CALIBRATION:

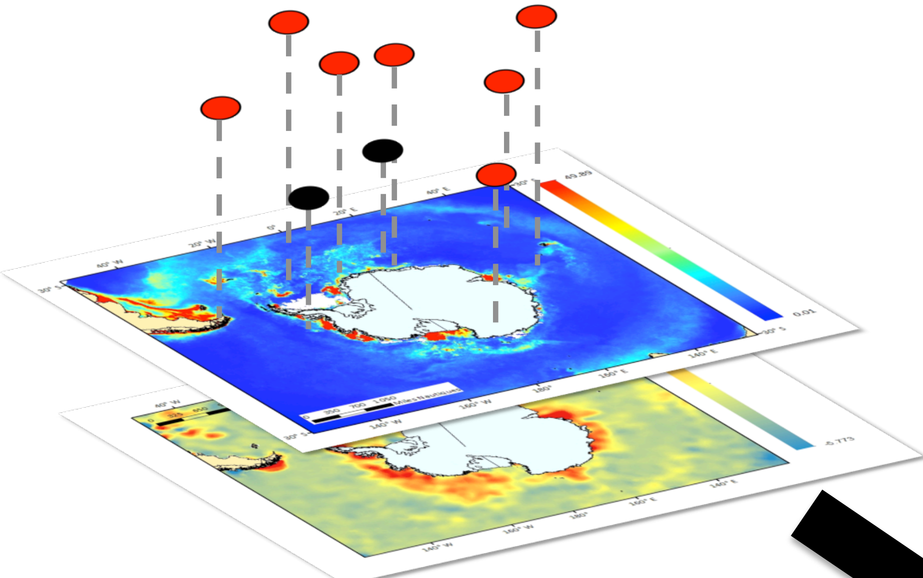
All you should think about and check before running a model !

Tuesday 3rd, September

Guillaumot Charlène
charleneguillaumot21@gmail.com

SPECIES DISTRIBUTION MODELS principle

[presence + absence records]

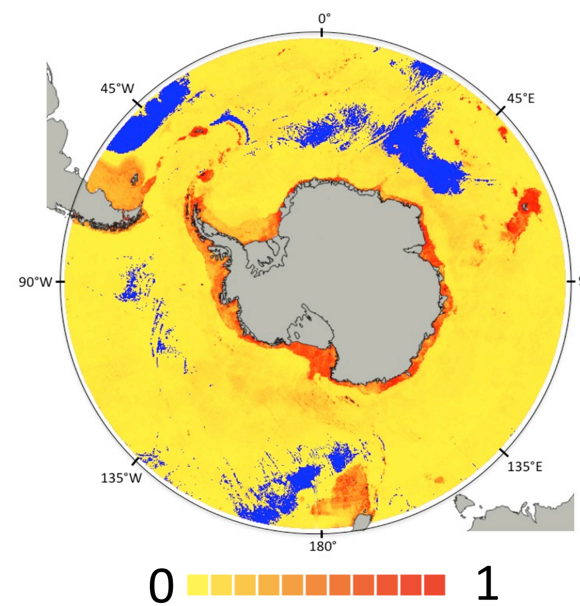


[set of environmental variables]

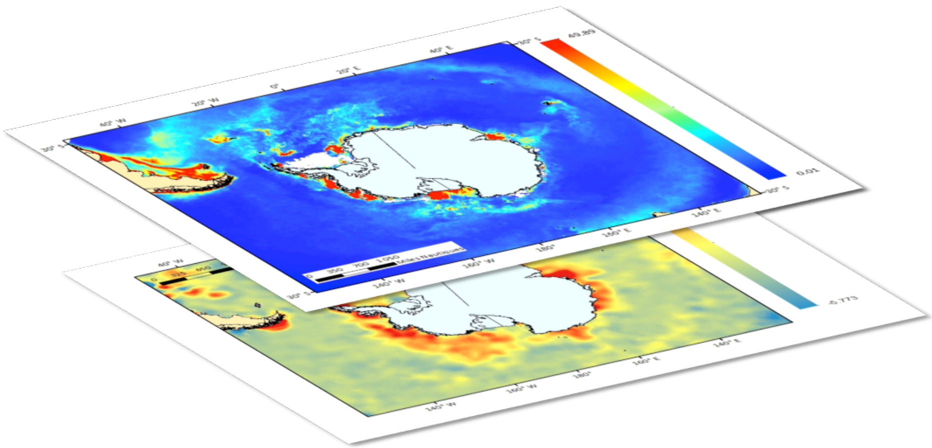
Presence / absence?	Layer 1 e.g. Depth	Layer 2 e.g. T°	Layer 3 e.g. Salinity
1	-351	0.2	32.4
1	-150	-1.4	32.1
0	-1053	-2	32.8
1	-3042	0.3	31.9
…	…	…	…

