Marine Ecological Modelling
Global Climate Change

Evaluating predictive performance and setting decision thresholds

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

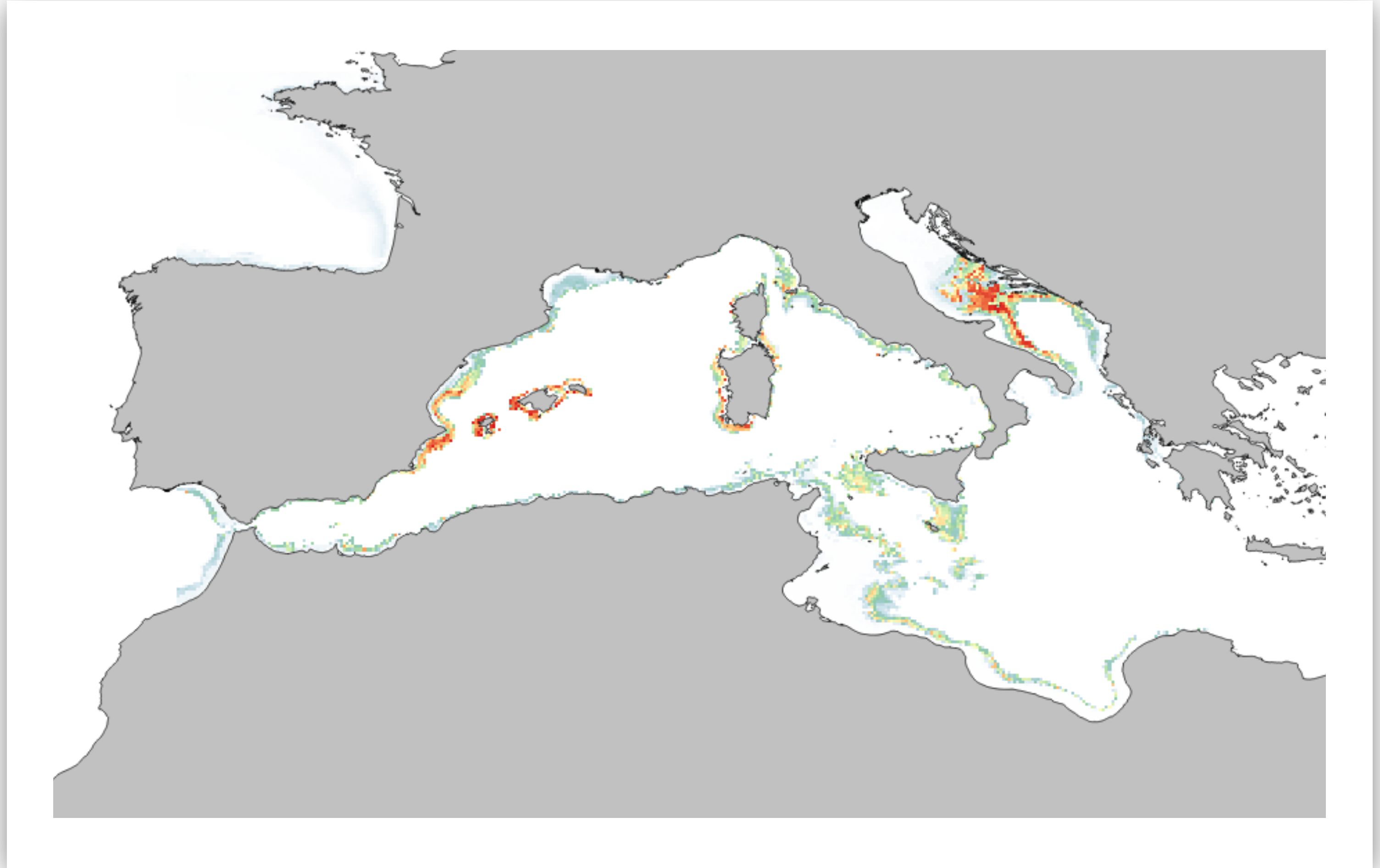
Model evaluation

Also called ‘validation’ or ‘performance’, is crucial to

(1) **verify if predictions are consistent with the observations;**

(2) **assess for the potential for transferability;**

(3) **assess for ecological realism.**



Is the model acceptable for the purpose?

Prediction errors inferred with:

‘false positives’, when the **model predicts occurrence** in places where the **species was not observed**;

‘false negatives’, when the **model predicts absence** in places where the **species was 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 can be used to compute **evaluation criteria** that **measure the performance of the model.**

Sensitivity: proportion of **presences correctly predicted [0-1]**;

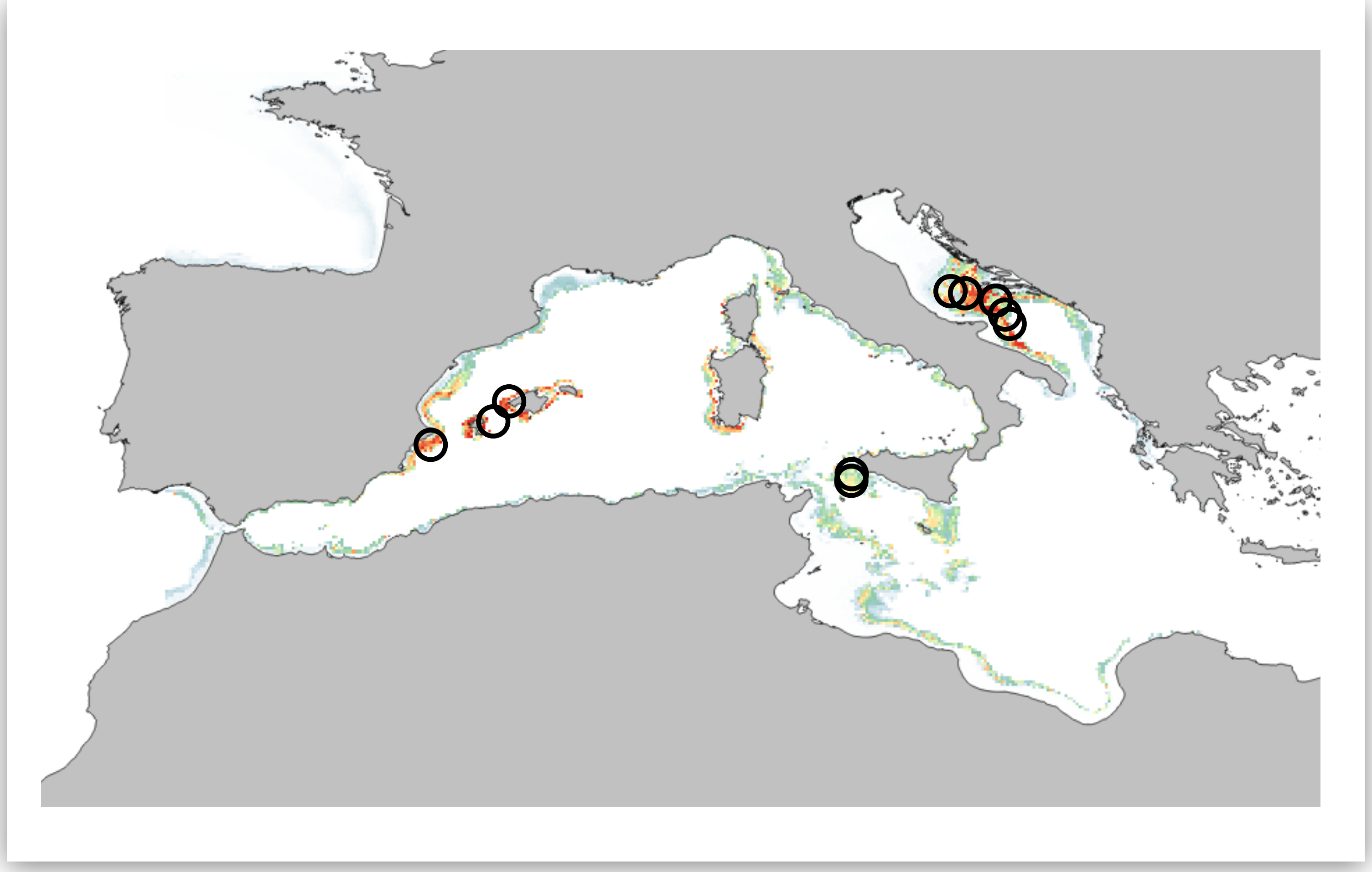
Specificity: proportion of **absences correctly predicted [0-1]**;

True Skill Statistics: **Sensitivity + Specificity - 1** (describes how well the model predicts presences and absences) [0-1];

Area Under the Curve of the Receiver Operating Characteristic.

