Conformal Prediction quantifies the uncertainty of Species Distribution Models

Timothée Poisot [1]

Abstract: Providing accurate estimates of uncertainty is key for the analysis, adoption, and interpretation of species distribution models. In this manuscript, through the analysis of data from an emblematic North American cryptid, I illustrate how Conformal Prediction allows fast and informative uncertainty quantification. I discuss how the conformal predictions can be used to gain more knowledge about the importance of variables in driving presences and absences, and how they help assess the importance of climatic novelty when doing future predictions.

Introduction

The ability to predict where species may be found is a cornerstone of biogeography and macroecology (Elith, 2019). Techniques from the field of applied machine learning (ML hereafter) are now routinely used alongside ecological approaches to train generalizable species distribution models (SDMs hereafter) (Beery et al., 2021). SDMs generate, by increased order of refinement, a binary response (predicted presence/absence of the species, alternatively framed as predicted suitability/unsuitability of local habitat), a probability of habitat suitability, and a distribution of this probability (which can be, in its simplest expression, a measure of variance around the prediction).

Proper communication of the uncertainty associated to the prediction of a SDM is important, since we usually seek to apply these models to look both forward and backwards in time (Franklin, 2023) – this process is usually called "transfer" (Zurell et al., 2012), in that the model trained under extant condition is transferred to past/future values of the same predictors. Even when predictions are not projected in time, spatial knowledge of the uncertainty is valuable information as it provides more information about where the model outcomes are trustworthy. Current checklists on the reproductibility of SDMs emphasize the consequences of data uncertainty (Feng et al., 2019). Yet, predictions also have inherent uncertainty, which is usually not adequately communicated; this can be, for example, because of genuine uncertainty about (or inability to capture through the model) the actual response of the species to combination of predictors (Parker et al., 2024).

A common way to capture information about the variability of SDMs is to rely on non-parametric bootstrapping (Valavi, Guillera-Arroita, et al., 2021), wherein models trained on subsets of the data are compared to estimate the distribution of the response under incomplete sampling. This approach captures more than one type of variability (Thuiller et al., 2019), and provide valuable information about the range of performances that can be expected from a model. Other methods are built into the predictor itself, as is the case for *e.g.* BARTs (Carlson, 2020), which estimate their own uncertainty. But either situation comes with drawbacks. Bootstrapping requires to train and evaluate the model hundreds of times, and on partial datasets, which is computationally inefficient. Using built-in methods limits one to the classifier for which these methods are available, which prevents for example the use of a new algorithm with the same estimation of uncertainty.

In this manuscript, I illustrate how the ML technique of conformal prediction (Lei & Wasserman, 2013; Papadopoulos et al., 2002) allows to identify instances (combinations of environmental variables) for which a trained and calibrated model cannot confidently make predictions. By way of contrast to *e.g.*

bootstrapping, it does not involve retraining the same model many times over, but instead wraps the model into an additional prediction step, and returns estimates of credibility based on the distribution of model predictions. This is an important difference, as the variability measured through conformal prediction (CP hererafter; a brief introduction is given later in this manuscript) is inherent to the model, and is not a measure of variability coming through the distribution of data. Conformal prediction provides what is essentially (for classification problems) a confidence interval around the presence or absence of a species in a given location, which is weighted according to how likely this outcome is. This is a particularly important feature, as it ties machine learning back to some fundamental concepts in frequentist statistics (Neyman, 1937).

One of the reasons why CP is particularly promising for uncertainty quantification in SDMs is that it is a distribution-free method: it requires neither assumptions about the model nor prior knowledge of the outcome distribution to provide confidence intervals of arbitrarily small coverage that are *guaranteed* to contain the true value (Vovk et al., 2018). This is particularly important when transferring a SDM to novel environments (Zurell et al., 2012), where we expect covariate shift (the joint distributions of predictors are different when training and predicting), a prediction task that CP is robust to (Fannjiang et al., 2022; Tibshirani et al., 2019).

Using occurrence data about an emblematic North American cryptid, I show how predictions under CP (i) identify areas where the species range is uncertain, (ii) estimate uncertainty differently from bootstraping methods, (iii) can be explained using Shapley values analysis, and (iv) quantify the accumulated uncertainty when transferring the SDM to future conditions. I conclude by highlighting ways in which using CP can both simplify the process of training SDMs, and provide information that make their discussion and analysis more informative.

Dataset

Occurrence data

The occurrence data used in this article are geo-referenced observations of the Sasquatch (Lozier et al., 2009). Although these observations are likely to be mis-categorized American black bears (Foxon, 2024), they nevertheless share many features of the data that are used to train SDMs: high autocorrelation, uneven sampling effort, and clear association with several bioclimatic variables that is enough to train a predictive model.

Pseudo-absences generation

The dataset of observations is composed only of presences. In order to establish a baseline of absences to train a binary classifier, there is a need to generate a number of pseudo-absences, which simulates locations at which the species, if not absent, has not been observed. In order to do so, the presence data were first spatially thinned to be limited to one for each cell, at a 5.0 minutes of arc resolution. Cells that had no observation were potential candidates for a pseudo-absence, and were further selected by drawing a number of them, without replacement, where the probability of inclusion in the sample was proportional to h_{min}^{-1}, where h_{min} is the Haversine (great arc) distance to the nearest cell with an observation, measured in meters. In other words, cells that were close to an observation were unlikely to be included, and cells that were further away were more likely to be so.

The number of pseudo-absences was arbitrarily set to three times the number of presences. Although Barbet-Massin et al. (2012) recommend to use the same number of presences and pseudo-absences for classifiers, using an imbalanced dataset is not a problem: stratified k-folds cross-validation is perfectly able to handle the moderate class imbalance we introduce (Szeghalmy & Fazekas, 2023), and the model performance (as will be established in a later section) is sufficient. Moreover, most real-world applications of classification will have to deal with problems with class imbalance (this is particularly likely to be true of SDM application from sampling data, where presences may be the minority of outcomes); it is therefore important to ensure that we do not establish a testing scenario that is too optimistic about the prevalence of presences. In all cases, class imbalances is a feature of data that must be dealt with in order to get the more predictive models (Benkendorf et al., 2023).

