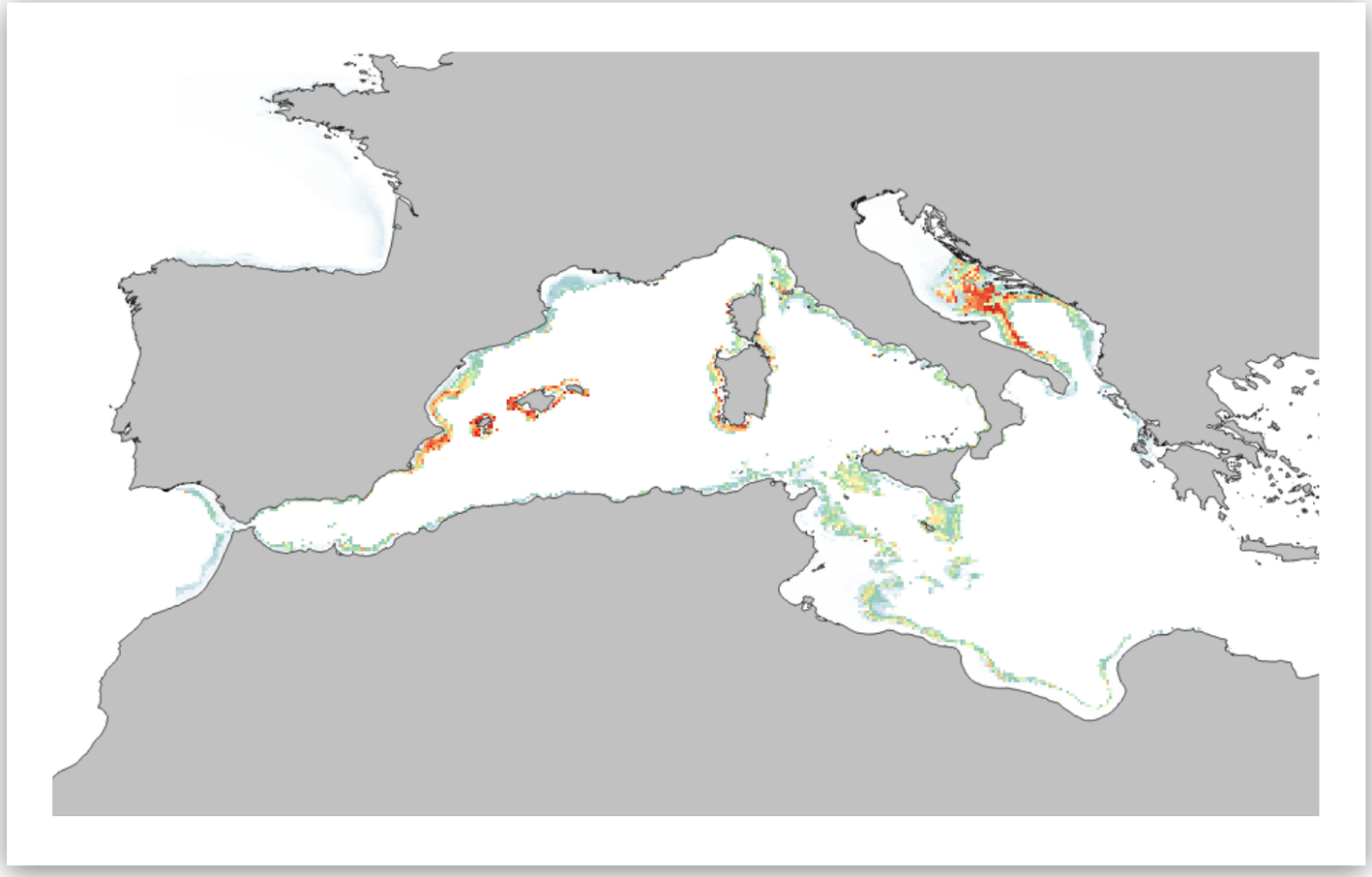
Marine Ecological Modelling
Global Climate Change

Evaluating predictive performance and setting decision thresholds

Jorge Assis, PhD // theMarineDataScientist, jmassis@ualg.pt
2020, Centre of Marine Sciences, University of Algarve

Model evaluation

Also called ‘validation’ or ‘performance’, is crucial to (1) **verify if predictions** (in terms of presence or absence) **are consistent with the observations**, (2) **assess for potential for transferability** and (3) **ecological realism**.



Is the model acceptable for the purpose?

Model evaluation

Prediction errors can be of ‘false positives’, when the model predicts occurrence in places where the species has not been observed, **and ‘false negatives’**, when the model predicts absent in places where it has been observed. Can be summarized in contingency / confusion matrices.

Contingency table or confusion matrix

Types of prediction errors

		Observation	
		Presence	Absence
Prediction	Presence	True Positive	False Positive
	Absence	False Negative	True Negative

Perfect models only retrieve true positives and true negatives.

Evaluation criteria

The elements of the contingency table are used to compute various evaluation criteria that measure the performance of the model.

Sensitivity: proportion of presences correctly predicted;

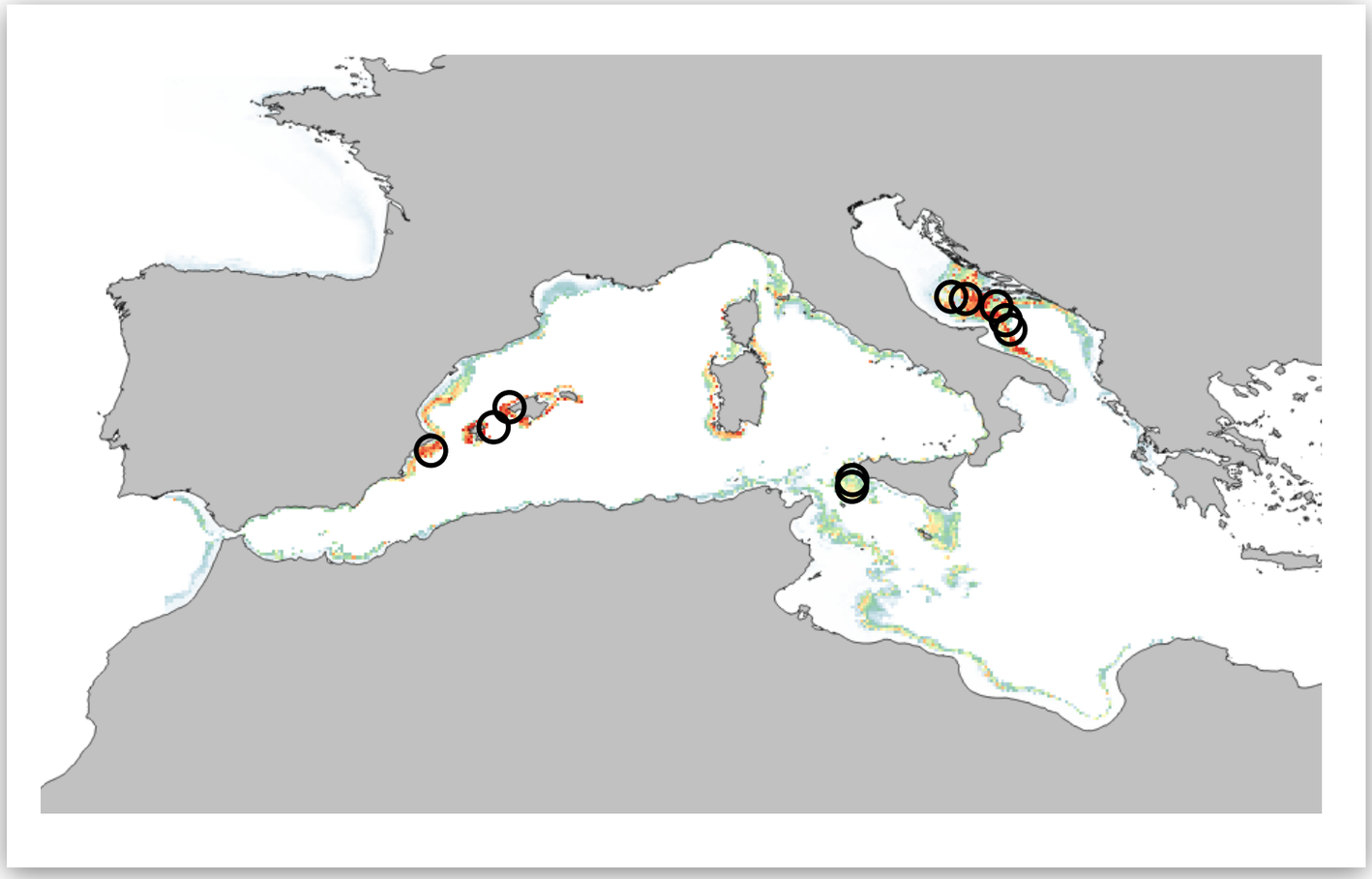
Specificity: proportion of absences correctly predicted;

True Skill Statistics: 1 - Sensitivity + Specificity (describes how well the model predicts presences and absences);

Area Under the Curve of the Receiver Operating Characteristic (based on sensitivity only, is the most widely used index).

