* **35,** 684-692.

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.

Pradervand, J.N., Pellissier, L., Rossier, L., Dubuis, A., Guisan, A. & Cherix, D. (2011) Diversity of bumblebees (Bombus Latreille, Apidae) in the alps of the canton Vaud (Switzerland). *Mitteilungen der Schweizerischen Entomologischen Gesellschaft,* **84,** 45-66.

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Scherrer, D., D'Amen, M., Fernandes, R.F., Mateo, R.G. & Guisan, A. (2018) How to best threshold and validate stacked species assemblages? Community optimisation might hold the answer. *Methods in Ecology and Evolution,* **9,** 2155-2166.

Scherrer, D., Mod, H.K. & Guisan, A. (2019) Data from: How to evaluate community predictions without thresholding? *Methods in Ecology and Evolution,* doi: 10.5061/dryad.8sf7m0ch5

Sexton, J.P., McIntyre, P.J., Angert, A.L. & Rice, K.J. (2009) Evolution and Ecology of Species Range Limits. *Annual Review of Ecology, Evolution, and Systematics,* **40,** 415-436.

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.

Staniczenko, P., Sivasubramaniam, P., Suttle, K.B. & Pearson, R.G. (2017) Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecology Letters,* **20,** 693-707.

Strobl, C., Malley, J. & Tutz, G. (2009) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods,* **14,** 323–348.

Swets, J.A. (1988) Measuring the Accuracy of Diagnostic Systems. *Science,* **240,** 1285-1293.

Thuiller, W., Georges, D., Engler, R. & Breiner, F. (2016) biomod2: Ensemble Platform for Species Distribution Modeling.

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

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K.C. (2015) So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution,* **30,** 766-779.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.-A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.-C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews,* **88,** 15-30.

Wolda, H. (1981) Similarity indices, sample size and diversity. *Oecologia,* **50,** 296-302.

Table 1: Overview of the different community evaluation approaches. The sensitivity to detectability indicates how strongly the approach is influenced by uncertainty in the initial absence data. For details on the classification of sensitivity to detectability, see Appendix 2.

Approach	**Principe**	**Sensitivity to detectability**
Community AUC	The ability to distinguish species presence/absence based on the ranking of the predicted probabilities of occurrence of all species in a given site.	low
Deviation in SR	The probability that the observed species richness is not different from the expectation based on the probabilities of occurrence of various species.	high
MaxSørensen MaxJaccard	Maximisation of the Sørensen/Jaccard indices by optimal per site threshold selection.	low
probSørensen probJaccard	Based on the sum of probability of the species observed in contrast to the species observed and falsely predicted to be present.	low
Improvement over null models	Ratio to correctly get the SR/composition based on the predicted probabilities compared to null-models.	high

Table 2. Statistics of the used community datasets. Prevalence and species richness (SR) were calculated after removing the species with 10 occurrences or less.

taxonomic group	*n* of species (orig.)	*n* of modelled species	% of modelled species	*n* of sites	prevalence (mean ± sd; across species)	SR (mean ± sd; across sites)
Plants	795	296	37.2	909	0.08 ± 0.08	24.3 ± 14.0
Butterflies	140	78	55.7	208	0.22 ± 0.14	17.5 ± 8.6
Grasshoppers	41	21	51.2	202	0.25 ± 0.19	5.2 ± 3.4
Bumble bees	29	20	68.9	202	0.25 ± 0.16	5.1 ± 2.7

Box 1: Example of an observed species community and those predicted by the SDMs and the different null models. The presented example has a species pool (*N*) of five species, an average species richness ($\overline{SR}$) of two species and the five species have a prevalence (*Prev*) of 0.4, 0.6, 0.1 0.2, 0.7. Improvement is counted as the probability of obtaining the composition$_{SDM.prediction}$ / probability to obtain the composition$_{null.model}$

	Sp1	Sp2	Sp3	Sp4	Sp5	Probability to get composition	Improvement
Observation	0	1	0	0	1	-	
SDM prediction	0.1	0.8	0.3	0.2	0.9	0.9*0.8*0.7*0.8*0.9 = **0.363**	
null.SR (eqn. 4)	0.4	0.4	0.4	0.4	0.4	0.4^2 * 0.6^3 = **0.035**	**10.5**
null.Prev (eqn. 5)	0.4	0.6	0.1	0.2	0.7	0.6*0.6*0.9*0.8*0.7 = **0.181**	**2**

