Marine Ecological Modelling Global Climate Change

Principles of Ecological Niche Modelling

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

Ecological Niche Modelling**

Process of using **computer algorithms to estimate** the **relationship between biodiversity observations and the environment** (i.e., observations within the prevailing environmental conditions).

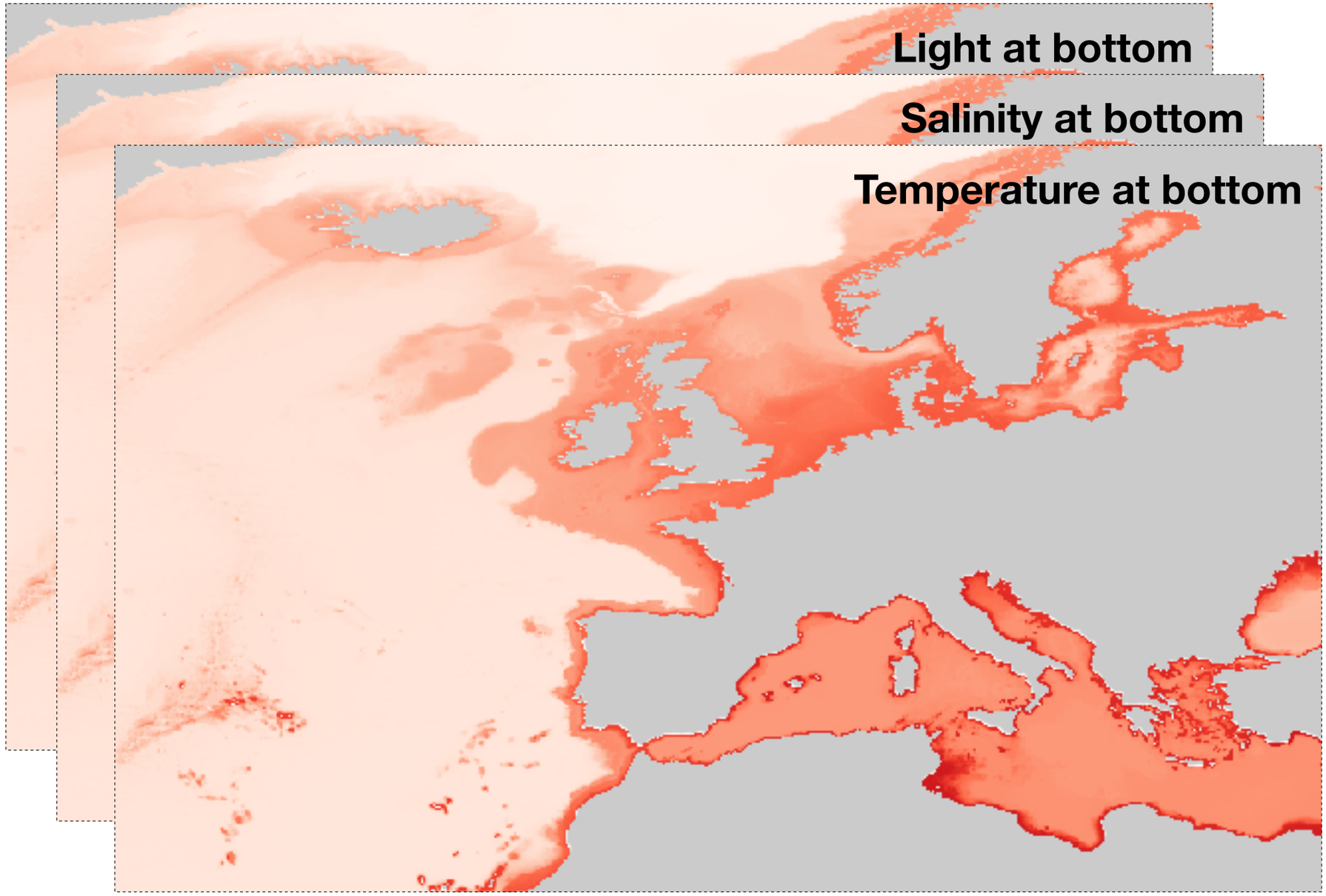
Provides insights into **species environmental tolerances or habitat preferences**, and allows **making spatial predictions** to available environmental layers (potential geographical distributions).

** also known as environmental niche modelling, species distribution modelling, habitat distribution modelling, …

Main approaches in ENM

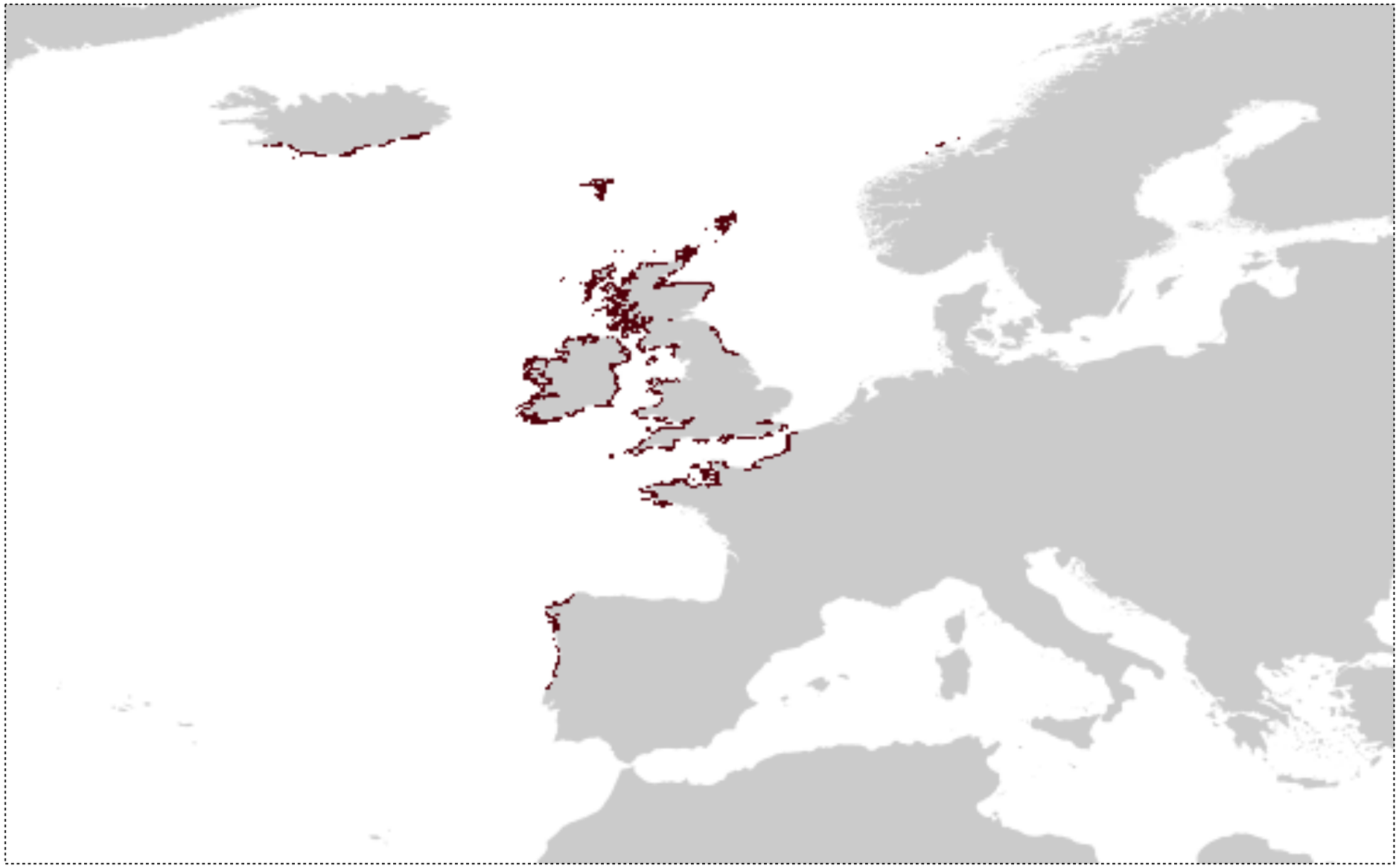
Mechanistic modelling specifically incorporates detailed data on the physiological response of species to environmental conditions (e.g., maximum temperature in which a species can survive - data is often not available).

Correlative modelling is based on the assumption that the distribution of a species is an indicator of its ecological requirements (niche theory).



