Conformal Prediction quantifies the uncertainty of Species Distribution Models

Timothée Poisot — Département de Sciences Biologiques, Université de Montréal, Montréal QC, Canada

timothee.poisot@umontreal.ca

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 projecting the models under future climate change scenarios.

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 a binary response (corresponding to the prediction that the species is likely present/absent under given environmental conditions) or a quantitative score most often as a probability of presence or habitat suitability, indicating how strongly we believe that the species may be present at the location.

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 projection if the model to different times is usually called “transfer” (Zurell et al. 2012), whereby a model trained under historical (baseline) conditions is applied to past/ future projections of the same predictors. The projection of SDMs can also happen in space (Petitpierre et al. 2016), to predict where species may invade or be naturalized. Even when predictions are not projected, spatial knowledge of the uncertainty is valuable information: it can be used to identify areas where the model predictions 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 et al. 2021), wherein models trained on random 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 (CP) allows to identify instances (combinations of environmental variables) for which a trained and calibrated model cannot confidently make predictions (Gammerman et al. 1998). A brief introduction to CP is provided in this manuscript, but the topic is covered in more depth by Shafer & Vovk (2007) for the mathematical foundations, by Fontana et al. (2020) for a historical perspective, and by

Angelopoulos & Bates (2023) for concrete recommendations. By way of contrast to *e.g.* bootstrapping, CP does not necessarily 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 past model predictions compared to ground-truthed data. This is an important difference, as the variability measured through conformal prediction is inherent to the model, and is not a measure of variability coming through the distribution of data (Lei & Wasserman 2013). Conformal prediction provides what is essentially (for classification problems) a confidence interval around the presence or absence of a species in a given location. This is a particularly important feature, in that CP achieves this in a way that creates several analogues between ML prediction and 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 that are as small as possible while being *guaranteed* to contain the true value under a set risk level (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 context that CP is robust to (Fannjiang et al. 2022, Tibshirani et al. 2019).

Using occurrence data about an emblematic North American cryptid, I provide a template for the adoption of CP as a natural way to quantify uncertainty of species distribution models. In particular, 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.

Methods

Data

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 auto-correlation, uneven sampling effort, and clear association with several bioclimatic variables that is robust enough to train a predictive model. The recorded locations, as well a background points, are presented in Figure 1.

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 kilometers. 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. To avoid sampling pseudo-absences too close to presences, the pixels less than 10 kilometers away from known observations were excluded from the background data.

The number of pseudo-absences was arbitrarily set to two 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 2.5 minutes of arc. Preliminary analyses using 0.5, 2.5, 5,

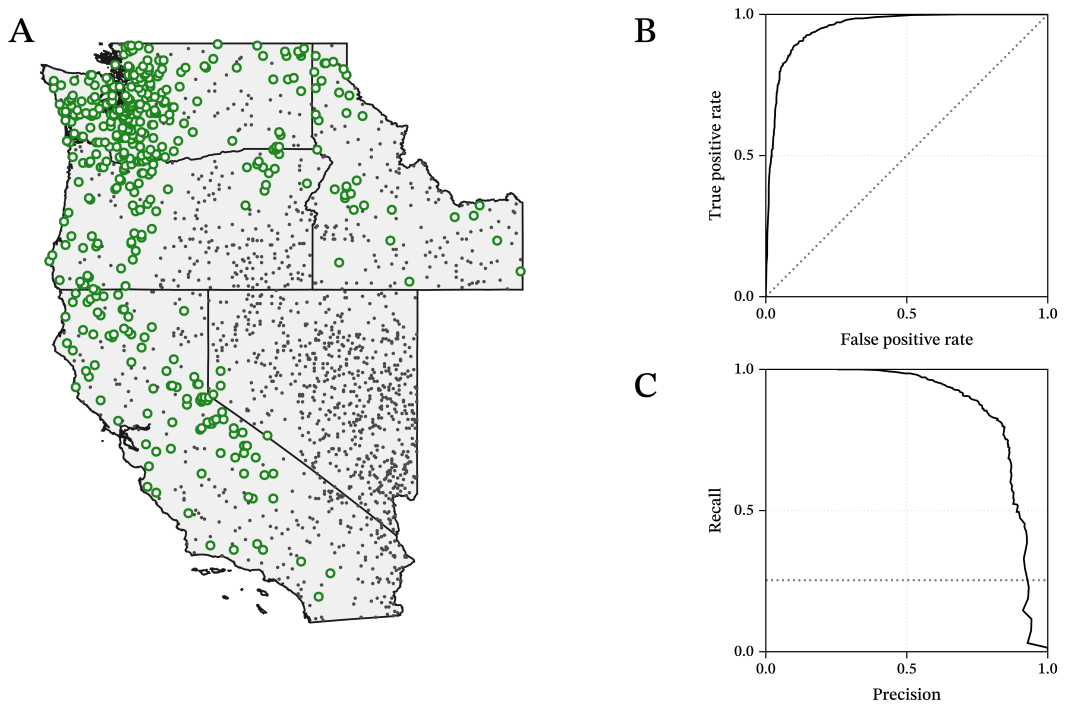

Figure 1: Overview of the occurrence data (green circles) and the pseudo-absences (grey points) for the states of, clockwise from the bottom, California, Oregon, Washington, Idaho, and Nevada (A). The underlying predictor data are at a resolution of 2.5 minutes of arc, and represented in the World Geodetic System 1984 CRS (EPSG 4326). The panels on the right column show the ROC curve (B) and PR curve (C), with the random classifier indicated by a dotted line. The area under the ROC curve is ≈ 96%.

and 10 minutes of arc show that the qualitative results presented hold. For the projection of the model under climate change, I only report the future data under the SSP370 scenario ("business as usual"), for the MRI ESM2-0 GCM, over the period 2081-2100.

