---article under review---

Title: Theoretical solutions for the evaluation of discrimination capacity of species distribution models

Authors: Boris Leroy[1*], Robin Delsol[1,2], Bernard Hugueny[1], Christine N. Meynard[3,4], Chéïma Barhoumi[1,2,5], Morgane Barbet-Massin[2] & Céline Bellard[6]

Short title: Species distribution models discrimination capacity

Word count: 4714 (without references) ; 5913 (with references)

*Corresponding author, leroy.boris@gmail.com

[1] *Unité Biologie des Organismes et Ecosystèmes Aquatiques (BOREA, UMR 7208), Muséum national d'Histoire naturelle, Université Pierre et Marie Curie, Université de Caen Basse-Normandie, CNRS, IRD, Sorbonne Universités, Paris, France.*

[2] *Ecologie, Systématique & Evolution, UMR CNRS 8079, Univ. Paris-Sud, Orsay Cedex FR-91405, France*

[3] *Virginia Institute of Marine Science, College of William & Mary, P.O. Box 1346 Gloucester Point, VA 23062*

[4] *INRA, UMR CBGP, Montpellier, France*

[5] *Institut des Sciences de l'Evolution de Montpellier, UMR CNRS 5554, Univ. De Montpellier, Montpellier Cedex 05 FR-34095, France*

[6]*Department of Genetics, Evolution and Environment, Center for Biodiversity and Environment Research, University College of London, UK*

ABSTRACT

The discriminating capacity (i.e., ability to correctly classify presences and absences) of species distribution models (SDMs) is commonly evaluated with metrics such as the Area Under the Receiving Operating Characteristic Curve, the Kappa statistic and the True Skill Statistic (TSS). AUC and Kappa have been repeatedly criticised, but the TSS has fared relatively well since its introduction, mainly because it has been considered as independent of prevalence. In addition, discrimination metrics have been contested because they should be calculated on presence-absence data, but are often used on presence-only or presence-background data. With examples and simulations, we demonstrate here that TSS is misleading because of its dependence on prevalence. We then introduce an alternative set of metrics –based on similarity indices– that represents a theoretical solution to evaluate model discrimination capacity. In the ideal situation where species and sample prevalence (respectively, true and sampled ratio of sites occupied by species) are equal –corresponding to a perfectly random presence-absence sampling or evaluation of virtual species, similarity indices are appropriate to evaluate model discrimination capacity. In situations where sample and species prevalence are not equal –corresponding to non-random presence-absence samplings or presence-background samplings, similarity metrics can correctly evaluate discrimination capacity provided that an estimate of species prevalence can be obtained. However, estimates of species prevalence are challenging to obtain for presence-absence schemes and unlikely to be possible for presence-pseudoabsence schemes. Our recommendations are therefore to use similarity metrics in the specific case of virtual species, and, until a robust framework to estimate species prevalence is developed, focus on reliability metrics in other cases.

INTRODUCTION

During the last decades, species distribution models (SDMs) have become one of the most commonly used tools to investigate the effects of global changes on biodiversity. Specifically, SDMs are widely used to explore the potential effects of climate change on the distribution of species of concern (Gallon et al., 2014), to anticipate the spread of invasive species (Bellard et al., 2013), but also to prioritise sites for biodiversity conservation (Leroy et al., 2014). Therefore, conservation managers increasingly rely on SDMs to implement conservation strategies and policies to mitigate the effects of climate change on biodiversity (Guisan et al., 2013). There are various methodological choices involved in the application of SDMs (e.g., data type and processing, variable choice, choice of resolution, algorithms, protocols, global climate models, greenhouse gas emission scenarios), which make them particularly difficult to interpret, compare, or assess. However, evaluation of their predictive accuracy is probably a common step to most SDM studies across methodological and technical choices. This evaluation allows us to quantify model performance in terms of how well predictions match observations, which is a fundamental and objective part of any theoretical, applied or methodological study.

To evaluate model predictive performance, the occurrence dataset is often partitioned into two subsets (one for calibrating models, and one for testing) and predictions are assessed in terms of whether or not they fit observations using various accuracy metrics (Araújo et al., 2005), a method called cross-validation. Other approaches include calibrating on the full dataset and testing on an independent dataset, or, when the modelled species is a virtual, *in silico*, species (e.g., for testing methodological aspects), directly comparing the predicted distribution with the known true distribution (Leroy et al., 2015). Accuracy metrics can be divided into two groups: discrimination vs. reliability metrics (Liu et al., 2009; Pearce et al., 2000). Discrimination metrics measure classification rates, i.e. the capacity of SDMs to distinguish

correctly between presence and absence sites. Reliability metrics measure whether the predicted probability is an accurate estimate of the likelihood of occurrence of the species at a given site. Here, we focus on the issues of discrimination metrics, since they are often used in the SDM literature to test model robustness; however we stress the importance of evaluating reliability (see Meynard & Kaplan 2012 as well as Liu *et al.* 2009), for example with the Boyce index which is probably the most appropriate reliability metric (Boyce et al., 2002; Cola et al., 2016; Hirzel et al., 2006).

Discrimination metrics rely on the confusion matrix, i.e., a matrix comparing predicted versus observed presences and absences (Table 1). Such discrimination metrics have largely been borrowed from other fields of science, such as medicine and weather forecasting, rather than being specifically developed for SDM studies (Liu et al., 2009). Three classification metrics stand out in the SDMs literature: Cohen's Kappa, the Area Under the receiver operating characteristic curve (AUC), and the True Skill Statistic (TSS). The AUC was introduced in ecology by Fielding & Bell (1997) (2,821 citations on Web of Science in June 2017), but has since repeatedly been criticised (Jiménez-Valverde, 2012; Lobo et al., 2010, 2008) because of its dependence on prevalence (i.e., the proportion of recorded sites where the species is present). Cohen's Kappa has also been repeatedly criticised for the same reason (Allouche et al., 2006; Lobo et al., 2010; McPherson et al., 2004). TSS, on the other hand, has fared relatively well since its introduction by Allouche *et al.* (2006) (719 citations in June 2017), mainly because it had been shown as independent of prevalence. However, this claim has recently been questioned because of a flawed testing design (Somodi et al., 2017). More recently, all of these metrics have been contested because they should be calculated on presence-absence data, but are often used on presence-only or presence-background data, i.e. data with no information on locations where species do not occur (Jarnevich et al., 2015; Somodi et al., 2017; Yackulic et al., 2013). In these cases, False Positives and True Negatives

(Table 1) are unreliable, which led Li & Guo (2013) to propose alternative approaches, specifically designed for presence-background models. They proposed the use of F_{pb}, a proxy of an *F*-measure ("the weighted harmonic average of precision and recall", Li & Guo (2013)) based on presence-background data, and F_{cpb}, a prevalence-calibrated proxy of an *F*-measure based on presence-background data. Despite the apparent relevance of Li & Guo's (2013) metrics (13 citations in June 2017), the field is still dominated by metrics that have been repeatedly criticised, such as AUC and Kappa, or TSS (*e.g.*, D'Amen *et al.* 2015; Jarnevich *et al.* 2015; Mainali *et al.* 2015).