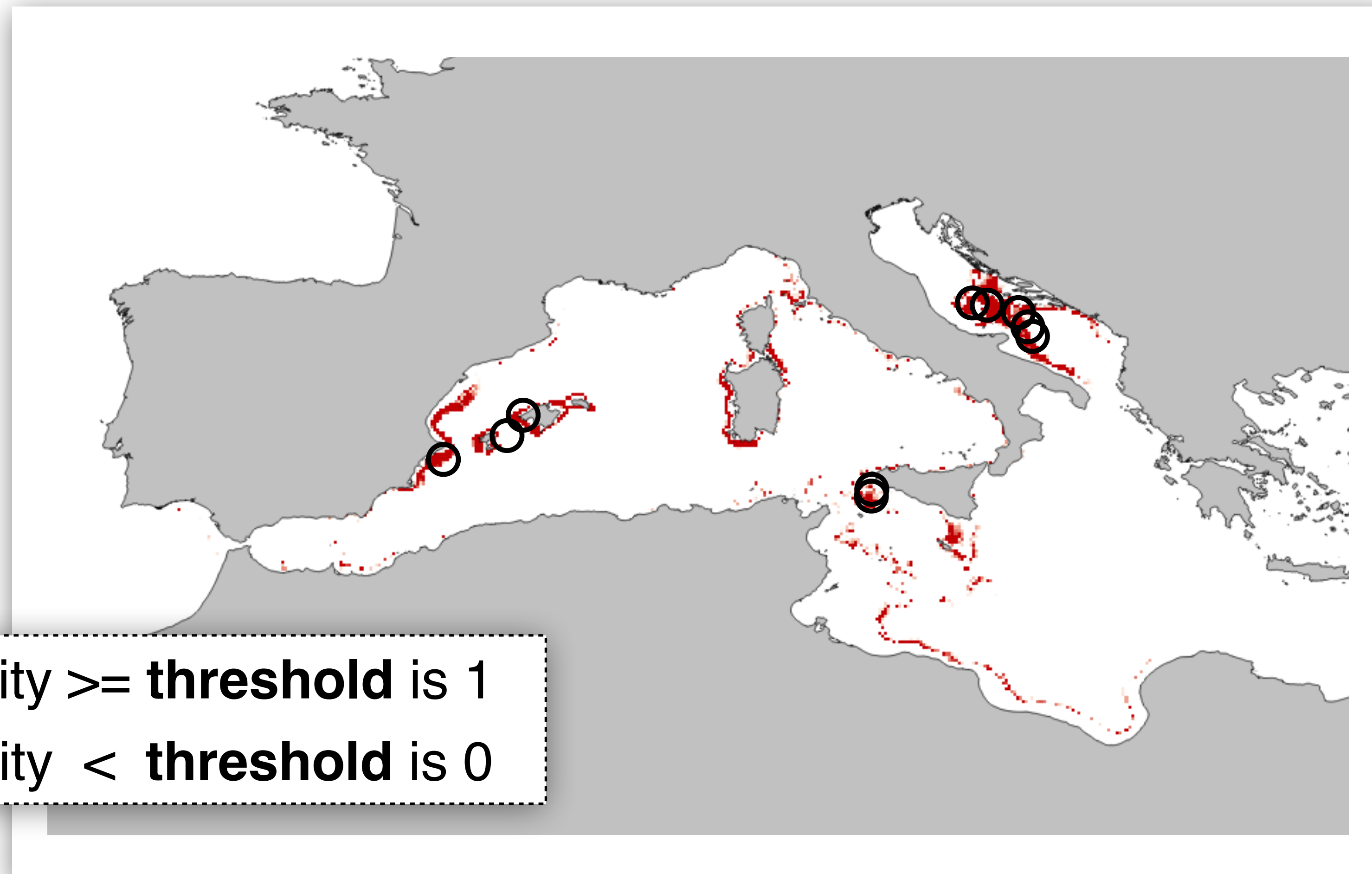
Evaluation criteria

Predictions are continuous surfaces (e.g., probability or suitability; from 0 to 1); To assess the **proportion of presences / absences correctly predicted** one **cannot compare one observation** (e.g., 1 for presence) with its corresponding **model output (**e.g., P: 0.7).



Evaluation criteria

Predictions are continuous surfaces (e.g., probability or suitability; from 0 to 1); To assess the **proportion of presences / absences correctly predicted** one **cannot compare one observation** (e.g., 1 for presence) with its corresponding **model output (**e.g., P: 0.7).



