Supplement 4: Performance metrics

Peter A. Vesk, William K. Morris, Will C. Neal,

Karel Mokany & Laura J. Pollock

February 2020

Relations between performance metrics and prevalence

Area under the Receiver Operator Curve

Prevalence of taxa in target regions varied log-normally over three orders of magnitude, from < 0.0005 to > 0.5. This variation in prevalence strongly influences metrics of predictive performance. The role of varying prevalence on the AUROC is shown in Fig. S4.1. This illustrates a triangular relationship, whereby at low prevalence, wide variation in AUROC is found. With increasing prevalence, AUROC values converge to a value just above 0.5. In particular, the upper end of the range declines, such that very few high-prevalence species achieve high AUROC values. This pattern is similar between taxa occurring in the Grampians and as well as target regions (grey symbols) as well as taxa only found in target regions (black symbols).

The same pattern of AUROC and prevalence can be seen when plotted by regions (Fig. S4.2).

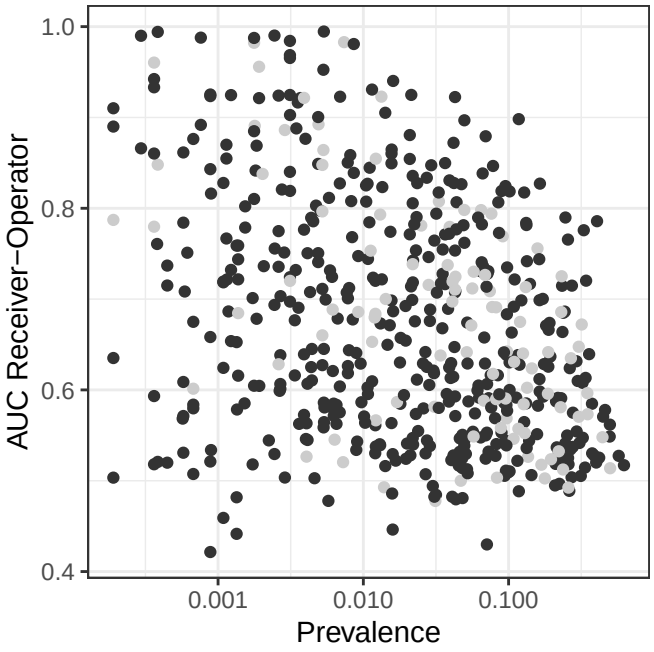

Figure S4.1: Area under the receiver operator curve vs. taxon within-region prevalence in southeast Australia. Light grey circles are taxa that occur in the Greater Grampians region as well as at least five target regions in the broader south-east. Darker circles are all other south-eastern taxa.

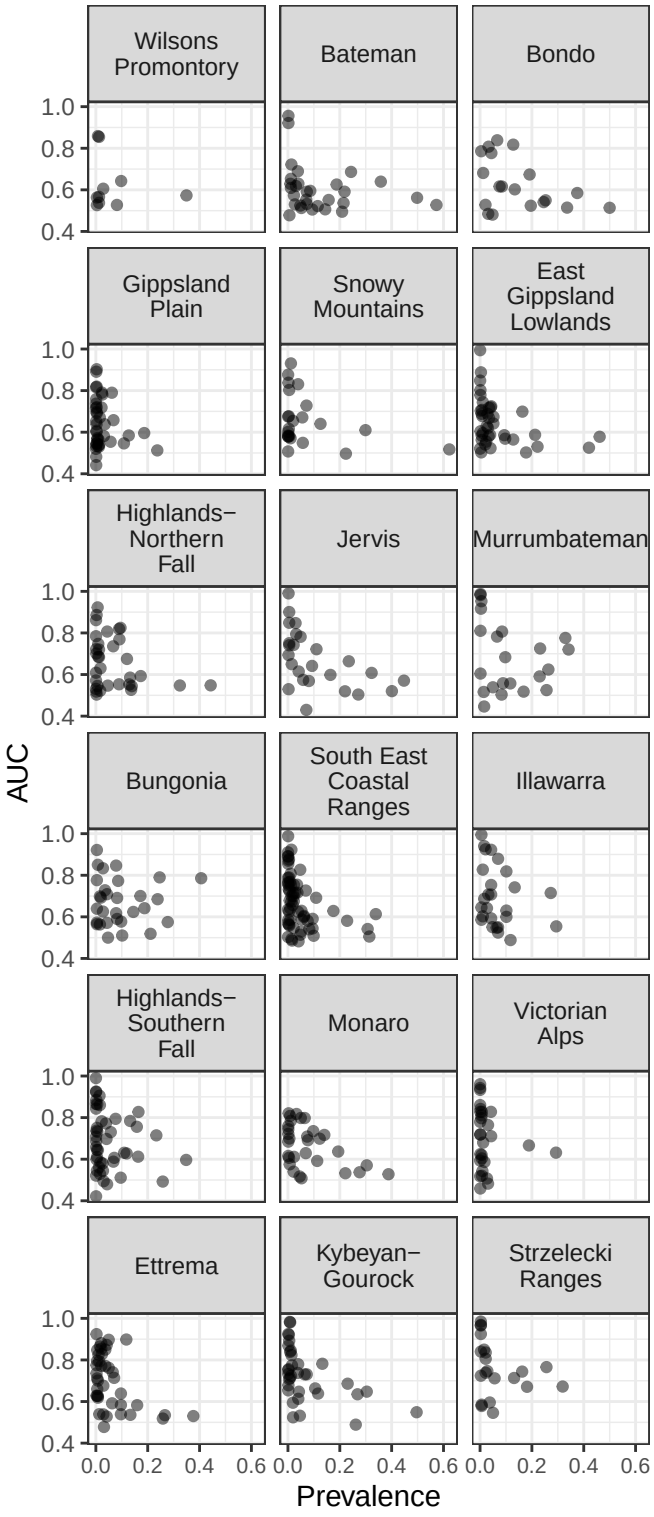

Figure S4.2: Area under the receiver operator curve vs. taxon prevalence in regions of southeast Australia. Regions are ordered by increasing median AUROC.