The climatic novelty of the baseline *v.* future data is estimated through the Euclidean distance (Fitzpatrick et al. 2018), specifically by assigning as a novelty score for each pixel in the future the distance to its closest baseline analogue. This novelty is measured on de-meaned predictors with unit variance.

Species distribution model

All analyses are conducted using the `SpeciesDistributionToolkit` package (Poisot et al. 2025) for *Julia* 1.11.

Model structure

The model used here is a logistic regression, with interactions terms up to a maximum degree of two (preliminary analyses with random forests, naive Bayes classifiers, and rotation forests gave similar results). When trained on a vector of features $\boldsymbol{x}_i$ (with null means and unit variances), the model will return a probability p_+, which correspond to the probability of these environmental conditions being associated to the presenceof the species. This probability is turned into a presence/absence decision by comparing it to a threshold, as explained in a later section. Because this logistic regression is a deterministic classifier, the prediction $p_i +$ statisfies $0 \leq p_i + \leq 1$, and we use $p_- = 1 - p_+$ as the probability that the species is absent from the location.

Tuning

We tune this 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_+ is considered to be positive (turning the prediction of presence into $p_+ \geq \tau$). 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 validation. 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).

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 were retained as part of the final set of predictors if adding them increased the MCC for the model once retrained with this new variable.

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 $p_+ \geq \tau$ – a process known as moving threshold classification (Liu et al. 2013, 2016). 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 picking the value out of 200 linearly spaced value between the smallest and largest prediction made on the training set. The value of τ that maximizes the MCC during cross-validation was selected as the optimal threshold for the classifier. Note that even though our decision rule for the presence of the species is $p_+ \geq \tau$, we will keep the information about p_- as is it required for conformal prediction.

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). Models 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* result in 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.

An introduction to conformal prediction

Conformal prediction differs from regular prediction in that, rather than a single point prediction, it returns sets corresponding to the ensemble of *credible* outcomes given an input $\boldsymbol{x}$ representing environmental conditions at which we seek to make the prediction. Given the observed quantiles of the model output on validation data, these sets are obtained through a simple calibration step.

Therefore, CP requires 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 how they are obtained), and the notion of *coverage* , which is a measure of tolerance to error.

Understanding conformal predictions

By contrast to the non-conformal SDM, the conformal classifier returns, for an input of environmental predictors $\boldsymbol{x}$, a set C containing the "credible outcomes" for this prediction. This set is termed the *credible set*, and under a binary classification task (the species is either present or absent), there are four possible combinations for the content of credible sets: $C = \{+\}$, $C = \{-\}$, $C = \{+, -\}$, and $C = \varnothing$.

The first two cases are simple: if the credible set contains a single output, the model can confidently make a prediction that excludes the other class. In the case of $C = \{+\}$, for example, the point prediction for the presence score p_+ is high enough that the outcome of absence can be ruled out given the known predictions on training examples. In some cases, the credible set may contain both classes, as in $C = \{+, -\}$. Although they may not be *equally likely* (there is no guarantee that $p_+ \approx p_-$), the scores are close enough to not confidently exclude one of the outcomes from the model prediction. In the specific cases of SDMs, these correspond to areas of true uncertainty, where the known training examples credibly support both the presence or absence of the species. The final situation, $C = \varnothing$, corresponds to pathological cases where *neither* outcome can be credibly supported. Given the training data (and the distribution of presences and absences), the model is not able to make a prediction for this input. The increased frequency 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.

These situations correspond to four different outcomes in terms of the SDM certainty about the distribution of the species. The most intuitive situation is $C = \{+\}$ or $C = \{-\}$, 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. Throghout this manuscript, I will

refer to these predictions as “sure presences” and “sure absences”, as they convey the information that there is no reason to expect that the prediction is uncertain. The second situation, $C = \{+, -\}$, corresponds to inputs for which the presence and the absence of the species are credible, and I will refer to them as “unsure”. The rare cases where $C = \varnothing$ will be “undetermined” predictions.

Obtaining conformal predictions

There are several ways to decide whether a point prediction from the model results in which credible set. A core assumption of CP is that the data used for training should be exchangeable, or in other words, their joint probability distribution should be (close to) invariant under finite permutations (Aldous 1985). This will almost never be the case for data with a spatial structure; nevertheless, this does not rule out the use of CP for species distribution modeling, as Oliveira et al. (2024) show that CP is acceptably robust to lack of exchangeability.

The central idea of CP is to associate a conformal score to a point prediction. This can be achieved by applying the softmax function to the values for p_+ and p_-, giving

$$s_+ = \frac{\exp p_+}{\exp p_+ + \exp(1 - p_+)}, s_- = \frac{\exp(1 - p_+)}{\exp p_+ + \exp(1 - p_+)} \tag{1}$$

The conformal score associated to a prediction is $1 - s_\cdot$, where $\cdot$ is the prediction ($+$ or $-$) made by the model. We call the distribution of conformal scores $\mathcal{S}$. Note that this can be done without using the softmax function, but it is included here as it is best practice for classification.

The next step is to identify a critical value $\hat{q}$ above which a conformal score indicates that the prediction it describes is credible. This critical value is picked by examining the empirical quantile distribution of the conformal scores calculated over n training examples, and an acceptable level of risk α (explained in depth in the next sub-section), and specifically by identifying the q_i-th quantile, where

$$q_i = \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \tag{2}$$

The corresponding value of S below which a proportion q_i of values lies is $\hat{q}$. In other, more intuitive words, the value q_i indicates what proportion of wrong classification events we must accept before we have accumulated enough evidence to be confident about a prediction. When performing the prediction, we calculate the score of a new prediction according to Equation 1. For every possible class x, if $s_x \geq (1 - \hat{q})$, this class is retained as part of the credible set.