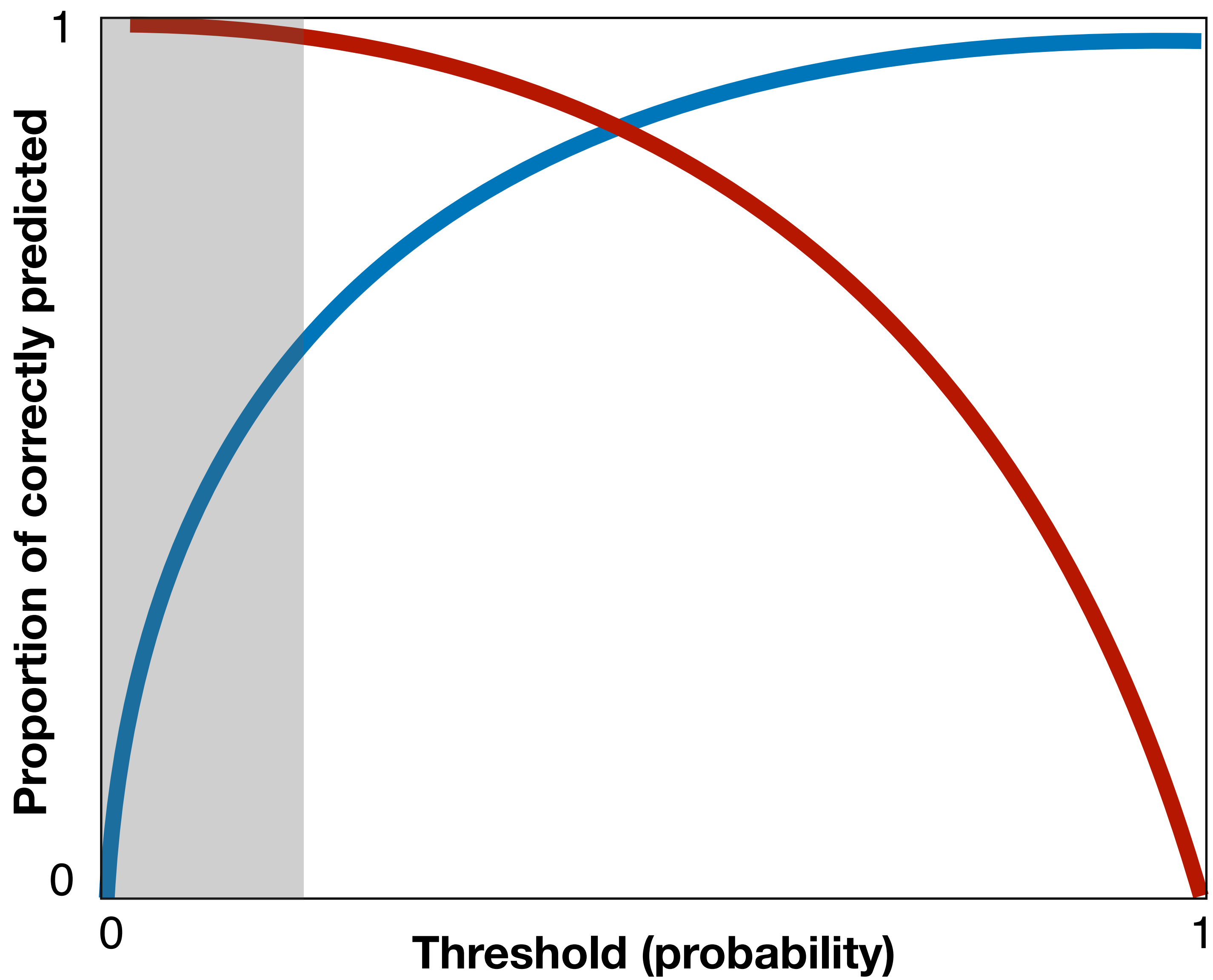
Probability >= **threshold** is 1

Probability < **threshold** is 0

Predictions need to be reclassified as binomial responses (model output **from 0 to 1**) **for comparison with the observed data**.

Different thresholds leading to different accuracy scores

Sensitivity and specificity vary with the thresholds used (0 to 1) to reclassify the models as binomial outputs.



Probability >= **threshold** is 1
Probability < **threshold** is 0

Effect on the indices

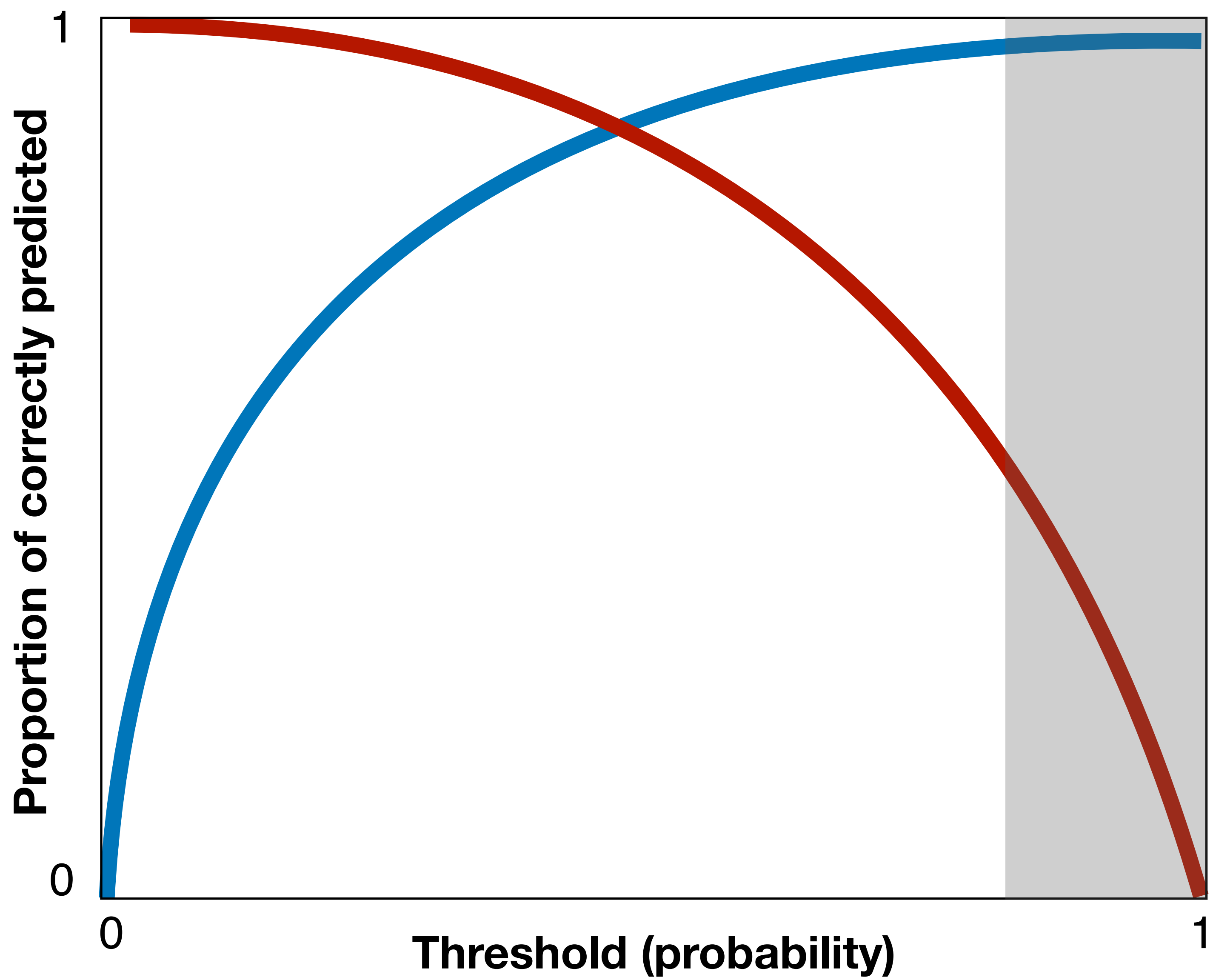
Sensitivity (presences)

Specificity (absences)

With a **low threshold**, most cells (in the map) will return 1, so **high sensitivity (true positive rate, presences accurately predicted)**.

Different thresholds leading to different accuracy scores

Sensitivity and specificity vary with the thresholds used (0 to 1) to reclassify the models as binomial outputs.