Area Under the Precision Recall Curve

We also evaluated the performance of model in the target region with the area under the performance-recall curve (AUPRC). The key advantage of AUPRC over AUROC is that it ignores true negatives. While mathematically AUROC is independent of prevalence, in species distribution modelling practice, AUROC is higher in rare species and declines with increasing prevalence (Morán-Ordóñez *et al.* 2017; Sofaer *et al.* 2019). When the true negative rate is high, as with most low-prevalence or spatially-restricted species, performance may be exaggerated by the large number of true negatives (Sofaer *et al.* 2019). AUPRC uses the precision of prediction (true presences as a fraction of predicted presences, TP+FP) against the recall (sensitivity; true positives as a fraction of observed presences, TP+FN). AUPRC is held to be particularly useful in reflecting model performance when surveys are directed to sites ranked higher by the model. This maps on to a context where a practitioner wishes to know where (along some environmental gradients) a focal species is more likely to occur. The problem is that AUPRC is explicitly related to prevalence (Sofaer *et al.* 2019). This confounds interpretation of a given AUPRC value and can be seen in our data in Fig. S4.7.

Two partial solutions to the influence of prevalence on AUPRC are that one can calculate the mathematically minimum AUPRC (dotted line in Fig. S4.3) and one can use this (and the maximum) to calculate a relative AUPRC. Also, it is possible with known prevalence to calculate the AUPRC expected from a random classifier (p=0.5). This enables a reference point for comparison. AUPRC values ranged over three orders of magnitude, ranging $1.50 \times 10^{-}5$–0.622. Relative AUPR varied over five orders of magnitude: $1.12 \times 10^{-}6$–0.523. Figs. S4.4 & S4.5 illustrate the variation in absolute

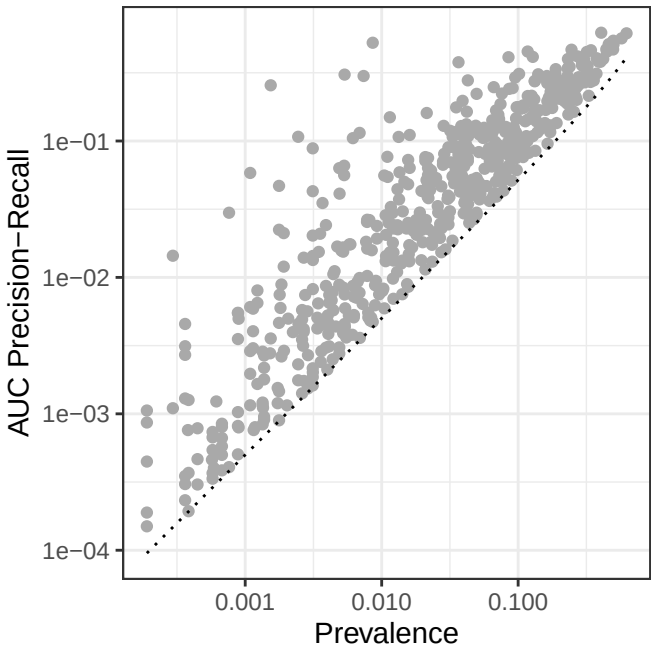

Figure S4.3: Area under the precision-recall curve vs. taxon prevalence within test regions of southeast Australia. The dotted line is the minimum feasible AUPRC

and relative AUPRC, and their relationships with the expected performance of a random classifier. The great variation in these values can be understood as a function of prevalence when considering that the numerator in the formula for precision is the number of true positives. The number of true positives is less than equal to the number of occurrences, so for a given prediction, precision can only every be high when the number of occurrences is large. The predictions were worse than random in a large number of cases, but also extend to much higher values of AUPRC other cases. The performance of a random classifier of p=0.5 indicates again the strong influence of prevalence on the AUPRC and Relative AUPRC.

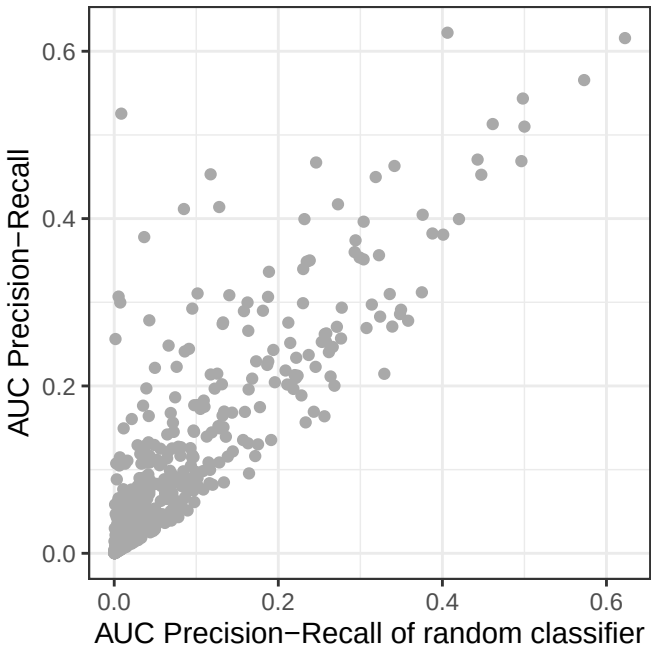

Figure S4.4: Area under the precision-recall curve vs. performance of a random classifier.