SDM

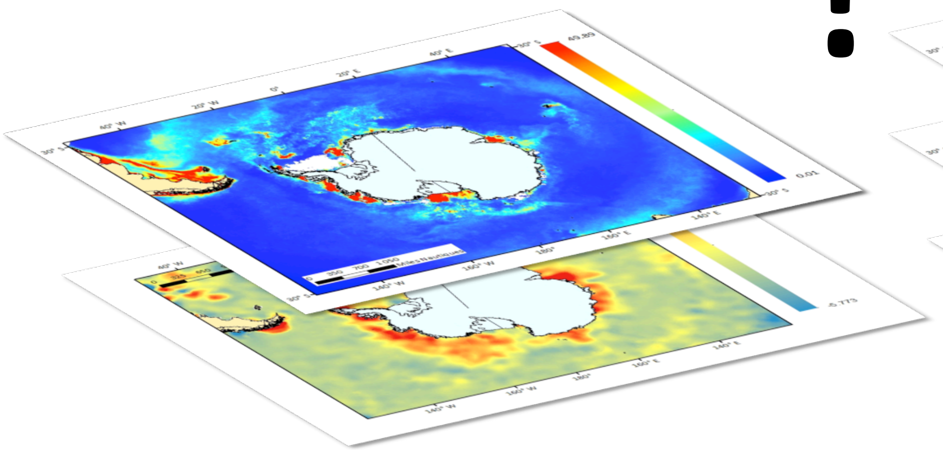
[Predicted distribution]