The value of $\hat{q}$ can be obtained either through using a holdout set for training (Split Conformal Prediction), by retraining the model in a way aking to Leave-One-Out cross-validation (Full Conformal Prediction), through the use of quantile regression (Romano et al. 2019), or through taking the median of several estimates of $\hat{q}$ after cross-validation (Vovk et al. 2018). In this manuscript, I employ the later method, as it provides a rapid and statistically acceptable estimate of $\hat{q}$, without requiring too much computing time.

To summarize, the output of the conformal classifier is, in a sense, a point estimate of the credible outcomes of a model, using the value estimated for p_+ as well as knowledge about which of these were associated to the correct label in the training data. 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) like 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 logisitc regression model, we can further divide the "unsure" predictions as a function of whether they would be within the range as predicted by the SDM, which I will call "unsure presences"; the other unsure predictions are referred to as "unsure absences".

The coverage level

CP allows users to set a desired error rate, α, which appeared in Equation 2. Intuitively, what CP does, is inform the user on whether the credible set contains the true value with probability $1 - \alpha$, which allows to directly interpret this value as a true 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.

Chaning the risk level α leads to different estimates of how commonly multiple classes will be accepted as a credible outcome. Using a low level of risk ($\alpha \approx 0$) yields usually leads to all outcomes being credible ($\hat{q} \approx 1$), at the cost of a very high uncertainty. When values of α get too large ($\hat{q} \approx 0$), no class can be confidently predicted, and the model will eventually always return $C = \emptyset$. Although this later situation is more difficult to make sense of intuitively, a value of inefficiency that gets smaller than 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). Conformal prediction can therefore inform us on the acceptable risk levels we can operate under given a trained predictive model.

In the rest of this analysis, I will 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. Recall that the CP credible sets are estimated based on the model output, and therefore even when aiming for full coverage, there may be non-ambiguous combinations of environmental predictors.

Measure	Validation	Training	Ensemble
MCC	0.75	0.76	0.76
NPV	0.93	0.93	0.94
PPV	0.82	0.83	0.82
κ	0.75	0.76	0.76
TSS	0.74	0.75	0.76
Accuracy	0.91	0.91	0.91

Table 1: Overview of measures of model performance for the validation and training sets of the SDM, as well as the same measures for the ensemble model (measured on the out-of-bag models only). The values of κ and the true-skill statistic are generally comparable to the MCC, but are included as they are commonly reported in the SDM litterature (Allouche et al. 2006). The high values of the negative and positive predictive values indicate that the model is suitable to detect both presences and absences.

Results

Performance of the baseline model

In panels B and C of Figure 1, we report the ROC and PR curves for the model. As evidenced by both these diagnostic tools, the model achieves a very high predictive accuracy. In Table 1, we report additional measures of performance for the training and validation set of the model (so as to ensure that the model is not performing better on training data), as well as a measure of the performance of the ensemble, to show that it can make valid predictions in addition to quantifying variability. These results confirm that the model is able to identify areas that are suitable to the species, and can be used for CP.

Before applying CP, it is useful to examine the output of the SDM in space. The predictions of the model for the entire region are given in Figure 2, alongside information about the model variability. Areas of lowest variability (according to the IQR based on non-parametric boostrap results from the ensemble) seem to be associated with the absence of the species, with the variability mostly increasing within the predicted range.

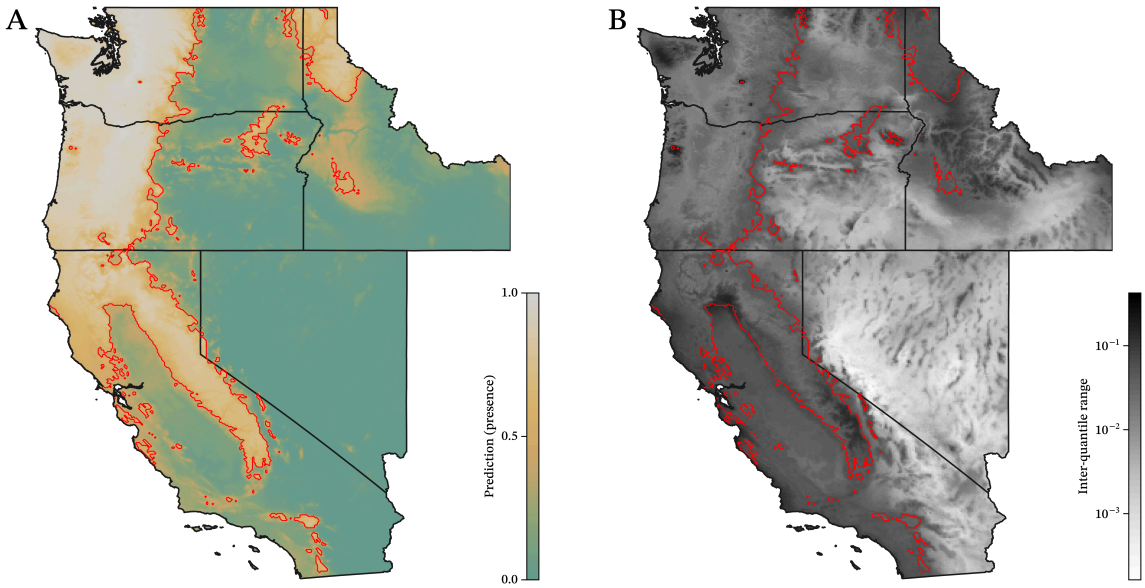

Figure 2: Overview of the probability p_+ returned by the model (A), and the inter-quantile range of the non-parameteric bootstrap model predictions (B). The range, *i.e.* the limit of cells for which $p_+ \geq \tau$, is indicated by a solid red line; I maintain this convention for all subsequent figures. Note that the scale of the variability is logarithmic, as the model shows good performance and therefore has low variability overall.