Evaluation criteria

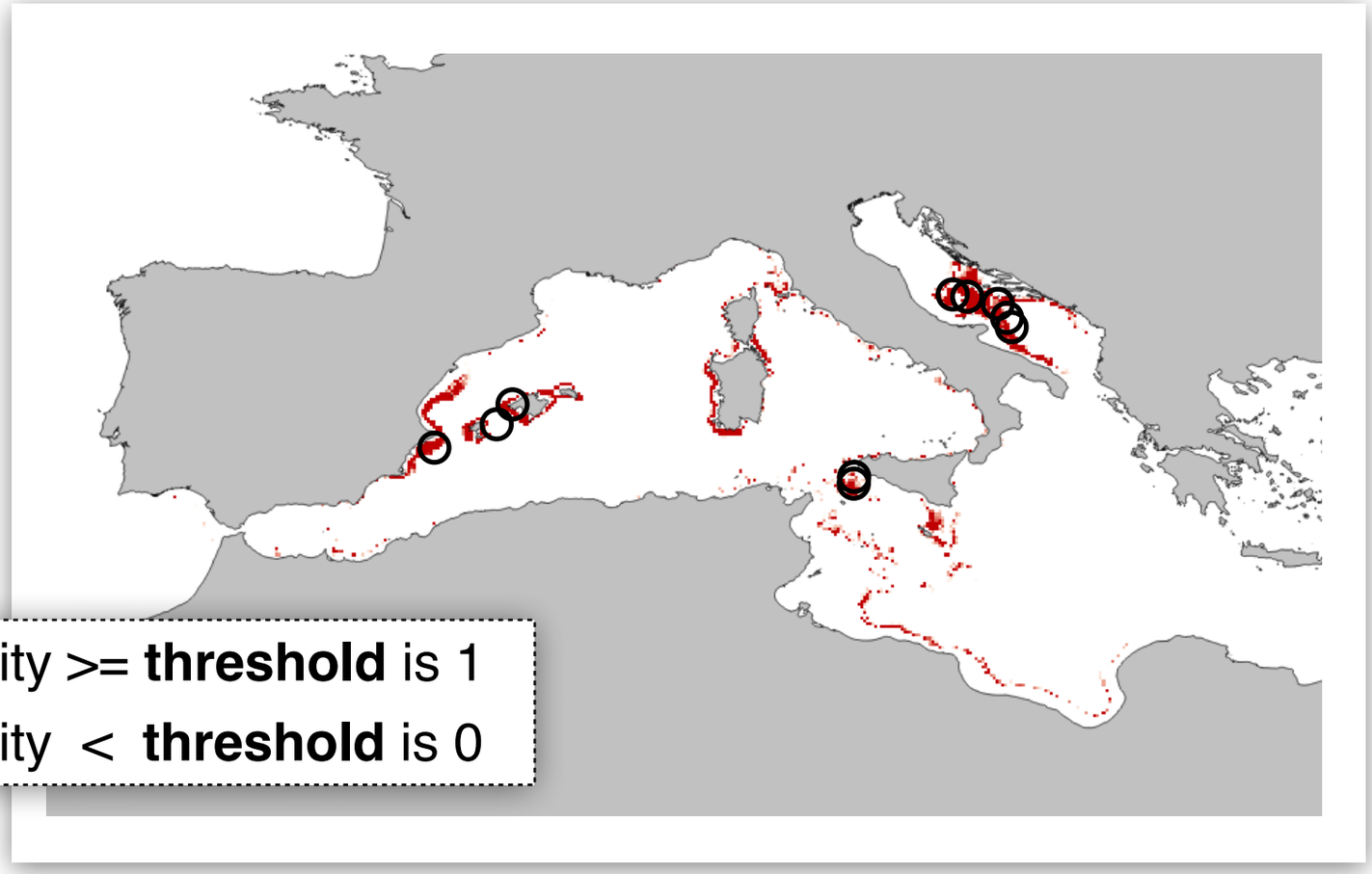
Predictions are continuous surfaces (e.g., probability or occurrence or suitability; from 0 to 1); This, one cannot assess accuracy between a presence (1) and the model output (e.g., P: 0.7).



Predictions need to be reclassified as binomial responses for comparison with the actual observed data.

Evaluation criteria

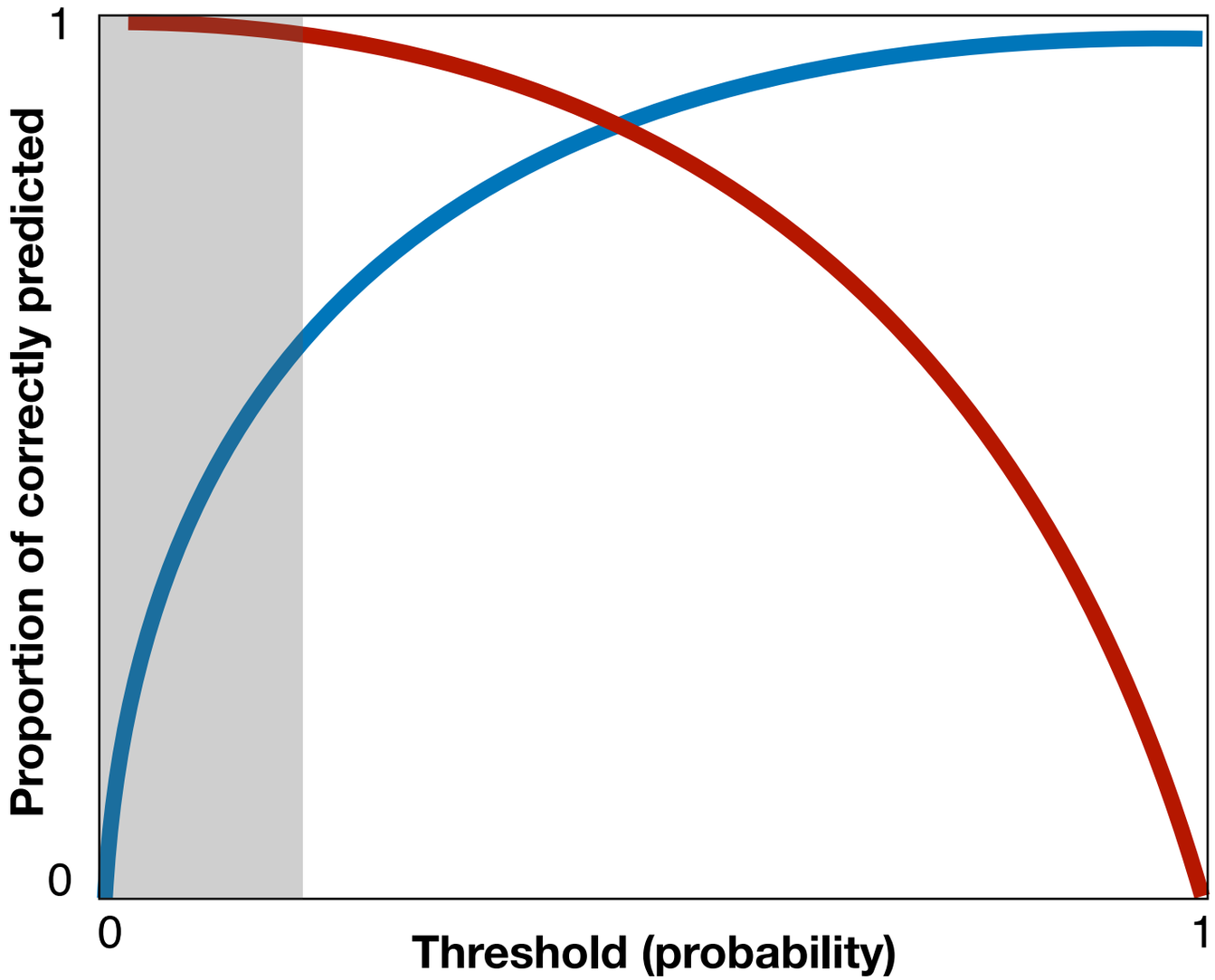
Predictions are continuous surfaces (e.g., probability or occurrence or suitability; from 0 to 1); This, one cannot assess accuracy between a presence (1) and the model output (e.g., P: 0.7).



Predictions need to be reclassified as binomial responses for comparison with the actual observed data.

Evaluation criteria

Sensitivity and specificity vary with the potential range of thresholds (0 to 1) that reclassify models to binomial outputs.



Probability >= **threshold** is 1

Probability < **threshold** is 0

Effect on the indices

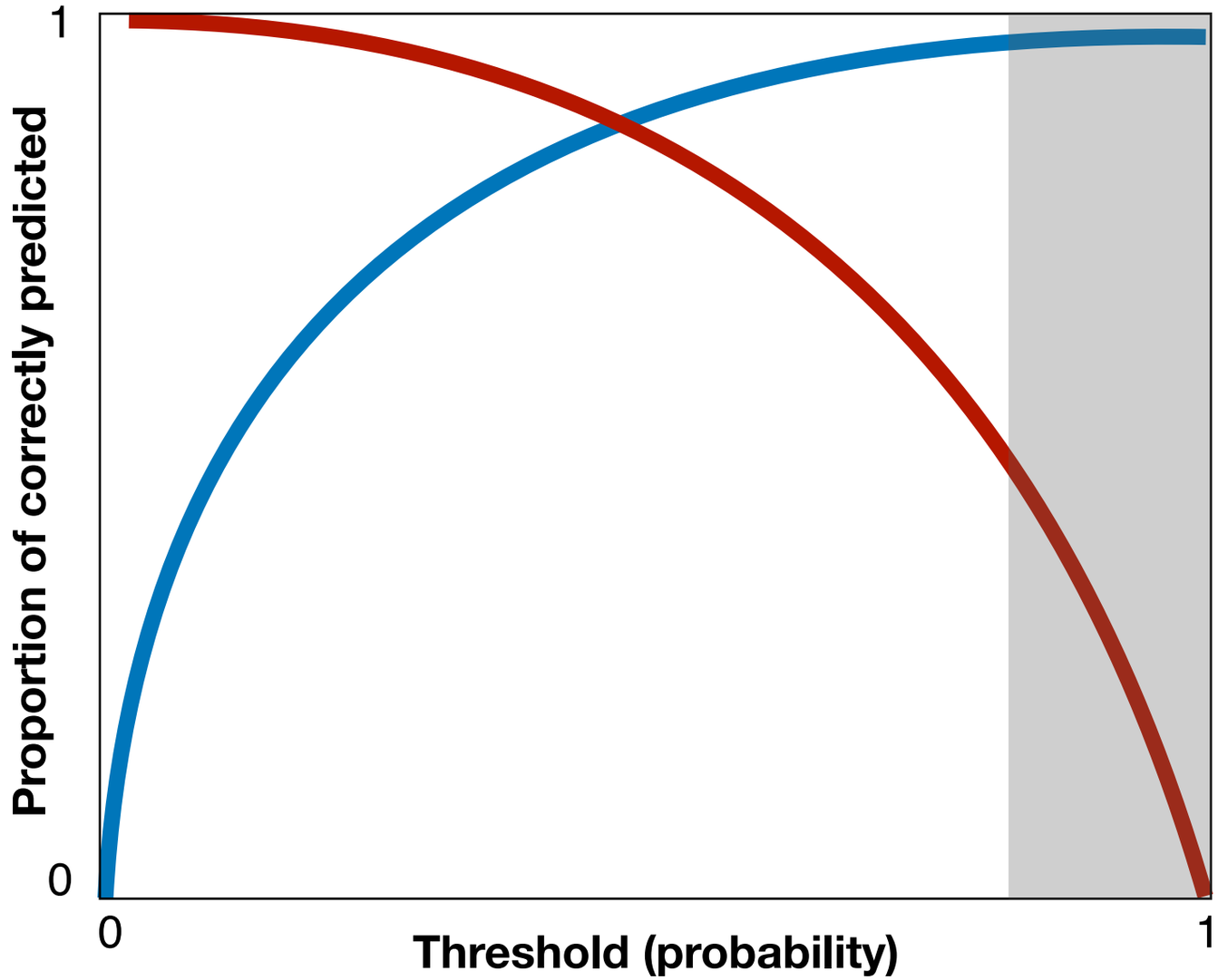
Sensitivity (presences)

Specificity (absences)

With a low threshold most cells (in the map) will return 1.

Evaluation criteria

Sensitivity and specificity vary with the potential range of thresholds (0 to 1) that reclassify models to binomial outputs.