Mechanistic distribution models

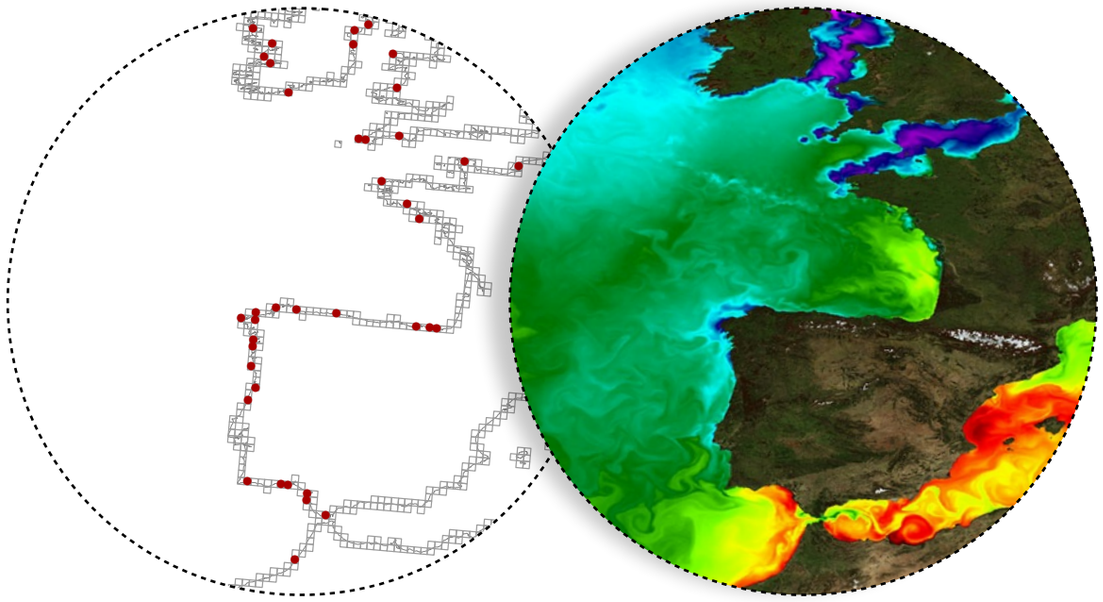
Built by **reclassifying environmental gradients with tolerance limits inferred from empirical physiological experiments**. Can include information on different life stages (e.g., reproduction / growth).



Mechanistic distribution models

A straightforward prediction of the potential distribution of species.

Light at bottom >= 50 E.m2.year-1 5°C <= Temperature <= 20.5°C	Presence

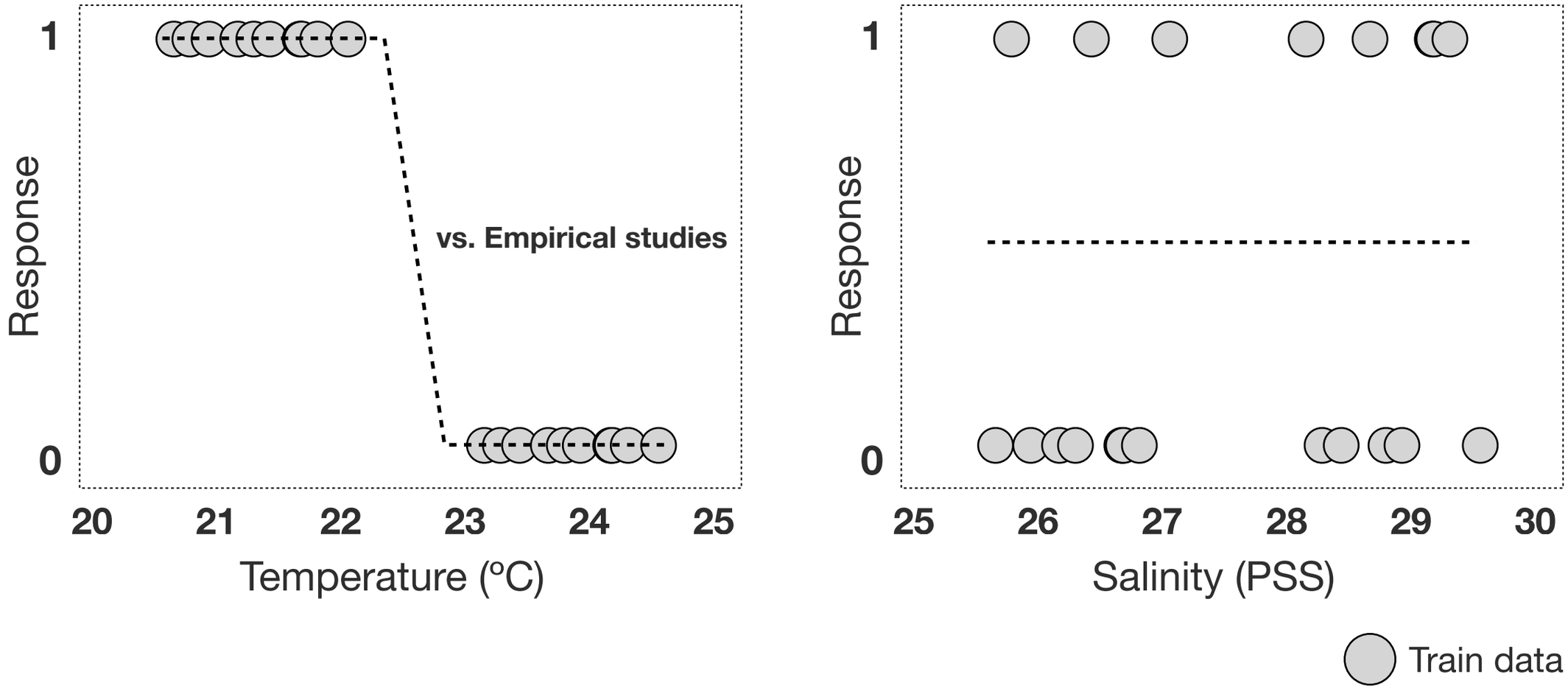


Resp	TempMax	Nitrate	Salinity
1	21	3	27
1	22	2	28
1	21	3	30
1	20	3	26
1	21	2	26
1	22	2	26
0	23	1	27
0	24	0	30
0	23	0	28
0	25	1	27
0	23	0	26
0	23	0	26

Training data for modelling

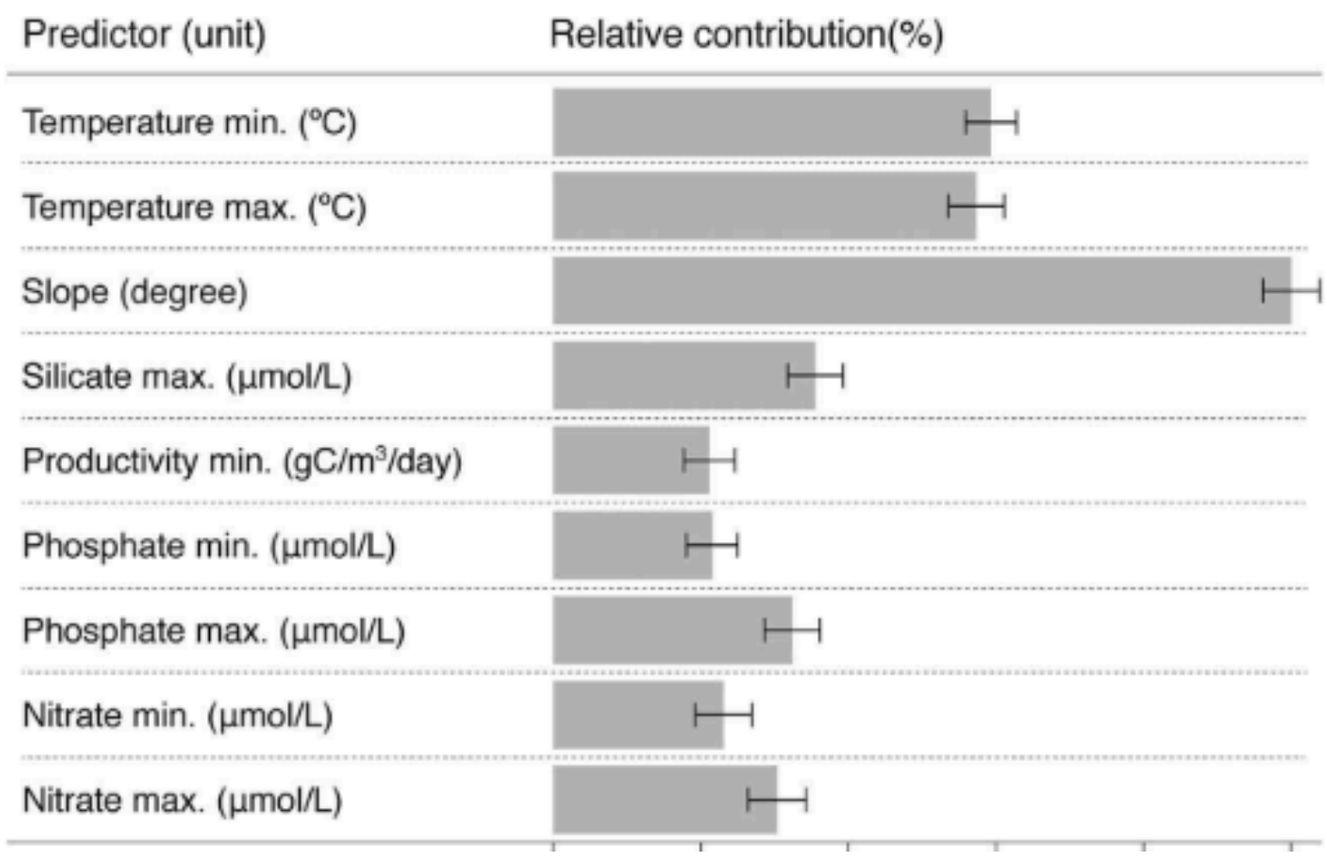
Correlative distribution models

Describe the **statistical relationship between distribution records and environmental variables.** The models should always be evaluated for “ecological realism”, that is, consistency with prior ecological knowledge of limiting factors and species response curves.