Conformal prediction of the species range

Before discussing the spatial output of running the conformal model, it is worth considering why the thresholding step as visualized in Figure 2 is not really providing us with a set of certain presences and absences. When optimizing the threshold τ above which a prediction p_+ from the non-conformal model is determined to be a presence, we inherently establish a sort of certain presences and certain absences, specifically by ignoring the possibility that there can be uncertain predictions. 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_+ 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.

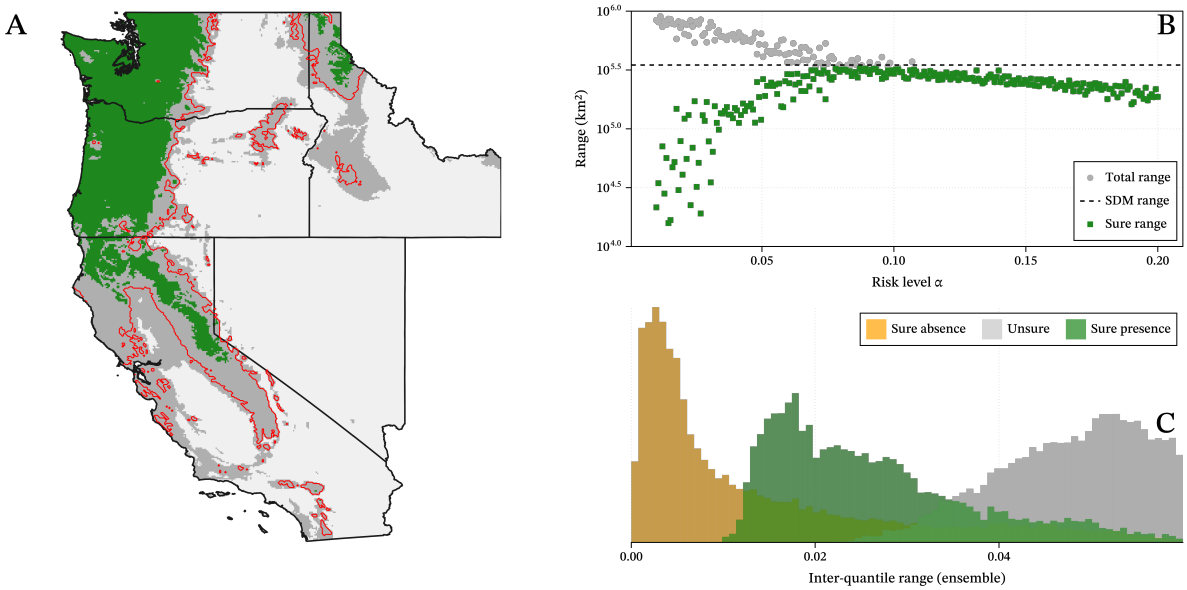

Figure 3: Overview of areas where the presence of the species is certain according to the CP model under a risk level $\alpha = 0.05$ (A). The certain areas are in dark green, and the uncertain areas, wherein both presence and absence are credible, are in dark grey. (B) Surface covered by the sure absence and total range (including the superficy of the unsure area) for different risk levels. Note that for $\alpha \approx 0.1$, the total predicted range starts being lower than the range predicted by the SDM, and the uncertain range collapses. (C) Distribution of variability from Figure 2B by type of CP model outcome.

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_+ is just above the threshold) and for areas that are outside of it (for which p_+ 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.

In Figure 3, we show that this prediction indeed stands: the range as predicted by the SDM (fig. 3A) falls within the range of unsure predictions. We also see that lowering the risk level α leads to a contraction of the area (in km^2) considered to be credibly associated to only the presence of the species ($C = \{+\}$), while the range that is ambiguous ($C = \{+, -\}$) increases (Figure 3B). 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, which I discuss in the next section) are *potentially* part of the range of the species: some certainly, some uncertainly. Depending on the purpose for which the SDM is produced, the uncertain areas can be treated differently. As Prescott et al. (2025) argue, when dealing with invasive species, it may be more reasonable to interpret SDMs by erring on the side of caution, which here would mean considering that unsure presence area should be considered part of the species's range. On the other hand, when SDMs are meant to guide conservation actions that are costly or should be focused on areas of high certainty of suitability for the target species (Pěknicová & Berchová-Bímová 2016), it may make sense to ignore the unsure presences.

Relationship between variability and uncertainty

Note that the relationship between the certainty associated to CP, and the variability under the ensemble model presented in Figure 2B is nuanced: in fig. 3C, it appears that although areas identified as unsure using CP tend to have higher variability, there is considerable overlap between the categories. 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. There are a number of points classified as sure presences for which the IQR is very high, **i.e.** points whose certainty is not affected by undersampling the training data. Notably, the results in fig. 3C show that it is not possible to find a cutoff in the measure of bootstrap variability that would identify areas of model 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 known dataset.

Identification of undetermined areas

In Figure 3B, we see that there is a risk level above which the total predicted range starts to get lower than the range predicted bu the SDM. We can explain this behavior through the lens of the number of undetermined predictions, *i.e.* the number of inputs for which the CP model returns $C = \varnothing$.

In fig. 4A, we see that above $\alpha \approx 0.1$, the inefficiency of the classifier starts to fall under 1 - this indicates that *on average*, the model is returning fewer than one output for each prediction. In a sense, this creates an upper limit to the risk we can accept: the model trained on this dataset does not support conformal prediction for larger risk levels. In fig. 4B, we see that this change of behavior in the model is indeed resulting in an increase in the range for which the model makes no prediction, which gets larger when the risk level is too high. The spatial distribution of undetermined areas is shows in fig. 4C for $\alpha = 0.2$: these areas are concentrated around the range limit as identified by the SDM. This suggests that using a risk level that it too high would result to no conformal predictions being made for the areas where our need to accurately quantify uncertainty are the most important, and calls for a cautious investigation of the appropriate risk level.