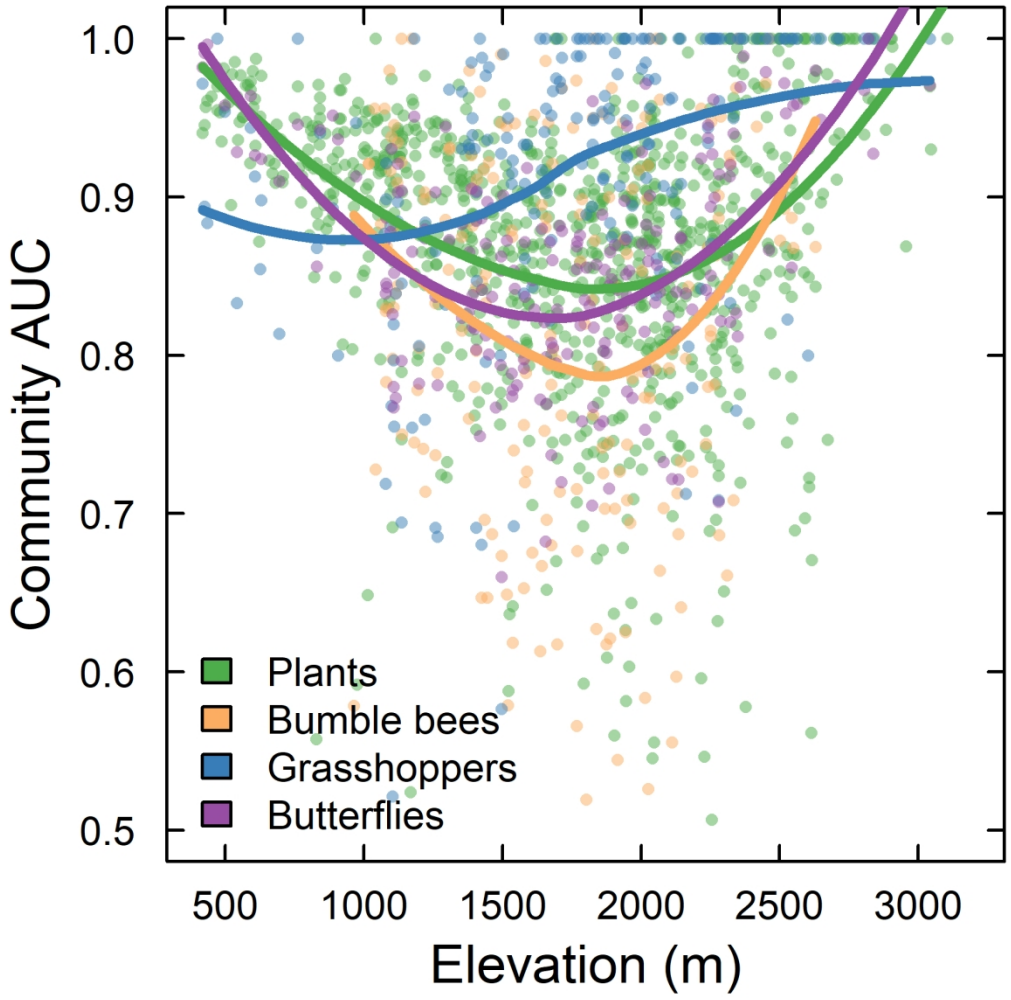

Figure 1: Community AUC for the species groups along the elevation gradient of the study area. Each dot represents a site, and the solid lines are the smoothed mean.