Probability >= **threshold** is 1

Probability < **threshold** is 0

Effect on the indices

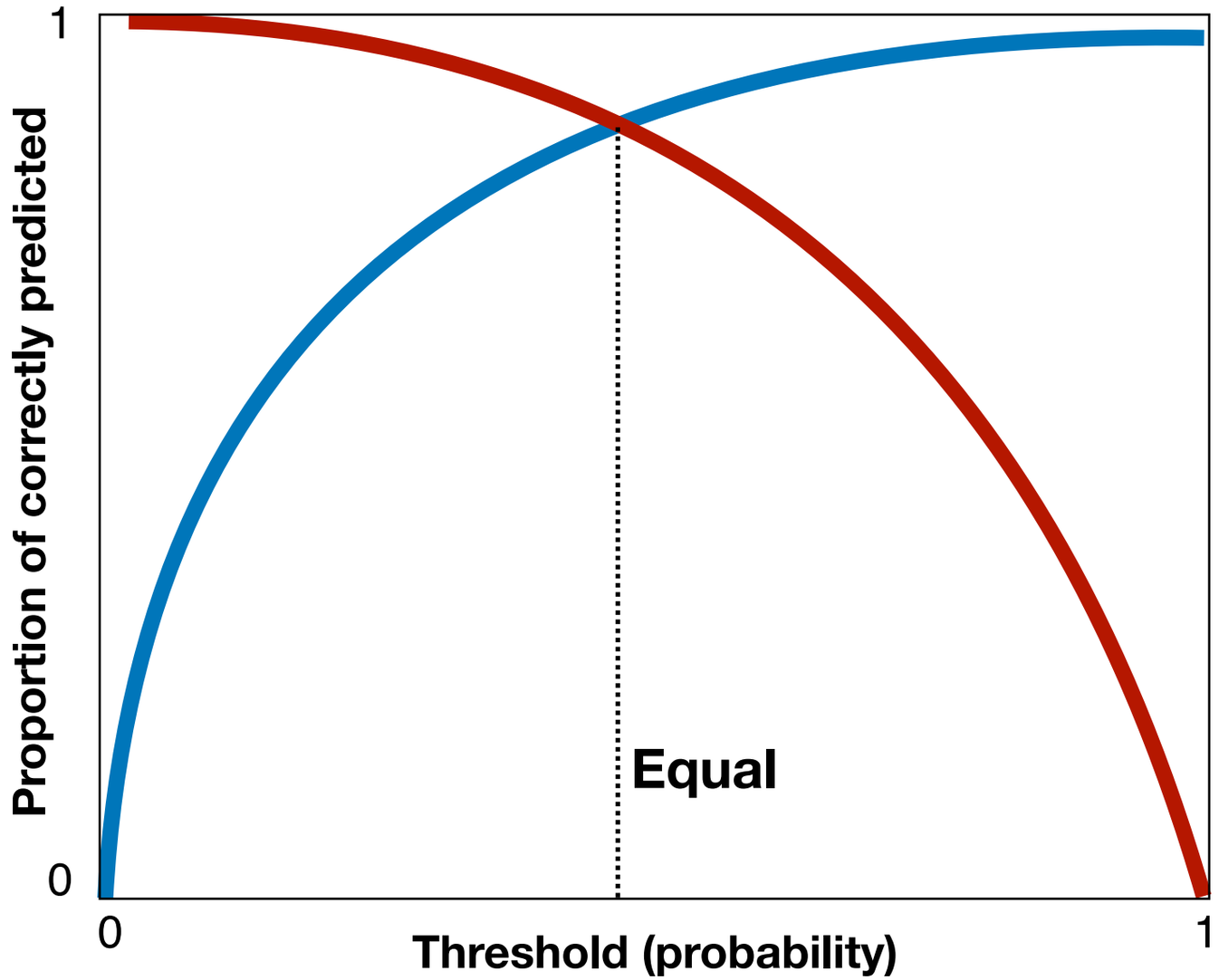
Sensitivity (presences)

Specificity (absences)

With a high threshold most cells (in the map) will return 0.

Evaluation criteria

There are threshold rules design to maximize the agreement between observed data and the produced binomial surface.



Probability >= **threshold** is 1

Probability < **threshold** is 0

Effect on the indices

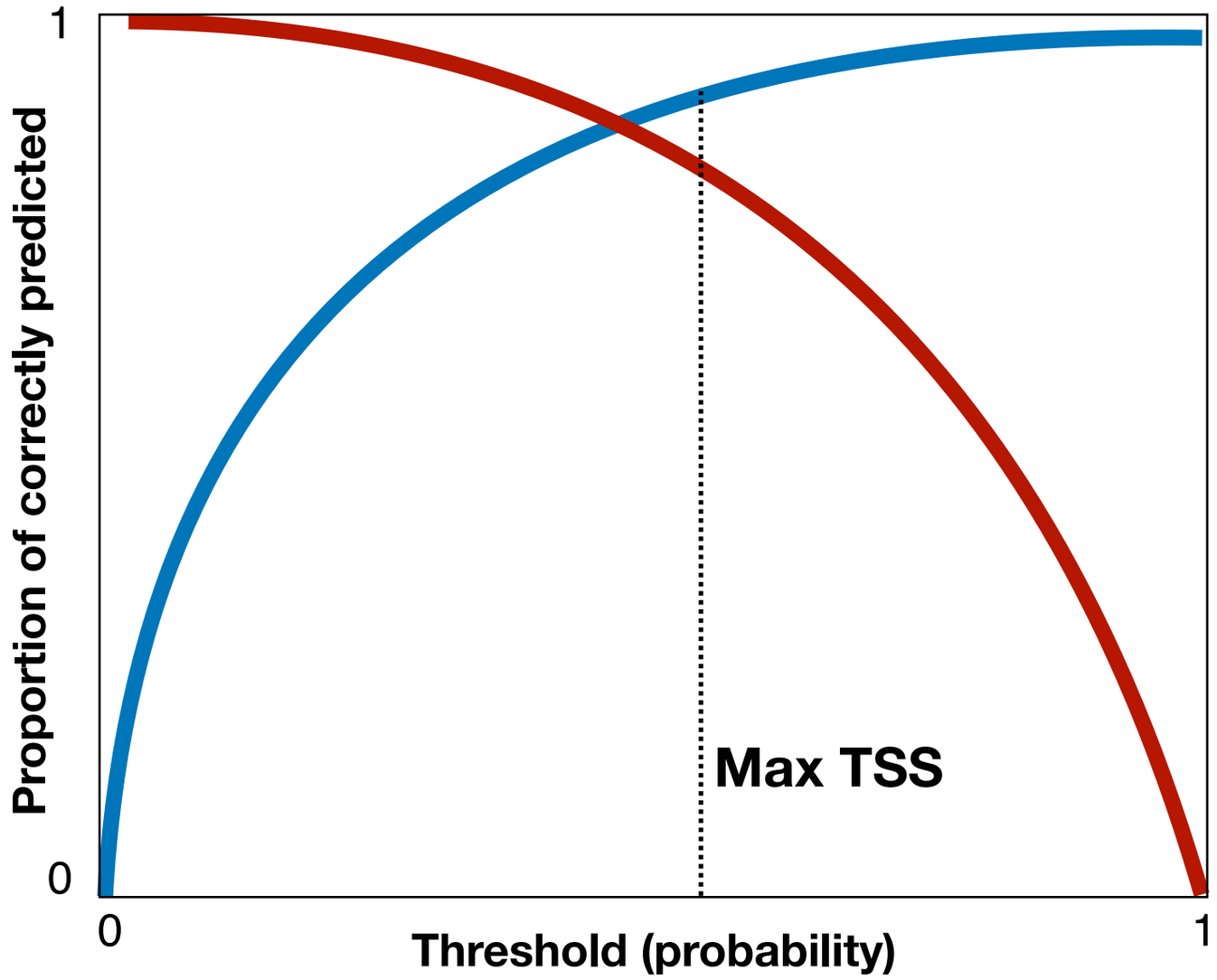
Sensitivity (presences)

Specificity (absences)

(1) **Equal sensitivity and specificity**

Evaluation criteria

There are threshold rules design to maximize the agreement between observed data and the produced binomial surface.



Probability >= **threshold** is 1

Probability < **threshold** is 0

Effect on the indices

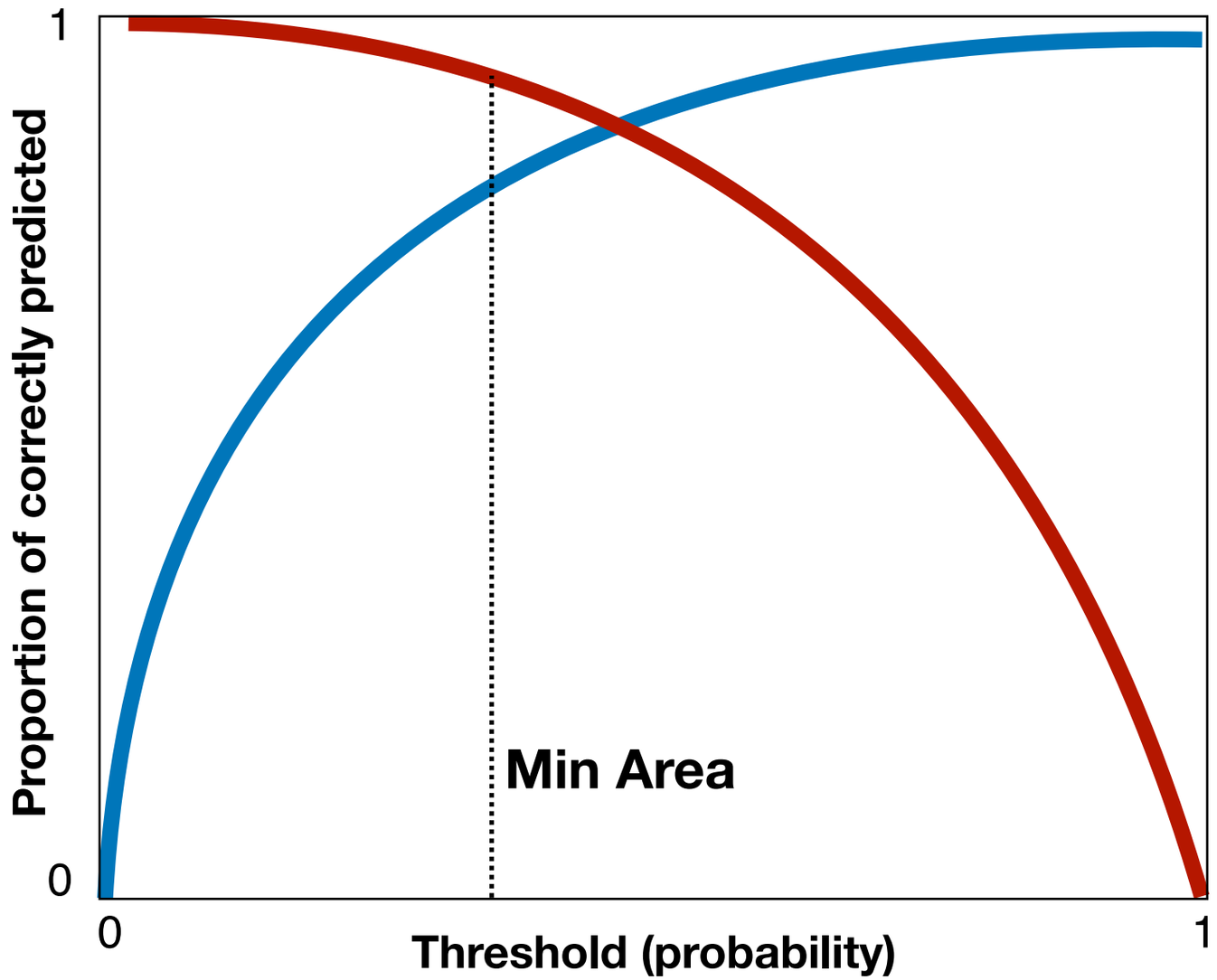
Sensitivity (presences)

Specificity (absences)

(2) **Maximization of sensitivity + specificity**

Evaluation criteria

There are threshold rules design to maximize the agreement between observed data and the produced binomial surface.