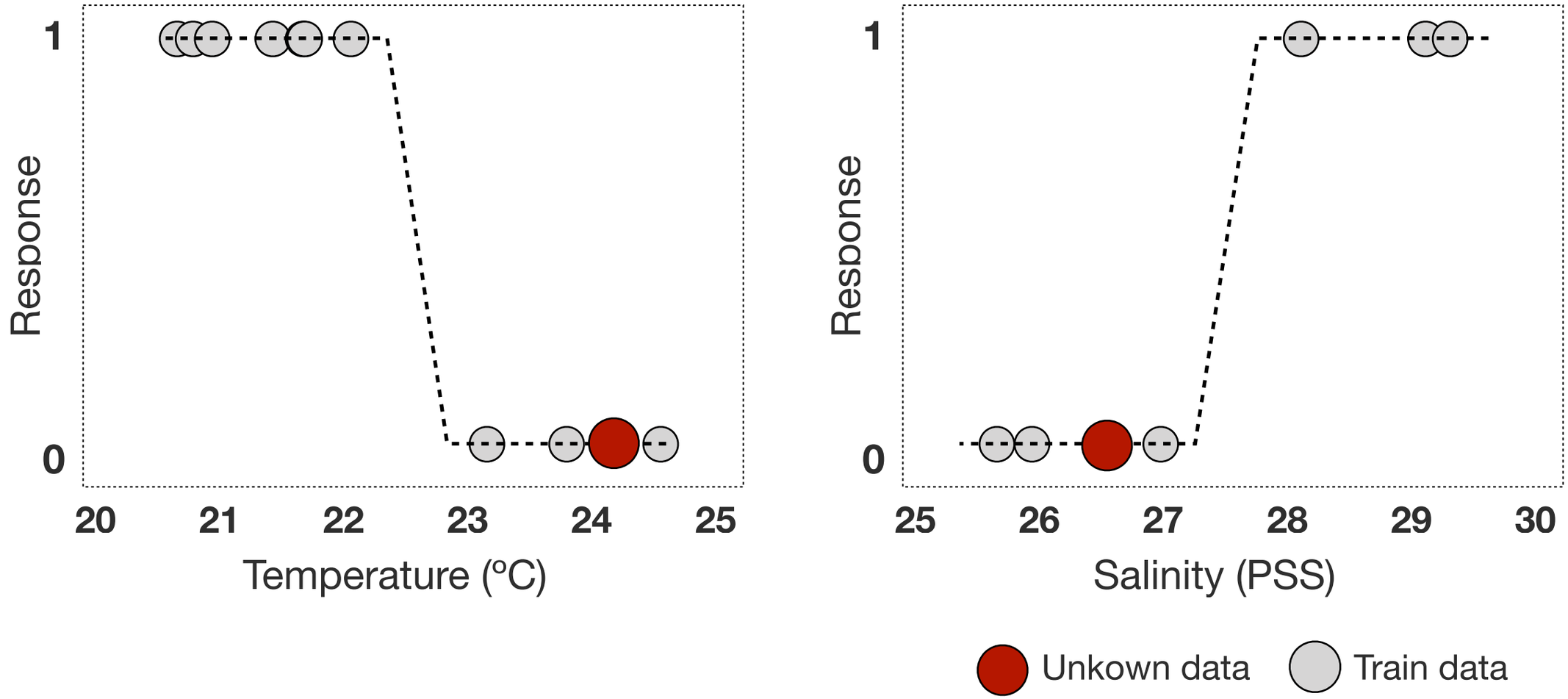
Correlative distribution models

Describe the **statistical relationship between distribution records and environmental variables**. The models should be evaluated for "ecological realism", that is, consistency with prior ecological knowledge of limiting factors and species response curves.



Correlative distribution models

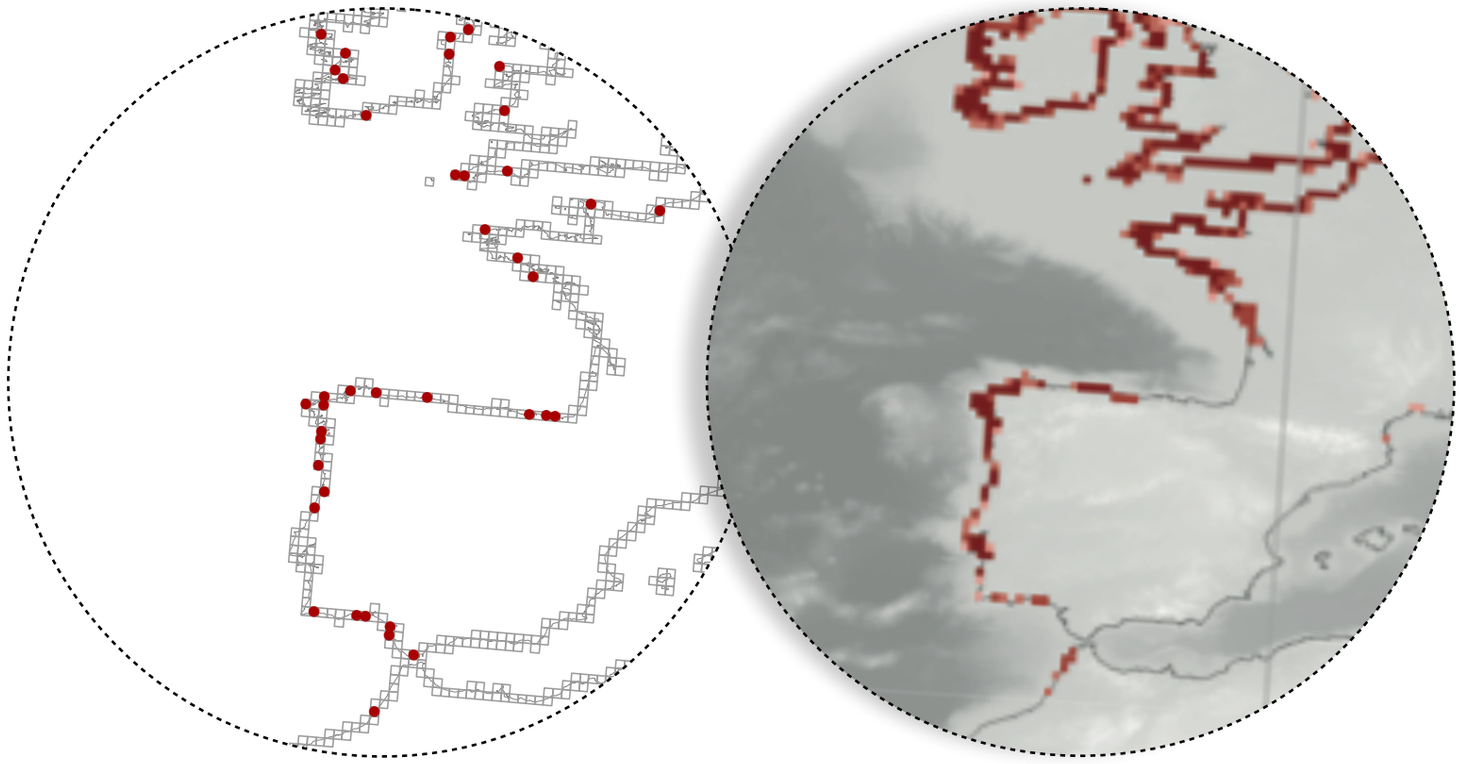
Allow to statistically infer and explain the patterns of the observed data (which drivers shape the distribution of a species and in what relative contribution).



Correlative distribution models

When models can explain the relationship between distribution records and environmental variables, predictions can be made for unknown samples.