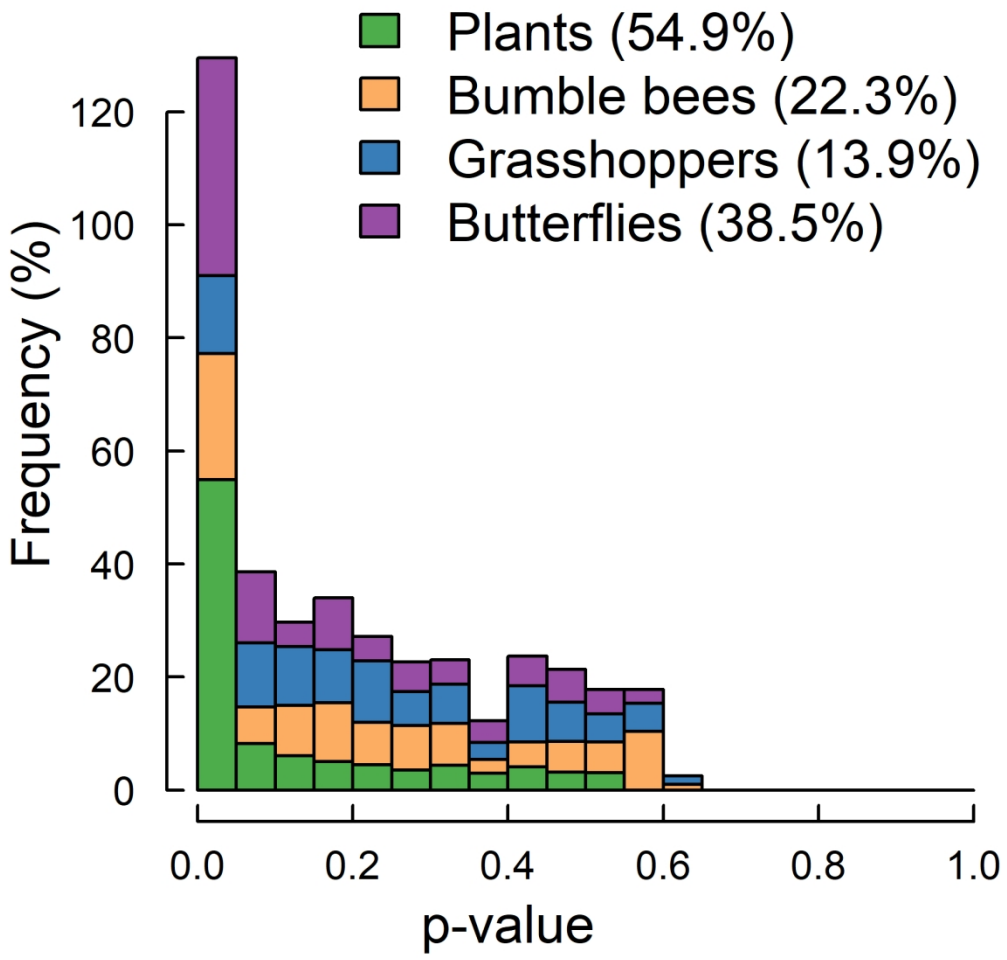

Figure 2: Histogram of the p-values for the probability that a value equal or higher/lower than the observed SR is the result of the predicted probabilities of occurrence. The percentage numbers in the brackets indicate the proportion of sites with significant differences between the observed and expected values ($p<0.05$; H0 no difference between observed and expected SR).

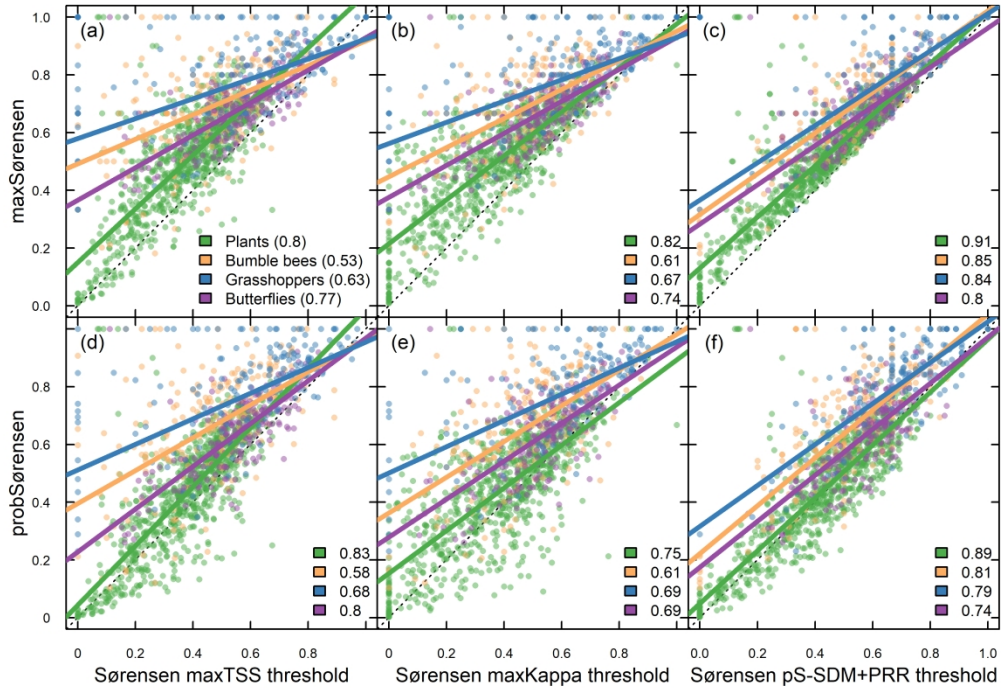

Figure 3: Correlation of the Sørensen similarity metrics based on binary data with threshold-independent Sørensen similarity metrics. The top row shows the correlation of binary metrics with maxSørensen, and the bottom row shows the correlation of binary metrics with probSørensen. The binary Sørensen similarity of panels a and d is based on a "species-specific" maxTSS threshold, of panels b and e on a "species-specific" maxKappa threshold, and of panels c and f on a "site-specific" probability ranking rule. The numbers in the legends indicate the Spearman correlation coefficient between binary and threshold independent metrics.