Probability >= **threshold** is 1
Probability < **threshold** is 0

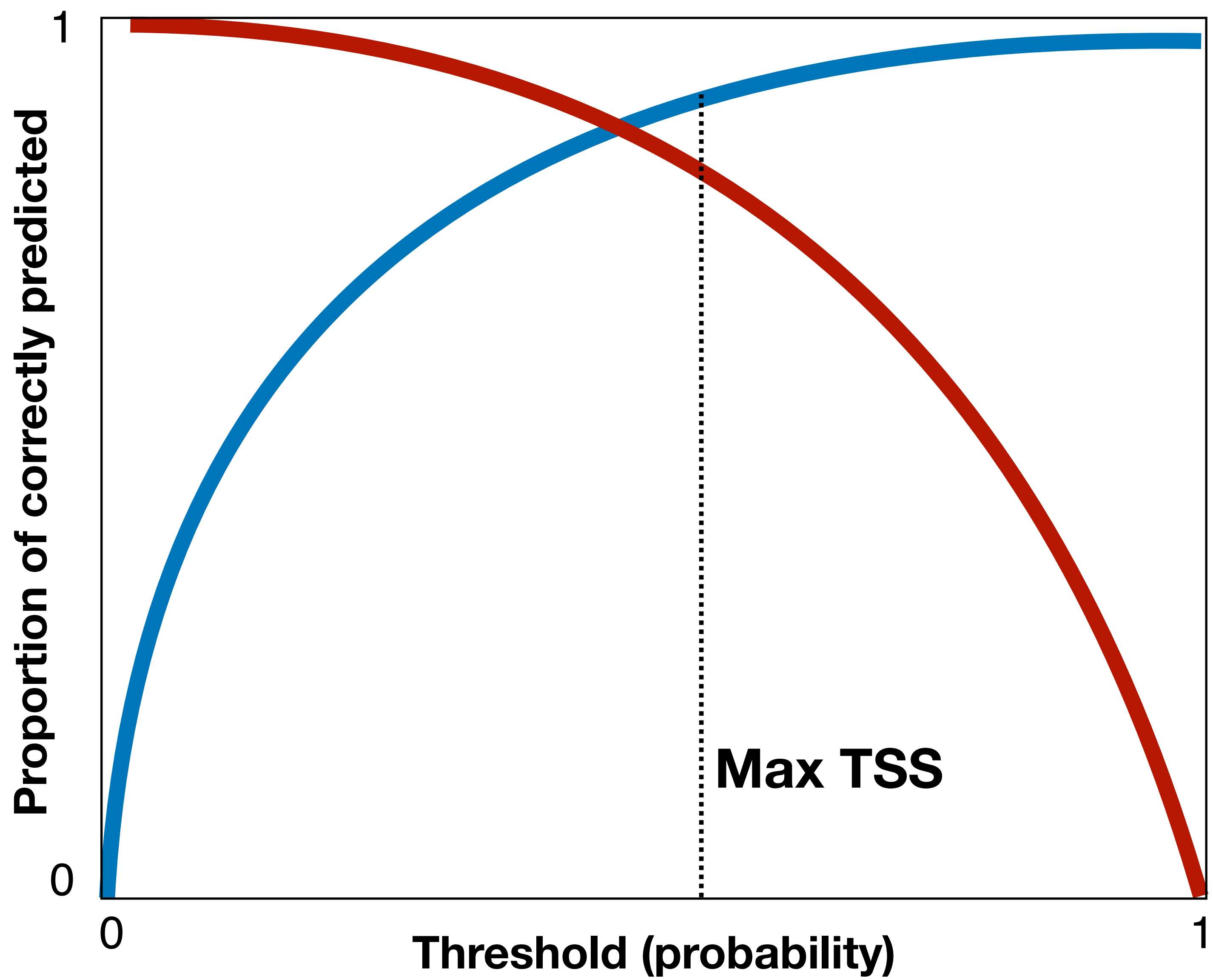
Effect on the indices

Sensitivity (presences)
Specificity (absences)

With a **high threshold**, most cells (in the map) will return 0, **so high specificity (true negative rate, true absences well predicted)**.

There are threshold rules to maximize the agreement between observed data and the predicted reclassified binomial surfaces.

Threshold allowing the maximization of sensitivity + specificity



Probability >= **threshold** is 1

Probability < **threshold** is 0

Effect on the indices

Sensitivity (presences)

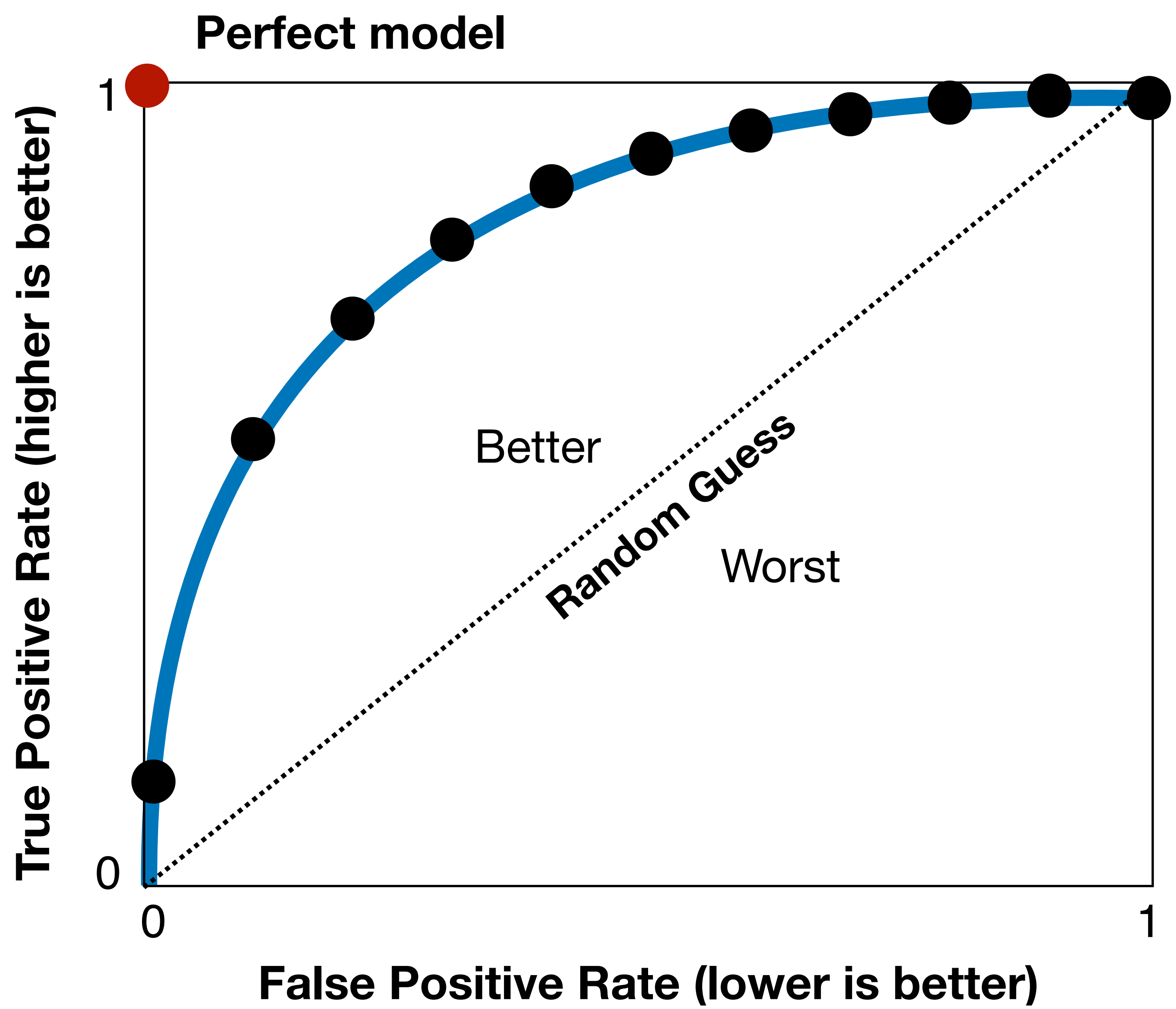
Specificity (absences)

Threshold-dependent estimates of accuracy

Methods like “Max. TSS” allow to:

(1) **assess the agreement between observed data and the model output (e.g., probability of occurrence);**

(2) **convert probability maps into binary maps** - using the threshold maximising the agreement between observed and model output.



Threshold-independent estimates of accuracy

Area **U**nder the **C**urve compares the **true positive rate** against **false positive rate across the range of all possible thresholds** (0 to 1).

The closer the curve from y-axis, the larger the **AUC**, and the more accurate the model.

Performance of indices

Indices like AUC, sensitivity, specificity and TSS can be interpreted:

1.0 - 0.9 : Excellent model
0.9 - 0.8 : Good model
0.8 - 0.7 : Fair model
0.5 - 0.7 : Poor model

(>= 0.8 is a good, usable model)

Model evaluation

What to report in model evaluation.