Bioclimatic data

The model was trained, validated, and applied on the 19 WorldClim2 BIOCLIM variables (Fick & Hijmans, 2017), at a spatial resolution of 5 minutes of arc. This data resolution is coarser than many applications, but this choice has been made in order to speed up the computation. Preliminary analyses using 2.5 minutes of arc, and 30 seconds of arc, show that the qualitative results presented hold (but the computation times for the most demanding steps, like Shapley values analysis, increases by several orders of magnitude).

Training of the non-conformal model

Conformal Prediction requires a well-trained model to serve as a baseline before it can be applied. For this reason, in this first section, we go into some detail into the training and validation of a suitable model, and further derive a first approximation of its uncertainty by relying on bagging to create a homogeneous ensemble. The model we use is a Boosted Regression Tree (BRT). BRTs are highly flexible, make few assumptions about the data, efficiently model non-linear relationship between variables, and use an ensemble of shallow trees to avoid overfitting. Indeed, BRTs are excellent classifiers for species distribution models (Elith et al., 2008). When trained on a vector of features **x**, a BRT will return a vector of predictions $\mathbf{p} = \{p_\oplus, p_\ominus\}$, which correspond to the probability of these environmental conditions being associated to, respectively, the presence and the absence of the species. Because the BRT as we initially train it is a deterministic classifier, $p_\oplus + p_\ominus = 1$, and $0 \leq p_\oplus, p_\ominus \leq 1$. These assumptions are not true when using conformal prediction, which estimates the confidence in the presence and absence as two distinct features. The outcome of a BRT prediction **p** can be turned into a binary response (corresponding to the presence of a species) through $p_\oplus \geq p_\ominus$, which is equivalent to $p_\oplus \geq 0.5$.

[Figure 1 about here.]

We optimize the initial model by (i) iteratively forward selecting the best set of predictor variables, and (ii) optimizing the threshold τ above which a site with a probability for the positive class $p_\oplus$ is considered to be positive (turning the prediction of presence into $p_\oplus \geq \tau$). In both cases, the cross-validation strategy is the same: the dataset is split in 10 random folds, 9 of which are used for training and one for evaluation. All folds are used for evaluation, providing exhaustive cross-validation. The folds are stratified so that the relative number of present cases in the training set is similar to that of the entire dataset. The performance on each set, for the purpose of defining the set of variables to include of the threshold to use, is measured as the average of the Matthews Correlation Coefficient (MCC) across each of the ten folds. The MCC is the most accurate representation of a binary classifier performance (Chicco & Jurman, 2023), and avoids the pitfalls of several other validation measures.

For all steps of model training and validation, the identity of instances composing the different folds remains fixed. This ensure that the changes in MCC are only due to the addition of the variable, and not to the random sampling of a training/validation set with different properties. Although some authors encourage the use of spatially-stratified cross-validation (Soley-Guardia et al., 2024), this is not a desirable strategy for this use-case. The area in which the predictions will be made is entirely

delimited by the bounding box of observed presences, and there is therefore no risk of covariate shift when shifting from validation to prediction (outside of the situation of temporal transfer of the SDM). Because BRTs establish their baseline prediction (the first tree) as the prevalence of presences in the training dataset (Valavi, Guillera-Arroita, et al., 2021), we used stratified ten-fold cross-validation, in which the ten folds all have a the same number of instances, with correct representation of the relative frequency of presences and absences.

Variable selection

The predictors included in the model have been decided through the use of forward selection. This is an important step in order to perform dimensionality reduction (which generally increases the predictive accuracy), but also to ensure that the set of retained variables is reduced enough that it can be interpreted. Variables where retained as part of the final set of predictors if adding them increased the MCC for the model once retrained with this new variable.

An initial attempt to cross-validate the model using all variables resulted in a MCC that was close to the model using an optimal set of predictors. Nevertheless, minimizing the number of inputs to a model is generally a good idea. First, it makes the assessment of the contribution of variables far more efficient and informative; second, it decreases the risk of covariate shift when predicting (by lowering the number of covariates); finally, it makes the training more efficient, by having less variables to split during the training of the BRTs (while maintaining the number of trees, leading to a better fit).

Thresholding

One of the most efficient ways to increase the performance of binary classifiers is to change the decision rule leading to a positive (here, presence) prediction, so that presences are assigned when $p_{\oplus} \geq \tau$ – a process known as moving threshold classification (Liu et al., 2013, 2015). The value of τ is an hyper-parameter of the model, which is chosen to maximize the value of a measure of model performance (here the MCC) when evaluated over many different values. In this instance, we optimized the value of τ by cross-validation 30 meta-models (models that only differ in their hyper-parameters), with different values chosen through Latin hypercube sampling (McKay et al., 1979). The value of τ that maximizes the MCC was selected as the optimal threshold for the BRT.

Estimation of bootstrap variability

Bagging (bootstrap aggregating) is often used as a measure of uncertainty to the underlying data when training SDMs (Beale & Lennon, 2012). When performing bagging, the model is trained on samples drawn with replacement from the training set (which leaves out approx. 37% of the dataset). Trees are then evaluated on samples that were not used as part of their training, usually using cross-validation (Bylander, 2002) or measures of the out-of-bag error (Janitza & Hornung, 2018). Although ensemble models *can* get to a better predictive performance compared to single models (Drake, 2014), this is not a guarantee (and depends on the structure of the bias/variance trade-off for the specific model and its training set). The many models trained on the bagging dataset form an homogeneous ensemble, which is to say a set of models that share the same algorithm and hyper-parameters, and only make different predictions as the result of having been trained on different subsets of the full training set.