Model explanation

In this section, I 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. I rely on the common Monte-Carlo approximation of Shapley

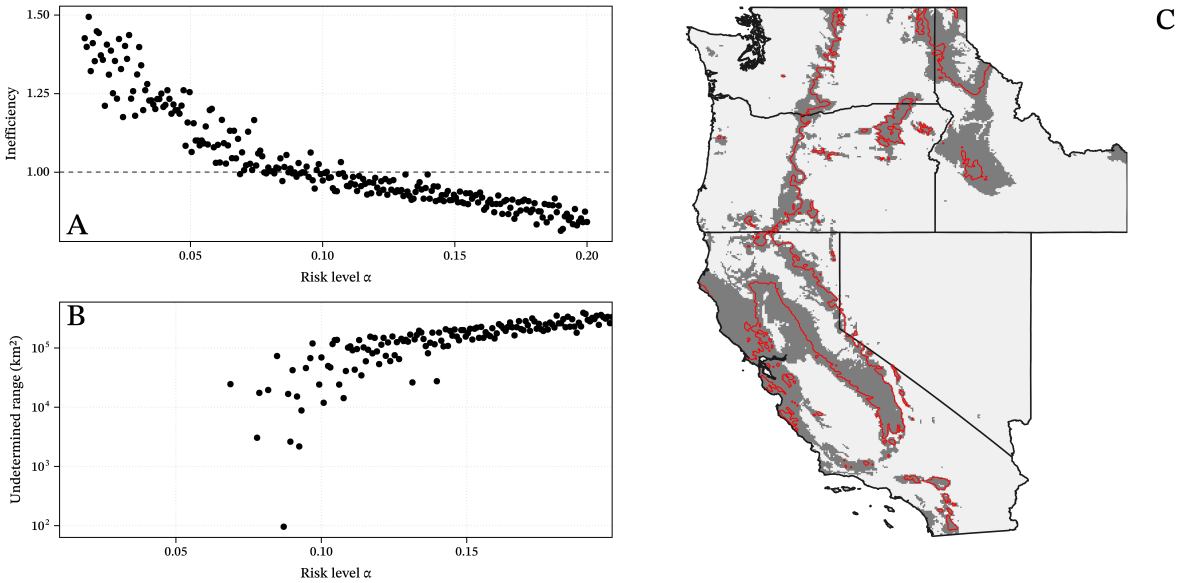

Figure 4: Inefficiency (average number of classes in the credible set) for various levels of α (A); above $\alpha \approx 0.1$, the conformal prediction starts returning empty credible sets. This results in an increase in the spatial area for which no prediction can be made (B). For $\alpha = 0.2$, these areas are distributed around the limit of the predicted range, showing that the areas in which uncertainty quantification are most important cannot be predicted.

values (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 $\varphi_i \in [-1, 1]$, 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, $p = \hat{p} + \sum_i^{\text{variables}} \varphi_i$. Because of this additive property, the importance of variables across many predictions is usually measured as the average of $|\varphi|$, 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 and towards its actual value. 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

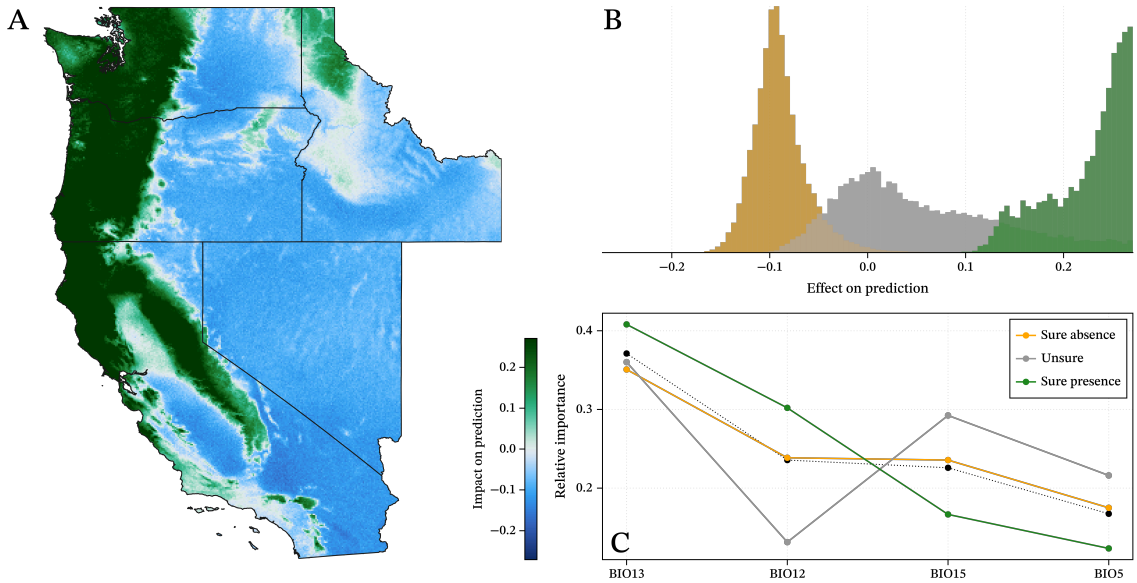

Figure 5: Overview of the effect of the most important predictor (A); areas with high values indicate that the value of BIO13 at this location make the presence of the species more likely. These values are associated to different prediction certainties (B), with predictions within the unsure range being centered around 0 (*i.e.* not moving the needle on the average prediction one way or another). Nevertheless, the contribution of the variables in different uncertainty categories are different (C), suggesting that Shapley values can help create explanations of where uncertainty originates.

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 investigate whether the factors leading to a species being present or absent are the same. An example of the spatial contribution of a variable is given in Figure 5A.