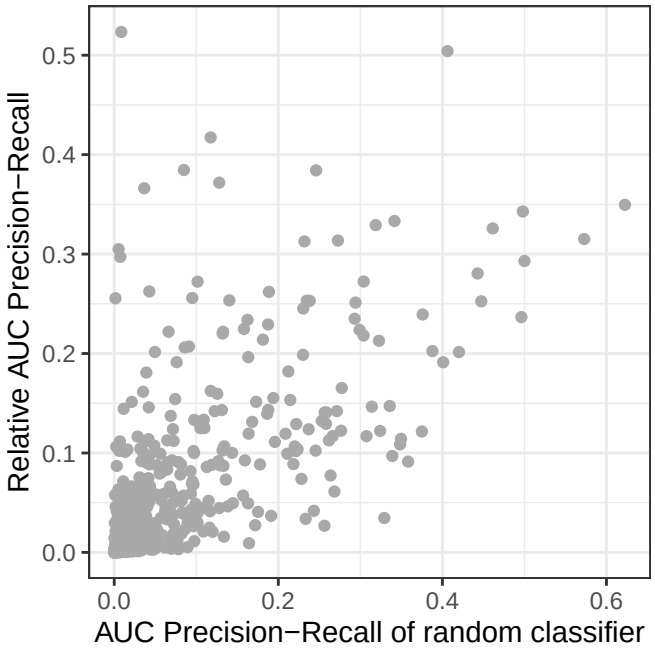

Figure S4.5: Relative Area under the receiver operator curve vs. performance of a random classifier.

Performance within the reference region

Within the Grampians, predictive performance of the trait-SDM was equal or higher when the prediction included the taxon random effect included (AUROC median = 0.82, range 0.60–1.0, RelAUPRC median = 0.2, range 0.02–1.00) than without the random effect (i.e., based on their traits only) (AUROC median = 0.68, range 0.47–0.93, RelAUPRC median = 0.12, range 0.006–0.45). But taxa varied considerably in how well traits predicted their responses (Fig. S4.6). Some taxon distributions (at top right of Fig. S4.6 were well predicted by the environment **and** traits explained those environmental responses (e.g., *E. verrucata* and *E. camaldulensis subsp. camaldulensis*). The farther taxon points lay to the left of the 1:1 line, the lower the predictive capacity of traits (e.g., *E. pauciflora subsp. parvifructa* and *E. arenacea*); environment predicted those taxon's occurrences, but the traits did not explain those environmental responses. But the trait-only prediction was sometimes comparable (e.g., *E. melliodora* and *E. camaldulensis subsp. camaldulensis*). For taxa at the bottom left of S4.6, predictive performance was low, but we cannot know if traits were potentially useful, because our fitted environmental covariates were not useful predictors. One of the worst predicted taxa according to AUROC was *E. obliqua*, which has high prevalence and little response to any of the modelled gradients (Fig. S3.3). Trait-based model performance for two taxa fell below AUROC=0.5, implying worse than random. Those two taxa fell at extremes—*E. aromaphloia* occurred in only three plots and *E. falciformis* was widespread on valley floors. *E. aromaphloia* also performed poorly according to RelAUPRC.

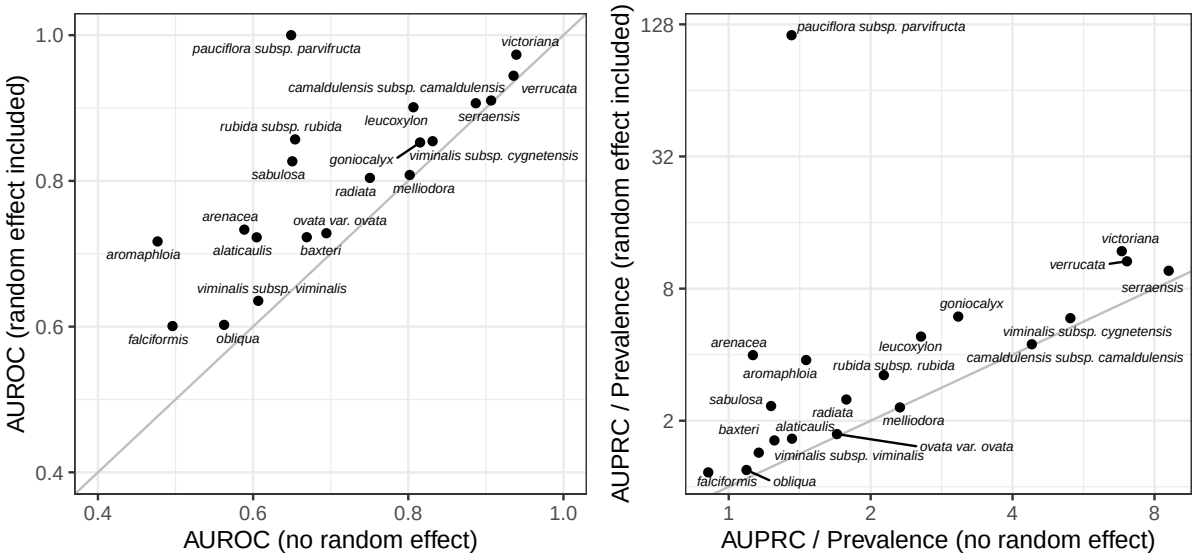

Figure S4.6: Taxon-specific area under the receiver operator curve (AUROC) and area under the precision recall curve (AUPRC) divided by prevalence based on predictions within the Grampians (reference) region, made with and without taxon-level random effect model terms.