Measures of whether the different models composing the homogeneous ensemble agree can provide a measure of the effect of data and parameter uncertainty (Petropoulos et al., 2018), or what Davies et al. (2023) termed the "SDM uncertainty". The best model identified after thresholding was evaluated on a hundred bootstrap samples, yielding an homogeneous ensemble model from which we estimate bootstrap variability (Chen et al., 2019). Because the model is kept constant in this analysis, the measure of variability we will derive from the ensemble model is an estimate of how sensitive the estimation of the model parameters is to small perturbations (specifically spatially homogeneous under-sampling) to the training data.

Performance of the baseline model

The optimal threshold found through Latin hypercube sampling is $\tau \approx 0.35$; although this is quite far away from the untuned threshold of $1/2$, the quantitative effect on the behavior of the model, *i.e.* the effect on the predictions as measured by the MCC, is quite small. The MCC after the threshold optimization is only increased by 0.01, and the MCC of the ensemble model is lower than the thresholded BRT (though not by a lot, and not enough to preclude the use of this model to evaluate uncertainty). The prediction made by the BRT, as well as the range at the optimal threshold, are given in Figure 1.

Table 1: Comparison of the performance of the BRT (trained and cross-validated on the same folds) before and after optimizing the threshold. The out-of-bag performance of the ensemble model (trained on 100 bootstrap samples) is also reported. All three models are roughly equivalent in terms of their predictive ability. The value considered ideal for each measure is bolded. The MCC is used as the criteria to evaluate the best model for variable selection and thresholding.

Measure	BRT	BRT with threshold	Ensemble
True positive rate	0.731	0.763	**0.783**
True negative rate	**0.928**	0.909	0.909
False positive rate	**0.072**	0.091	0.091
False negative rate	0.269	**0.215**	0.217
F score	0.751	**0.763**	0.761
Balanced accuracy	0.829	**0.847**	0.846
MCC	0.672	**0.683**	0.679

Training of the conformal model

The trained model from Figure 1 can be used for conformal prediction. Conformal prediction differs from the regular prediction in that it creates sets (or, for quantitative responses, intervals) given an input value. Given the observed quantiles of the model output on the validation data, these sets are obtained through a simple calibration step. Therefore, CP can be applied on an already trained model, and is agnostic to the process through which this model is trained. In this section, I highlight two important features of CP: the notion of *credible sets*, and the *coverage* statistic, which is a measure of tolerance to error. An in-depth introduction to CP is found in Angelopoulos & Bates (2023).

Understanding conformal predictions

By contrast to the non-conformal SDM, the conformal classifier returns, for an input of environmental predictors **x**, a set C containing the "credible outcomes" for this prediction. This set is termed the *credible set*, and there are three scenarios for its membership. First, if both the presence and absence are credible for this prediction, the credible set will be $C = \{\hat{p}_{\oplus}, \hat{p}_{\ominus}\}$. Note that because the credibility of either outcomes is expressed relatively to the estimation of their distribution, there is no guarantee that $\hat{p}_{\oplus} = 1 - \hat{p}_{\ominus}$. Second, the credible set can have a single outcome in it, either $C = \{\hat{p}_{\oplus}\}$ or $C = \{\hat{p}_{\ominus}\}$. In this case, one of the outcomes is credible, but the other is not. Finally, there is a chance that $C = \emptyset$, in which case the conformal model has not enough evidence to include *either* outcome

credibly.

These situations correspond to four different outcomes in terms of the SDM certainty about the distribution of the species. The most intuitive situation is $C = \{\hat{p}_\oplus\}$ or $C = \{\hat{p}_\ominus\}$, in which case the conformal model predicts that the absence (resp. presence) of the species is *not* a credible outcome for the environmental conditions given as an input. We term these predictions "sure presences" and "sure absences", as *for a given value of the coverage statistic* α, there is no reason to expect that the prediction is uncertain. The second situation, $C = \{\hat{p}_\oplus, \hat{p}_\ominus\}$, corresponds to inputs for which the presence and the absence of the species are credible (they may not be *equally* credible, as the score for one may be larger than the score for the other), and we term these predictions "unsure". The final situation corresponds to $C = \emptyset$, which means that neither absence or presence can be credibly predicted – given the training data (and the distribution of presences and absences), the model is not able to make a prediction for this input. The multiplication of such predictions is most likely a strong sign that the risk level is too high (the confidence interval is too broad) for the training data given to the conformal model.

To summarize, the output of the conformal classifier is, in a sense, a point-specific stand-in for the application of a threshold. A location is defined as included in the range is the positive outcome is included within the credible set returned by the conformal classifier, and as excluded from the range when it is not. Because the conformal classifier can identify that both outcomes are credible based on the training data (while giving them different weights), predictions in which both the positive and negative outcomes are included in the credible set can be seen as "uncertain" at this given risk level. How frequently a specific prediction is uncertain is termed the inefficiency of the classifier, which is defined as the average cardinality of all credible sets. The inefficiency is bounded upwards by the number of classes (two for binary classification); when the inefficiency is ≈ 1, the conformal classifier behaves (essentially) as a deterministic classifier, by returning a single class for each instance. An inefficiency close to unity is not desirable: smaller sets can hide our actual uncertainty (Sadinle et al., 2018). Because the conformal models wraps the BRT model, we can further divide the "unsure" predictions as a function of whether they would be within the range as predicted by the BRT (*i.e.* $C = \{\hat{p}_\oplus, \hat{p}_\ominus\}, p_\oplus \geq \tau$), which we call "unsure presences"; the other unsure predictions are referred to as "unsure absences".

Understanding the effect of the coverage level

CP allows users to set a desired error rate, α: the conformal prediction is that the credible set contains the true value with probability $1 - \alpha$, which allows to directly interpret this value as a confidence interval. This error rate is usually referred to as the *marginal coverage*, in that it captures the probability of success marginalized over the known validation points. Because the estimate of uncertainty involves the original model, it is important to apply CP on a model with adequate performance.