Probability >= **threshold** is 1

Probability < **threshold** is 0

Effect on the indices

Sensitivity (presences)

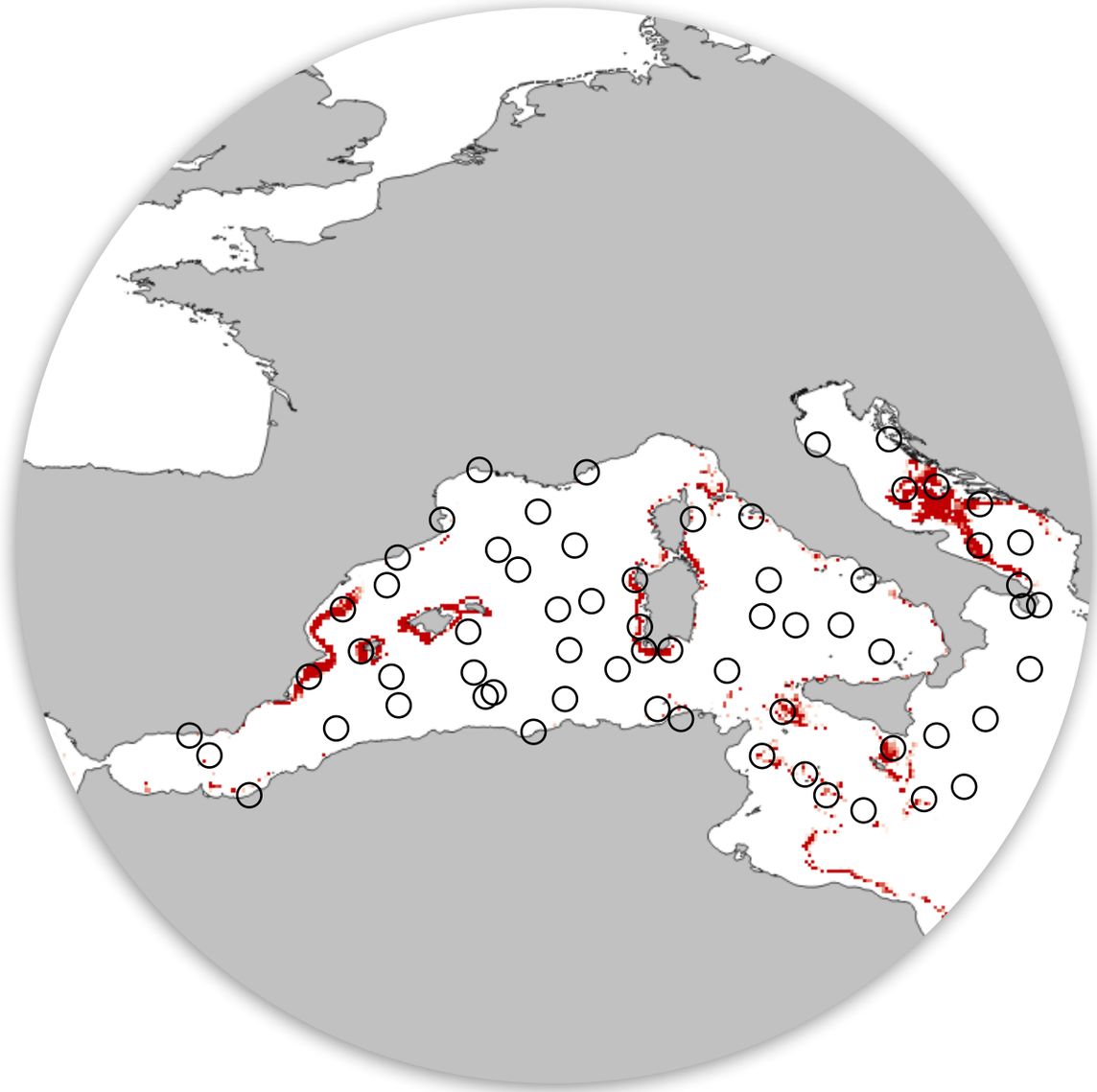
Specificity (absences)

(3) **Minimum predicted area with high sensitivity (e.g., >= 0.95)**

Threshold-dependent measures of accuracy

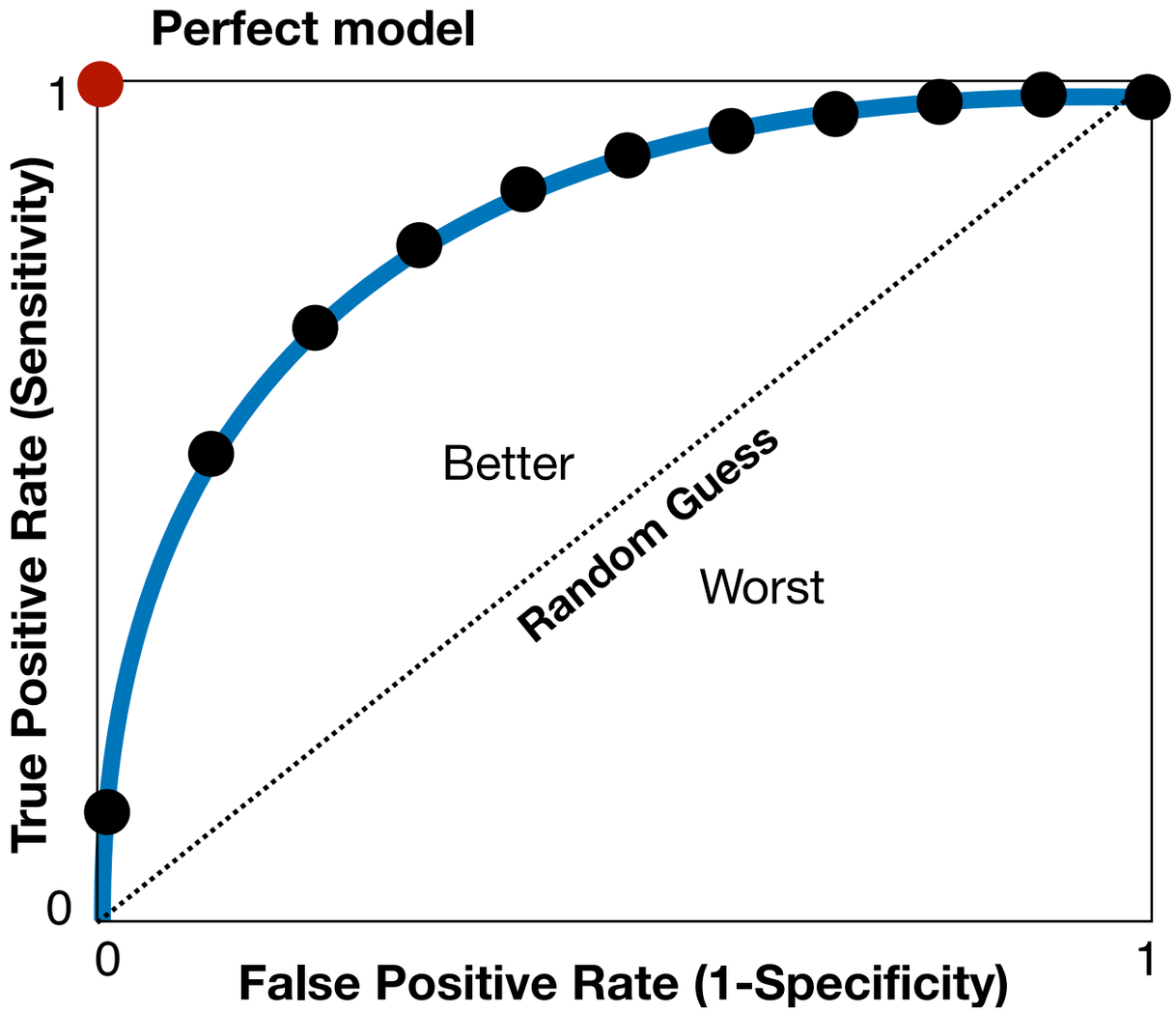
Thresholds like “Max. TSS”, “Min. Area” and “Equal” allow **converting probability maps into binary maps** and also **assess the agreement between observed data (0 / 1) and the corresponding raw model output (e.g., probability of occurrence)**.

** Although not suggested, fixed thresholds (default of 0.5) can be used in logistic regression and related modelling methods.



Random pseudo-absences

In presence-only methods with random pseudo-absences (e.g., GLM) or background information (e.g., MaxEnt), **specificity** (absences correctly predicted) will be **lower than expected because these data also occur in suitable regions** (also TSS).



T	Sens	1-Spe
0	0.1	0
0.1	0.6	0.1
0.2	0.75	0.2
0.3	0.77	0.3
0.4	0.79	0.4
0.5	0.82	0.45
0.6	0.86	0.6
0.7	0.88	0.65
0.8	0.90	0.7
0.9	0.95	0.9
1	1	1

Receiver Operating Characteristic Plot is threshold-independent, based on sensitivity only.

True Positive rate (sensitivity) against **False Positive rate (1-Specificity)** across the range of all possible thresholds. The closer the curve from y-axis, the larger the **Area Under the Curve**, and thus the more accurate the model.

The area under the curve is the evaluation index.