With this forum, our aim is twofold: (1) illustrate with examples and simulations that, contrary to early claims, TSS is dependent on prevalence as well as other metrics, and (2) evaluate an alternative set of metrics based on similarity indices, in line with the approaches proposed by Li & Guo (2013). Similarity indices assess the similarity of observed and predicted distributions, and can be partitioned into two components to evaluate model characteristics: Over Prediction Rate (OPR) and Unpredicted Presence Rate (UPR). We compare the performance of TSS and similarity-derived metrics on three modelling situations corresponding to the most common modelling setups, depending on the interplay between species and sample prevalence (see below). We finally discuss the applicability of these discrimination metrics in SDM studies and provide recommendations.

SPECIES AND SAMPLE PREVALENCE

Here we will define *species prevalence* as the ratio between the species area of occupancy (AOO, i.e., the area within the distribution of a species that is actually occupied) and the total study area (see Rondinini et al. 2006 for definitions). For example, if the study area encompasses Europe and we have divided the study area into 1-km grid cells, and if we are studying a species that occupies only 15% of those grid cells its prevalence would be 0.15 (or 15%). Notice that species prevalence will vary depending on the resolution of the gridded data and on the extent of the study area. In practice, however, species prevalence is never known, because the true AOO is generally not known, except for the specific case of virtual species (Leroy et al., 2015). Hence, for real species, only the *sample prevalence* is known, which is the proportion of sampled sites in which the species has been recorded. Meynard and Kaplan (2012) showed with virtual species that sample prevalence should be similar to species prevalence to produce accurate predictions. However, in practice, we expect sample prevalence to be different from species prevalence, unless the sampling of presences and absences is perfectly random throughout the entire study area. Indeed, samplings of species presences are generally spatially biased (Phillips et al., 2009; Varela et al., 2014). For example, ecologists look for their species of interest in sites where they have a sense a priori that they will find it, which will inevitably result in a mismatch between sample and species prevalence. Furthermore, a substantial proportion of SDM studies rely on presence-only modelling techniques, which requires to sample ‘pseudo-absence’ or ‘background’ points (hereafter called pseudo-absences). In such cases the sample prevalence is artificially defined by the number of chosen pseudo-absences, and is thus unlikely to be equal to species prevalence.

Neither species prevalence nor sample prevalence should influence accuracy metrics. In the following, we investigate three different cases corresponding to the most common situations

of SDM evaluation. First, we investigate the ideal 'presence-absence' case where species prevalence is equal to sample prevalence; this case corresponds to the evaluation of SDMs based on virtual species, or on well-designed presence-absence samplings. Second, we investigate 'presence-absence' situations where sample prevalence differs from species prevalence. Last, we investigate 'presence only' situations where sample prevalence differs from species prevalence.

PRESENCE-ABSENCE, SPECIES PREVALENCE = SAMPLE PREVALENCE

In this first case, we define the sample confusion matrix as perfectly proportional or equal to the true confusion matrix, i.e. the entire predicted species distribution is compared to the true species distribution. In practice, this case occurs when samplings are perfectly random throughout the landscape and species detectability is equal to one, or when evaluating SDM performance with virtual species (e.g., Qiao et al., 2015). With this first case we can therefore analyse the sensitivity of discrimination metrics to species prevalence only.

The unexpected dependence of TSS on prevalence

We demonstrate here that TSS is dependent on prevalence with a simple set of illustrations. When species prevalence is very low (and so is sample prevalence), we expect the number of True Negatives (Table 1) to be disproportionately high. In these cases, specificity will approximate zero, and therefore TSS values will be approximately equal to sensitivity (Table 2). As a result, TSS values can be high even for models that strongly overpredict distributions. Figure 1 represents graphically some examples of how overprediction and underprediction play into TSS performance. For example, Fig. 1a shows a model that strongly overpredicts the distribution, therefore producing 300% more False Positives than True Positives, and yet TSS is close to 1 (Fig. 1a, TSS=0.97). Such a high value can in turn be produced by a model which correctly predicts the true distribution with few overpredictions (Fig. 1b, TSS = 1.00). In

addition, the over-predicting model (Fig. 1a) will also have higher TSS values compared to a model that only missed 15% of presences (Fig. 1c, TSS=0.85). Furthermore, for identically-performing models, if sample prevalence decreases (from 0.25 to 0.01), then the proportion of True Negatives is increased, and consequently TSS values increased from 0.60 to 0.70 (Fig. 1d-f). TSS values can therefore be artificially increased by decreasing sample prevalence. As an unexpected consequence, for two species with different AOO in the study area (thus different sample prevalence), the species with the smaller distribution will be considered better predicted than the one with a larger distribution (Fig. 1d-f).

To summarise, TSS values can be misleading in situations where the number of True Negatives is high by (i) not penalising overprediction and (ii) assigning higher values to species with smaller prevalence for identical discrimination accuracy. These flaws can be strongly problematic for ecologists, and during SDM performance evaluation it is generally preferable to assume that overprediction should be equivalent to underprediction (e.g., Lawson et al., 2014). We therefore conclude that TSS is prone to similar shortcomings as AUC and Kappa when it comes to its dependence on sample prevalence and AOO.

Similarity metrics as an alternative

To avoid these shortcomings, we propose to focus the evaluation metrics on only three parts of the confusion matrix (Table 1): True Positives, False Positives and False Negatives, neglecting the True Negatives that could be easily inflated. In particular, we seek to maximise True Positives, and minimise both False Positives and False Negatives with respect to True Positives. This definition exactly matches the definition of similarity indices from community ecology, such as Jaccard and Sørensen indices (Table 2). Furthermore, this definition matches the indices identified by Li & Guo (2013) as potential presence-background metrics. The F_{pb} index is in fact equal to twice the Jaccard index (eqn. 13 in Li & Guo 2013), while the F index is equal to the Sørensen index of similarity (eqn. 4 in Li & Guo 2013).