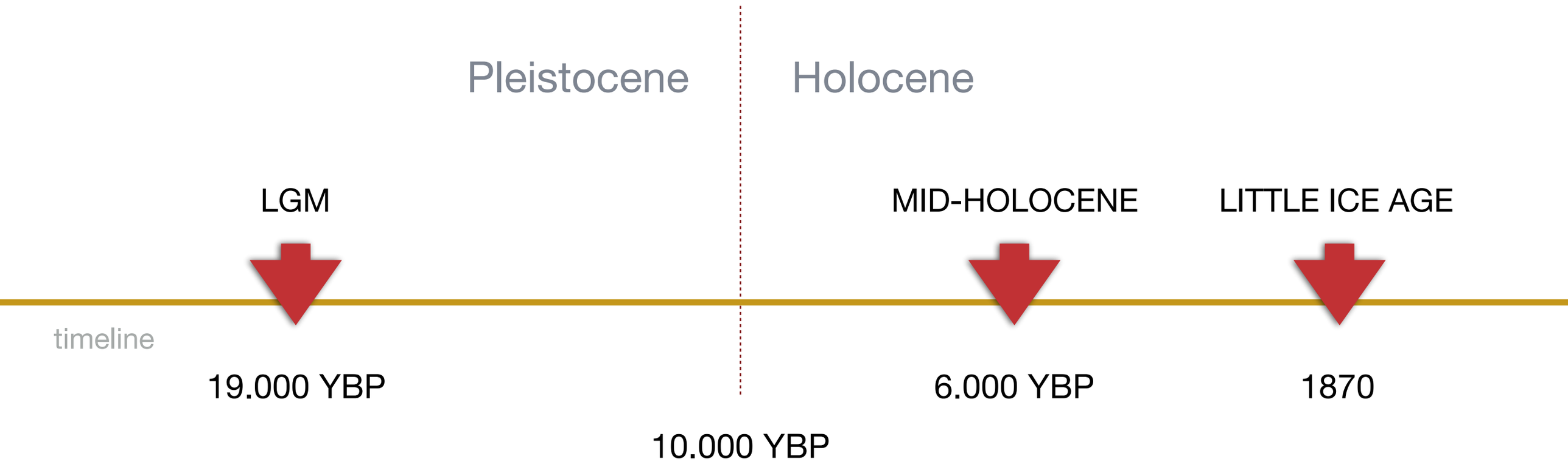
e.g., Temp. = 24.5ºC & Sal. = 26.5 PSS, response is 0 (i.e., absence).



Predictive model-based interpolation

Predictions made to **new sites within the range of environments sampled by the training data** and within the same time window in which the sampling occurred. **Scattered occurrence records produce continuous distribution surfaces**.

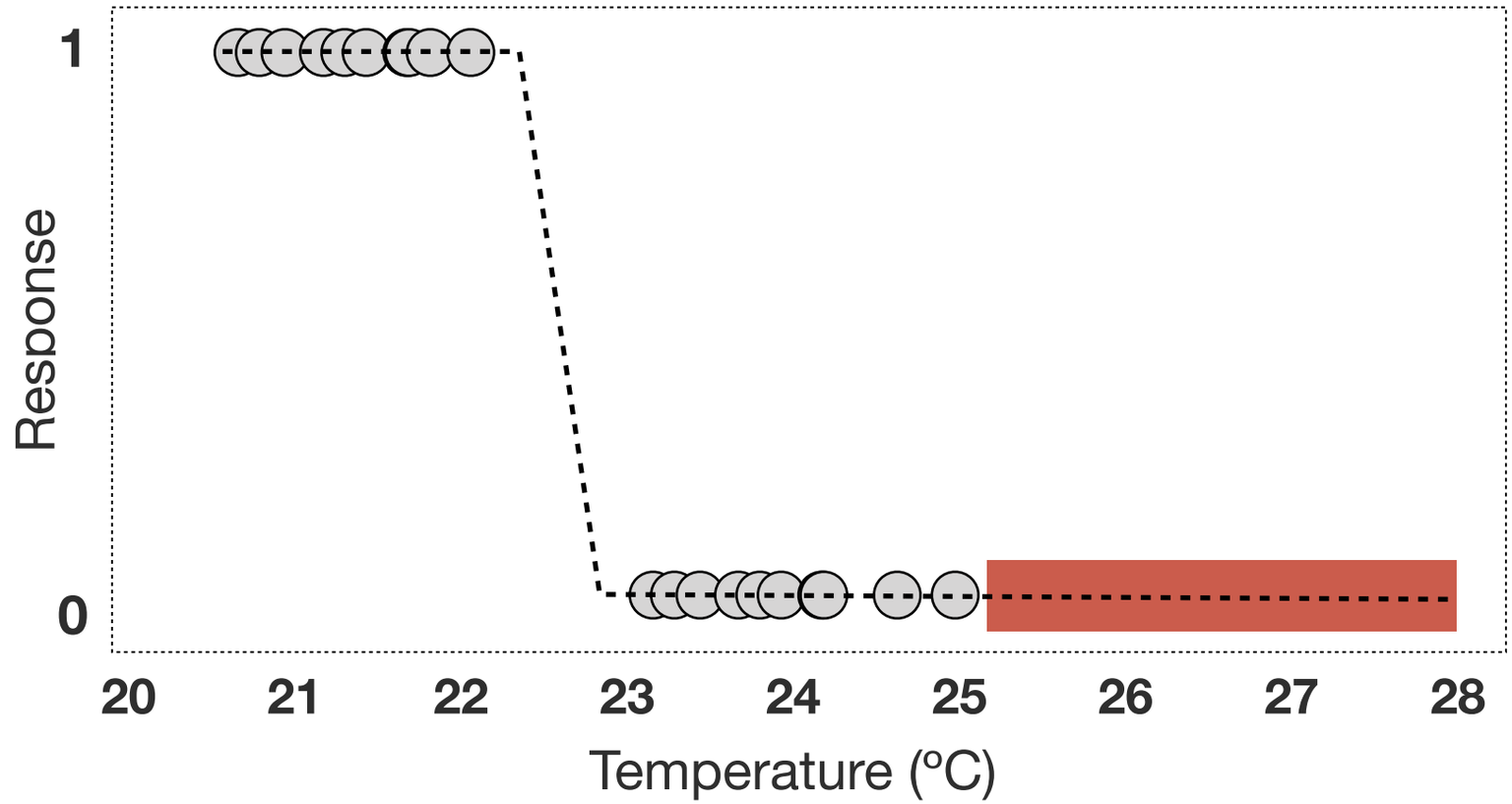
Typical applications include analyses of species distributions (present traits) and mapping within a region for conservation planning.



Predictive model-based transferability

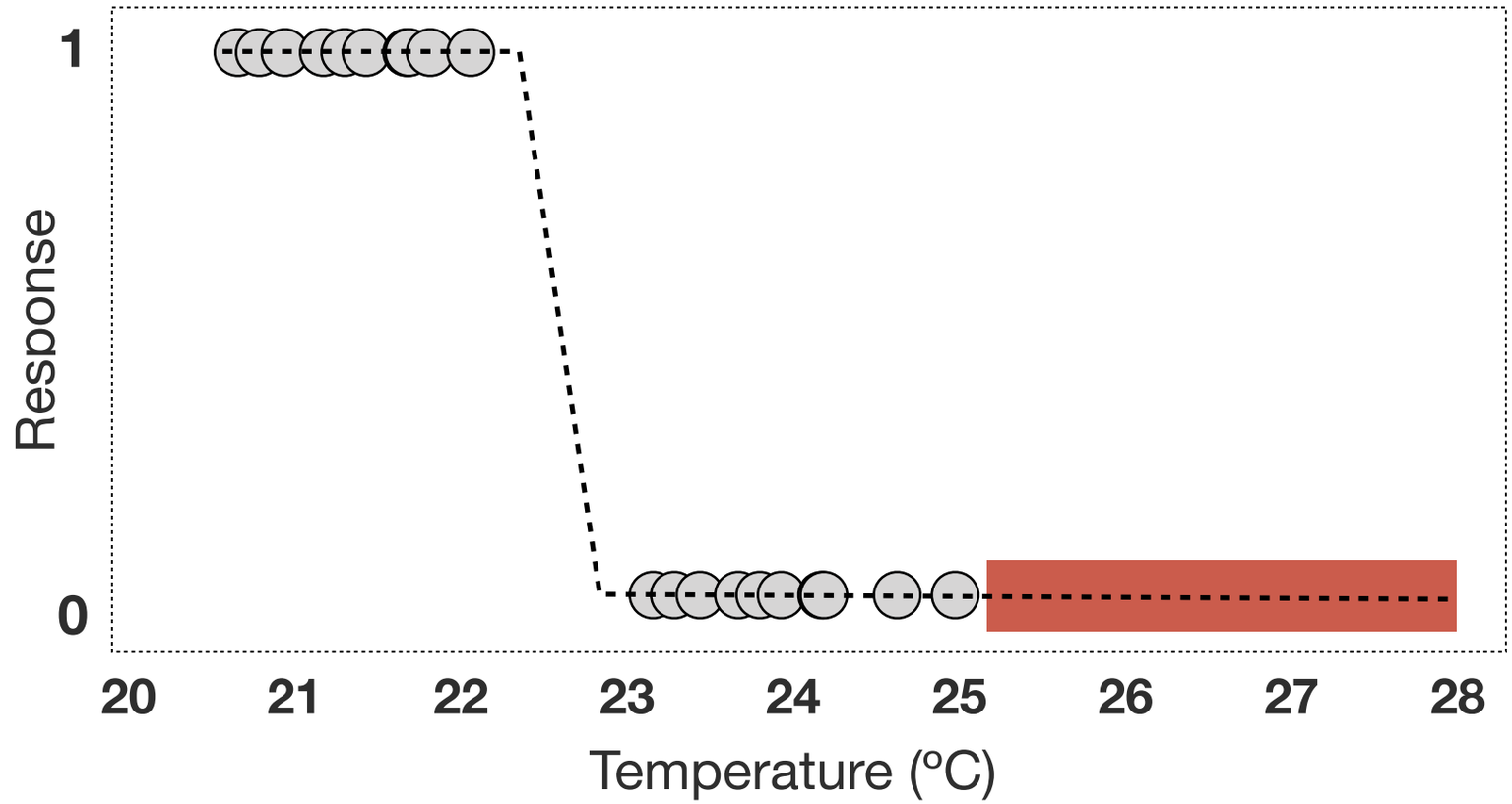
Predictions made to **unsampled geographic and temporal domains**. No information on the environmental similarities and differences between where the model fits and where predictions were made.