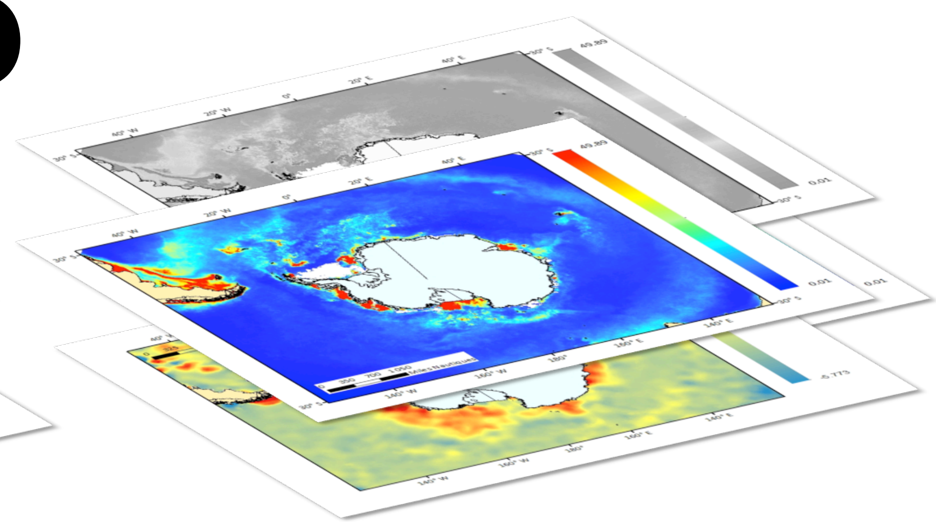
CALIBRATION: Environmental variables



CALIBRATION: Environmental variables

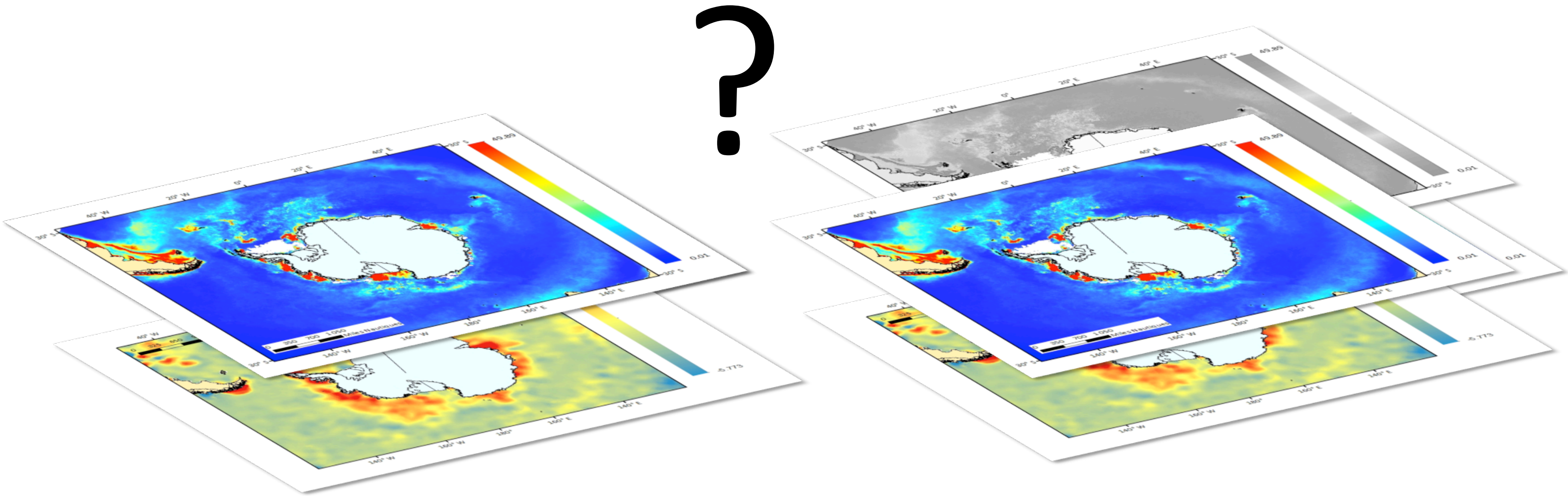


?



- Number of environmental variables?