AUC performance

Accuracy indices like AUC can be interpreted (debatable) as follow:

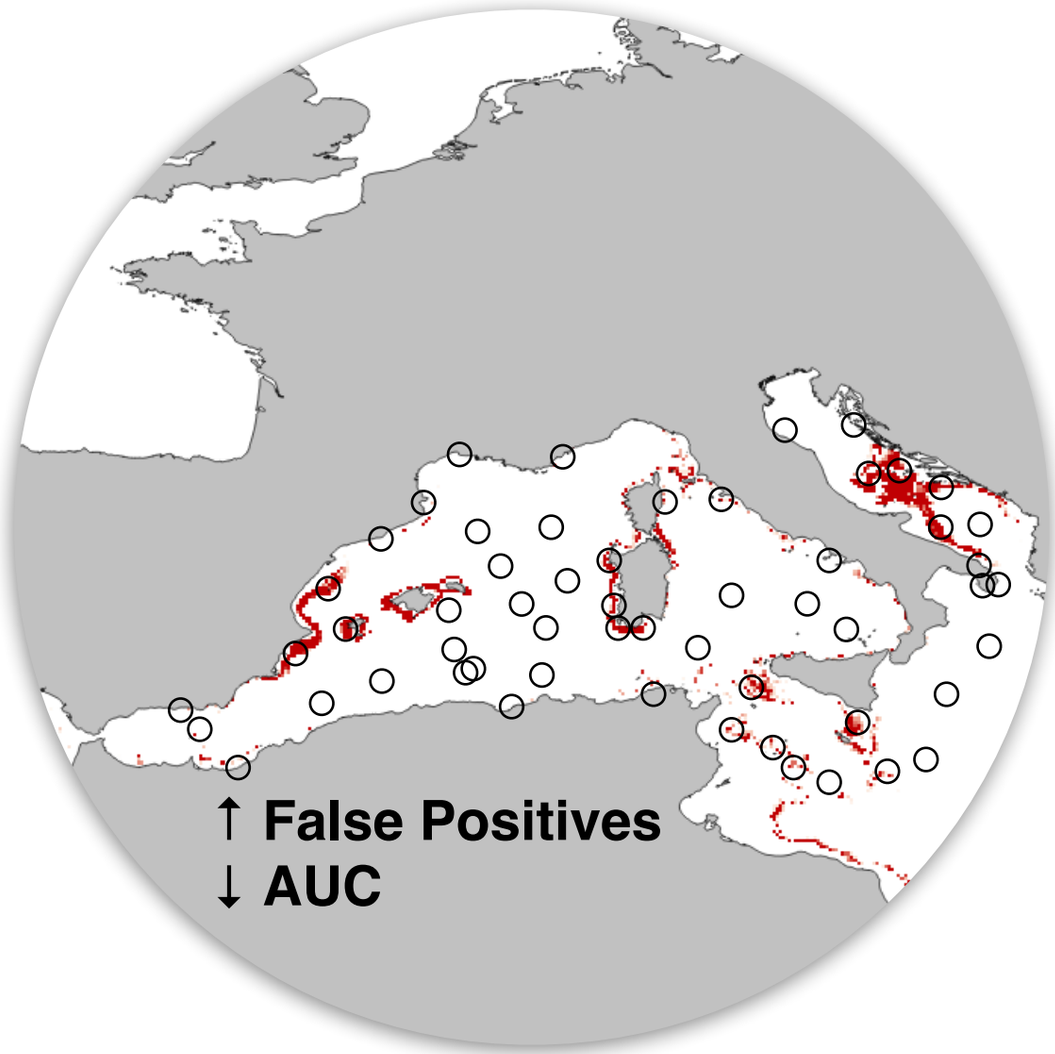
1 - 0.9 : excellent model
0.9 - 0.8 : good model
0.8 - 0.7 : fair model
0.5 - 0.7 : poor model

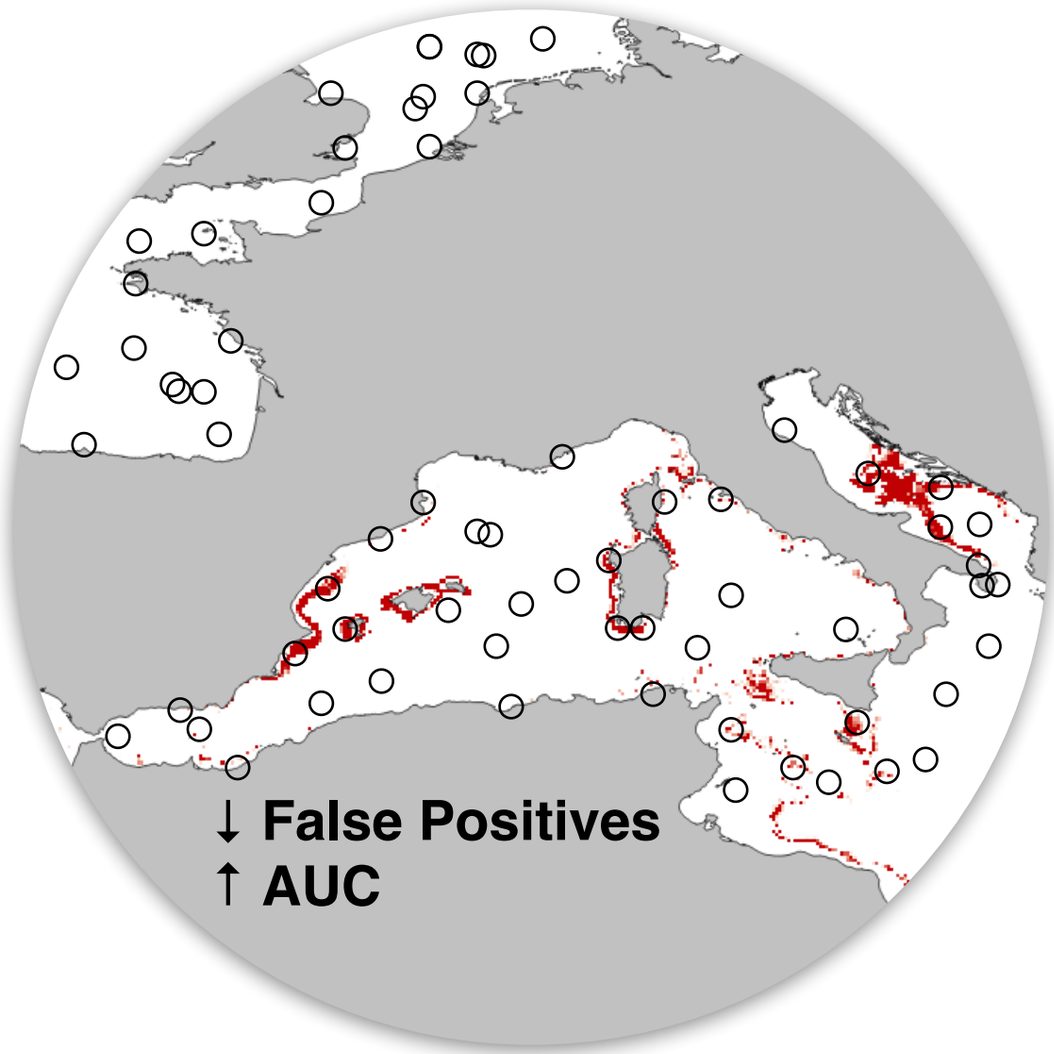
AUC recommend because:

A well-known standard in SDM;
Relies only on presence records (good for models based on random pseudo-absences or background information).

AUC not recommend because:

Ignores goodness of fit (no measure of how well models fitted data);
The geographic extent to which models are carried out influences false positive rates.
(…)





Influence of geographic in AUC

Larger extent have reduced false positives (higher AUC) for the same number of pseudo-absences or background information.

Model evaluation

What to look at and what to report in model evaluation.

AUC
Sensitivity
TSS
Goodness of fit (Adjust. R^2, deviance explained, etc.)

The importance of Sensitivity

When low AUC / TSS but high Sensitivity (True positive rate), discuss the potential role of absence data in model evolution.

When evaluating models we should also consider:

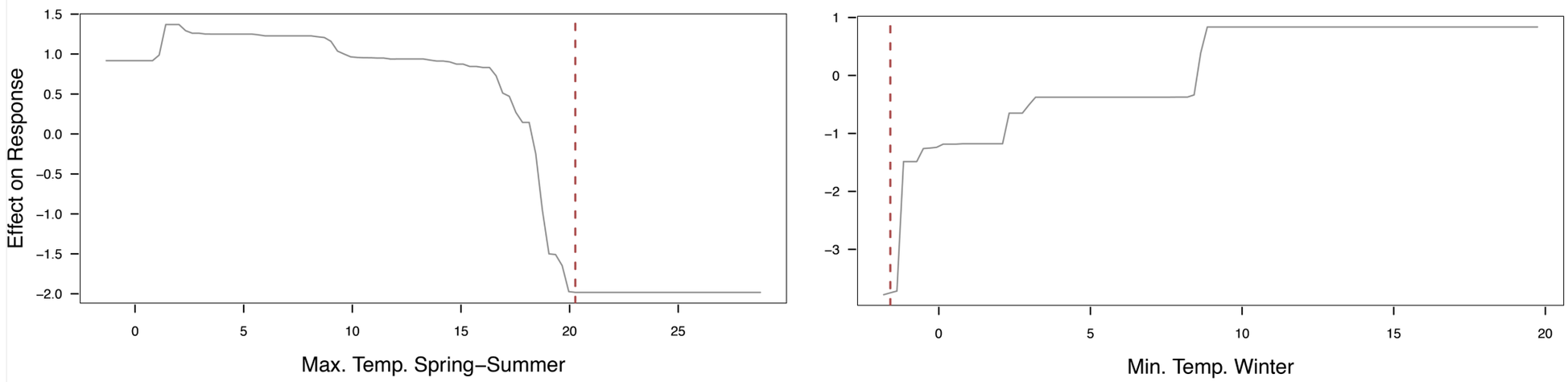
Does the model fit the expectations of ecological theory?

Is accuracy inferred directly linked to the potential of transferability? High accuracy meaning a model potentially transferable?

Partial dependency plots

Show the relationship between the probability of occurrence and each environmental variable.

For each plot, the r**esponse is modelled for one environmental variable while the other environmental variables are held constant at their mean.** The x-axis represents the range of values of the environmental variable, and the y-axis gives the probability of occurrence on a scale from 0 (low probability) to 1 (high probability).



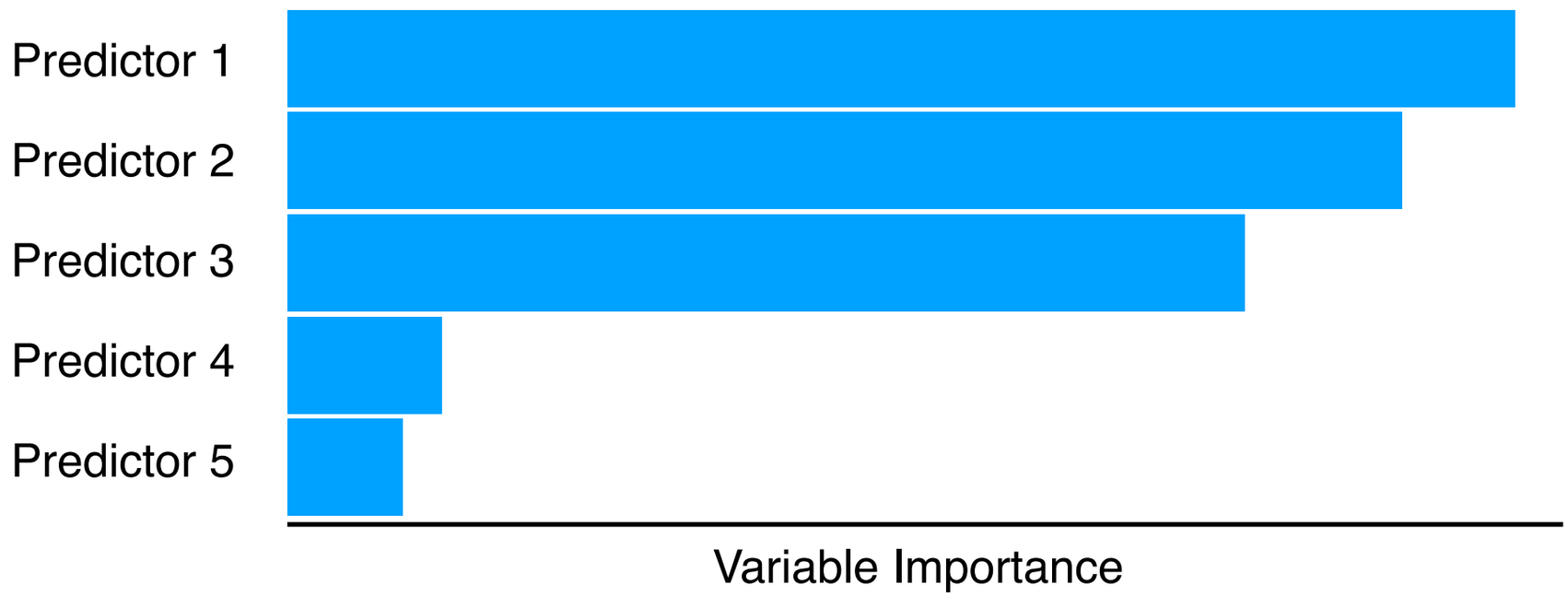
Does the model fit the expectations of ecological theory?