Transferability to new environments **may lead to extrapolation**. This creates uncertainty because no records of occurrence are available to support predictions.



Extrapolation refers to **predictions for** environmental **values** that are **beyond the range of the data used to fit the model**.

e.g., Model used records that spanned a temperature range of 20-25ºC. If predicting in a different region or climatic scenario where temperatures reach > 25ºC, then the model will extrapolating. No prior information exists for the probability of occurrence at > 25ºC, so the predictions will be uncertain.



Avoid extrapolation in favor of interpolation

But when extrapolation exists (e.g., future climate changes), model interpretation should be treated with a great deal of caution.

1. Avoid predicting with complex functions;
2. Use a parsimonious models (to the minimum number of predictors).
3. Interpret models with sound ecological knowledge;

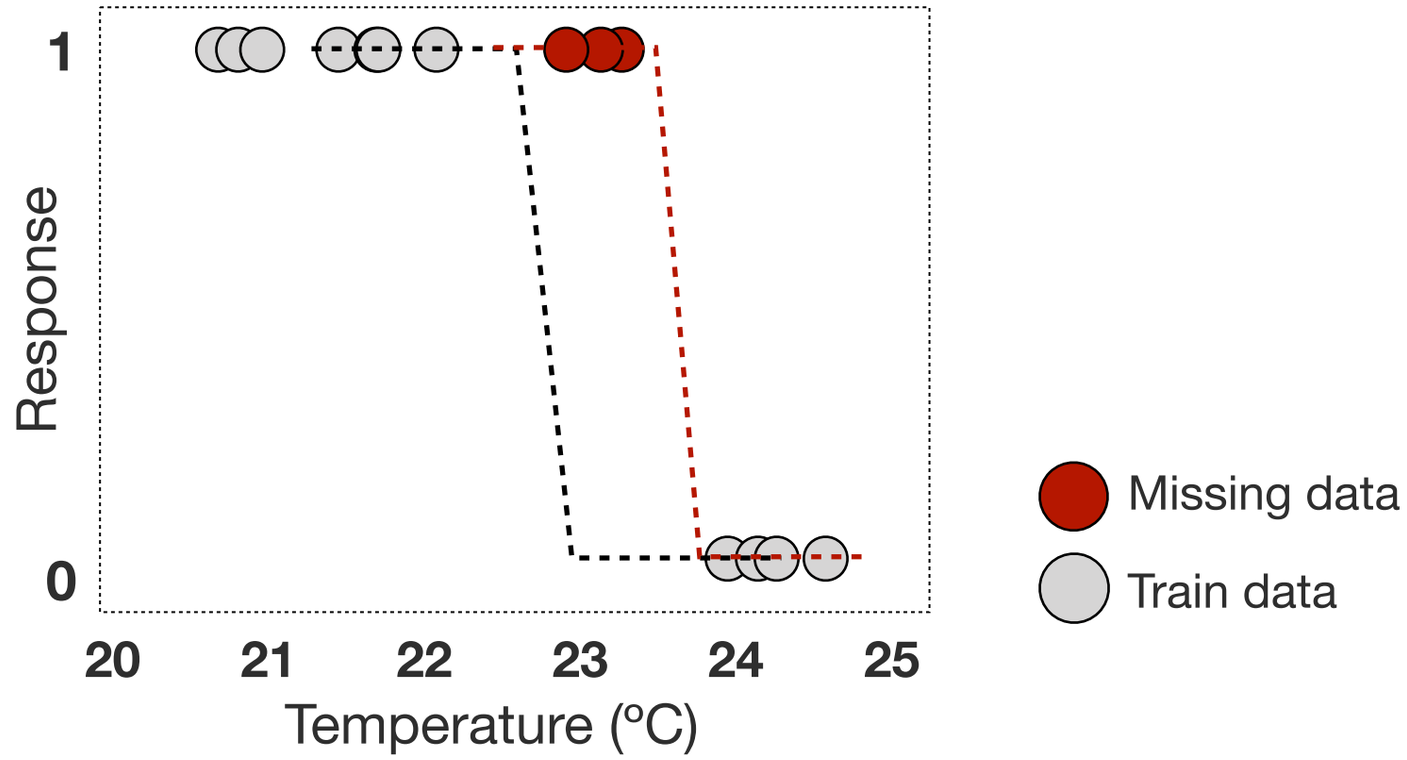
Which niche is being modelled?

Mechanistic models are rooted in the fundamental niche.

What is being modelled in corrective models? The fundamental niche, the realized niche or something in between?

Corrective models identify the **realized niche if the records represent the full distribution of a species** - data used on the actual species distribution so the model extrapolates in the geographical space where the species realizes its occurrence**.** When mapped in space, it represents the **potential distribution or habitat suitability.**

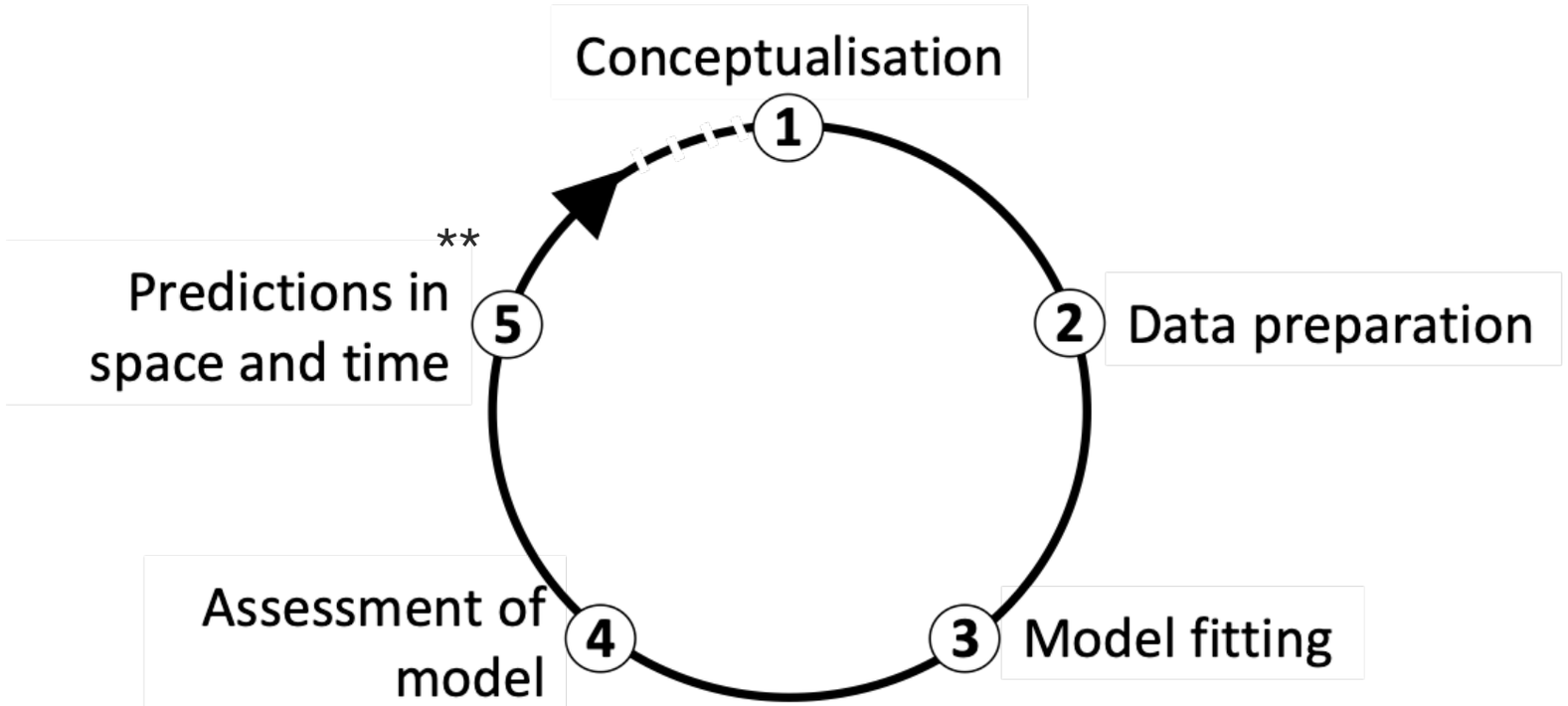
If insufficient distribution records feed the model, it identifies neither the realized niche nor the fundamental niche; **the model fits only to the portion of the niche that is represented by the observed records** (truncated niche).