In Figure 2 we show how changing the risk level (α) leads to different estimates of the range size of the species. Using a low level of risk ($\alpha \approx 0$) yields the largest possible range, but at the cost of a very high uncertainty - this is evidenced by the value of inefficiency getting closer to 2 (the maximum value, as the outcomes of the classification are either positive or negative). For values larger than $\alpha \approx 0.12$, there is a situation in which the inefficiency of the conformal prediction (which is to say, the average number of outcomes in the credible set) is less than one; this corresponds to a situation where some instances are impossible to assign to either outcome. Although this situation is more difficult to make sense of intuitively, a value of inefficiency that gets further away from unity should be interpreted as a model that accumulates more uncertainty (at a given risk level) than the data can support (Romano et al., 2020).

[Figure 2 about here.]

In the rest of this analysis, we set $\alpha = 0.05$. As noted by Angelopoulos & Bates (2023), this corresponds to estimating whether a specific prediction falls within, or outside of, the 95% confidence interval across all predictions, which is a convenient callback to frequentist statistics' usual risk tolerance. From Figure 2, this level of risk would represent an inefficiency of about 1.2, meaning that 20% of the predictions would have both presence and absence in their credible set. Note that even when setting the risk at $\alpha = 0.0$, the inefficiency does not climb up to 2 (the theoretical maximum); there would be a number of pixels (about 15%) that only have either presence or absence in their credible set. Recall that the CP credible sets are estimated based on the model output, and therefore even when aiming for full coverage, there are non-ambiguous combinations of environmental predictors.

Analysis of the predicted species range

Before discussing the spatial output of running the conformal model, it is worth considering why the thresholding step applied in Figure 1 is not really providing us with a set of certain presences and absences. When optimizing the threshold τ above which a prediction $p_{\oplus}$ from the non-conformal

model is determined to be a presence, we establish a sort of certain presences and certain absences; indeed, the space covered by positive predictions is usually interpreted as the (potential) distribution of the species. But this prediction conveys a false sense of certainty, that has to do with the very nature of the threshold we optimize. By definition, the threshold is the value that finds the best balance between the false/true positive/negative cases on the validation data; this is in fact why the optimal threshold is the point closest to the corners of the ROC and PR curves indicating a perfect classifier (Balayla, 2020). When a prediction $p_{\oplus}$ gets closer to the threshold, a small perturbation to the environmental conditions locally could bring it on the other side of the threshold, and therefore flip the predicted class using the non-conformal classifier. Around the threshold is where we expect uncertainty to be the greatest.

To bring these considerations into a spatial context: we expect the areas where the score for the present class are closer to the threshold (the limits of the predicted range of the species) to be the most uncertain. Importantly, this is true *both* for areas that are inside the range (for which $p_{\oplus}$ is just above the threshold) and for areas that are outside of it (for which $p_{\oplus}$ is just below the threshold). CP is perfectly suited to solving this issue, by identifying the areas where one class is predicted, but the other class is also credible. In this section, we will project the areas with uncertain predictions, and compare the uncertainty quantified by the conformal model to the uncertainty derived from the ensemble model.

Identification of areas with uncertainty

As far as ecologists are concerned, the areas in which the credible set only has a score for the absence of the species are the easiest to make sense of: they correspond to regions where the model is certain (under the specified risk level) that the species is absent. All other areas (assuming that there are no predictions for which the credible set is empty) are *potentially* part of the range of the species: some certainly, some uncertainly. In Figure 3, we present the result of the conformal prediction under $\alpha = 0.05$, by showing the class attributed to the present class ($\hat{p}_{\oplus}$), as well the type of prediction: sure presence, unsure presence, unsure absence, and sure absence. This information can be conveyed in a number of ways. For example, what is the threshold α for which a pixel is included into the range of a species, either certainly or uncertainly? This question is not explored here, but shows the possible versatility of CP.

[Figure 3 about here.]

Uncertain areas are different from bootstrap estimates of variability

In Figure 4, we present a map of the uncertainty as estimated through the variance of the 100 models used in the homogeneous ensemble. This measure of uncertainty represents the potential effect of sampling the training data; it is, as expected, higher in areas that are not very close to either 0 or 1 in Figure 1. Intriguingly, the overlap between areas that are uncertain according to the conformal classifier, and areas that are uncertain according to the bootstrap model, is imperfect. Although there are, predictably, a large number of points in uncertain areas that have a very high bootstrap variance, there are also a number of points for which the variance is ≈ 0, *i.e.* points whose uncertainty is not a consequence of undersampling the training data.

[Figure 4 about here.]

Nevertheless, CP captures some of the underlying model uncertainty: in Figure 4, predictions that are uncertain but within the range predicted by the BRT had an over-representation of very low (≈ 0) uncertainty, whereas predictions that are uncertain but likely out of range had an over-representation of high (≈ 1) uncertainty. This suggests that the classification of predictions as certain/uncertain according to the conformal prediction is in part reflecting genuine uncertainty in the underlying data, but also contributing novel information about the fact that some instances are more difficult to call.

These results can be better understood by contrasting what "uncertain" means in the context of CP, and how it differs from the uncertainty in the ensemble model. The uncertainty derived from the ensemble model represents whether many models trained on small perturbations of the full training dataset would agree on a specific prediction task, represented by an array of environmental predictors. Therefore, the uncertainty from the ensemble originates in the estimation of the parameters, and its sensitivity to being able to access the full information within the training data. Uncertainty in the conformal classifier is coming from comparing the prediction to all other predictions under an estimation of the distributions for the conditions leading to the prediction of the presence (or absence) outcome. Therefore, the uncertainty from the conformal predictors accounts for all the predictions the model can make, and accounts for the variability *across* predictions within a fully accessible dataset.

Model explanation

In this section, we perform an analysis of Shapley values of the conformal predictor, in order to (i) assess the importance of variables and (ii) provide explainable results about the relationships between predictors and response. Although initially a game-theoretic concept, we rely on the common Monte-

Carlo approximation (Roth, 1988; TOUATI et al., 2021). Monte-Carlo Shapley values represent, for each prediction, how much the ith variable contributed to moving the prediction away from the average prediction. The Shapley value associated to variable i is $\phi_i \in [-1, 1]$, which measures how much this variable modified the *average* prediction for this class. Shapley values have a number of desirable properties regarding the explanation of prediction of responses for environmental studies (Wadoux et al., 2023), including their additivity: for any given prediction, $p = \bar{p} + \sum_i^{\text{variables}} \phi_i$. Because of this additive property, the importance of variables across many predictions is usually measured as the average of $\|\phi\|$, where both positive (the class is more likely) and negative (the class is less likely) are counted. This measure of variable importance represents the relative impact that each variable had on the process of moving all predictions away from the average prediction. Because Shapley values are both additive and independent, they can be measured and aggregated for any arbitrary stratification of the data (which allows reporting them conditional on the uncertainty status of the prediction).