AUC
Sensitivity
Specificity

When evaluating models we should also consider:

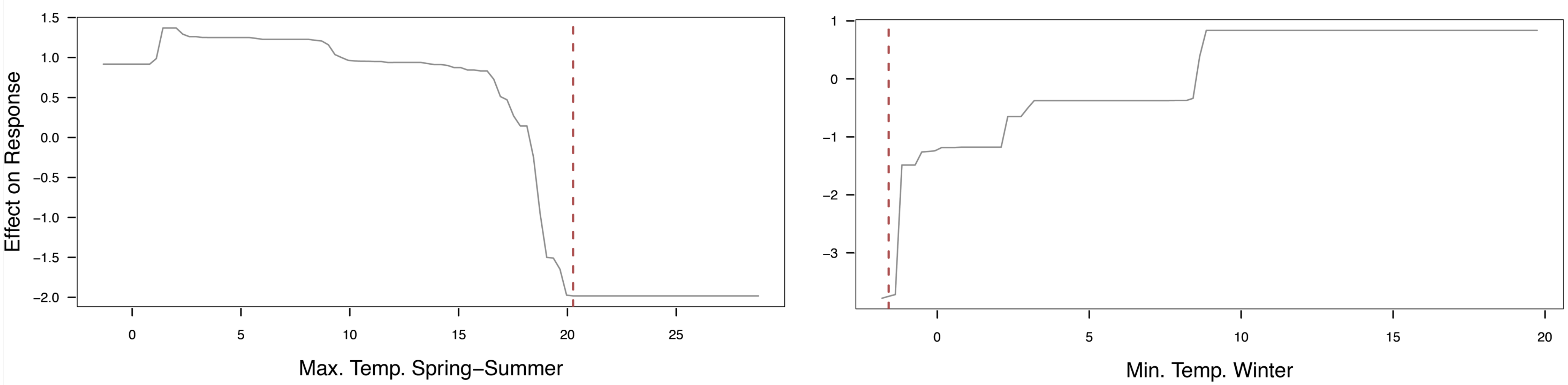
Does the model fit the expectations of ecological theory?

Is accuracy inferred directly linked to the potential of transferability?

Model evaluation

Partial dependency plots

Show the effect of each environmental predictor on the output of the model (probability of occurrence).



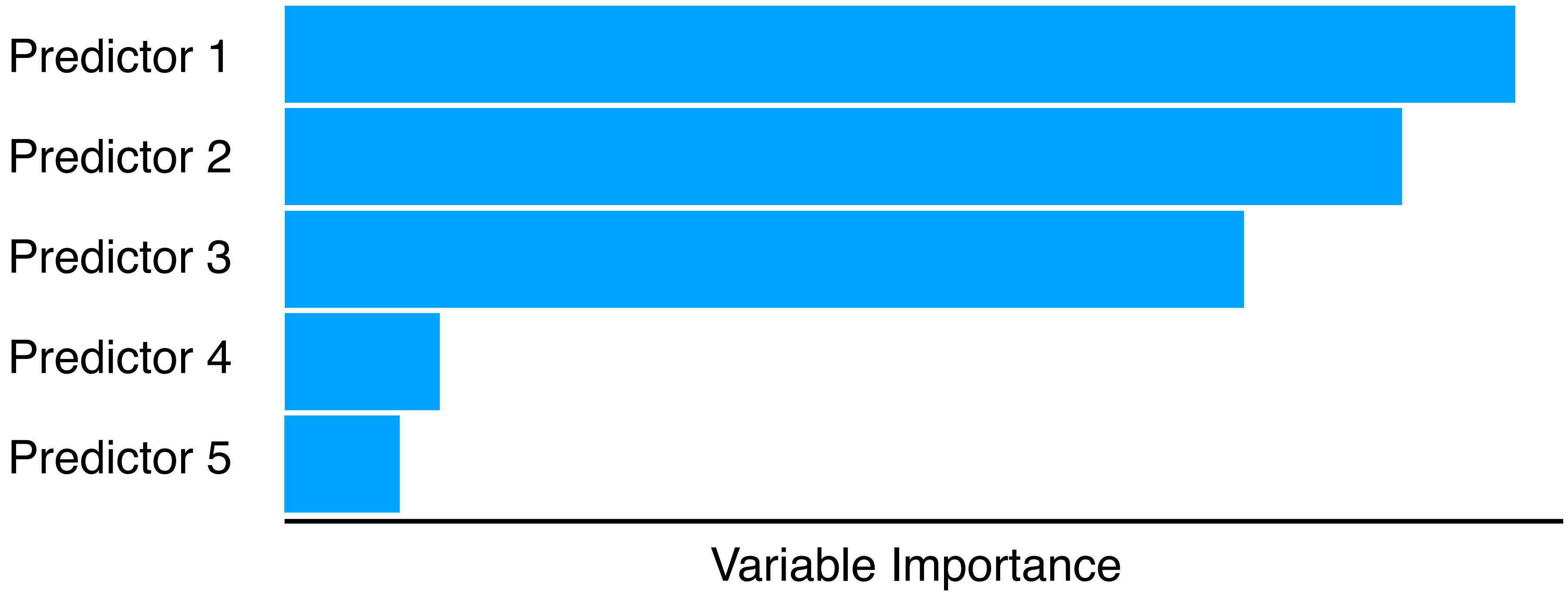
Does the model fit the expectations of ecological theory?

Curve and limiting points match physiological data, reliable model!

Model evaluation

Relative importance of predictors to the model

The approach is to fit models with and without each variable, in order to determine the potential increase in performance. **Without an important variable, a model should reduce its performance (e.g., lower AUC)**.

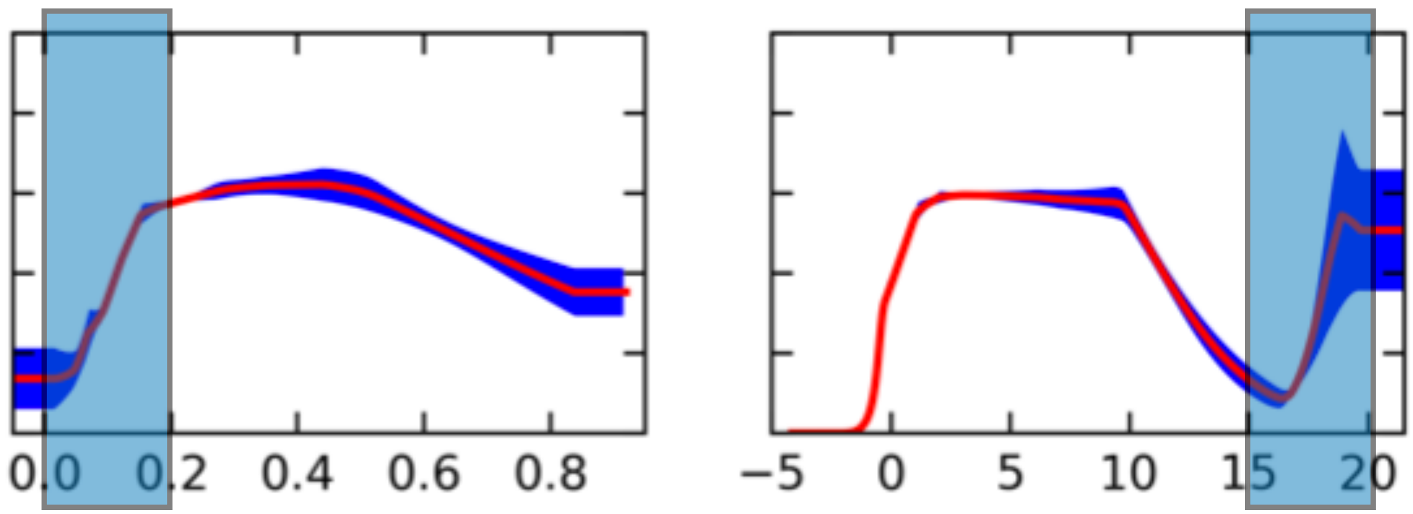


Does the model fit the expectations of ecological theory?