Corrective models identify the **realized niche if the records represent the full distribution of a species** - data used on the actual species distribution so the model extrapolates in the geographical space where the species realizes its occurrence**.** When mapped in space, it represents the **potential distribution or habitat suitability.**

If insufficient distribution records feed the model, it identifies neither the realized niche nor the fundamental niche; **the model fits only to the portion of the niche that is represented by the observed records** (truncated niche)**.**

Steps for proper model building



Model building is an iterative process and there is much to learn on the way (a cycle rather than a workflow with a pre-defined termination point).

You may want to revisit and improve certain steps (e.g., improve biodiversity data collection or remove surplus environmental layers).

** not always part of ENM studies but depends on the model objective.

1. Conceptualisation

Formulation of the **main research objectives** and decide on the **study setup based on sound ecological knowledge on the species and study system.**

Is there comprehensive biodiversity and environmental data available or there is the need to gather more data?

(may require an appropriate sampling design or data may be unavailable for the objectives proposed).

2. Data preparation

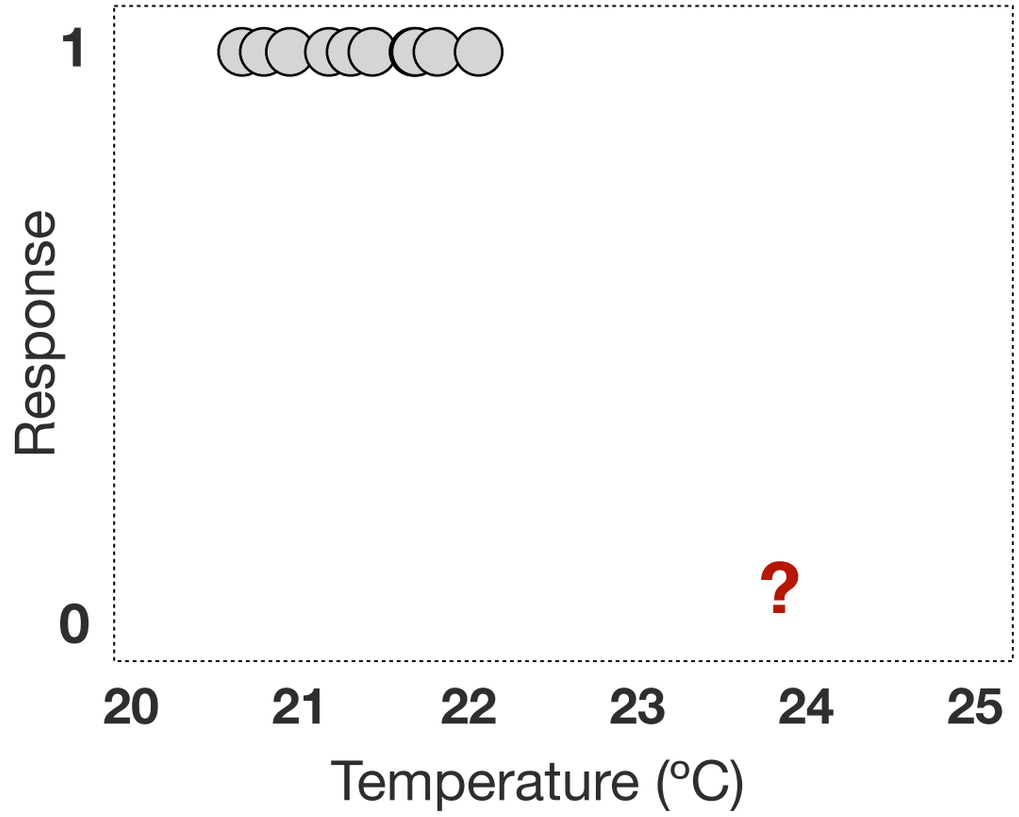
To gather and process biodiversity and environmental data to develop models and making predictions (e.g., model transferability).

How many biodiversity records for modelling?

Sample size has been found to be positively related to the performance of models. Experiments using artificial data point to at least **50 records are needed** to estimate response functions.

But the absolute number of observations is less important than having **observations that are well-distributed throughout the environmental space that the species occupies** (including observations defining range limits!). Thus, compiling biodiversity data should be guided by the best knowledge possible on distributions.

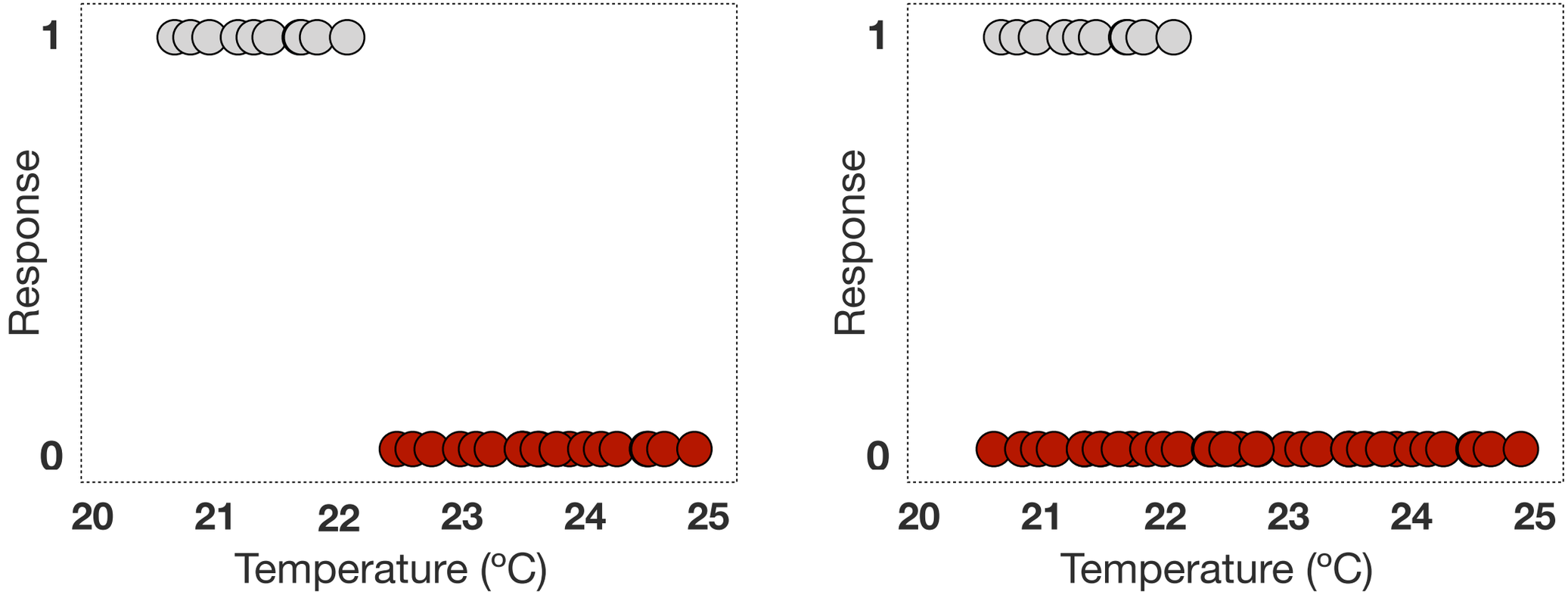
Coudon and Gegout, 2007



Absence is often unknown leading to **presence-only datasets**.

Broad use of **models with presence-only data justified by the lack of systematic surveys and the demand for making predictions**. Predictions by presence-only models **pointed as more robust than with presence-absence data**.