As the predictions of the conformal model can be split by whether they are certain or uncertain, they offer a unique opportunity to delve into the mechanisms that *generate* this uncertainty. Namely, if the relative importance of variables is different across these classes of predictions, this is strongly suggestive of the fact that there are certain environmental conditions (represented by combination of values for each variables) that create or reduce uncertainty. Furthermore, because we can split the certain predictions into a presence and absence class, this is a unique opportunity to generate whether the factors leading to a species being present or absent are the same. In Figure 5, we show the importance of the selected variables for all predictions, but also sub-divide the relative importance of these variables for classes of prediction certainty.

We find that the certain absences follow the same variable importance as the full prediction (which is expected as the range of this species is a small part of the total study area, therefore absences contribute disproportionately to the total predictions). None of the other classes did so, with, notably, the uncertain presences and absences having markedly different variable importance when compared to the certain prediction *and* to one another. For example, the BIO10 variable (mean temperature of warmest quarter) was much more important for predictions classified as uncertain absences.

[Figure 5 about here.]

Model projection

Zurell et al. (2012) highlight the importance of uncertainty when transferring the model to novel climate data: there is a chance that the future climate condition will not have occurred in the training

dataset, and therefore our confidence in the model outcome may be lowered. This covariate shift is well documented to decrease the performance of models (Mesgaran et al., 2014), and CP offers an opportunity to quantify this phenomenon.

Using the data from the CanESM5 model (Swart et al., 2019) under the SSP370 scenario for the year 2090, it is possible to split the landscape as a function of (i) climatic novelty defined as values of the bioclimatic variables not observed in the training data and (ii) status of the range for the species. These results are presented in the table below:

Climatic novelty	Sure absence	Unsure	Sure presence
Yes	50.46%	48.36%	1.16%
No	54.54%	37.27%	8.17%
(difference)	4.07%	11.09%	7.01%

These results show that *on average*, the areas with climatic novelty had more uncertain outcomes, which is in line with ecological expectations.

Conclusion

Conformal prediction, like most SDM methods, is not quite delivering a true estimate of the probability of presence (Phillips & Elith, 2013). Nevertheless, it brings valuable information, in the form of a quantified measure of whether a prediction comes with uncertainty (are both presence and absence in the credible set?) in a way that is directly comparable with the non-conformal prediction. "Class overlap", where both presences and absences are observed under the same values of the predictions, decreases the predictive performance of models (Valavi, Elith, et al., 2021) – CP is naturally suited at handling this, by assigning the area where overlap occurs to uncertain predictions.

Transparent communication of uncertainty, meaning, it is both spatially explicit, quantified, and expressed under a risk set by the user, is important: we do not expect a fully trained model to always be certain, as some areas are genuinely more difficult to predict. For example, small organisms are more inherently stochastic (Soininen et al., 2013); any form of stochastic event will drive species distribution in the general case (Mohd et al., 2016); these stochastic events can appear even in areas that are close to the species' environmental optimum (Dallas et al., 2020).

CP contributes to dispel what Messeri & Crockett (2024) called the "illusion of understanding", which is

often associated with ML models: it generates an understanding of the uncertainty from observations of a pre-trained model, and expresses this uncertainty both in absolute (is the "presence" event in the credible set?) and relative (is the conformal score for presence larger than for absence?) terms. Because this technique is computationally efficient and works on pre-trained models, it opens up the opportunity for more systematic uncertainty quantification (Zurell et al., 2020) in SDMs. CP, in short, can deliver the "maps of ignorance" that Rocchini et al. (2011) argued for: how difficult is it to make a prediction for the range at a given risk level is, in and of itself, an important information to frame the reliability of the results. Finally, CP can provide guidance on the feedback loop between SDM training and field validation (Johnson et al., 2023) – areas where the range is certain are a much lower priority for sampling. Looking back at Figure 1, the uncertain areas are much smaller than the certain ones, which provides actionable guidance for field-based validation.

References

Angelopoulos, A. N., & Bates, S. (2023). *Conformal prediction: A gentle introduction*. https://doi.org/10.1561/9781638281597

Balayla, J. (2020). Prevalence threshold (ϕe) and the geometry of screening curves. *PLOS ONE*, *15*(10), e0240215. https://doi.org/10.1371/journal.pone.0240215

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210x.2011.00172.x

Beale, C. M., & Lennon, J. J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1586), 247–258. https://doi.org/10.1098/rstb.2011.0178

Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species distribution modeling for machine learning practitioners: A review. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. https://doi.org/10.1145/3460112.3471966

Benkendorf, D. J., Schwartz, S. D., Cutler, D. R., & Hawkins, C. P. (2023). Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models. *Ecological Modelling*, *483*, 110414. https://doi.org/10.1016/j.ecolmodel.2023.110414

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, *48*, 287–297.