Curve and limiting points match physiological data (reliable model!)

Model evaluation

Relative variable importance

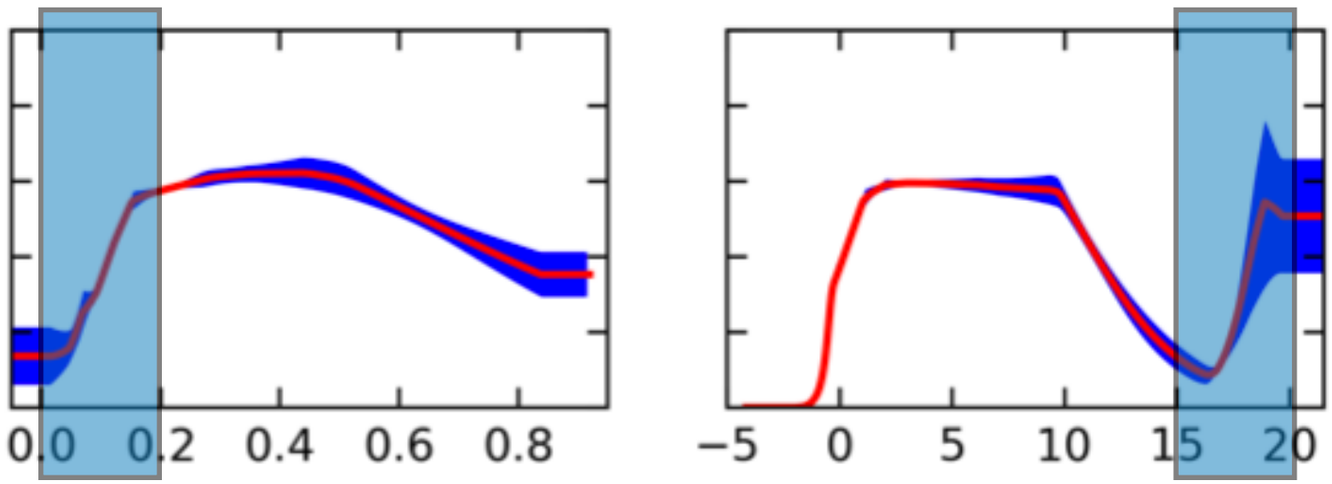
Assist in understanding the contribution of environmental predictor variables to the model outputs. The approach is to fit models with and without each variable, in order to determine the potential increase in model performance. Without an important variable, a model should reduce performance.



Does the model fit the expectations of ecological theory?

e.g.,

Partial dependency plots for an **intertidal algae** distributed in the N Atlantic Ocean modelled with MaxEnt to predict future range shifts.



Light attenuation **Minimum SST**

Good accuracy (AUC > 0.85)

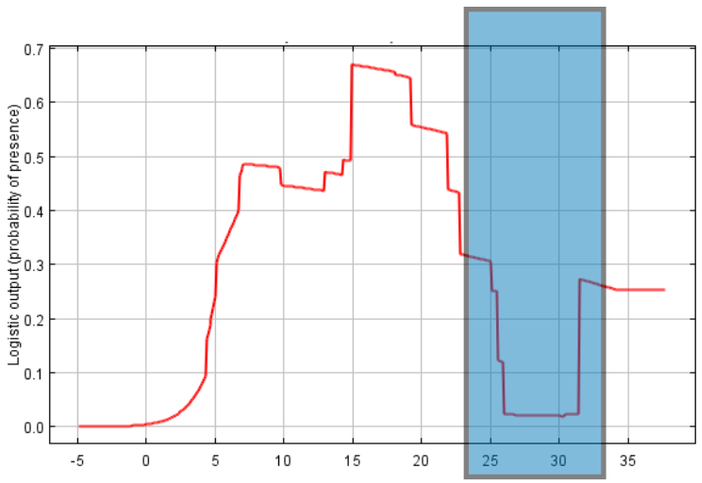
Low light attenuation (high transparency waters) limiting in its lower values for an intertidal species?

Minimum temperatures > 15ºC are unsuitable and > 20ºC suitable?

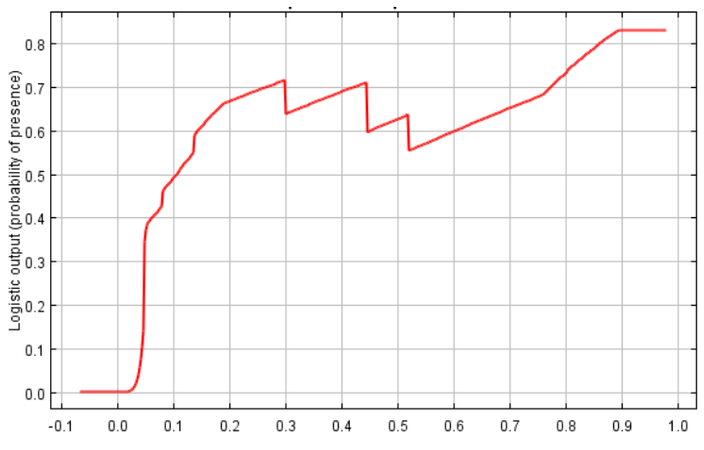
How will it perform when transferred to future warmer conditions?

e.g.,

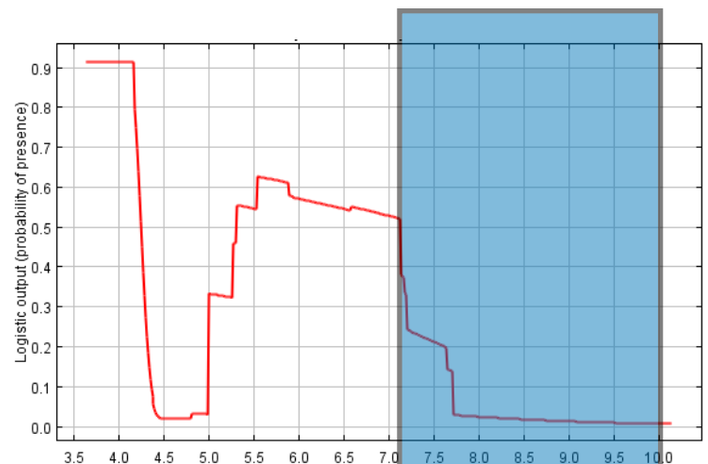
Partial dependency plots for an **subtidal algae** distributed in the N Atlantic Ocean modelled with MaxEnt to predict future range shifts.



Maximum temperatures



Light attenuation



Dissolved Oxygen

High accuracy (AUC > 0.95)

Higher suitability at 35ºC than 25ºC?

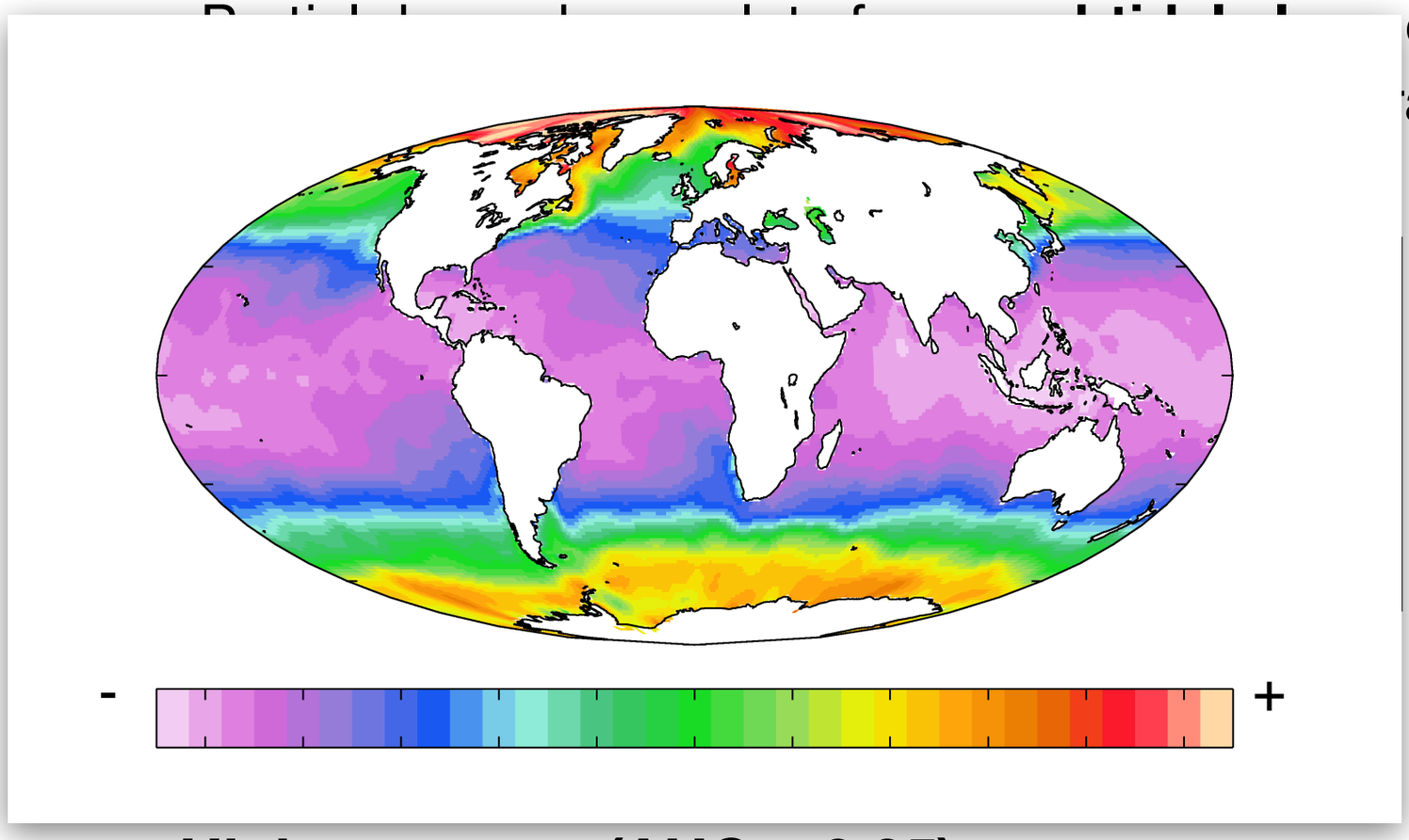
Light attenuation increasing probability of occurrence? A proxy predictor differentiating tropical (clear waters) from higher latitudes (more productive, lower penetration)?