Absence records may introduce confounding information as they indicate that (1) the habitat is unsuitable or (2) the habitat is suitable but is unoccupied due to inaccessibility or low species detectability.



ENM based on ‘pseudo-absences’ used in place of real absence data. Generated from the study area where occurrences do not exist. Any regression algorithm can be implemented (e.g., GLM).

ENM based on ‘background’ environmental data from the entire study area. Focus on how the environment where the species is known to occur relates to the environment across the rest of the study area (e.g., MaxEnt).

How many absence records?

Absences must cover beyond the range of environmental values where the species is present to allow describing the relationship between distributions and the environment.

Completeness, the degree to which the range of the environmental variables is covered by species absence / presence records. Incomplete absence records, negatively affect the performance of models.

Kadmon et al., 2003

How many absence records?

Take into account **prevalence**, which is the **proportion of presence records relative to the number of absences**.

A default ratio of 1:1 (same number of pseudo-absences as presences) often performs well for presence-absence data.

In contrast with species natural prevalence, the sampled ratio is under the control of the modeller and is a function of the number of designated pseudo-absence or background locations.

When generating **pseudo-absence or background** locations **to complement presence-only** species data, **a very large number of background sites may be required to fully describe variation in the environment** - can be several orders of magnitude larger than the number of presences.

Manly et al., 2002; Phillips & Dudik, 2008

How many absence records?

It is recommend the use of a large number (e.g. 10 000) of pseudo-absences / background locations when using regression techniques (e.g. generalised linear models, generalised additive models and maximum entropy models);

Averaging several runs (e.g. 10) with fewer pseudo-absences (e.g. 100) with equal weighting for presences and absences with multiple adaptive regression splines and discriminant analyses;

The same number of pseudo-absences as available presences (averaging several runs if few pseudo-absences) for classification techniques such as Boosted Regression Trees models, Classification Trees models and Random Forest models.

Barbet-Massin et al., 2012

Which environmental predictors for modelling?

Different strategies for selecting predictors:

Large datasets, an approach with stronger criticism.

Preselected, corresponding directly to known physiological rules. The choice of predictors (resource variables, direct variables and indirect variables) is **guided by the research objectives and the hypotheses raised** regarding the species-environment relationship.

How many environmental predictors are needed?

Also depends on the scale of the question addressed, the complexity of the species' ecology and the availability of data.

The higher the scale, the more predictors needed (e.g., local scale models for intertidal species with shade, wave exposure, temperature).

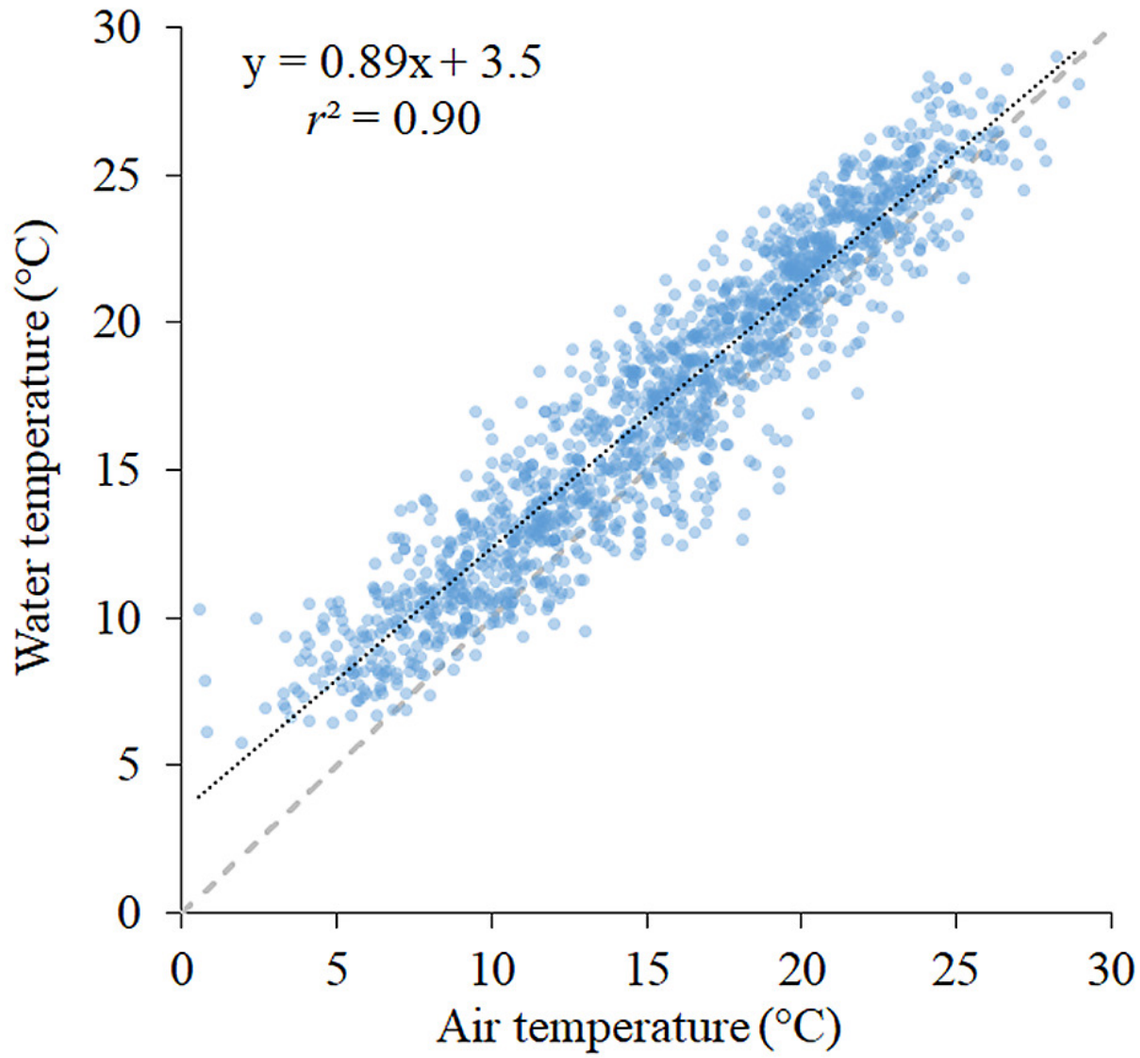
The complexity of ecological niches varies between species (e.g., intertidal species responding to both air and temperature).

The importance of predictors varies between species (e.g., light importance for intertidal Vs. Depth distributions for macroalgae species).

How many environmental predictors are needed?

When **few predictors are used**, the models risk missing important information: under characterization of niches, likely to produce broad potential distributional areas because the model misses information.

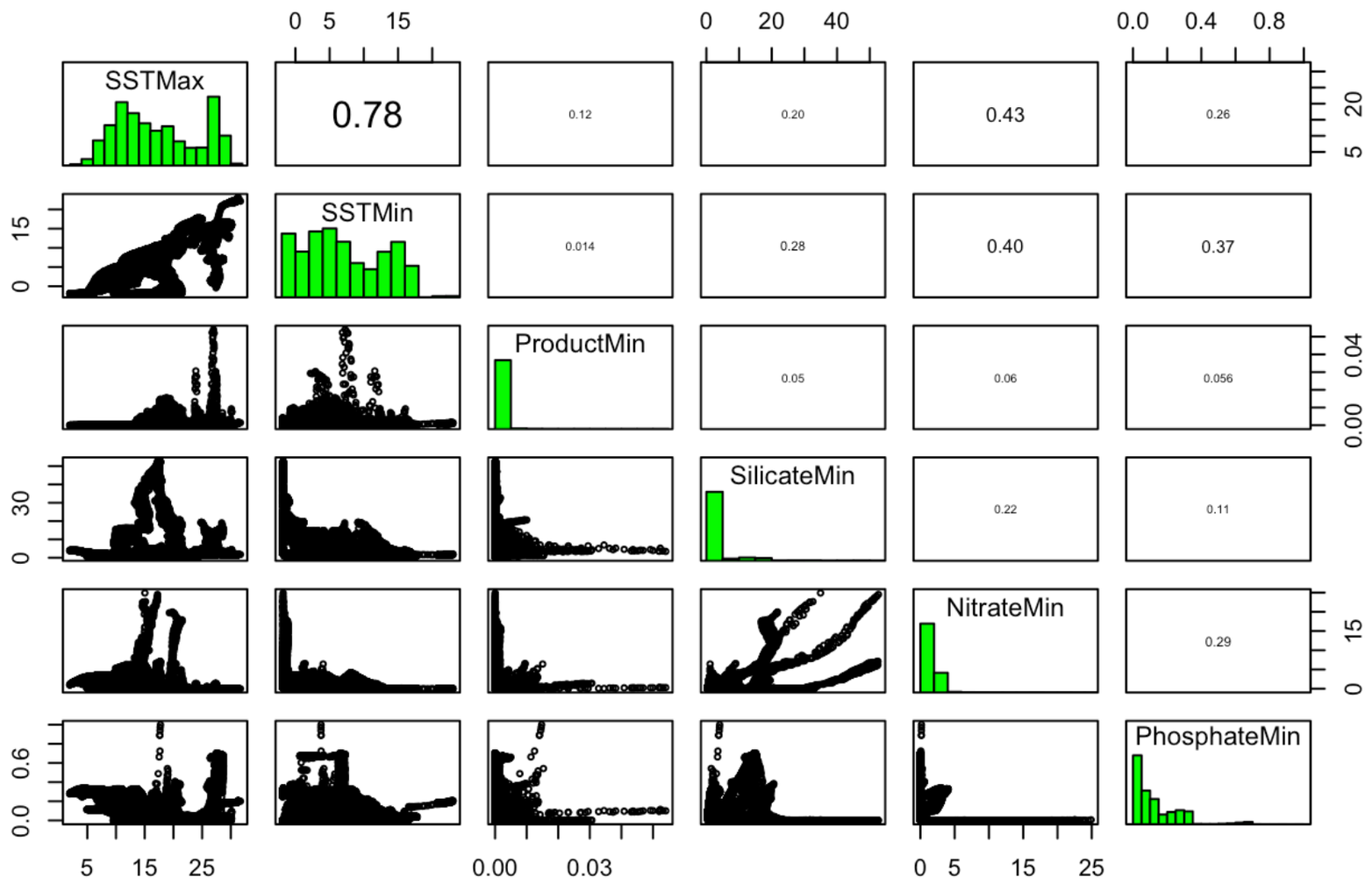
Over-dimensionality (i.e., **excessive numbers of predictors**) can lead to collinearity issues that often impede the characterization of the niche in ecological terms.