CALIBRATION: Environmental variables

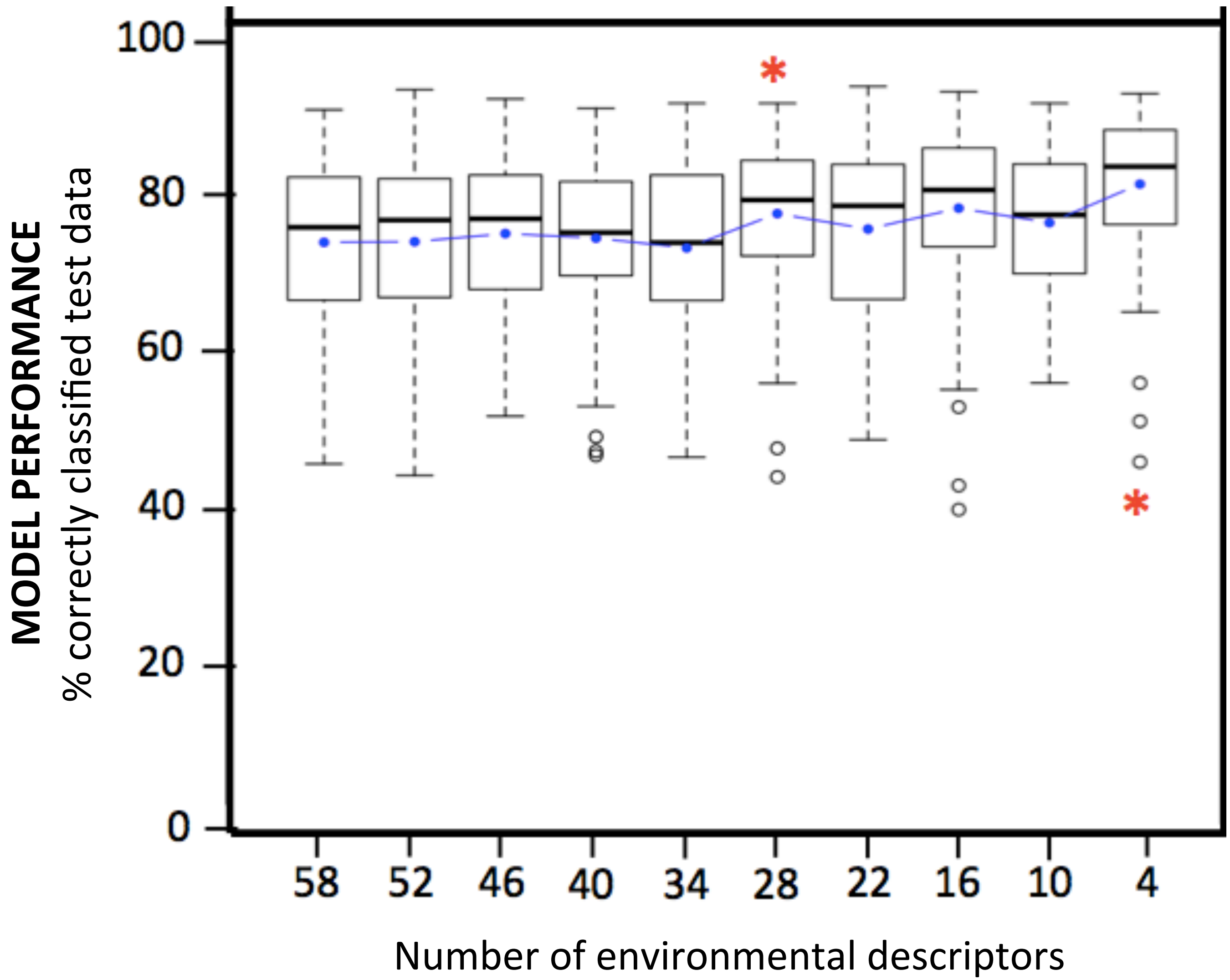


- Number of environmental variables?

➔Ecological relevance vs. parcimony

➔New algorithms can deal with redondant/useless information

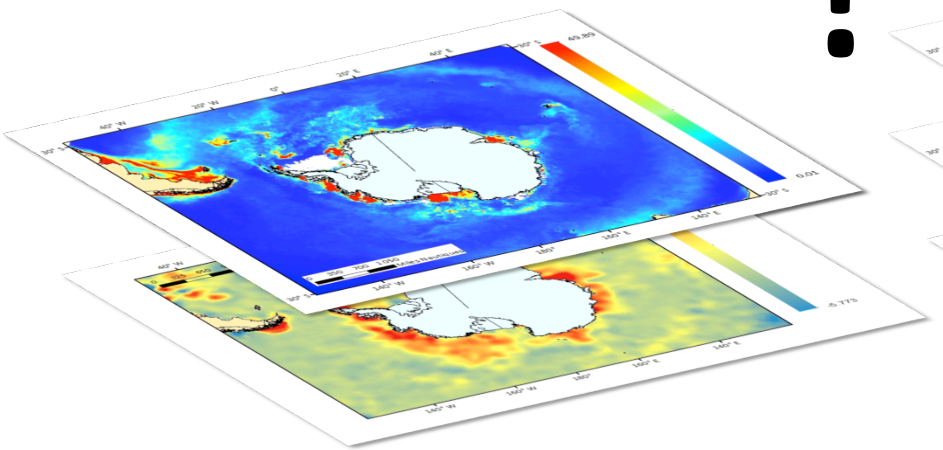
BRT

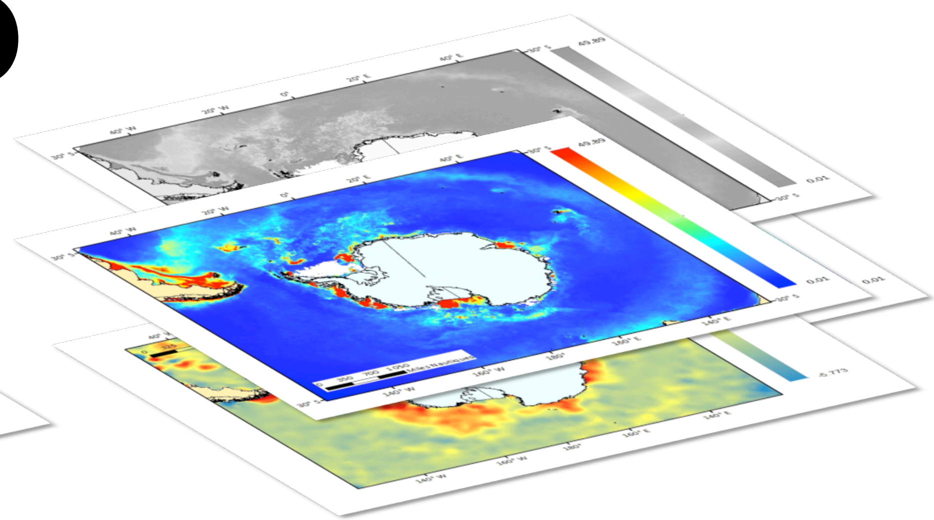


Guillaumot et al. in press

CALIBRATION: Environmental variables

?





- Number of environmental variables?