We find that, for the most important variable (*i.e.* the one with the largest $\sum|\varphi|$), the contribution of this variable tracks the status of the prediction: it tends to be negative when the absence is certain, positive when the presence is certain, and around zero when the prediction is unsure (fig. 5B). This is a fairly remarkable result, in that it ties Shapley values (a tool to help with ML models interpretation) to CP (a technique to accurately convey uncertainty). In Figure 5C, I present the relative contribution of all selected variables split by the status of the prediction; this reveals that the Shapley values for sure presences and unsure areas are distributed in different ways. Notably, BIO15 is far more important in areas of high model uncertainty than in areas of either sure presences or absences. This suggests that the division of the prediction according to CP status can provide information about which sets of environmental conditions are driving the

uncertainty, thereby providing useful information to guide future sampling or model interpretation.

Conformal prediction and climate-induced range shifts

In a recent contribution, Smith & Levine (2025) suggest that because of issues around the use of thresholds, projections of SDMs under climate change scenarios may benefit from a more continuous perspective. In this section, I present a comparison of the conformal prediction of the range under a climate change scenario (SSP370. 2081-2100), to illustrate how the future conformal range can convey information about the certainty of some types of range shift. These results are presented in Figure 6.

Based on the comparison between the baseline (fig. 2A) and projected (fig. 6A) ranges, we can establish a series of transitions and their interpretations as range change scenarios, which are presented in fig. 6B. Areas that are certain both now and in the future, $\{+\} \rightarrow \{+\}$, can safely be assumed to be conserved. Areas that where unsure and become surely negative, $\{+, -\} \rightarrow \{-\}$ are *possible* losses, as they may have been presences in the baseline data, but are considered lost in the future. The reverse scenario, $\{+, -\} \rightarrow \{+\}$, corresponds to *possible* gains. Sure losses of range

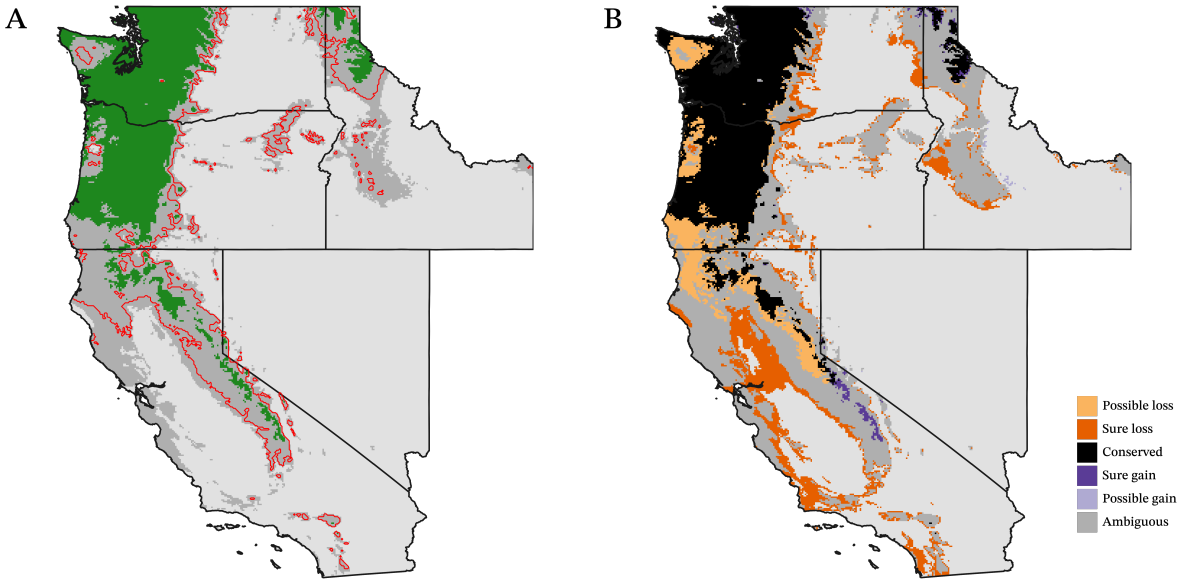

Figure 6: Overview of the conformal prediction of the range for the future climate data, equivalent to fig. 2A (panel A). Spatial distribution of areas where loss and gain are expected to be possible *v.* certain, as explained in main text (B).

correspond to the transition $\{+\} \rightarrow \{-\}$, and sure gains of range correspond to $\{-\} \rightarrow \{+\}$. Other situations are considered ambiguous.

By applying these rules on the predicted changes in presence/absence status, we can identify large areas that are confidently loss towards the Southern edge of the species's range, with very limited areas of either possible or sure gain, strongly suggesting that this species would undergo range contraction. Note that the area corresponding to ambiguous transitions is relatively large, which provides a good understanding of the possible variation to be expected under this climate change scenario.

Conformal prediction and climatic novelty

Zurell et al. (2012) highlight the importance of fully considering uncertainty when transferring the model to novel climate data: there is a chance that the future climate conditions will not have occurred in the training dataset, and therefore our confidence in the model outcome should be lowered. This covariate shift is well documented to decrease the performance of models (Mesgaran et al. 2014), and CP offers an opportunity to shine a different light on this phenomenon.

This task is particularly crucial given that entirely novel climatic conditions are likely to become the norm (Mahony et al. 2017), which in turn will drive the emergence of a novel biosphere globally (Kerr et al. 2025, Ordonez et al. 2024). In this section, I compare the results of conformation prediction to measures of climatic novelty, by partitioning the climate novelty according to the type of range shifts from fig. 6B. The study area shows higher novelty in parts of the range that are currently predicted to be habitable by the species; nevertheless, this does not translate to an association between types of prediction transition and the distribution of novelty within the regions undergoing this transition. In other words, the projected uncertainty under conformal prediction contributes different information when compared to measures of climatic novelty; specifically, it conveys the uncertainty tied to the model itself.