Carlson, C. J. (2020). embarcadero: Species distribution modelling with Bayesian additive regression

trees in r. *Methods in Ecology and Evolution*, *11*(7), 850–858. https://doi.org/10.1111/2041-210x.13389

Chen, X., Dimitrov, N. B., & Meyers, L. A. (2019). Uncertainty analysis of species distribution models. *PLOS ONE*, *14*(5), e0214190. https://doi.org/10.1371/journal.pone.0214190

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1). https://doi.org/10.1186/s13040-023-00322-4

Dallas, T. A., Santini, L., Decker, R., & Hastings, A. (2020). Weighing the evidence for the abundant-center hypothesis. *Biodiversity Informatics*, *15*(3), 81–91. https://doi.org/10.17161/bi.v15i3.11989

Davies, S. C., Thompson, P. L., Gomez, C., Nephin, J., Knudby, A., Park, A. E., Friesen, S. K., Pollock, L. J., Rubidge, E. M., Anderson, S. C., Iacarella, J. C., Lyons, D. A., MacDonald, A., McMillan, A., Ward, E. J., Holdsworth, A. M., Swart, N., Price, J., & Hunter, K. L. (2023). Addressing uncertainty when projecting marine species' distributions under climate change. *Ecography*, *2023*(11). https://doi.org/10.1111/ecog.06731

Drake, J. M. (2014). Ensemble algorithms for ecological niche modeling from presence-background and presence-only data. *Ecosphere*, *5*(6), 1–16. https://doi.org/10.1890/es13-00202.1

Elith, J. (2019). *Species distribution modeling*. https://doi.org/10.1093/obo/9780199830060-0226

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., & Jordan, M. I. (2022). Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, *119*(43). https://doi.org/10.1073/pnas.2204569119

Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C., & Papeş, M. (2019). A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology & Evolution*, *3*(10), 1382–1395. https://doi.org/10.1038/s41559-019-0972-5

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Foxon, F. (2024). Bigfoot: If it's there, could it be a bear? *Journal of Zoology*, *323*(1), 1–8. https://doi.org/10.1111/jzo.13148

Franklin, J. (2023). Species distribution modelling supports the study of past, present and future biogeographies. *Journal of Biogeography*, *50*(9), 1533–1545. https://doi.org/10.1111/jbi.14617

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLOS ONE*,

13(8), e0201904. https://doi.org/10.1371/journal.pone.0201904

Johnson, S., Molano-Flores, B., & Zaya, D. (2023). Field validation as a tool for mitigating uncertainty in species distribution modeling for conservation planning. *Conservation Science and Practice*, *5*(8). https://doi.org/10.1111/csp2.12978

Lei, J., & Wasserman, L. (2013). Distribution-free Prediction Bands for Non-parametric Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 71–96. https://doi.org/10.1111/rssb.12021

Liu, C., Newell, G., & White, M. (2015). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, *6*(1), 337–348. https://doi.org/10.1002/ece3.1878

Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, *40*(4), 778–789. https://doi.org/10.1111/jbi.12058

Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *Journal of Biogeography*, *36*(9), 1623–1627. https://doi.org/10.1111/j.1365-2699.2009.02152.x

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239. https://doi.org/10.2307/1268522

Mesgaran, M. B., Cousens, R. D., & Webber, B. L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, *20*(10), 1147–1159. https://doi.org/10.1111/ddi.12209

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Mohd, M. H., Murray, R., Plank, M. J., & Godsoe, W. (2016). Effects of dispersal and stochasticity on the presence–absence of multiple species. *Ecological Modelling*, *342*, 49–59. https://doi.org/10.1016/j.ecolmodel.2016.09.026

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, *236*, 333–380.

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). *Inductive confidence machines for regression* (pp. 345–356). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-36755-1_29

Parker, E. J., Weiskopf, S. R., Oliver, R. Y., Rubenstein, M. A., & Jetz, W. (2024). Insufficient and biased representation of species geographic responses to climate change. *Global Change Biology*, *30*(7). https://doi.org/10.1111/gcb.17408

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does

bagging for time series forecasting work? *European Journal of Operational Research*, *268*(2), 545–554. https://doi.org/10.1016/j.ejor.2018.01.045

Phillips, S. J., & Elith, J. (2013). On estimating probability of presence from use–availability or presence–background data. *Ecology*, *94*(6), 1409–1419. https://doi.org/10.1890/12-1520.1

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., & Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography: Earth and Environment*, *35*(2), 211–226. https://doi.org/10.1177/0309133311399491

Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 3581–3591.

Roth, A. E. (1988). *Introduction to the shapley value* (pp. 1–28). Cambridge University Press. https://doi.org/10.1017/cbo9780511528446.002

Sadinle, M., Lei, J., & Wasserman, L. (2018). Least Ambiguous Set-Valued Classifiers With Bounded Error Levels. *Journal of the American Statistical Association*, *114*(525), 223–234. https://doi.org/10.1080/01621459.2017.1395341

Soininen, J., Korhonen, J. J., & Luoto, M. (2013). Stochastic species distributions are driven by organism size. *Ecology*, *94*(3), 660–670. https://doi.org/10.1890/12-0777.1

Soley-Guardia, M., Alvarado-Serrano, D. F., & Anderson, R. P. (2024). Top ten hazards to avoid when modeling species distributions: a didactic guide of assumptions, problems, and recommendations. *Ecography*, *2024*(4). https://doi.org/10.1111/ecog.06852

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., … Winter, B. (2019). The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Szeghalmy, S., & Fazekas, A. (2023). A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors*, *23*(4), 2333. https://doi.org/10.3390/s23042333

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., & Zimmermann, N. E. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09519-w

Tibshirani, R. J., Barber, R. F., Candes, E. J., & Ramdas, A. (2019). *Conformal prediction under covariate shift*. https://doi.org/10.48550/ARXIV.1904.06019

TOUATI, S., RADJEF, M. S., & SAIS, L. (2021). A Bayesian Monte Carlo method for computing the Shapley value: Application to weighted voting and bin packing games. *Computers & Operations Research*, *125*, 105094. https://doi.org/10.1016/j.cor.2020.105094

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731–1742. https://doi.org/10.1111/ecog.05615

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2021). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, *92*(1). https://doi.org/10.1002/ecm.1486

Vovk, V., Shen, J., Manokhin, V., & Xie, M. (2018). Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, *108*(3), 445–474. https://doi.org/10.1007/s10994-018-5755-8

Wadoux, A. M. J.-C., Saby, N. P. A., & Martin, M. P. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. *SOIL*, *9*(1), 21–38. https://doi.org/10.5194/soil-9-21-2023