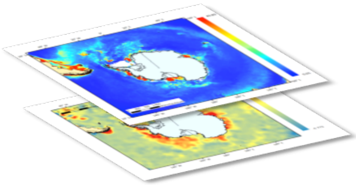
➔Ecological relevance vs. parcimony

➔New algorithms can deal with redondant/useless information

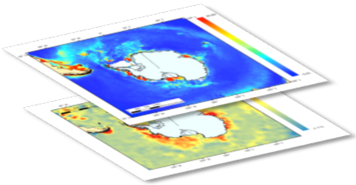
- Be careful with average information

➔(relevance of average environment ? vs. amplitude/min/max?)

CORRELATION BETWEEN ENVIR. VARIABLES

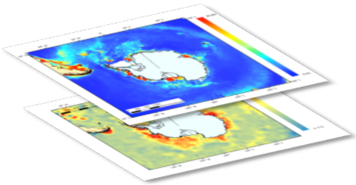


CORRELATION BETWEEN ENVIR. VARIABLES



-> situation where at least two variables are related in a statistical model

CORRELATION BETWEEN ENVIR. VARIABLES



-> situation where at least two variables are related in a statistical model

- Can biais modelling outputs
- Can inflate errors
- Generally removed before generating the models

STATISTICS TO DEAL WITH COLLINEARITY

- Spearman correlation/ correlation matrix

- Variance Inflation Factor (VIF) (threshold : 10 or 5 according to studies)