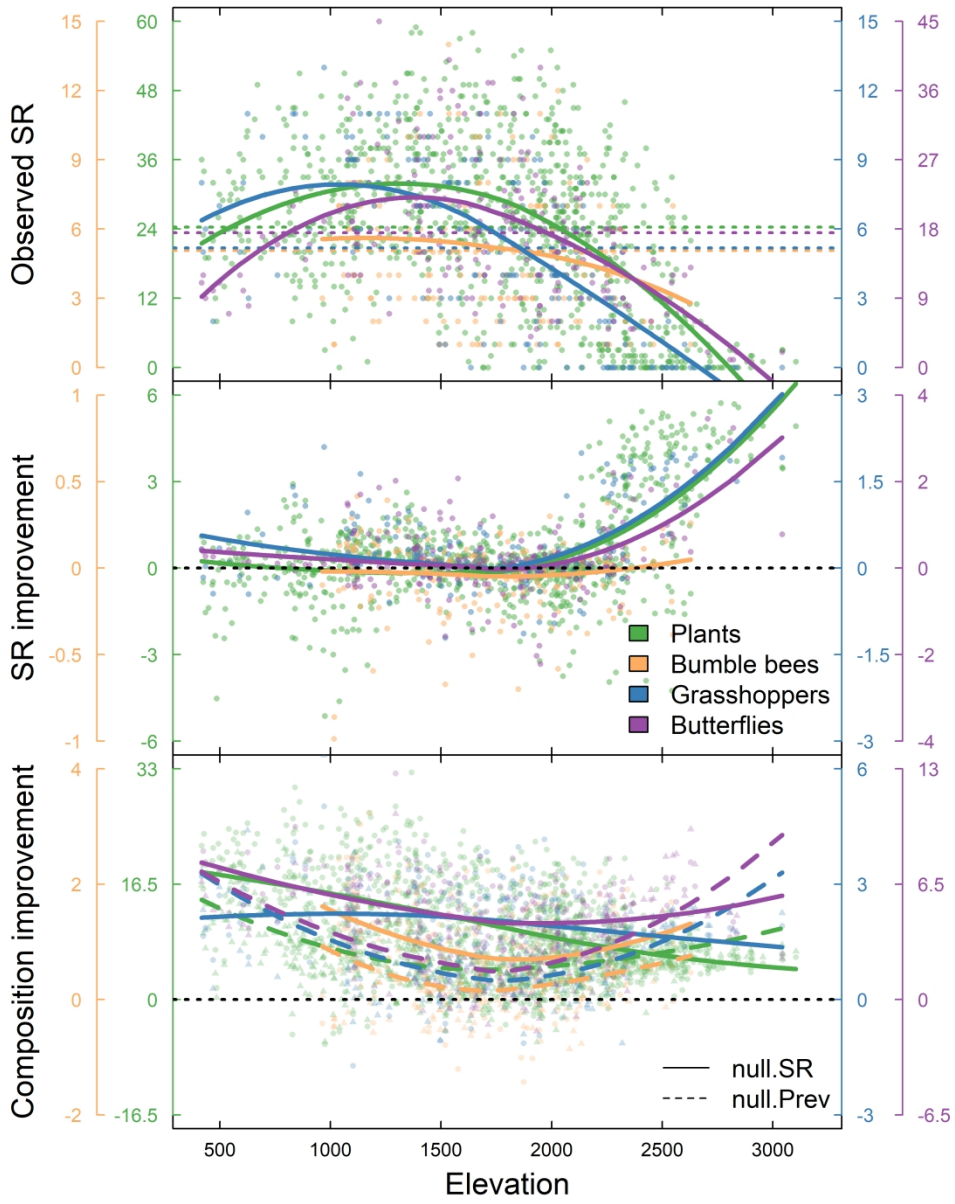

Figure 4: Observed species richness (SR, based on the modelled species pool; a) with dotted lines indicating the mean SR per taxa, improvement of SDM prediction over the null model of species richness (b) and improvement of SDM prediction over null models of species composition (c) along the elevation gradient for all taxa. Improvements are above one null model (null.SR) for SR and above the two null models (null.SR, null.Prev) for composition. All y-axes are logarithmic (log-fold change) with a base of 10.