Similarity indices have two main benefits. First, their conceptual basis is easy to understand: they measure the similarity between predictions and observations. A value of 1 means predictions perfectly match observations, without any False Positive or False Negative. A value of 0 means that none of the predictions matched any observation. The lower the similarity value, the higher the number of False Positives and False Negatives, proportionally to the number of True Presences. Second, as they do not include True Negatives, they are not biased by a disproportionate number of True Negatives. In return, they do not estimate the capacity of models to correctly predict absences. To illustrate this, we calculated Jaccard and Sørensen indices of similarity on the same examples as above. Both Jaccard and Sørensen indices accurately discriminated between highly over-predicting and well performing models (Fig. 1a-c). In addition, when species prevalence was artificially increased for identical models, both indices remained identical (Fig. 1d-f). The Sørensen index was always greater than Jaccard, because its formula puts greater emphasis on True Positives (Table 2).

Because the specific objectives of SDM studies can be very different (e.g., invasion monitoring versus habitat identification for threatened species), in a particular context we may be more interested to assess whether predictions tend to over- or underestimate observations. Such additional information can be obtained with similarity metrics by partitioning them into two components: overprediction rate and unpredicted presence rate (Table 2). The overprediction rate measures the percentage of predicted presences corresponding to false presences, and was already recommended for assessing model overprediction (Barbosa et al., 2013). The unpredicted presence rate measures the percentage of actual presences not predicted by the model, and is also called the false negative rate (Fielding and Bell, 1997). Taken together these metrics provide a full view of model discrimination accuracy and allow interpreting the results in the specific context of the study.

Demonstration based on simulations

To validate these theoretical demonstrations, we performed simulations of the metrics for three case studies with different performances: a first model with 40% overprediction and 40% underprediction, a second model with 40% underprediction and no overprediction, and a third model with 40% overprediction and no underprediction. The first case addresses a predicted range that is shifted with respect to the real one; the second and third cases address situations where the predicted range is, respectively, smaller or larger than the real one. For each model, we predicted the distribution range of theoretical species with different prevalence (from 0.01 to 0.60 with a step of 0.01) over an area of 100 000 pixels. Then, for each species, we randomly sampled 500 presences in the total area and a number of absences verifying the condition that the sample prevalence is equal to species prevalence. We repeated this procedure five times. For each repetition, we calculated the True Skill Statistic and the Jaccard index (R scripts available at https://github.com/Farewe/SDMMetrics).

Our results (Figure 2) showed that TSS had higher values at lower prevalence for cases that overpredicted species distributions, but not for cases that only underpredicted distributions (Figure 2a). This result confirms our expectation that TSS does not penalise overprediction at low prevalence. Jaccard values, on the other hand, remained similar regardless of species prevalence (Figure 2b). These results confirm that in the ideal situation where species prevalence = sample prevalence, the Jaccard index of similarity is a more appropriate metric of model discrimination capacity.

PRESENCE-ABSENCE, SPECIES PREVALENCE ≠ SAMPLE PREVALENCE

Effect of the mismatch between sample and species prevalence

When sample prevalence is different from species prevalence, the ratio of sampled absences over sampled presences is different from the ratio of true presences over true absences. For example, if too many absences are sampled (sample prevalence lower than species prevalence), then the numbers of False Positives and True Negatives will be too large

compared to True Negatives and False Positives. The major consequence of this mismatch is that any metric comparing sampled presences and absences will not reflect true model performance, unless it contains a correction factor for the mismatch between sample and species prevalence. Note, however, that metrics focusing only on sampled presences (omitting sampled absences) will not be affected by this bias (for example, sensitivity or rate of unpredicted presences will not be affected). We propose a correction factor for the Jaccard index in the following equations:

$$\mathrm{Jaccard}_{\mathrm{pc}} = \frac{TP}{TP + x \times FP + FN} \quad (1)$$

$\mathrm{Jaccard}_{\mathrm{pc}}$ is the prevalence-calibrated index of Jaccard, and x is the correction factor:

$$x = \frac{P}{A} \times \frac{1 - prev_{\mathrm{sp}}}{prev_{\mathrm{sp}}} \quad (2)$$

where P and A are the respective number of sampled presences and absences, and $prev_{sp}$ is an estimate of species prevalence (see section **ESTIMATIONS OF SPECIES PREVALENCE**).

Similarly, the over-prediction rate can be corrected:

$$\mathrm{OPR}_{\mathrm{pc}} = \frac{x \times FP}{TP + x \times FP} \quad (3)$$

Demonstration based on simulations

To validate our propositions, we compared the behaviour of $Jaccard_{\mathrm{pc}}$ with those of Jaccard and TSS for simulated data. We used the same case studies as in the first section and the same species in an area of 100 000 pixels. We sampled presences and absences with the condition of a mismatch between sample prevalence and true species prevalence. For each species, we sampled 500 presences and three sets of absences (500, 1 000 and 10 000), resulting in three sets of sample prevalence (0.50, 0.33 and 0.05). We repeated this procedure five times, and calculated the three indices for all the samples. To calculate $Jaccard_{\mathrm{pc}}$, we used the true species prevalence in order to provide a validation of the metric in ideal conditions.

As expected, our results showed that when sample prevalence was not equal to species prevalence, both the True Skill Statistic and uncorrected Jaccard index were dependent on species prevalence, with higher values at lower prevalence (Figure 3a-b). Metrics for cases with only underprediction were unaffected, because they do not compare sampled presences and absences. $\text{Jaccard}_{\text{pc}}$, on the other hand, showed no dependence on species prevalence across all simulations, except for a larger uncertainty at lower prevalence due to smaller sample size (Figure 3c). Two main conclusions emerge from these results. First, without correction for the mismatch between sample and species prevalence, none of the classical discrimination metrics can be used, because of their dependence on species prevalence. Second, model discrimination capacity can be accurately estimated with an index such as $\text{Jaccard}_{\text{pc}}$, provided that a correct estimate of species prevalence can be obtained.