Contributions match expectations for the species, reliable model!

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**

High accuracy (AUC > 0.85)

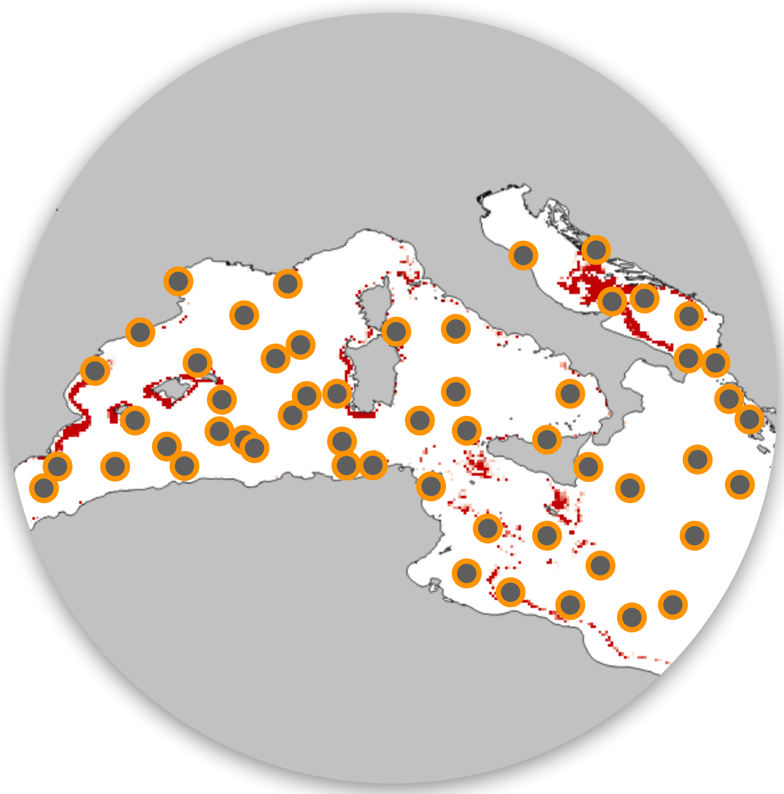
Low light attenuation (high transparency waters) limiting the distribution of an intertidal species? Why model an intertidal species with light conditions?

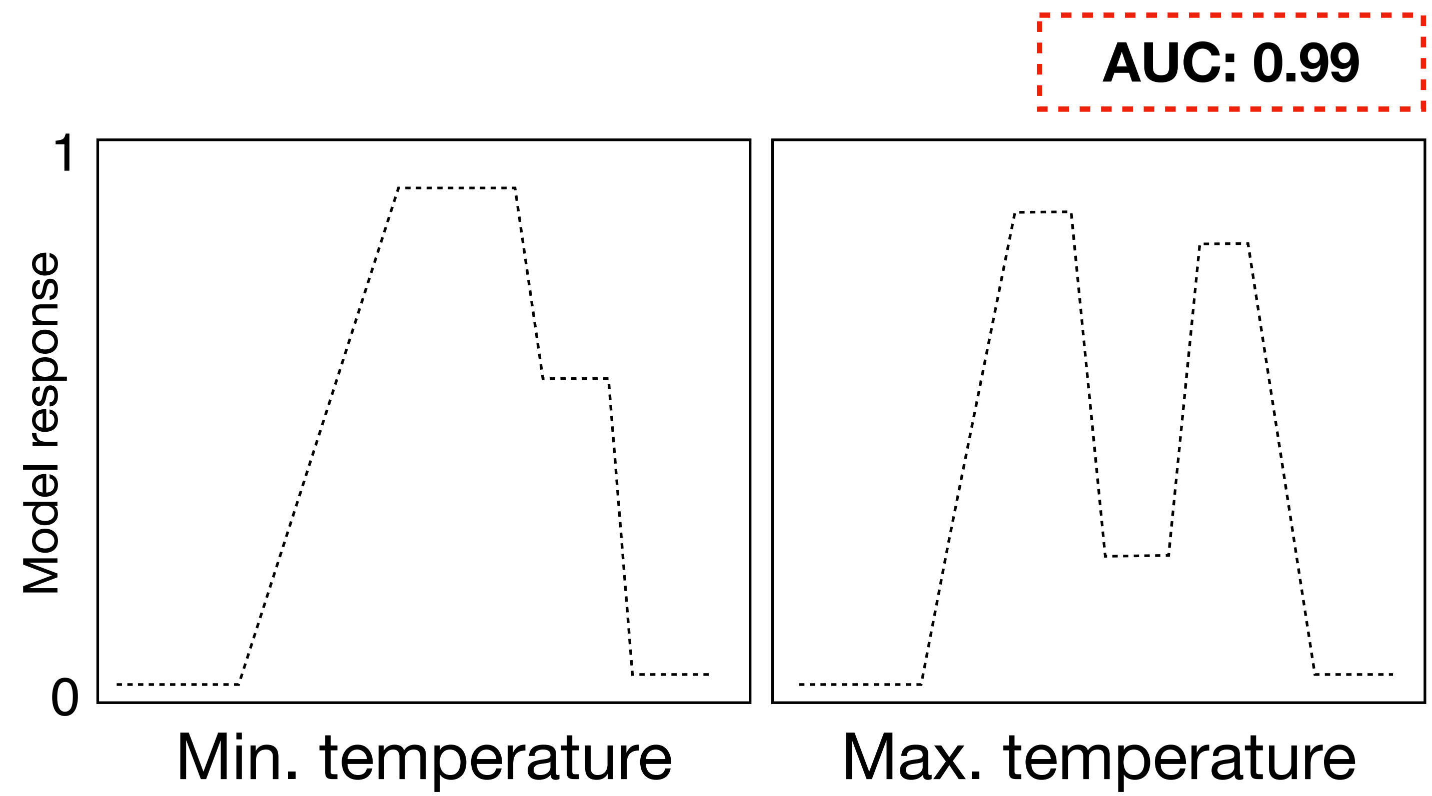
Minimum temperatures > 15ºC are unsuitable and > 20ºC suitable?

Model does not match ecological theory, despite high AUC.

High accuracy scores not linked to good transferability?

Depends on how it is measured.