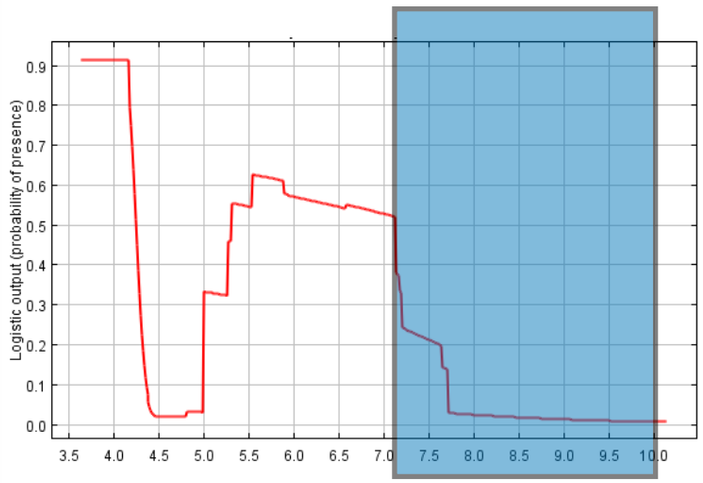
O_2 with unexpected pattern for a kelp. A proxy predictor differentiating tropical regions (low O2) from polar latitudes (higher O2)?

e.g.,

distributed in the N Atlantic

ange shifts.





Dissolved Oxygen

High accuracy (AUC > 0.95)

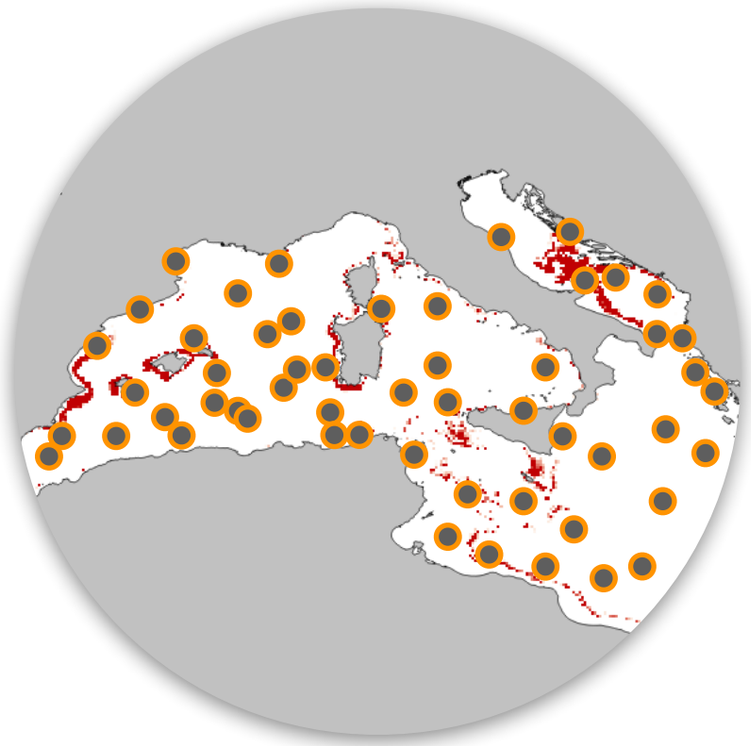
Higher suitability at 35ºC than 25ºC?

Light attenuation increasing probability of occurrence? A proxy predictor differentiating tropical (clear waters) from higher latitudes (more productive, lower penetration)?

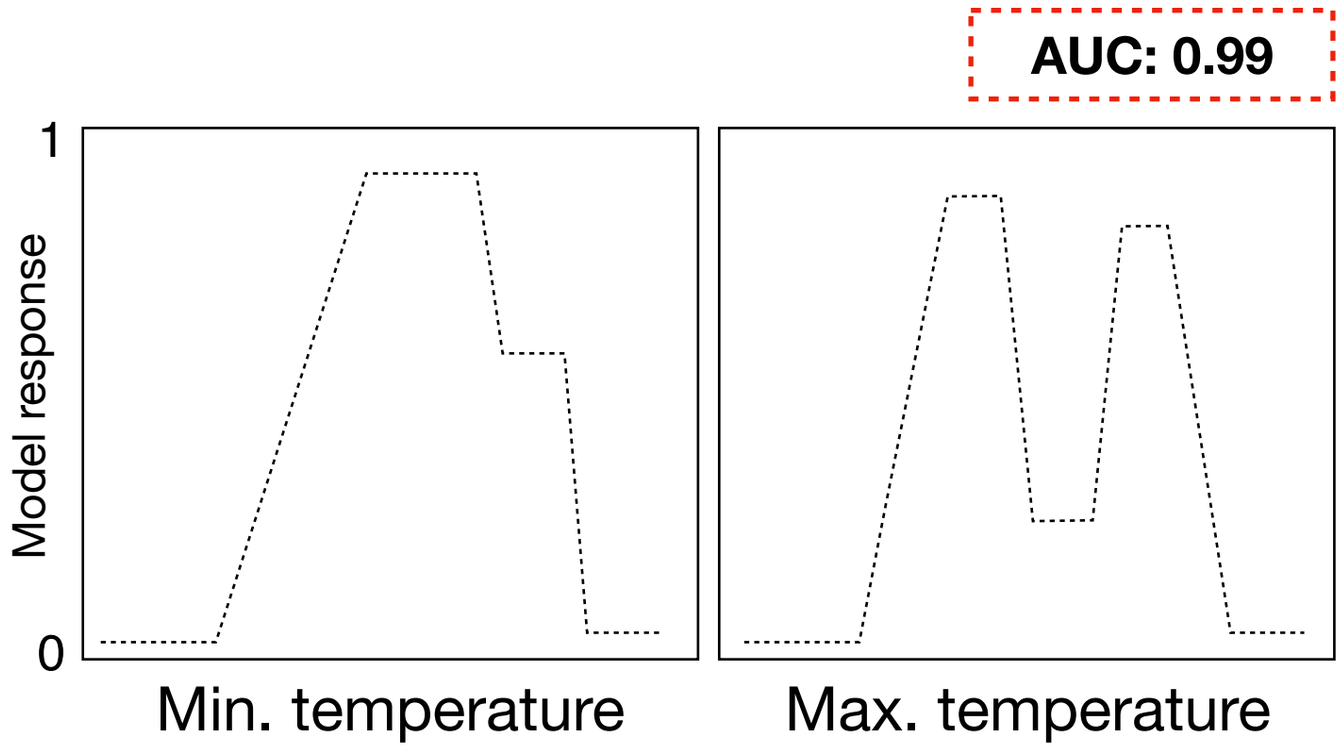
O_2 with unexpected pattern for a kelp. A proxy predictor differentiating tropical regions (low O2) from polar latitudes (higher O2)?

High accuracy scores not linked to good transferability?

Depends on how it is measured.



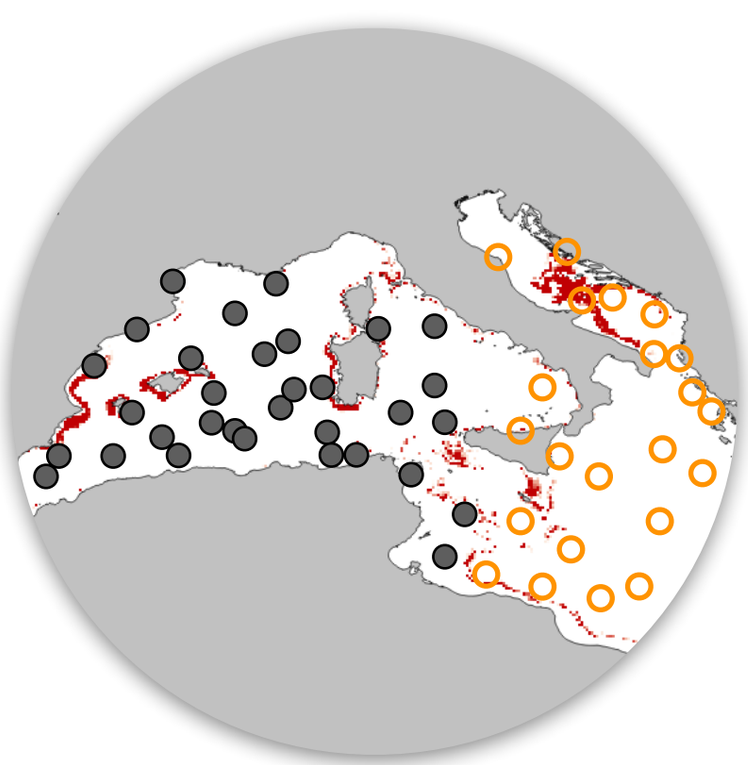
- Testing data
- Training data



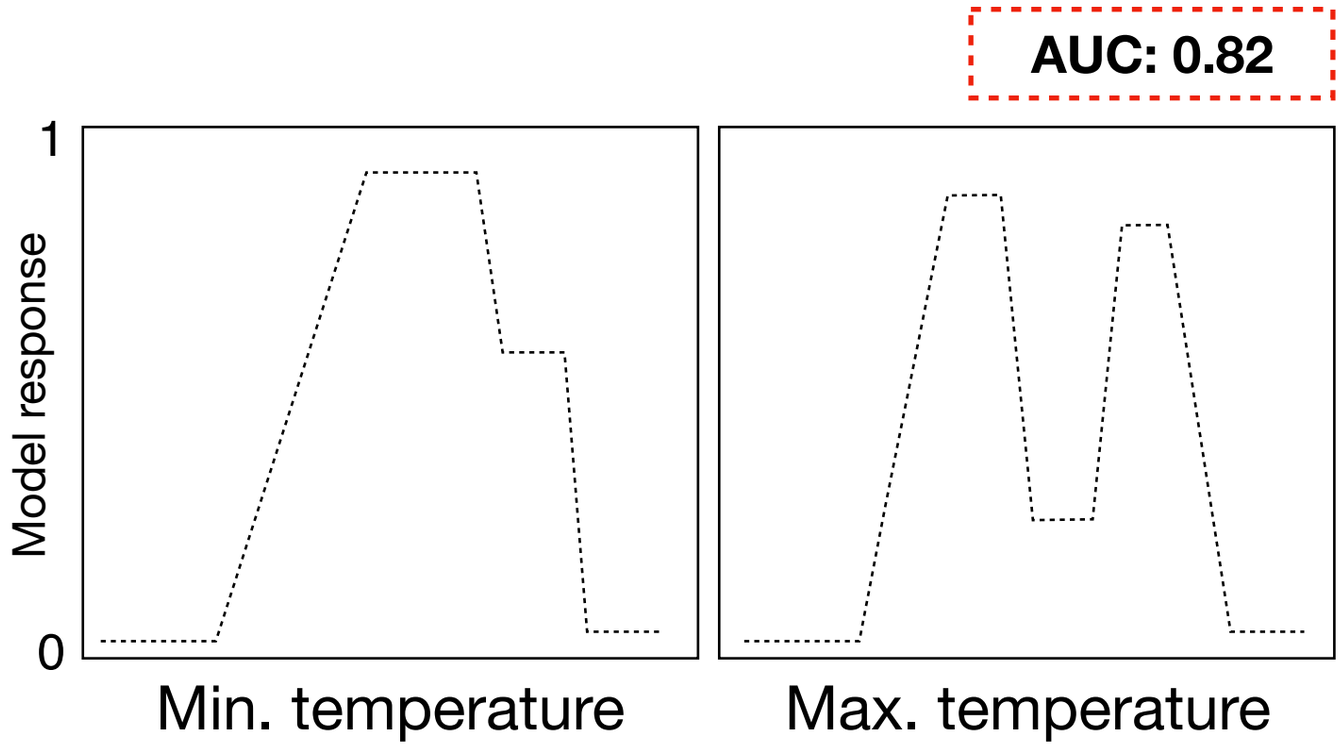
Testing accuracy with training data (resubstitution) leads to an overestimation of accuracy regardless of the model's potential for transferability to new observations or its ecological meaning.