Collinearity issues and the selection of predictors

Model characterization of the niche in ecological terms for species in Lake Taihu (China) can be misleading.

Solution: model with just one (which? read literature!) and discuss that additional effects cannot be discarded.



Collinearity issues and the selection of predictors

Also, many statistical algorithms (e.g., GLM) have problems fitting data if predictors are strongly correlated - multicollinearity. To avoid it, a pairwise correlation test is made to discard predictors that are strongly correlated. Correlations below $|r|<0.85$ are considered unproblematic (below $|r|<0.5$, more conservative threshold).

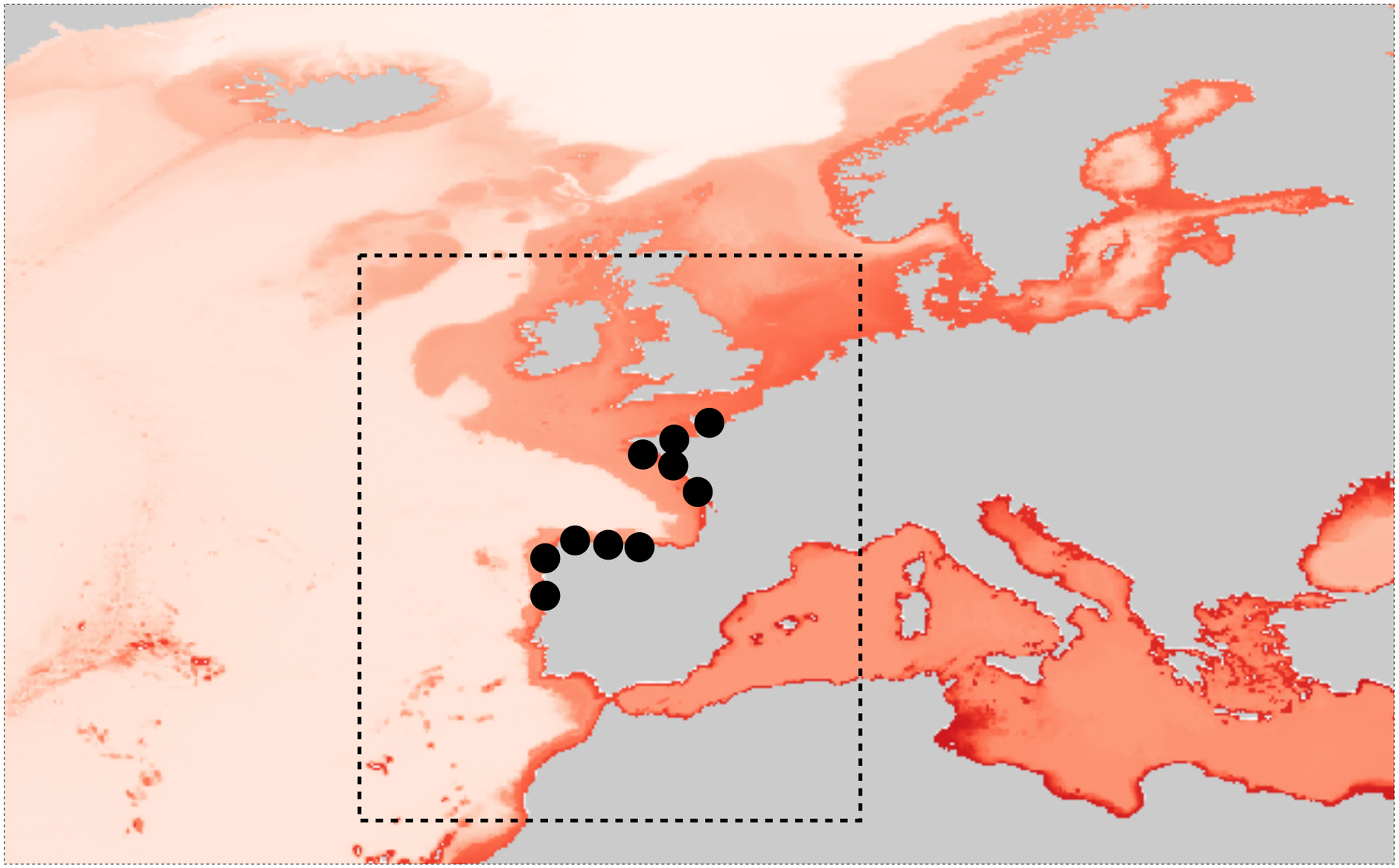
Elith et al. 2006

Which environmental predictors for modelling?

1. The minimum number of predictors that represent the relevant conditions for the species' ecology (ecological meaningful model).

2. Prior to modelling perform a correlation analysis between predictors to avoid excessive autocorrelation.

3. (additional rule of thumb) Use at least 20 times more records than predictor variables.



Which geographic extent for predictors?

Extent refers to the region **over which the model will run**. **Must include all presence records** and allows a sufficient large contiguous **absence area** to detect unfavorable environmental conditions. Should not consider additional areas where the environmental conditions are different (e.g., other oceans).

Model fitting

Key steps in good modelling practice include the following:

1. Gathering relevant biodiversity and environmental data and assessing its adequacy (the relevance and completeness of the predictors);

2. Deciding how to deal with correlated predictor variables;

3. Selecting an appropriate modelling algorithm;

4. Fitting the model to the training data and evaluating performance, including the realism of fitted response functions, the model's fit to data, and predictive performance on test data;

5. Mapping predictions to geographic space;

6. Apply thresholds if continuous predictions need reduction to a binary map;

7. Iterating the process to improve the model in light of knowledge gained throughout the process.

(Elith & Leathwick 2009)

Model fitting

Key steps in good modelling practice include the following:

1. Gathering relevant biodiversity and environmental data and assessing its adequacy (the relevance and completeness of the predictors);

2. Deciding how to deal with correlated predictor variables;

3. Selecting an appropriate modelling algorithm;

4. Fitting the model to the training data and evaluating performance, including the realism of fitted response functions, the model's fit to data, and predictive performance on test data;

5. Mapping predictions to geographic space;

6. Apply thresholds if continuous predictions need reduction to a binary map;

7. Iterating the process to improve the model in light of knowledge gained throughout the process.

(Elith & Leathwick 2009)

Main assumptions of ENM

1. Observed distribution is indicative of environmental tolerances and resource requirements.

2. Species are in (quasi-) equilibrium with environmental conditions - species occur in all suitable areas and is absent from all unsuitable areas.

The degree of equilibrium depends both on biotic interactions (for example, competitive exclusion from an area) and dispersal ability (organisms with higher dispersal ability are expected to be closer to equilibrium than organisms with lower dispersal ability).

Desired properties of ENM fitting

Deductive: develop hypotheses about the causes of the pattern explained and predicted by the models.

Distinct between patterns and process: distinguish between the patterns observed and the mechanisms involved.

Simplicity: highlight a few mechanisms without becoming entangled in complex interactions and correlations between many variables.

Parsimonious: preferring the simpler of two equally adequate models; favor simple explanations over complex.

Generality: aimed to achieve general broad ecological conclusions.