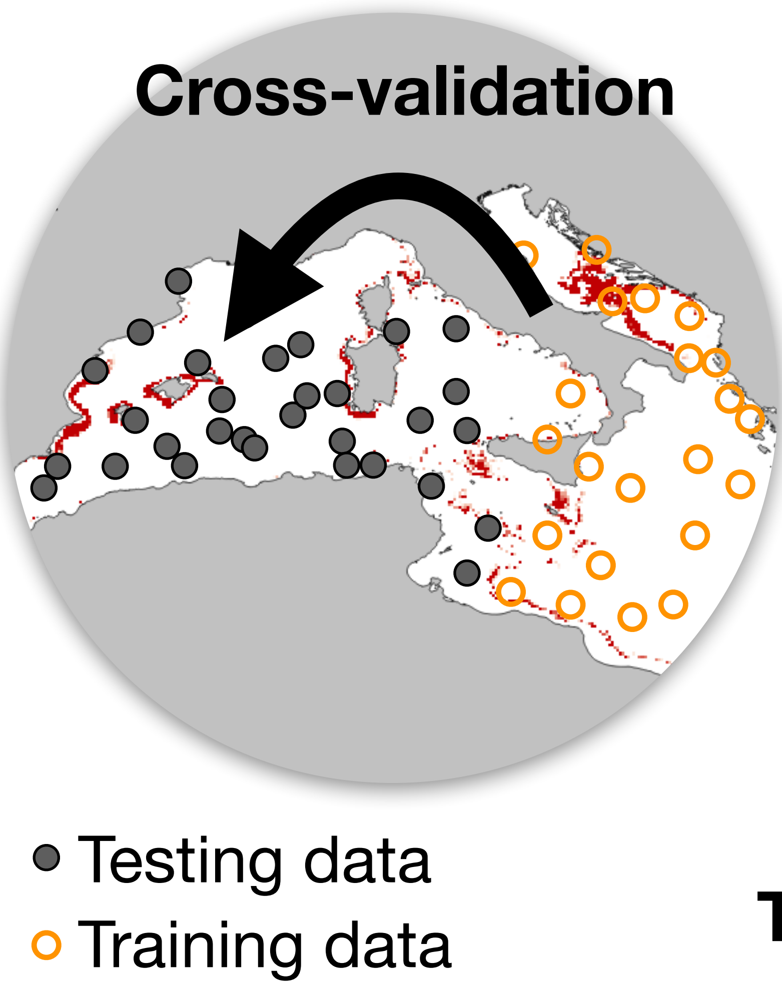


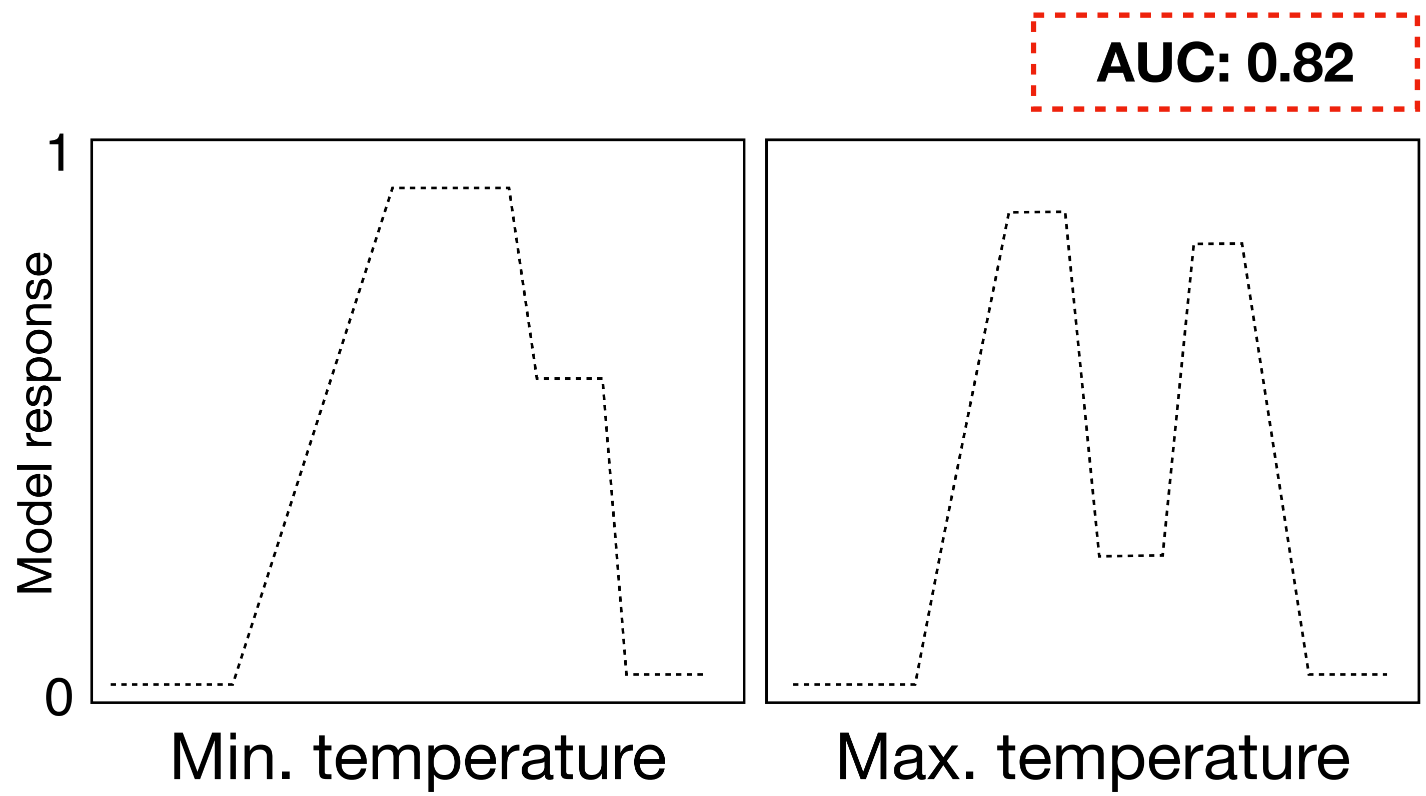
● Testing data
○ Training data

Testing accuracy with the training data leads to an overestimation of accuracy, regardless of the model's potential for transferability or if it ecologically sound.

High accuracy scores not linked to good transferability?

Depends on how it is measured.





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

This is the same as projecting to other places or times.

Leads to a lower accuracy score but a more reliable score.

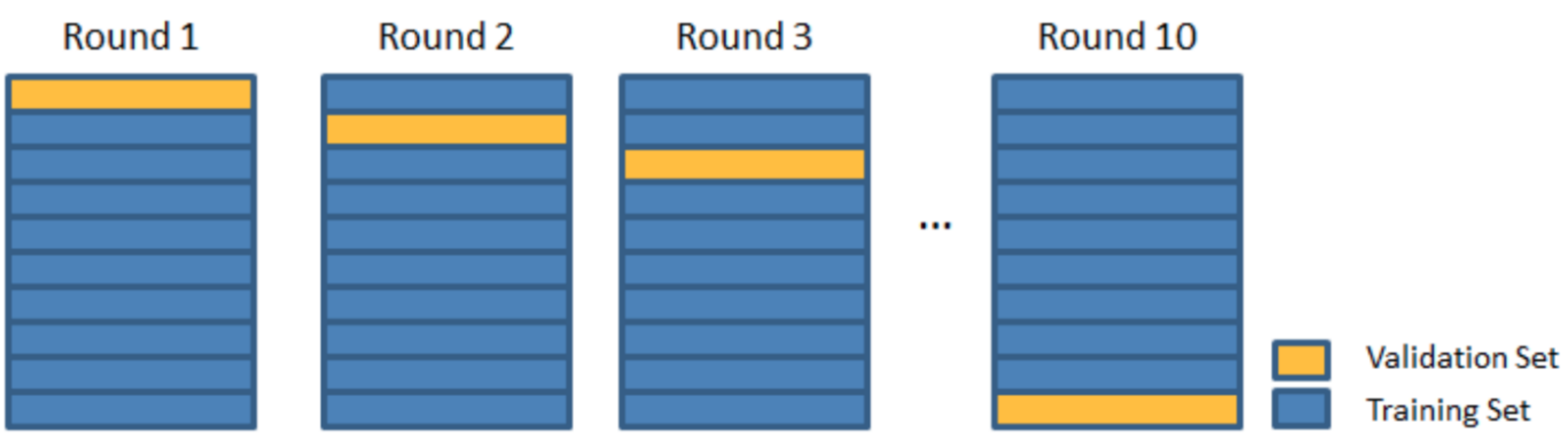
Missing independent data?