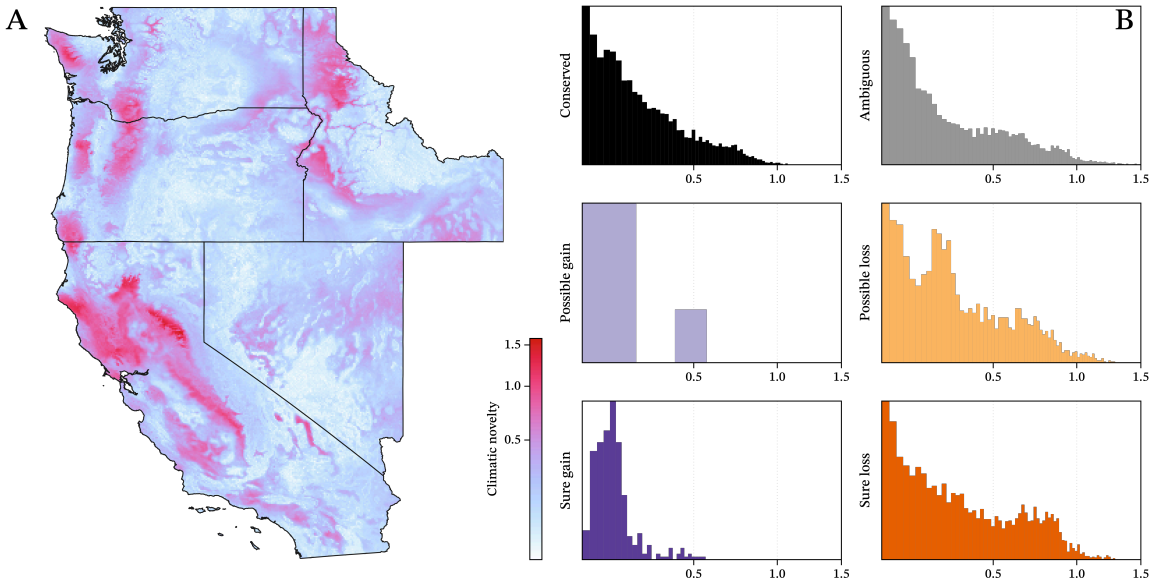

Figure 7: Climate novelty measured as Euclidean distance to the closest contemporary analogue (A); note that the scale is square-root transformed, as most areas show low novelty. Distribution of novelty values split by the expected transition in occupancy (B); colors are as in fig. 6B.

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 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 that 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 even when there is strong environmental signal (Mohd et al. 2016) these stochastic events can even manifest in areas that are close to the species’ environmental optimum

(Dallas et al. 2020). For these reasons, CP can produce interpretable estimates of uncertainty in species distribution models, and does not require the adoption of additional modeling tools or paradigms as it functions on an already trained model.

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 point estimate of the 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.

Bibliography

Aldous DJ. 1985. Exchangeability and related topics. In *Lecture Notes in Mathematics*, pp. 1–198. Berlin, Heidelberg: Springer Berlin Heidelberg

Allouche O, Tsoar A, Kadmon R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *The journal of applied ecology*. 43(6):1223–32

Angelopoulos AN, Bates S. 2023. *Conformal Prediction: A Gentle Introduction*. Hanover, MD: now

Balayla J. 2020. Prevalence threshold (ϕe) and the geometry of screening curves. *PloS one*. 15(10):e240215

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many?: How to use pseudo-absences in niche modelling?. *Methods in ecology and evolution*. 3(2):327–38

Beale CM, Lennon JJ. 2012. Incorporating uncertainty in predictive species distribution modelling. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. 367(1586):247–58

Beery S, Cole E, Parker J, Perona P, Winner K. 2021. Species Distribution Modeling for machine learning practitioners: A review. *arXiv [cs.LG]*

Benkendorf DJ, Schwartz SD, Cutler DR, Hawkins CP. 2023. Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models. *Ecological modelling*. 483(110414):110414

Bylander T. 2002. Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates. *Machine learning*. 48(1/3):287–97

Carlson CJ. 2020. embarcadero: Species distribution modelling with Bayesian additive regression trees in r. *Methods in ecology and evolution*. 11(7):850–58

Chen X, Dimitrov NB, Meyers LA. 2019. Uncertainty analysis of species distribution models. *PloS one*. 14(5):e214190

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):4

Dallas TA, Santini L, Decker R, Hastings A. 2020. Weighing the Evidence for the Abundant-Center Hypothesis. *Biodiversity informatics*. 15(3):81–91

Davies SC, Thompson PL, Gomez C, Nephin J, Knudby A, et al. 2023. Addressing uncertainty when projecting marine species' distributions under climate change. *Ecography*. 2023(11):

Drake JM. 2014. Ensemble algorithms for ecological niche modeling from presence-background and presence-only data. *Ecosphere (Washington, D.C)*. 5(6):1–16

Elith J. 2019. Species Distribution Modeling

Fannjiang C, Bates S, Angelopoulos AN, Listgarten J, Jordan MI. 2022. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences of the United States of America*. 119(43):e2204569119

Feng X, Park DS, Walker C, Peterson AT, Merow C, Papeş M. 2019. A checklist for maximizing reproducibility of ecological niche models. *Nature ecology & evolution*. 3(10):1382–95

Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS. *International journal of climatology: a journal of the Royal Meteorological Society*. 37(12):4302–15

Fitzpatrick MC, Blois JL, Williams JW, Nieto-Lugilde D, Maguire KC, Lorenz DJ. 2018. How will climate novelty influence ecological forecasts? Using the Quaternary to assess future reliability. *Global Change Biology*. 24(8):3575–86