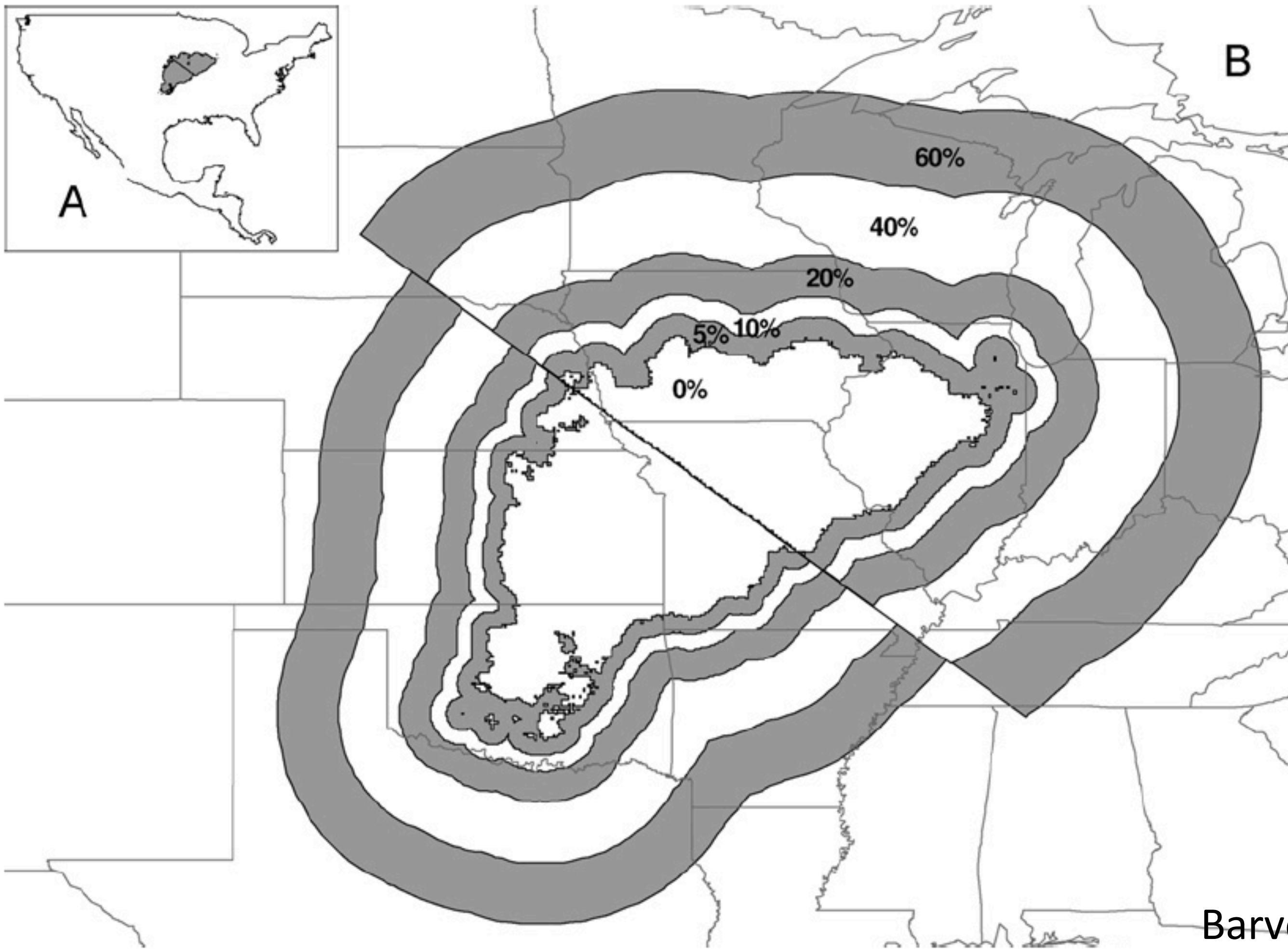
$$VIF= \frac{1}{1- R^2}$$

(more details in https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/)