PRESENCE-PSEUDOABSENCE OR PRESENCE-BACKGROUND, SPECIES PREVALENCE ≠ SAMPLE PREVALENCE

Pseudoabsences or background points

In presence-pseudoabsence schemes, the correction to $Jaccard_{pc}$ remains correct, but a further bias is added by the fact that pseudo-absence points may be actual presence points. The number of pseudo-absences corresponding to presence points can be estimated if pseudo-absences are drawn randomly, on the basis of an estimate of species prevalence. Pseudo-absence errors affect our estimation of False Positives in $\text{Jaccard}_{\text{pc}}$ by generating "False False Positives" (FFP), *i.e.* predicted presences corresponding to actual presences but sampled as pseudo-absences. We can estimate the total number of potential FFP of a model in the study area with the following equation:

$$\text{FFP}_{\text{total}} = (1 - \text{UPR}) \times (prev_{\text{sp}} \times N - P) \quad (4)$$

where UPR is the Underprediction Rate (Table 2) and therefore (1 – UPR) is the rate of predicted presences that corresponds to true presences, N is the total number of sites in the study area, and P the number of sampled presences. False Positives in pseudoabsence samples can then be corrected by subtracting the relative proportion of FFP:

$$FP_{\mathrm{pa}} = A \times \frac{\left(\frac{FP_{\mathrm{sample}}}{A} \times (N - P) - FFP_{\mathrm{total}}\right)}{(1 - prev_{\mathrm{sp}}) \times N} \quad (5)$$

where FP_{sample} is the sampled number of False Positives and P is the number of sampled presences. Finally, this correction allows calculating prevalence-pseudoabsence-calibrated Jaccard and OPR indices:

$$\mathrm{Jaccard}_{\mathrm{ppac}} = \frac{TP}{TP + x \times FP_{\mathrm{pa}} + FN} \quad (6)$$

$$\mathrm{OPR}_{\mathrm{ppac}} = \frac{x \times FP_{\mathrm{pa}}}{TP + x \times FP_{\mathrm{pa}}} \quad (7)$$

There are two important caveats with $\mathrm{Jaccard}_{\mathrm{ppac}}$ and $\mathrm{OPR}_{\mathrm{ppac}}$. First, they are based on the assumption that pseudoabsences are drawn randomly from the entire study area. Any deviation from random sampling (*e.g.* by defining areas from which pseudoabsences are not sampled such as buffers around presence points) will bias their calculation. Second, the subtraction in (5) implies that the corrected number of False Positives FP_{pa} can be occasionally negative by chance. Multiple repetitions of pseudoabsence samples are therefore necessary to correctly estimate the values of $\mathrm{Jaccard}_{\mathrm{ppac}}$ and $\mathrm{OPR}_{\mathrm{ppac}}$.

Demonstration based on simulations

We tested $Jaccard_{\mathrm{ppac}}$ based on the same set of case studies as in the “presence-absence, species prevalence ≠ sample prevalence section”, except that we sampled pseudoabsences instead of absences randomly across the entire area. We compared $\mathrm{Jaccard}_{\mathrm{ppac}}$ to the previous

metrics TSS, Jaccard and Jaccard_{pc} as well as to the prevalence-calibrated presence-pseudoabsence metric F_{cpb} proposed by Li and Guo (2013) (Table 2).

Our simulations showed that TSS, Jaccard and Jaccard_{pc} were all sensitive to prevalence in presence-pseudoabsence situations where sample prevalence is different from species prevalence (Figure 3a-c). For all three models, TSS, Jaccard and Jaccard_{pc} had highest values at lowest prevalence, whereas Jaccard_{pc} had higher values at both lowest and highest prevalence. Note that in this presence-pseudoabsence case, metrics fail for the model with only underprediction, contrary to the presence-absence case. This is due to the fact that some pseudoabsence points are actual presences that fall within the predicted range, and are thereby erroneously considered as False Positives. These results show that the dependence of classical metrics on species prevalence highlighted in presence-absence cases is further aggravated in presence-pseudoabsence cases.

More surprisingly, we found that the prevalence-calibrated F_{cpb} metric specifically designed for presence-background (Li and Guo, 2013) was also sensitive to very low species prevalence in our simulations (<0.10) (Fig. 3d), but behaved otherwise much better than the uncorrected TSS and Jaccard metrics. These effects were not detected with Jaccard_{ppac} (Fig. 3e). Consequently, Jaccard_{ppac} did not show substantial dependence on prevalence for any of the models, but did show a larger uncertainty at low prevalence due to smaller sample size.

These results illustrate that it is possible to estimate model discrimination capacity in presence-pseudoabsence situations. However, we stress the importance of the underlying assumptions and requirements: (a) an estimate of species prevalence is necessary (see section **ESTIMATIONS OF SPECIES PREVALENCE**); (b) pseudoabsences must be drawn randomly from the entire study area; (c) multiple repetitions are necessary to correctly estimate the metric value, especially at small sample sizes.

ESTIMATIONS OF SPECIES PREVALENCE

The corrections proposed to estimate model discrimination capacity in cases where sample prevalence is different from species prevalence require an estimate of species prevalence. In presence-absences schemes, species prevalence is usually estimated from the sample of presences and absences – however we assumed here that in many situations this estimate may be biased. Besides, in presence-pseudoabsence schemes this estimation is not available. An alternative approach consists in estimating species prevalence from the modelled species distribution (e.g., Li and Guo, 2013; Liu et al., 2013). A threshold of conversion from probability of presence into presence-absence must be defined, e.g. by maximising a specific metric based on the evaluation dataset. On the basis of this threshold, the species potential presence-absence distribution can be mapped onto the entire study area, from which a prevalence can be derived and used as an estimate of species prevalence. Li and Guo (2013) demonstrated that this approach yielded satisfactory results for presence-pseudoabsence based on the F_{pb} index. However, these results were later contested by Liu et al. (2016) who found that neither F_{pb}, nor a TSS-derived metric were able to correctly estimate species prevalence with presence-pseudoabsence data. This inability to estimate species prevalence from presence-pseudoabsence data was expected because an accurate estimation would require strong conditions which are unlikely to be met in reality (see Guillera-Arroita et al., 2015 for a demonstration). On the other hand, for presence-absence data, estimating species prevalence may be feasible, albeit challenging (Lawson et al., 2014). Consequently, a method or framework to reliably estimate species prevalence does not currently exist and, if it is possible, has yet to be established. This inability to estimate species prevalence will severely limit the applicability of metrics for presence-absence and presence-pseudoabsence models where sample prevalence is different from species prevalence.

CONCLUSION AND RECOMMENDATIONS

In this paper, we have demonstrated that evaluating model discrimination capacity (i.e., the capacity to accurately discriminate between presence and absence) depends on the interplay between sample and species prevalence. We studied three general situations that modelers are expected to encounter in their modeling exercises: (i) a presence-absence scheme where sample prevalence is equal to species prevalence – this situation corresponds to perfectly random presence-absence samplings with no detection bias, or to evaluations based on virtual species; (ii) a presence-absence scheme where sample prevalence is different from species prevalence – a likely situation for presence-absence modelling; and (iii) a presence-pseudoabsence scheme where sample prevalence is different from species prevalence – the general case for presence-pseudoabsence or presence-background modelling. We draw three major recommendations from this study.