High accuracy scores not linked to good transferability?

Depends on how it is measured.



- Testing data
- Training data



Testing accuracy with independent data is the approach to evaluate the model and transferability.

The same as projecting to other places or times.

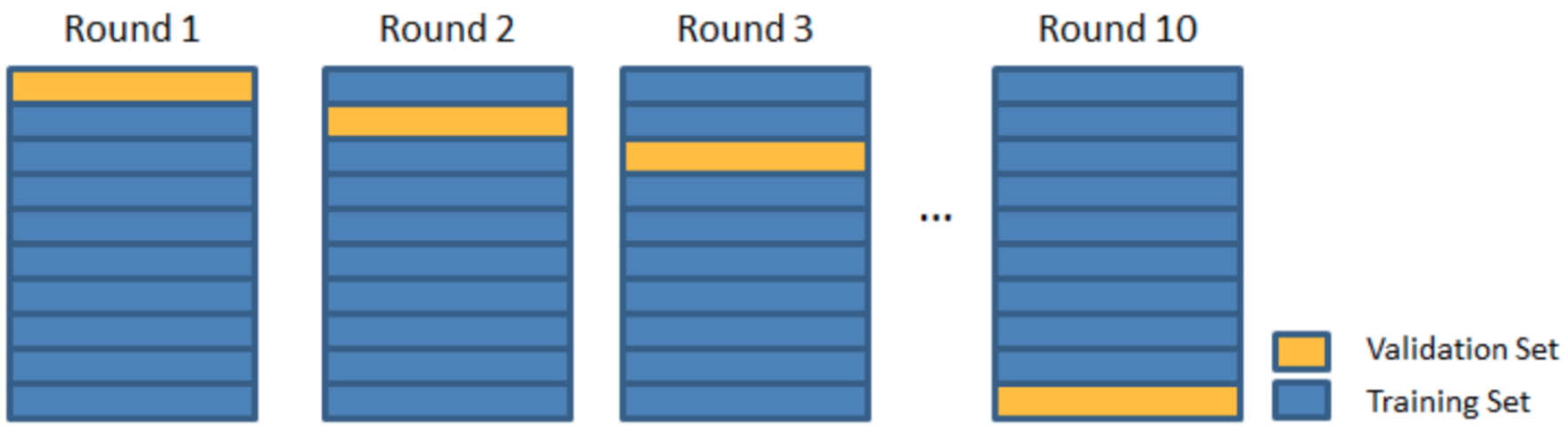
Generally leads to lower accuracy but more reliable accuracy indices.

Missing independent data?

Often it is not feasible to collect new independent data. Partitioning the data in k-fold cross-validation interactions, with data splits k times, yielding k estimates of accuracy that can be averaged.

e.g.,

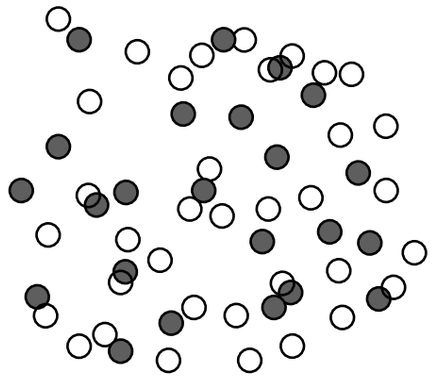
In 10-fold CV, 9/10 of the observations are used to train the model and the remaining 1/10 are used to estimate performance; this is repeated ten times and the estimated performance measures are averaged.



Missing independent data

Methods to produce independent datasets for cross-validation.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

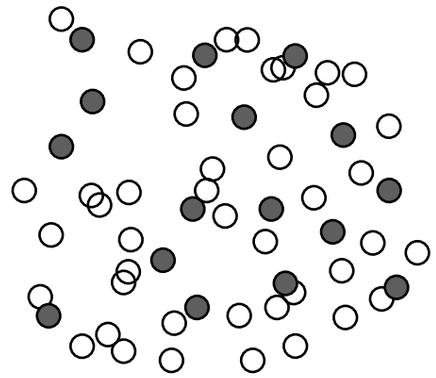
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

Methods to produce independent datasets for cross-validation.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

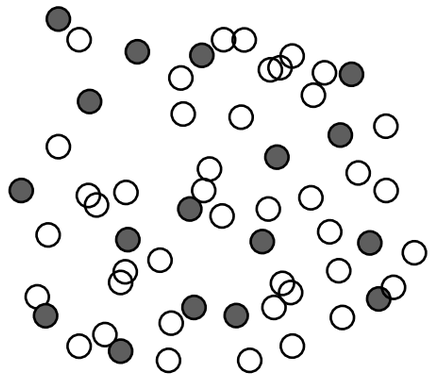
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

Methods to produce independent datasets for cross-validation.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

(latitudinal, longitudinal)

○ Training data ● Testing data

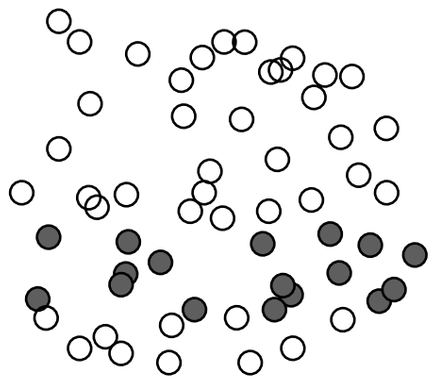
Missing independent data

Methods to produce independent datasets for cross-validation.

Some approaches provide more independent datasets than others.

Random (70/30)

(70/30 | k-fold)



Bands

(latitudinal, longitudinal)

Blocks

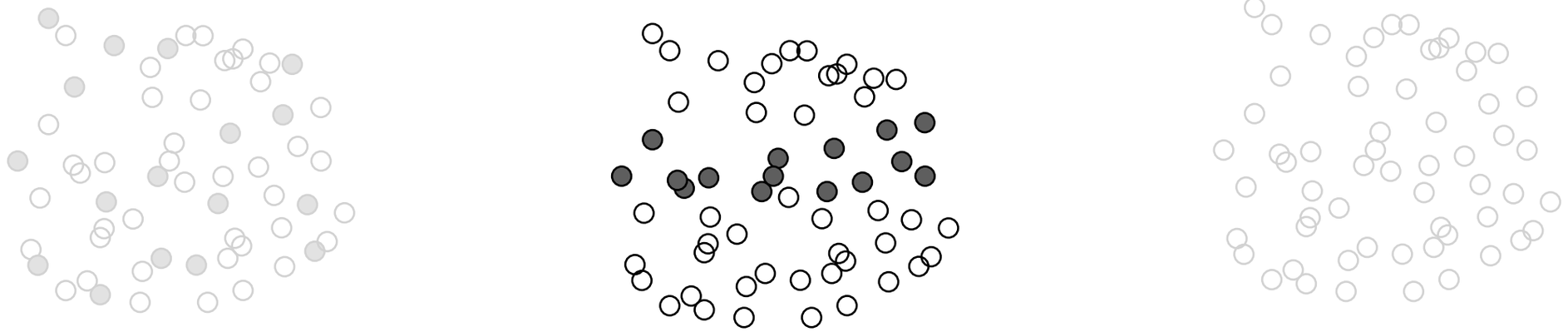
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

Methods to produce independent datasets for cross-validation.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

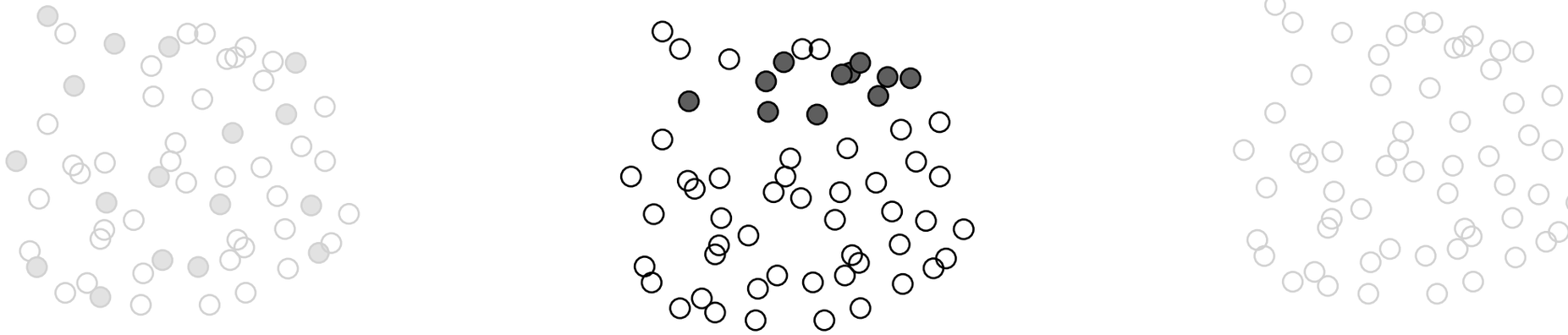
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

Methods to produce independent datasets for cross-validation.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

Methods to produce independent datasets for cross-validation.

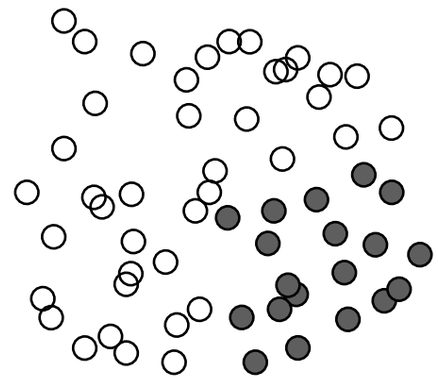
Some approaches provide more independent datasets than others.

Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)



Blocks

(latitudinal, longitudinal)

○ Training data ● Testing data