- Automatic removal by most machine learning approaches

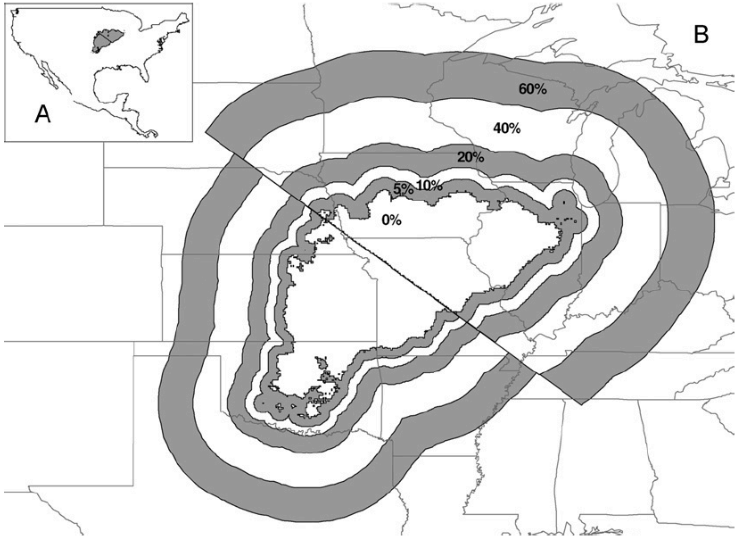
INFLUENCE OF SPATIAL RESOLUTION AND SCALE

INFLUENCE OF SPATIAL RESOLUTION AND SCALE



Barve et al. 2011

INFLUENCE OF SPATIAL RESOLUTION AND SCALE



Narrower niches
-> better predictive performances

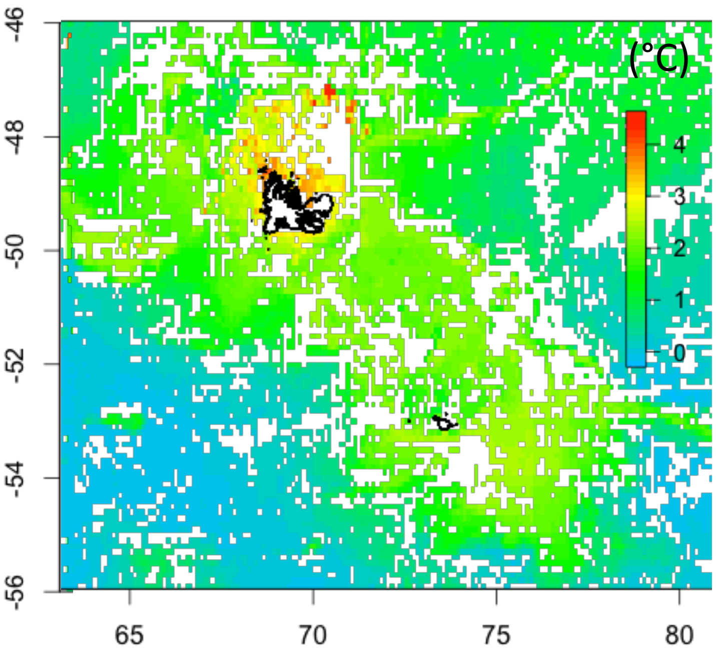
Barve et al. 2011

INFLUENCE OF MISSING DATA

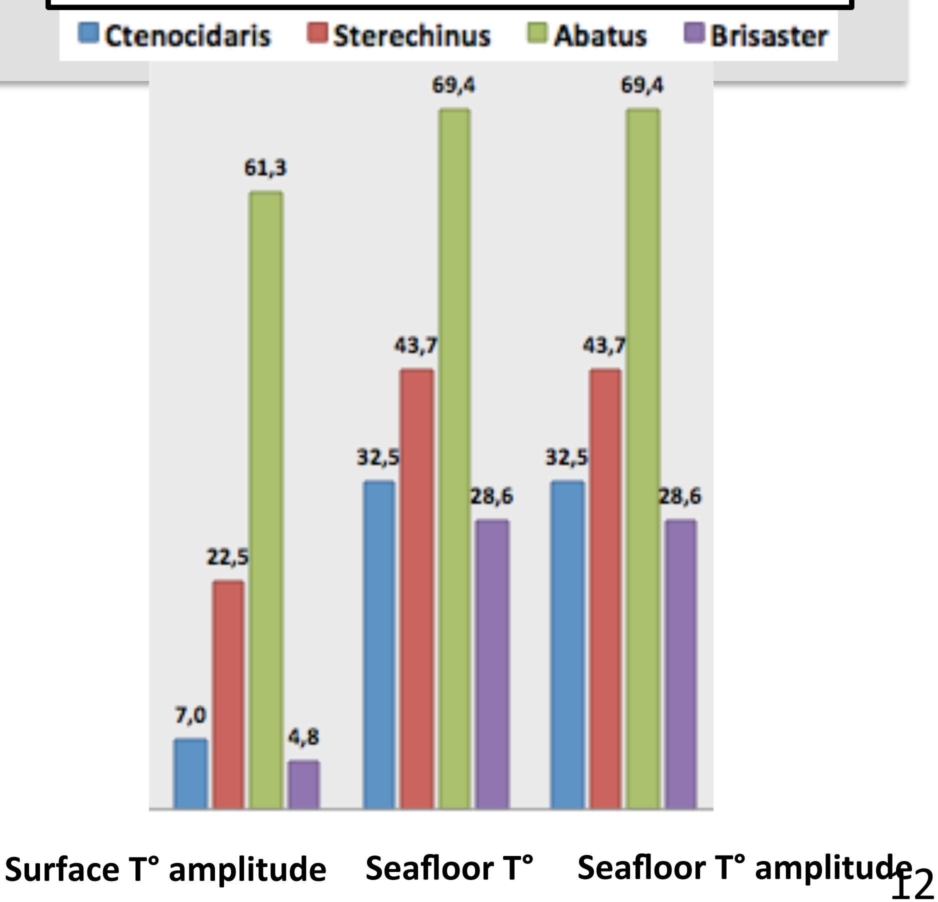
INFLUENCE OF MISSING DATA

- Partial coverage of the information -> interpolation or not / missing values

Seafloor T° on the Kerguelen Plateau



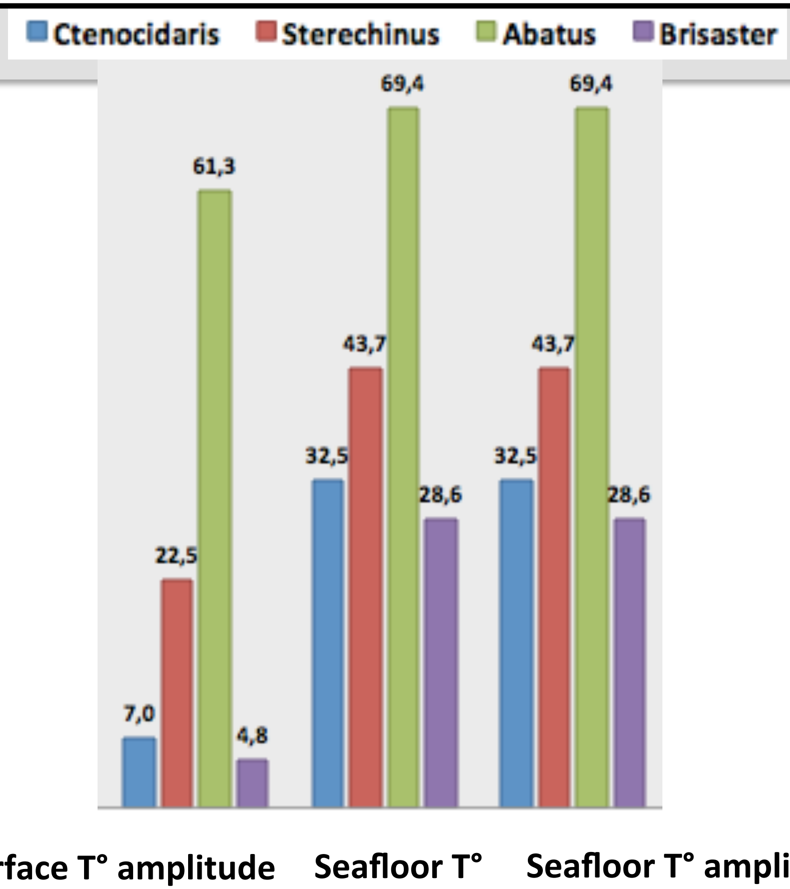
Presence data falling on missing values



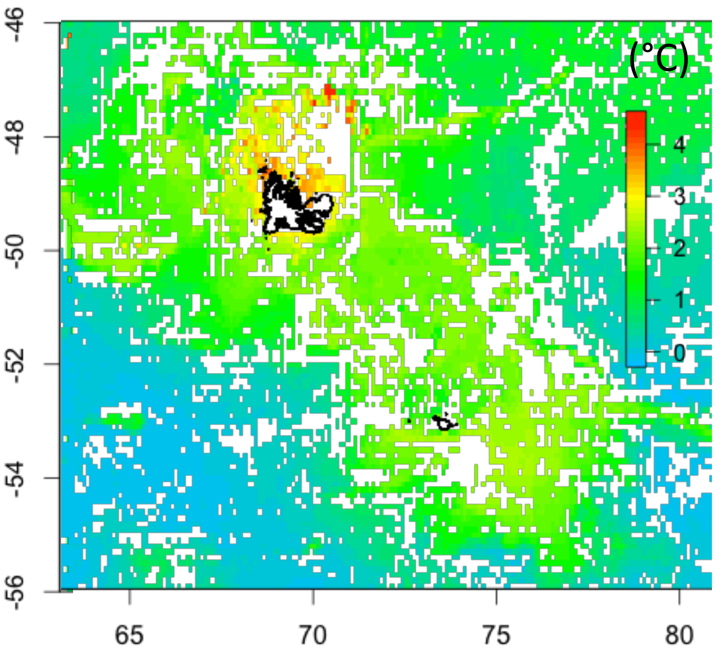