1. Regardless of the modelling scheme, TSS is biased by species prevalence and cannot be considered as an appropriate discrimination metric, which corroborates other conclusions on this metric (Barhoumi et al., unpublished; Somodi et al., 2017).

2. In presence-absence schemes where sample prevalence is equal to species prevalence, similarity indices (*e.g.* Jaccard) and their derived components (OPR, UPR) appeared unbiased by species prevalence and can thus be used as discrimination metrics. We therefore strongly recommend the use of these metrics in the specific case of virtual species, which are increasingly used to develop and test SDM methodologies (Hattab et al., 2017; Leroy et al., 2015; Li and Guo, 2013; Liu et al., 2016; Meynard and Kaplan, 2013; Miller, 2014; Ranc et al., 2016; Varela et al., 2014).

3. When sample prevalence is different from species prevalence, which is a likely case in presence-absence schemes and the general case in presence-pseudoabsence or presence-background schemes, all of the aforementioned discrimination metrics are biased by species prevalence. We showed that it is possible to reliably estimate discrimination capacity with

corrections. However, these corrections require to estimate species prevalence based on modelled species distributions, which is challenging for presence-absence schemes (Lawson et al., 2014) and unlikely to be done for presence-pseudoabsence schemes (Guillera-Arroita et al., 2015; Liu et al., 2016). Unless a sound framework to estimate species prevalence is developed, this impediment is likely to limit the usefulness of these corrected discrimination metrics in real studies. In this case, we therefore recommend avoiding discrimination metrics for SDM performance evaluation, and rather apply reliability metrics such as the Boyce index for presence-pseudoabsence data (Hirzel et al., 2006) or other continuous calibration metrics for presence-absence data (Lawson et al., 2014).

REFERENCES

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232. doi:10.1111/j.1365-2664.2006.01214.x

Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species–climate impact models under climate change. Glob. Chang. Biol. 11, 1504–1513. doi:10.1111/j.1365-2486.2005.001000.x

Barbosa, A.M., Real, R., Muñoz, a. R., Brown, J. a., 2013. New measures for assessing model equilibrium and prediction mismatch in species distribution models. Divers. Distrib. 19, 1333–1338. doi:10.1111/ddi.12100

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

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.., 2002. Evaluating resource selection functions. Ecol. Modell. 157, 281–300. doi:10.1016/S0304-3800(02)00200-4

Cola, V. Di, Broennimann, O., Petitpierre, B., Breiner, F.T., D'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R.G., Hordijk, W., Salamin, N., Guisan, A., 2016. ecospat: an R package to support spatial analyses and modeling of species niches and distributions. Ecography (Cop.). Early View.

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. Glob. Ecol. Biogeogr. Early View. doi:10.1111/geb.12357

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49. doi:10.1017/S0376892997000088

Gallon, R.K., Robuchon, M., Leroy, B., Le Gall, L., Valero, M., Feunteun, E., 2014. Twenty years of observed and predicted changes in subtidal red seaweed assemblages along a biogeographical transition zone: inferring potential causes from environmental data. J. Biogeogr. 41, 2293–2306. doi:10.1111/jbi.12380

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., Mccarthy, M.A., Tingley, R., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24, 276–292. doi:10.1111/geb.12268

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435. doi:10.1111/ele.12189

Hattab, T., Garzón-López, C.X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., Feilhauer, H., Honnay, O., Kempeneers, P., Schmidtlein, S., Somers, B., Van De Kerchove, R., Rocchini, D., Lenoir, J., 2017. A unified framework to model the potential and realized distributions of invasive species within the invaded range. Divers. Distrib. 1–14. doi:10.1111/ddi.12566

Hirzel, A.H., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Modell. 199, 142–152. doi:10.1016/j.ecolmodel.2006.05.017

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. Bull Soc. Vaud. Sci. Nat 44, 223–270.

Jarnevich, C.S., Stohlgren, T.J., Kumar, S., Morisette, J.T., Holcombe, T.R., 2015. Caveats

for correlative species distribution modeling. Ecol. Inform. 29, 6–15. doi:10.1016/j.ecoinf.2015.06.007

Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21, 498–507. doi:10.1111/j.1466-8238.2011.00683.x

Lawson, C.R., Hodgson, J.A., Wilson, R.J., Richards, S.A., 2014. Prevalence, thresholds and the performance of presence-absence models. Methods Ecol. Evol. 5, 54–64. doi:10.1111/2041-210X.12123

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

Leroy, B., Meynard, C.N., Bellard, C., Courchamp, F., 2015. virtualspecies, an R package to generate virtual species distributions. Ecography (Cop.). Early View.

Li, W., Guo, Q., 2013. How to assess the prediction accuracy of species presence-absence models without absence data? Ecography (Cop.). 36, 788–799. doi:10.1111/j.1600-0587.2013.07585.x

Liu, C., Newell, G., White, M., 2016. On the selection of thresholds for predicting species occurrence with presence-only data. Ecol. Evol. 6, 337–348. doi:10.1002/ece3.1878

Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. J. Biogeogr. 40, 778–789. doi:10.1111/jbi.12058

Liu, C., White, M., Newell, G., 2009. Measuring the accuracy of species distribution models□: a review. 18th World IMACS / MODSIM Congr. Cairns, Aust. 13-17 July 2009 4241–4247.

Lobo, J.M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography (Cop.). 33, 103–114.

doi:10.1111/j.1600-0587.2009.06039.x

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151. doi:10.1111/j.1466-8238.2007.00358.x

Mainali, K.P., Warren, D.L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B.B., Parmesan, C., 2015. Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. Glob. Chang. Biol. Early View. doi:10.1111/gcb.13038

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: Ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823. doi:10.1111/j.0021-8901.2004.00943.x

Meynard, C.N., Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. J. Biogeogr. 40, 1–8. doi:10.1111/jbi.12006

Meynard, C.N., Kaplan, D.M., 2012. The effect of a gradual response to the environment on species distribution modeling performance. Ecography (Cop.). 35, 499–509. doi:10.1111/j.1600-0587.2011.07157.x

Miller, J. a., 2014. Virtual species distribution models: Using simulated data to evaluate aspects of model performance. Prog. Phys. Geogr. 38, 117–128. doi:10.1177/0309133314521448

Pearce, J., Pearce, J., Ferrier, S., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell. 133, 225–245.