Performance across the test regions

The performance of models measured by relative AUPRC is seen to vary widely within regions, with relatively little variation between region medians (Fig. S4.7). Medians of relative AUPRC values for regions do not appear to decrease with any measure of distance, reflecting the result for AUROC. Because performance of a random classifier is equal to the prevalence, we divide by prevalence to reflect performance relative to random for our main results.

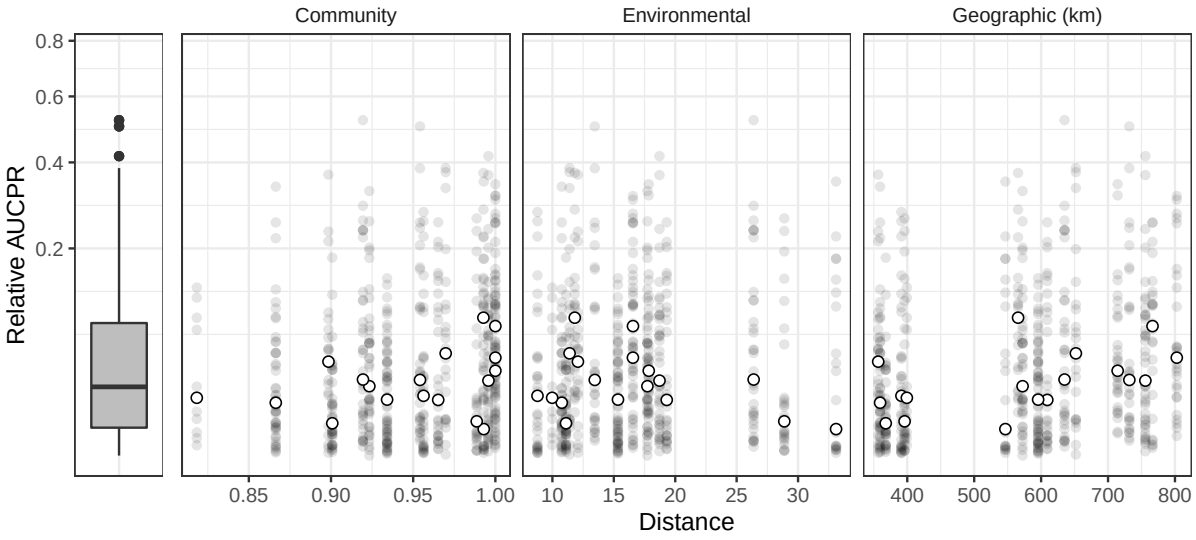

Figure S4.7: Relationship between within-region, taxon-specific, Relative Area Under the Precision-Recall Curve (Relative AUPRC) and the distance from the Grampians of each target region. Distance is measured as: Jaccard dissimilarity of communities, Kullback-Leibler distance of modelled environmental space, and distance in kms between centroids. White circles are the mean Relative AUPRC in each region. Boxplot in the left panel shows the distribution of within region taxon-specific Relative AUPRC across all the regions of the southeast. Note the y-axis has been scaled to aid visualisation.

Deviance

Deviance was calculated for all predictions. We initially calculated Explained Deviance in the usual way (i.e., as 1 - deviance of the fitted model/deviance of the null model). However, the choice of the null model is not obvious. Ordinarily, this is an intercept only model, equivalent to the prevalence (average of presences and absences) in the target region. That formulation seems inappropriate for this case, where nothing is known about the species in the region excepting the traits. And our model specifies no role of traits for prevalence. So we could just work with the deviance of the fitted model, but that would be sensitive to the number of plots and to the prevalence. So how to scale the deviances? What we do is to consider a null model using a probability of occupancy of 0.1, which is the approximate average prevalence across all species in all regions. This is equivalent to how common are species on average. That means all species in a region will have the same reference.

The explained deviance appears in Fig. S4.8. The values are very rarely good, corresponding to the unlikely event that the model predicts the prevalence reasonably. While the overall median is positive, often the explained deviance is negative, and strongly so (Fig. S4.8). There may be some indication of a decline in median performance across geographic distance to the target regions, but that pattern is an artefact of prevalence. Due to the strongly nonlinear, negative effect of prevalence, we do not consider explained deviance any further. .