Zurell, D., Elith, J., & Schröder, B. (2012). Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity and Distributions*, *18*(6), 628–634. https://doi.org/10.1111/j.1472-4642.2012.00887.x

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

Authorship information

Timothée Poisot timothee.poisot@umontreal.ca

Author for correspondance

[1] Département de Sciences Biologiques, Université de Montréal, Montréal, Canada

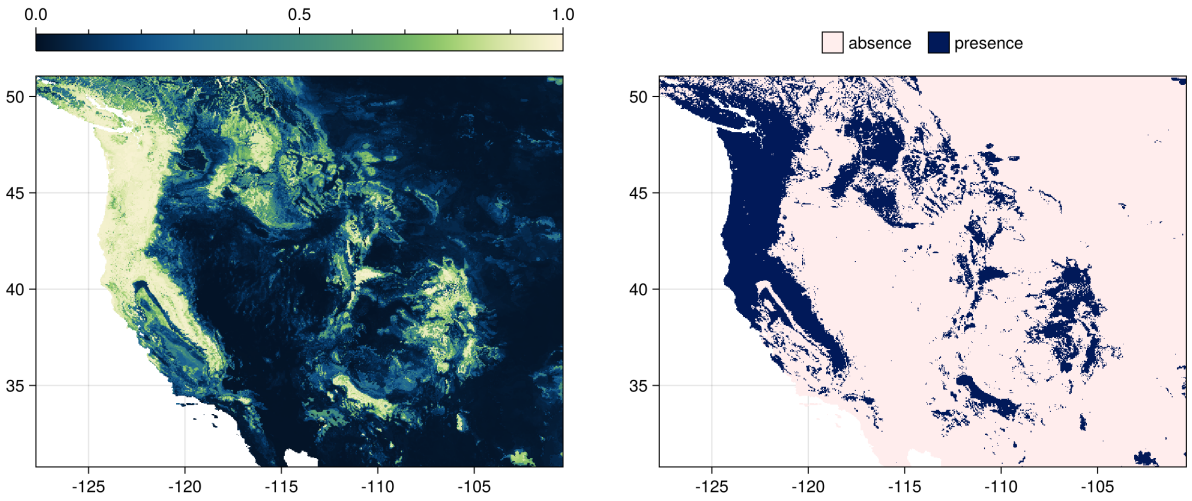

Figure 1: Overview of the prediction from the baseline Boosted Regression Tree (BRT) model, using the set of forward-selected variables. The left panels shows the score assigned to the positive class (presence), and the right panel shows (in black) the range, defined as $p_{\oplus} \geq \tau$, where τ is the threshold that maximizes the Matthews Correlation Coefficient (MCC).

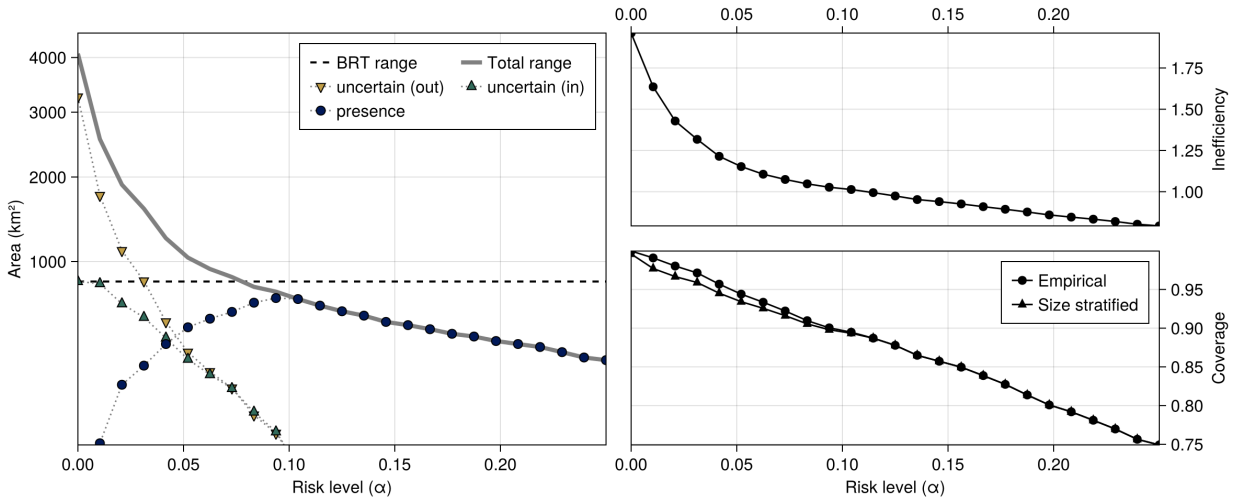

Figure 2: Effects of changing the value of α on the size of the range (left panel, split by uncertainty category) and conformal classifier performance (right column, top panel is inefficiency and bottom panel is coverage).