Peirce, C.S., 1884. The numerical measure of the success of prediction. Science (80-.). 4, 453–454. doi:10.1126/science.ns-4.93.453-a

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for

background and pseudo-absence data. Ecol. Appl. 19, 181–97.

Qiao, H., Soberón, J., Peterson, T. a., 2015. No Silver Bullets in Correlative Ecological Niche Modeling: Insights from Testing Among Many Potential Algorithms for Niche Estimation. Methods Ecol. Evol. n/a-n/a. doi:10.1111/2041-210X.12397

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., Maiorano, L., 2016. Performance tradeoffs in target-group bias correction for species distribution models. Ecography (Cop.). 1–12. doi:10.1111/ecog.02414

Rondinini, C., Wilson, K. a, Boitani, L., Grantham, H., Possingham, H.P., 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecol. Lett. 9, 1136–45. doi:10.1111/j.1461-0248.2006.00970.x

Somodi, I., Lepesi, N., Botta-Dukát, Z., 2017. Prevalence dependence in model goodness measures with special emphasis on true skill statistics. Ecol. Evol. 863–872. doi:10.1002/ece3.2654

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. Danske Vidensk. Selsk. 1–34.

Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography (Cop.). no-no. doi:10.1111/j.1600-0587.2013.00441.x

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H., Veran, S., 2013. Presence-only modelling using MAXENT: When can we trust the inferences? Methods Ecol. Evol. 4, 236–243. doi:10.1111/2041-210x.12004

Table 1. Confusion matrix used to calculate discrimination metrics.

		Sampled data	
		Presence	Absence
Predicted values	Presence	True Positives (TP)	False Positives (FP)
	Absence	False Negatives (FN)	True Negatives (TN)

Table 2. Existing discrimination metrics. TP = True Positives, FN = False Negatives, FP = False Positives, TN = True Negatives, *P* = number of sampled presences, *A* = number of sampled pseudoabsences, $prev_{sp}$ = estimate of species prevalence.

Metric	Calculation	References
Sensitivity	Sn = TP / (TP+FN)	Fielding & Bell (1997)
Specificity	Sp = TN / (TN+FP)	Fielding & Bell (1997)
True Skill Statistic	TSS = Sn + Sp -1	Peirce (1884), Allouche *et al*. (2006)
Jaccard's similarity index	Jaccard = TP / (FN+TP+FP)	Jaccard (1908)
Sørensen's similarity index	Sørensen = 2TP / (FN + 2TP + FP)	Sørensen (1948), *F* in Li & Guo (2013)
Proxy of *F* measure based on presence-background data	F_{pb} = 2 × Jaccard F_{cpb} = 2 × TP / (FN + TP + c × FP) where $c = P / (prev_{sp} \times A)$	Li & Guo (2013)
Overprediction Rate	OPR = FP / (TP+FP)	Barbosa *et al.* (2013)
Underprediction Rate	UPR = FN / (TP+FN) = 1 – Sn	False Negative Rate in Fielding & Bell (1997)

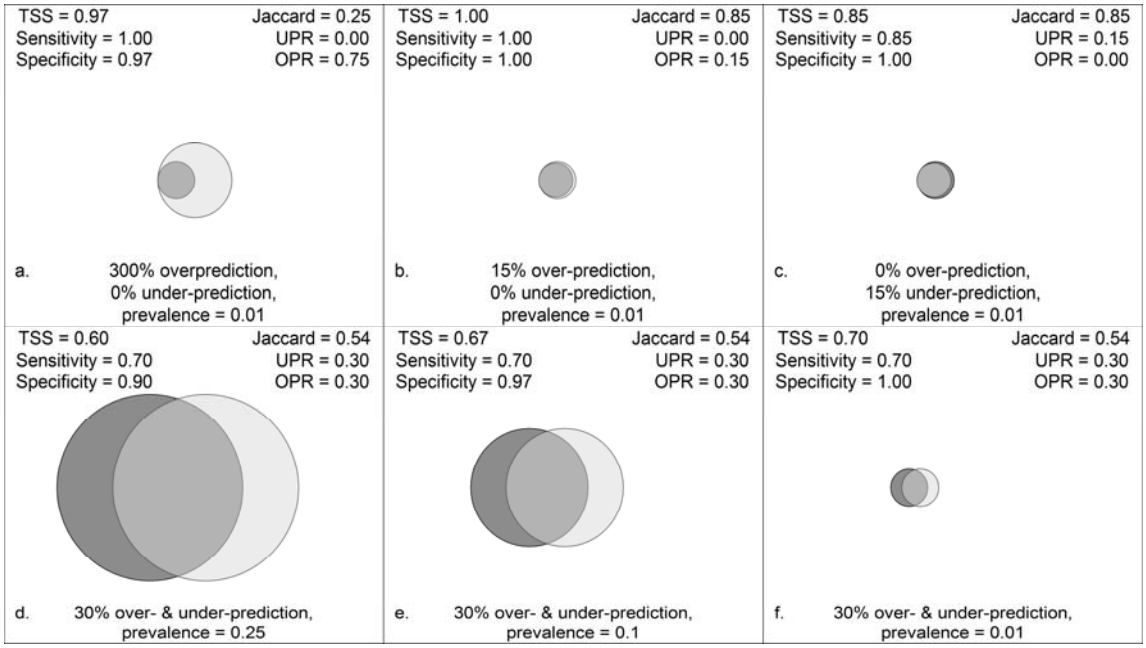

Figure 1. Examples of model performances and associated metrics. The dark grey filled circle represents the proportion of actual presences in the sample. The light grey filled circle represents the proportion of predicted presences in the sample. The overlap between the two circles therefore represents the proportion of actual presences correctly predicted as presences (‘True Positives’), whereas the white area represents the proportion of actual absences correctly predicted as absences (‘True Negatives’). At low prevalence (0.10), TSS does not penalise overprediction: a model that strongly overpredicts distribution (Fig.1a; 300% more False Positive than True Positives) can have a very high TSS (0.97), which is almost equivalent to a model with little overprediction (Fig. 1b, TSS = 1.00). TSS does penalise underprediction (Fig. 1c, TSS = 0.85) much more than overprediction (Fig. 1a-b). For identically-performing models (i.e., similar rates of over- and underprediction), if prevalence decreases (from 0.25 to 0.01) with increasing numbers of True Negatives, TSS values increased from 0.60 to 0.70 (Fig. 1d-f). In other words, for two species with different AOO in a given study area, the species with the smaller distribution have a higher TSS than the one with a larger distribution. Jaccard, on the other hand, accurately discriminates between highly over-predicting and well performing models (Fig. 1a-c). Similarity indices penalise identically over- and underprediction (Fig. 1b-c). In addition, when species prevalence is artificially increased for identical models, both indices remain identical (Fig. 1d-f).