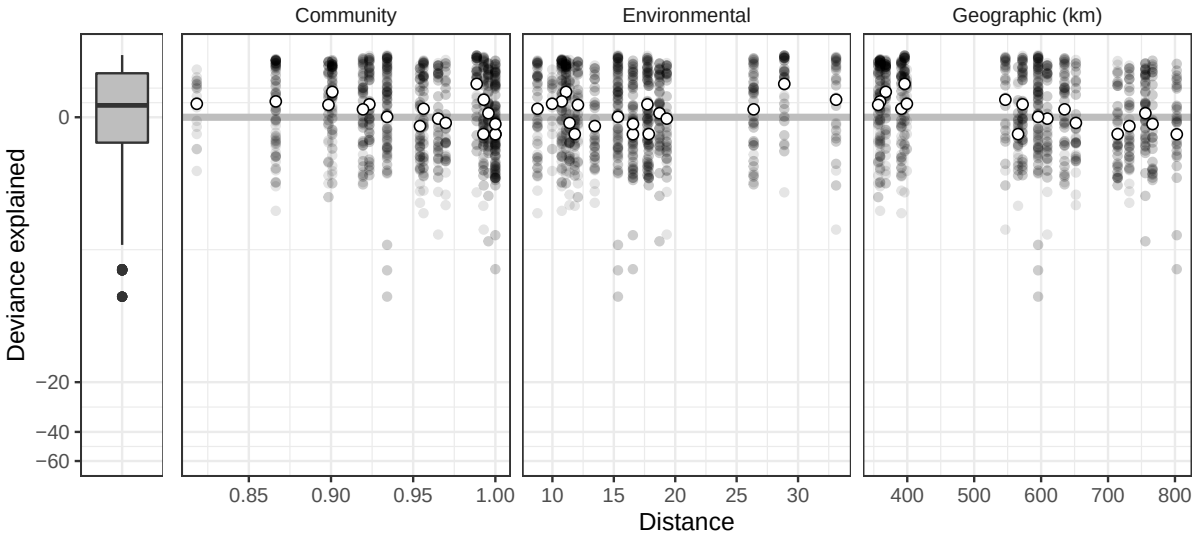

Figure S4.8: Relationship between within region, taxon-specific deviance explained and the distance from the Grampians of each target region. Distance is measured as: Jaccard dissimilarity of communities, Kullback-Leibler distance of modelled environmental space, and distance in kms between centroids. White circles are the mean deviance explained in each region. Boxplot in the left panel shows the distribution of within region taxon-specific deviance explained across all the regions of the southeast. Note the y-axis has been scaled to aid visualisation.

Models of performance measures

We fitted GLMMs to each of the performance metrics–AUPRC, AUROC–with linear effects of scaled distance and scaled log(prevalence). In all cases strong effects of prevalence were found, positive for AUPRC and negative for AUROC, but negligible effects of distance measures (Table S4.1). Note that distances measures and prevalence were scaled, so fixed effects reflect a one sigma change in the predictor variable and allows roughly direct comparison of fixed and random effects. Variation was least between regions, with residual variation large and taxon level variation intermediate (Table S4.2). The relatively large residual variation implies that it is not simply something about the taxon (e.g. unmeasured traits) which accounts for variable performance. It is in the specific combination of taxa in regions where most of the performance variation lies.

For AUROC the model had a mean of 0.66, with prevalence effects across four sigma leading to predictions of 0.75–0.58 from low to high prevalence taxa. This fixed effect of prevalence was 1.7 times that of remaining taxon and region level random effects, but 0.4 times the residual. Hence for AUROC, the residual variation (a taxon by region combination) was largest, followed by prevalence, and with taxon and region effects smallest (1/4 of the residual). at the mean, extra taxon variation could account for AUROC ranging from 0.61–0.71 from 2 sigma below to 2 sigma above the mean. Whereas the residual variation would be from 0.47–0.93. So the challenge to improving performance appears to be mainly in the combination of taxa in regions.

Table S4.1: Fixed effects

		AUROC		AUPRC/Prevalence	
Fixed Effect		μ	σ	μ	σ
Intercept		0.88	0.07	0.43	0.07
Distance	Geographic	0.01	0.07	-0.04	0.07
	Community	-0.05	0.07	-0.08	0.07
	Environmental	0.06	0.06	0.17	0.06
Error	MLQ	-0.01	0.04	-0.05	0.03
	TWI	-0.09	0.04	-0.19	0.03
	R1K	-0.04	0.04	-0.15	0.03
	TN	-0.02	0.04	-0.05	0.03

Table S4.2: Random effects

Random Effect	AUROC	AUPRC/Prevalence
Taxon	0.3	0.4
Region	0.2	0.2
Residual	0.8	0.7