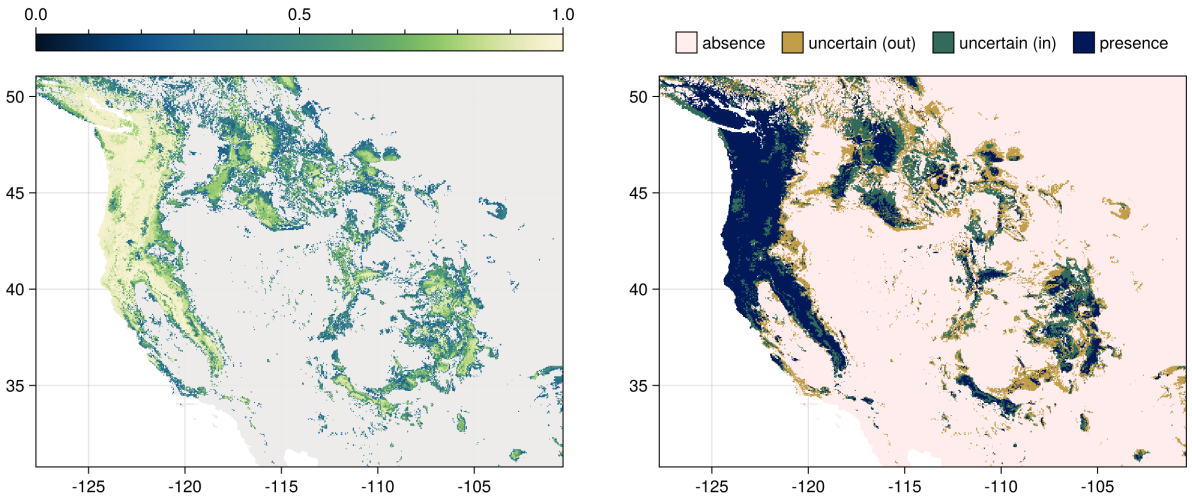

Figure 3: Prediction made by the conformal classifier at a risk level $\alpha = 0.05$. The left panel indicates score associated with the presence at this location. The left panels shows areas in gray where the negative class is associated in the credible set (the "uncertain" part of the range), and areas in black where the negative class is not part of the credible set (the "certain" part of the range). Changing the value of α would change the boundaries of the certain/uncertain range.

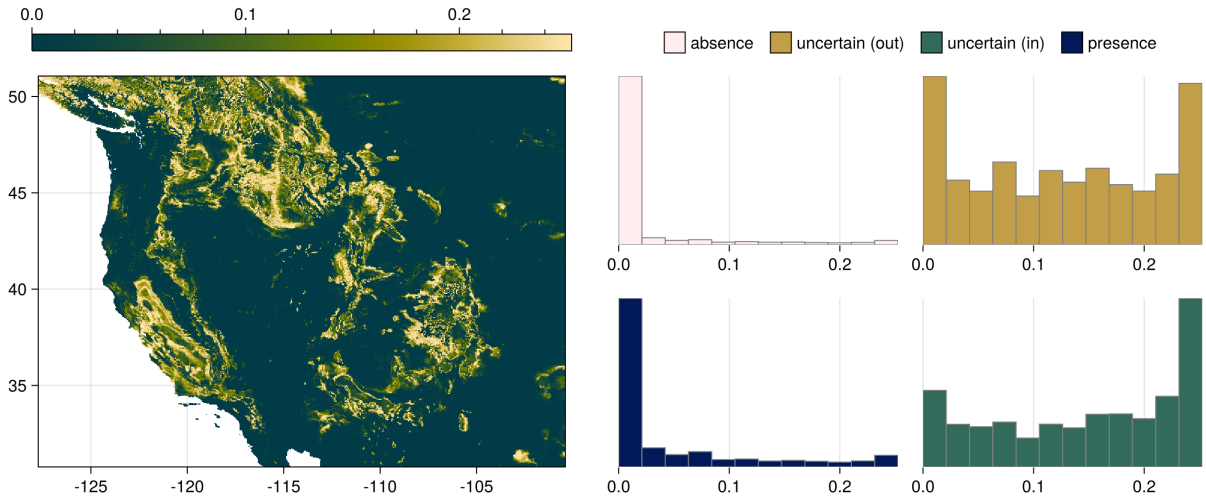

Figure 4: Variability in the predictions made by the thresholded BRT, based on the variance of 100 replicates of the bagging model (left). Splitting the values of uncertainties according to the type of conformal prediction (right) reveals that although certain presence/absence predictions are associated to low variance, the pixels classified as uncertain do not necessarily skew towards high bootstrap uncertainty.

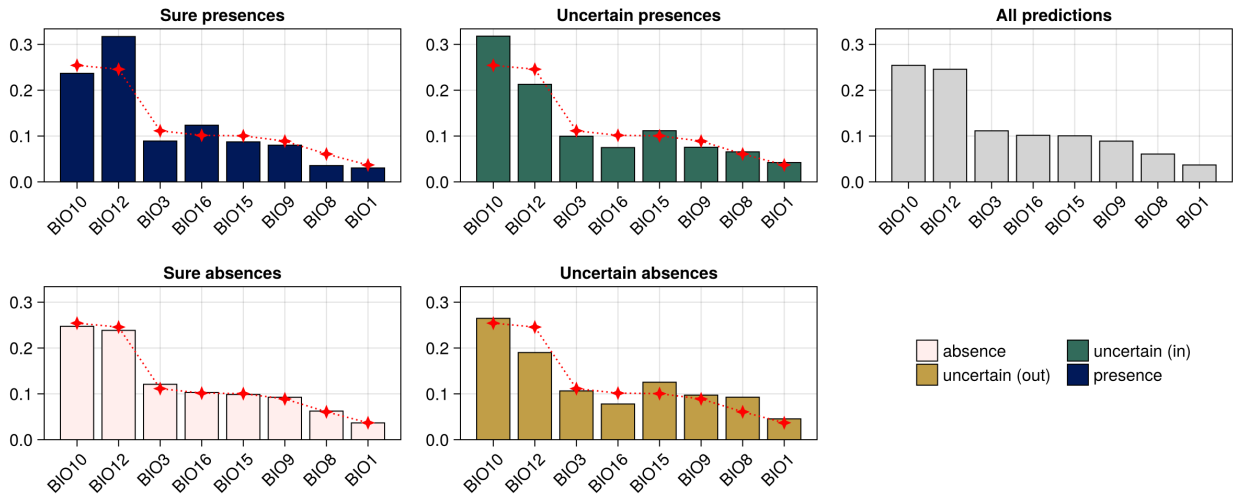

Figure 5: Relative variable importance (measured as the average absolute value of all Shapley values) to explain the conformal prediction ($\alpha = 0.05$) for the entire range (left), for the part of the range where absence is not part of the credible set (middle), and for the part of the range where it is (right). For the last two panels, the dots associated to each variable are the importance of this variable across the entire range. Note that the importance of variables is not accounting for the areas where the absence of the species is certain (*i.e.* presence is not part of the credible set). The difference in relative variable importance in the certain/uncertain area suggests that the conformal model is picking up on different relationships between predictors and response in areas of high *vs.* low certainty.