Fontana M, Zeni G, Vantini S. 2020. Conformal Prediction: A unified review of theory and new challenges. *arXiv [cs.LG]*

Foxon F. 2024. Bigfoot: If it's there, could it be a bear?. *Journal of zoology (London, England: 1987)*

Franklin J. 2023. Species distribution modelling supports the study of past, present and future biogeographies. *Journal of biogeography*. 50(9):1533–45

Gammerman A, Vovk V, Vapnik V. 1998. Learning by transduction. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. 148–55. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Janitza S, Hornung R. 2018. On the overestimation of random forest's out-of-bag error. *PloS one*. 13(8):e201904

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):e12978

Kerr MR, Ordonez A, Riede F, Atkinson J, Pearce EA, et al. 2025. Widespread ecological novelty across the terrestrial biosphere. *Nature ecology & evolution*. 1–10

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

Liu C, Newell G, White M. 2016. On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and evolution*. 6(1):337–48

Liu C, White M, Newell G. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of biogeography*. 40(4):778–89

Lozier JD, Aniello P, Hickerson MJ. 2009. Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *Journal of biogeography*. 36(9):1623–27

Mahony CR, Cannon AJ, Wang T, Aitken SN. 2017. A closer look at novel climates: new methods and insights at continental to landscape scales. *Global change biology*. 23(9):3934–55

Mesgaran MB, Cousens RD, Webber BL. 2014. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity & distributions*. 20(10):1147–59

Messeri L, Crockett MJ. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*. 627(8002):49–58

Mohd MH, Murray R, Plank MJ, Godsoe W. 2016. Effects of dispersal and stochasticity on the presence–absence of multiple species. *Ecological modelling*. 342:49–59

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(767):333–80

Oliveira RI, Orenstein P, Ramos T, Romano JV. 2024. Split conformal prediction and non-exchangeable data. *Journal of machine learning research: JMLR*. 25(225):1–38

Ordonez A, Riede F, Normand S, Svenning J-C. 2024. Towards a novel biosphere in 2300: rapid and extensive global and biome-wide climatic novelty in the Anthropocene. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. 379(1902):

Parker EJ, Weiskopf SR, Oliver RY, Rubenstein MA, Jetz W. 2024. Insufficient and biased representation of species geographic responses to climate change. *Global change biology*. 30(7):e17408

Petitpierre B, Broennimann O, Kueffer C, Daehler C, Guisan A. 2016. Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global ecology and biogeography: a journal of macroecology*. 26(3):275–87

Petropoulos F, Hyndman RJ, Bergmeir C. 2018. Exploring the sources of uncertainty: Why does bagging for time series forecasting work?. *European journal of operational research*. 268(2):545–54

Phillips SJ, Elith J. 2013. On estimating probability of presence from use-availability or presence-background data. *Ecology*. 94(6):1409–19

Poisot T, Bussières-Fournel A, Dansereau G, Catchen MD. 2025. A Julia toolkit for species distribution data. *EcoEvoRxiv*

Prescott VA, Marte J, Keller RP. 2025. Performance of alternative methods for generating species distribution models for invasive species in the Laurentian Great Lakes. *Fisheries*. vuaf12

Pěknicová J, Berchová-Bímová K. 2016. Application of species distribution models for protected areas threatened by invasive plants. *Journal for nature conservation*. 34:1–7

Rocchini D, Hortal J, Lengyel S, Lobo JM, Jiménez-Valverde A, et al. 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in physical geography*. 35(2):211–26

Romano Y, Patterson E, Candès E. 2019. Conformalized Quantile Regression. *Neural Information Processing Systems*. 32:3538–48

Romano Y, Sesia M, Candès EJ. 2020. Classification with valid and adaptive coverage. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 3581–91. Red Hook, NY, USA: Curran Associates Inc.

Roth AE. 1988. Introduction to the Shapley value. In , pp. 1–28. Cambridge University Press

Sadinle M, Lei J, Wasserman L. 2018. Least Ambiguous Set-Valued Classifiers With Bounded Error Levels. *Journal of the American Statistical Association*. 114(525):223–34

Shafer G, Vovk V. 2007. A tutorial on conformal prediction. *Journal of machine learning research: JMLR*. (12):371–421

Smith JR, Levine JM. 2025. Linking relative suitability to probability of occurrence in presence-only species distribution models: Implications for global change projections. *Methods in ecology and evolution*

Soininen J, Korhonen JJ, Luoto M. 2013. Stochastic species distributions are driven by organism size. *Ecology*. 94(3):660–70

Soley-Guardia M, Alvarado-Serrano DF, Anderson RP. 2024. Top ten hazards to avoid when modeling species distributions: a didactic guide of assumptions, problems, and recommendations. *Ecography*. 2024(4):

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 (Basel, Switzerland)*. 23(4):

Thuiller W, Guéguen M, Renaud J, Karger DN, Zimmermann NE. 2019. Uncertainty in ensembles of global biodiversity scenarios. *Nature communications*. 10(1):1446

Tibshirani RJ, Barber RF, Candes EJ, Ramdas A. 2019. Conformal Prediction Under Covariate Shift. *arXiv [stat.ME]*

Touati S, Radjef MS, 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

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. 2021. Modelling species presence-only data with random forests. *Ecography*. 44(12):1731–42

Vovk V, Nouretdinov I, Manokhin V, Gammerman A. 2018. Cross-conformal predictive distributions. *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*. 91:37–51

Wadoux AMJ-C, Saby NPA, Martin MP. 2023. Shapley values reveal the drivers of soil organic carbon stock prediction. *SOIL*. 9(1):21–38

Zurell D, Elith J, Schröder B. 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity & distributions*. 18(6):628–34

Zurell D, Franklin J, König C, Bouchet PJ, Dormann CF, et al. 2020. A standard protocol for reporting species distribution models. *Ecography*. 43(9):1261–77