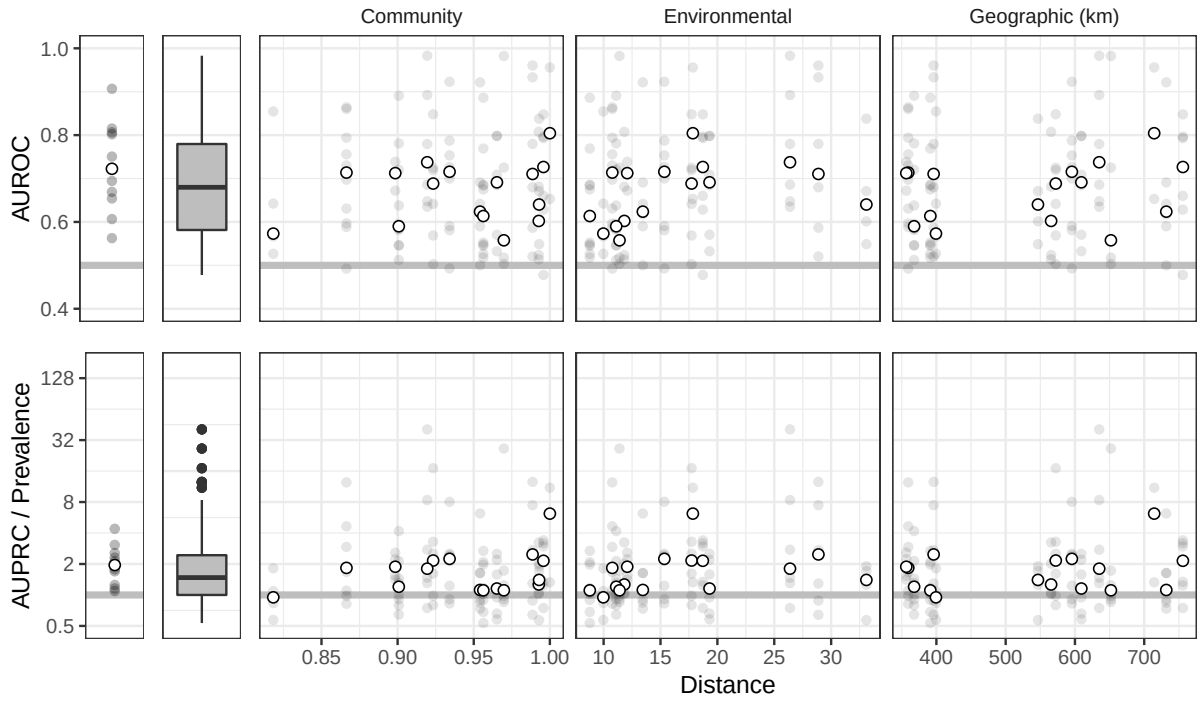

Figure S4.9: Relationship between within region, taxon-specific performance metrics (AUROC and AUPRC/prevalence) and the distance from the Grampians of each target region for taxon that are shared between the Grampians and the Southeast. Distance is measured as: Jaccard dissimilarity of communities, Kullback-Leibler distance of modelled environmental space, and distance in kms between centroids. White circles are the mean performance in each region. Leftmost panels show the performance metrics for the Greater Grampians. Boxplots show the distribution of within region taxon-specific performance across all the regions of the southeast.

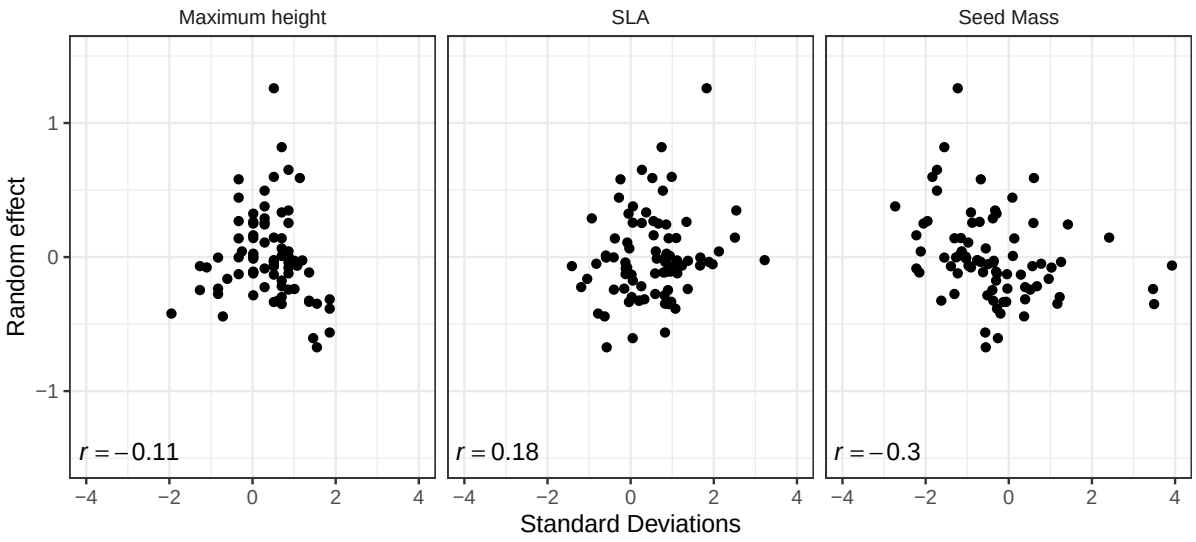

Figure S4.10: Taxon level random effect for model of AUPRC/Prevalence as function of bioregion distances and miscalibration metrics vs. species median trait values.

Probing predictive performance for some regions and environments

We illustrate predictive performance measured with AUPRC/prevalence for a subset of target regions chosen across the range of median model performance from least to best predicted, as well as the Grampians for reference (Fig. S4.11). On the right we see that the trait-SDM predicts taxon occurrences with similar performance to the Grampians—taxa vary in their predictive capacity in each of the regions. Most taxa are better predicted in the Victorian Alps than the Snowy Mountains and Jervis, but in each region some taxa are predicted well, with AUPRC > four times as good as random. Notably, the median AUPRC/prevalence is higher for Victorian Alps than the Grampians, where the model was trained. For AUROC, see Fig. S4.12.

Predicted response of taxa in regions along gradients

We compare the responses for two contrasting environmental covariates : moisture index (which varied widely between taxa but had limited interaction with traits) and; topographic wetness (with less variation between taxa but stronger interaction with traits (cf. Fig. **??**). In the Grampians (at top Fig. S4.11)) we can see that the trait-SDMs produced coefficients for Topographic Wetness similar in sign and magnitude to those from individual taxon regressions. Also taxa with high AUPRC/prevalence values tended to lie farther from the origin and closer to the 1:1 line, indicating that better predictions of occurrence (AUPRC) were associated with well-calibrated predictions of coefficients. Those patterns were not so evident for Moisture Index, where taxon regression responses varied widely but trait-SDM predictions did not capture that and varied little; (Fig. S4.11).