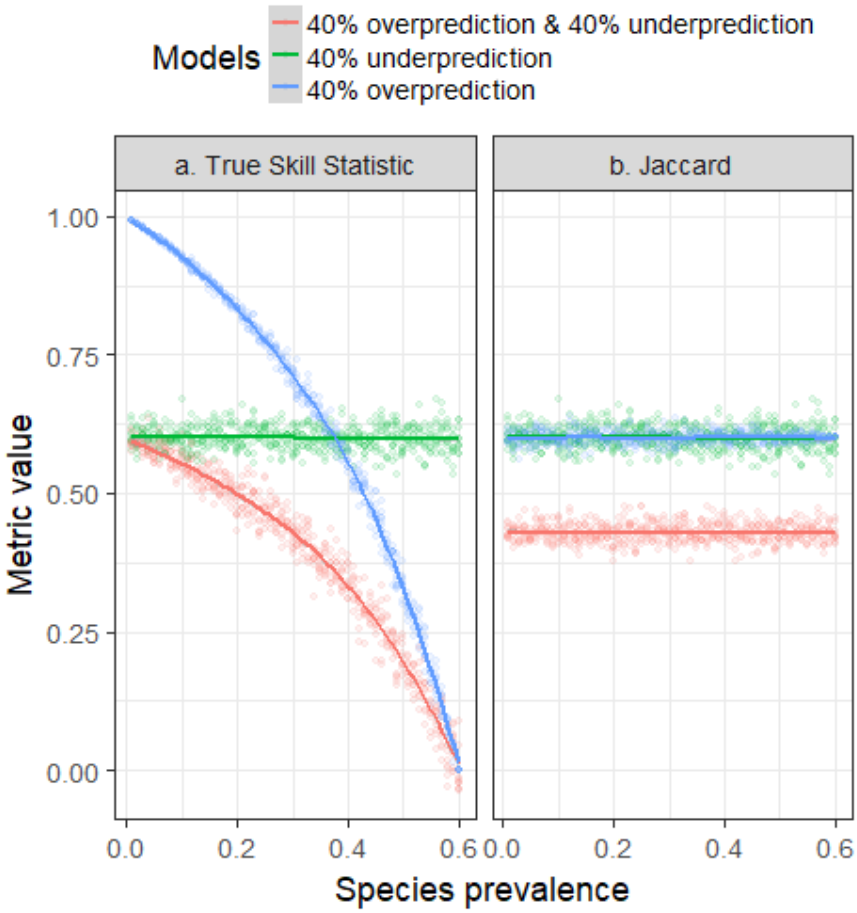

Figure 2. Simulations of the effect of species prevalence on species distribution model discrimination metrics ((a) TSS and (b) Jaccard, equations available in Table 2) in a presence-absence scheme where sample prevalence is equal to species prevalence. Three case studies with varying degrees of over- and underprediction are applied to theoretical species with prevalence ranging from 0.01 to 0.60 with a step of 0.01. The upper limit of 0.60 was chosen so that we can calculate values for models with 40% overprediction. For each species, an evaluation dataset was composed of 500 presences randomly sampled in the total area and a number of randomly sampled absences verifying the condition that the sample prevalence is equal to species prevalence, with 5 repetitions for each species (R scripts available at https://github.com/Farewe/SDMMetrics). These simulations showed that TSS attributes higher values at lower prevalence for case studies that overpredict species distributions, but not for case studies that have only underprediction (Figure 2a). Jaccard values, on the other hand, remain similar regardless of species prevalence (Figure 2b).