Often it is not feasible to collect independent data.

Partitioning the data in k-fold (k sets of data) to cross-validate in k interactions, with data partitioned k times.

e.g.,

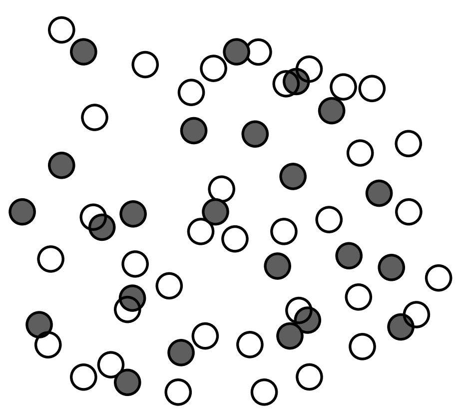
In 10-fold CV, 9 out of 10 of the observations are used to train the model and the remaining 1 out of 10 are held to estimate performance (e.g., AUC); this is repeated ten times and the estimate of performance is the average of the 10.



Missing independent data

There are different methods to produce independent datasets.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

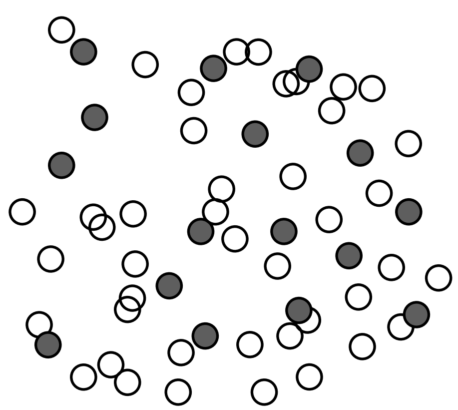
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

There are different methods to produce independent datasets.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

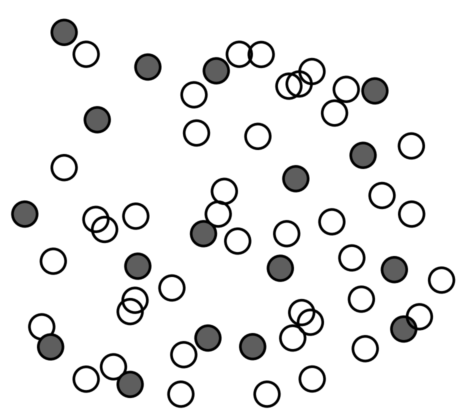
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

There are different methods to produce independent datasets.

Some approaches provide more independent datasets than others.



Random (70/30)

(70/30 | k-fold)

Bands

(latitudinal, longitudinal)

Blocks

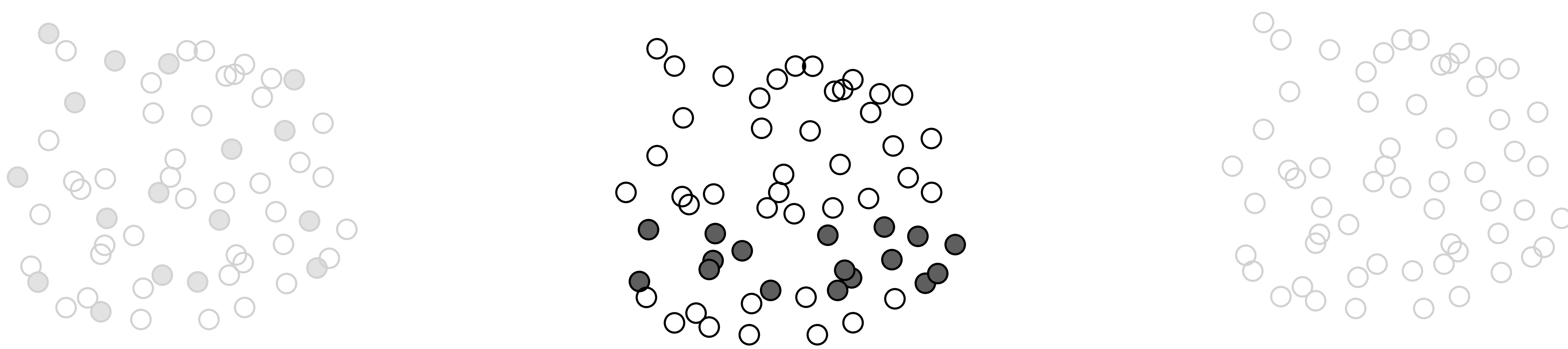
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

There are different methods to produce independent datasets.

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

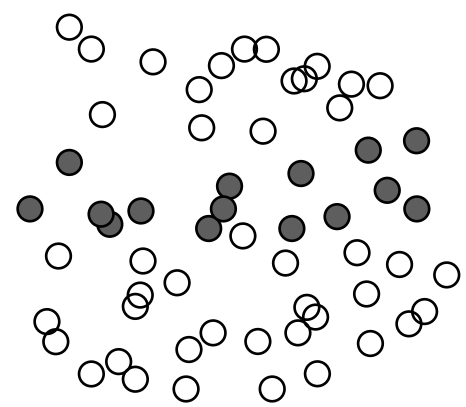
Missing independent data

There are different methods to produce independent datasets.

Some approaches provide more independent datasets than others.

Random (70/30)

(70/30 | k-fold)



Bands

(latitudinal, longitudinal)

Blocks

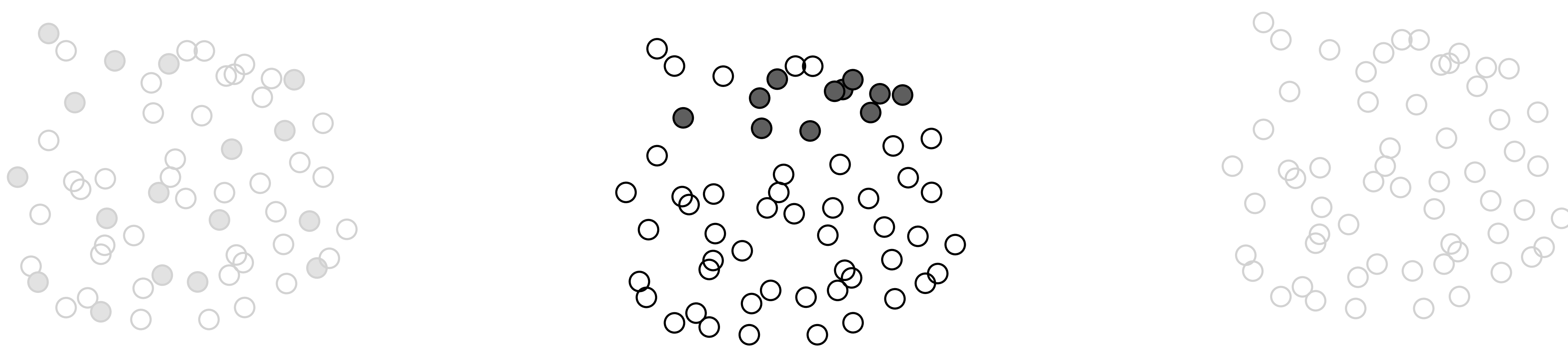
(latitudinal, longitudinal)

○ Training data ● Testing data

Missing independent data

There are different methods to produce independent datasets.

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

There are different methods to produce independent datasets.

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