The correlation between trait-SDM predicted responses to Topographic Wetness in target regions show that some taxon responses were well predicted (lying in top right and lower left quadrants, and close to the 1:1 line). Taxa with high AUPRC/prevalence values were not always close to the 1:1 line, because the plots indicate responses to a single gradient at a time, whereas AUPRC/prevalence measures overall model performance. Some taxon responses were poorly predicted (e.g., in Victorian Alps, the sign was often wrong; positive responses were predicted by the trait-SDM while taxon regressions resulted in negative responses).

Trait-SDM responses to Moisture (Fig. S4.11, left panels) were less correlated with those from taxon regressions. Still, most responses were in the correct quadrant (i.e., correct sign). High AUPRC/prevalence predictions were generally associated with

coefficients in the correct quadrant. In Jervis, it appears that taxa with low AUPRC/prevalence are dispersed widely in the taxon regression coefficients, without corresponding predictive coefficients. That is, taxa in Jervis varied widely in their responses to moisture index, but in a way that was not predicted by the trait-SDM from the Grampians.

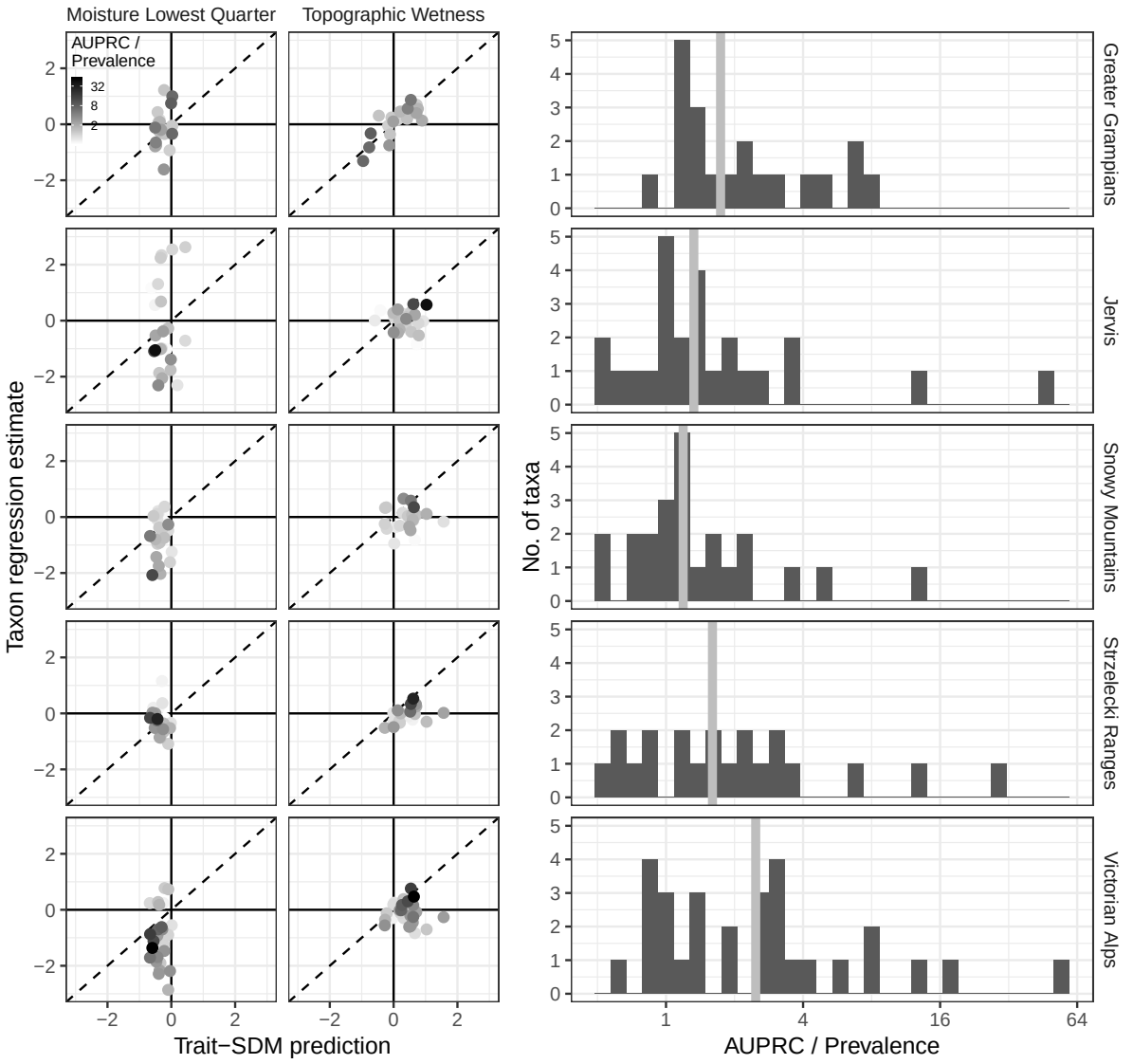

Figure S4.11: Left panels: Predicted responses from the trait-SDM versus taxon regression estimates. The top row of panels are the reference region, Greater Grampians. The four rows below are other regions in the southeast. Each point represents the response of a taxon within a given region. The position on the y-axis is the expected response predicted trait-SDM conditional on the median trait values. The position on the x-axis is the estimate of the response from taxon regressions of the taxa within the regions. Each point's black level indicates the area under the precision recall curve statistic (AUPRC) divided by the prevalence for the taxon in the region's plots based on the predicted probabilities of occupancy according to the trait-SDM. Right panels: Distribution of taxon-specific AUPRC divided by prevalence for predicted probabilities of occupancy conditional on traits for the regions. Grey line is the median AUPRC divided by prevalence value across the taxa in the region.