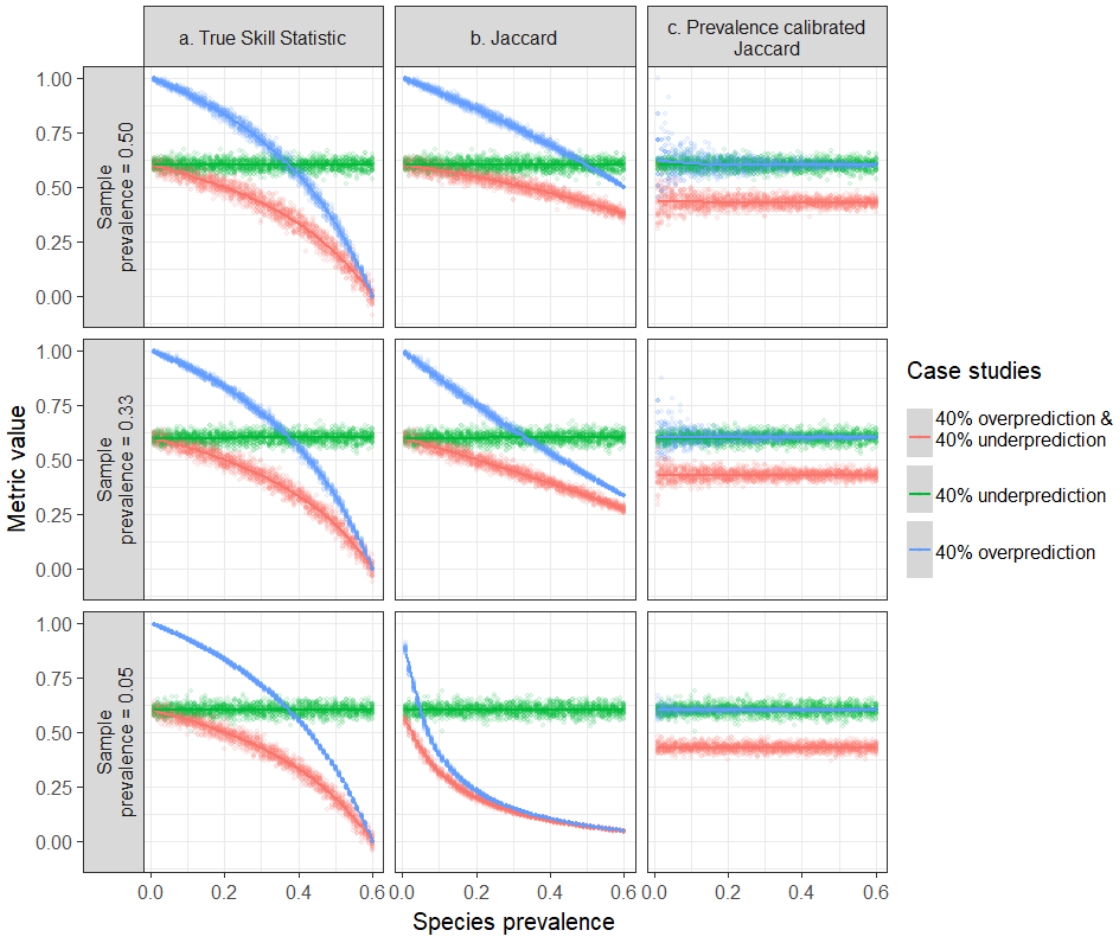

Figure 3. Simulations of the effect of species prevalence on species distribution model discrimination metrics ((a) True Skill Statistic, (b) Jaccard, equations available in Table 2, and (c) prevalence-calibrated Jaccard, equations (1) and (2) in main text) in a presence-absence scheme where sample prevalence is different from species prevalence. Three case studies with varying degrees of over- and underprediction are applied to theoretical species with prevalence ranging from 0.01 to 0.60 with a step of 0.01. The upper limit of 0.60 was chosen so that we can calculate values for case studies with 40% overprediction. For each species, an evaluation dataset was composed of 500 presences randomly sampled in the total area and 500 (top panel, prevalence = 0.50), 1 000 (middle panel, prevalence = 0.33) or 10 000 (bottom panel, prevalence = 0.05) randomly sampled absences, with 5 repetitions for each species (R scripts available at https://github.com/Farewe/SDMMetrics). These simulations showed that both TSS and Jaccard attribute higher values at lower prevalence for case studies that overpredict species distributions, but not for case studies that have only underprediction (Figure 3a-b). Conversely, the prevalence-calibrated Jaccard index was able to correctly estimate model discrimination capacity regardless of species prevalence (Figure 3c).

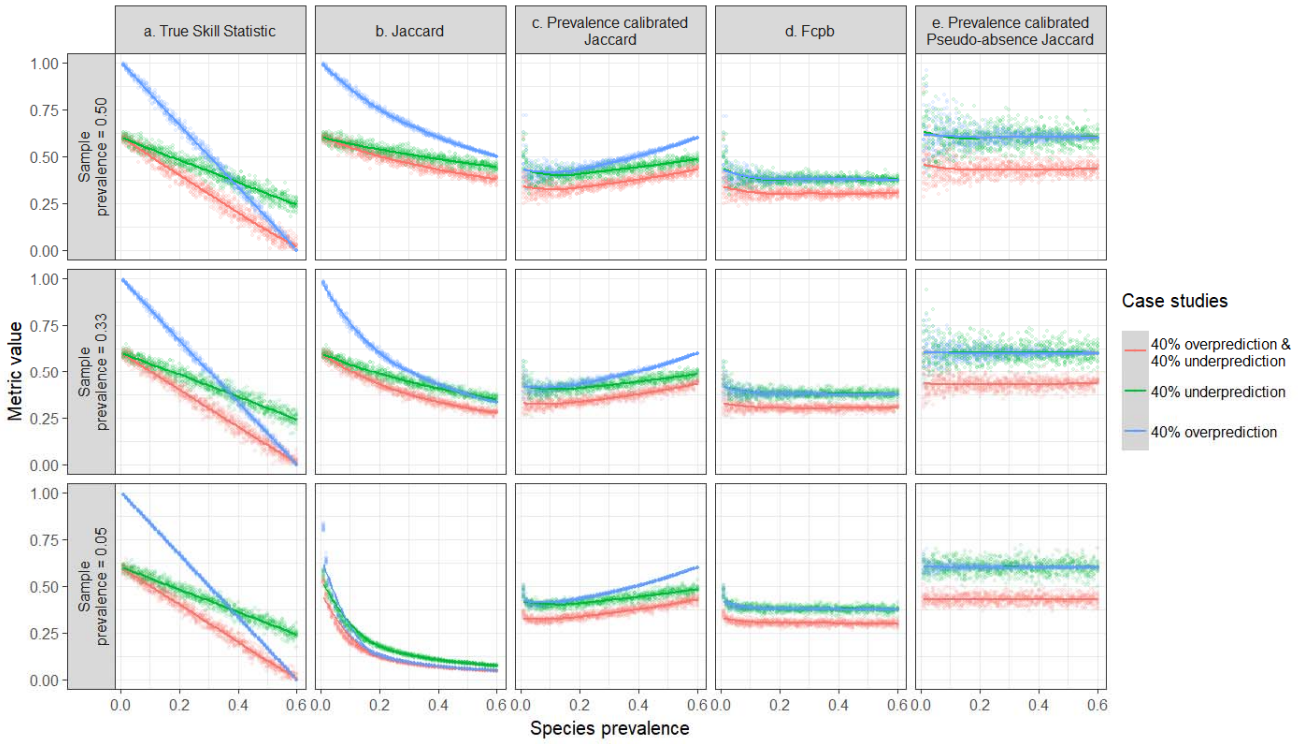

Figure 4. Simulations of the effect of species prevalence on species distribution model discrimination metrics ((a) the True Skill Statistic, (b) Jaccard, (c) prevalence-calibrated Jaccard, equations (1) and (2) in main text, (d) F_{cpb}, equations available in Table 2, and (e) prevalence-pseudoabsence-calibrated Jaccard, equations (4), (5) and (6) in main text) in a presence-pseudoabsence or presence-background scheme where sample prevalence is different from species prevalence. Three case studies with varying degrees of over- and underprediction are applied to theoretical species with prevalence ranging from 0.01 to 0.60 with a step of 0.01. The upper limit of 0.60 was chosen so that we can calculate values for case studies with 40% overprediction. For each species, an evaluation dataset was composed of 500 presences randomly sampled in the total area and 500 (top panel), 1 000 (middle panel) or 10 000 (bottom panel) randomly sampled pseudoabsences, with 5 repetitions for each species (R scripts available at https://github.com/Farewe/SDMMetrics). These simulations showed that both showed that TSS, Jaccard and Jaccard_{pc} are all sensitive to prevalence in this case, for all three case studies (Figure 3a-c). We found that the prevalence-calibrated F_{cpb} metric specifically designed for presence-background (Li and Guo, 2013) was also sensitive to very low species prevalence in our simulations (<0.10), but behaved otherwise much better than previous metrics (Fig. 3d). Jaccard_{ppac} did not show substantial dependence on prevalence for any of the case studies, except for a larger uncertainty at low prevalence due to smaller sample size (Fig. 3e).