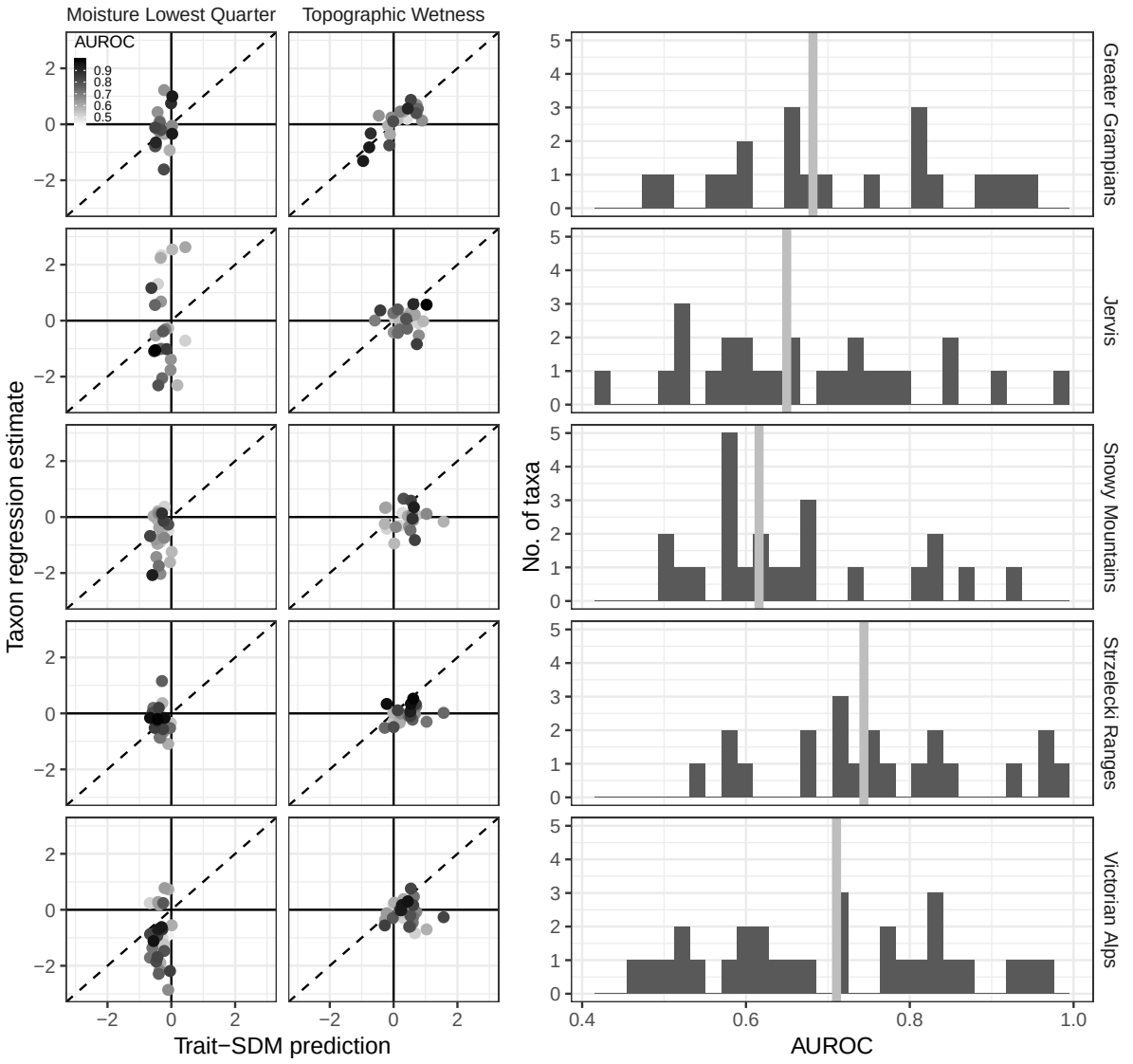

Figure S4.12: Left panels: Trait-SDM predicted coefficients versus taxon regression estimates. The top row of panels are for the reference region, Greater Grampians. The four rows below are other regions in the southeast. Each point represents the response of a taxon within a given region. The position on the y-axis represents the expected response predicted by the model conditional on median trait values. The position on the x-axis is the point estimate of the response coefficient from a logistic regression of the occupancy data of the taxon within the region. Each point's black level indicates the area under receiver-operator curve statistic (AUROC) for the taxon in the region's plots based on the predicted probabilities of occupancy according to the trait-SDM. Right panels: Distribution of taxon-specific AUROC for predicted probabilities of occupancy conditional on traits for the regions. Grey line is the median AUROC value across the taxa in the region.

References

Morán-Ordóñez, A., Lahoz-Monfort, J. J., Elith, J. & Wintle, B. A. (2017). Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence the reliability of predictions? *Global Ecology & Biogeography* 26, 371–384.

Sofaer, H. R., Hoeting, J. A. & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology & Evolution* 10, 565–577.