INFLUENCE OF MISSING DATA

- Partial coverage of the information -> interpolation or not / missing values
- Full night in winter -> no satellite data

Presence data falling on missing values



Seafloor T° on the Kerguelen Plateau



INFLUENCE OF MISSING DATA

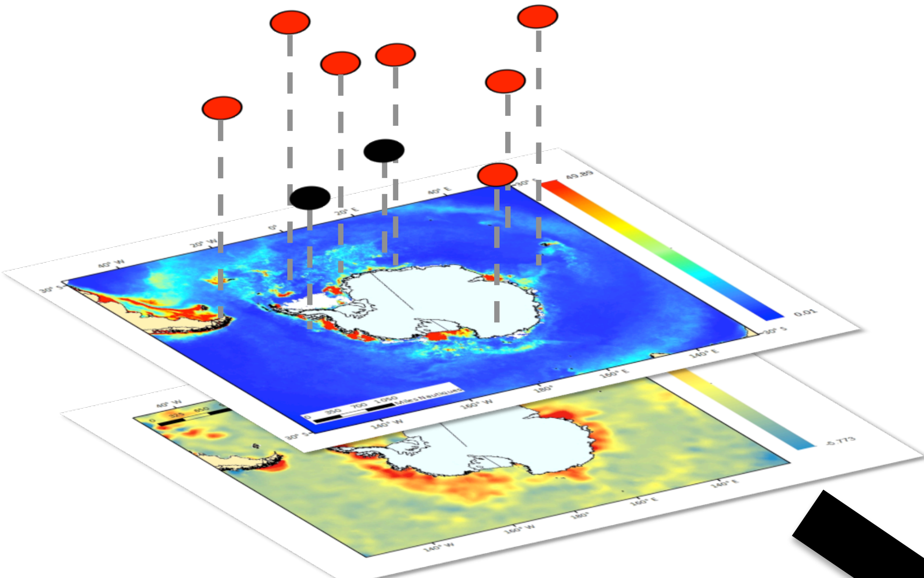
- Partial coverage of the information -> interpolation or not / missing values
- Full night in winter -> no satellite data
- Some algorithms cannot handle missing data !

→ See tomorrow's course
→ Need to interpolate the data
→ Be careful with the interpretation of your results

Questions on this part ???

SPECIES DISTRIBUTION MODELS principle

[presence + absence records]

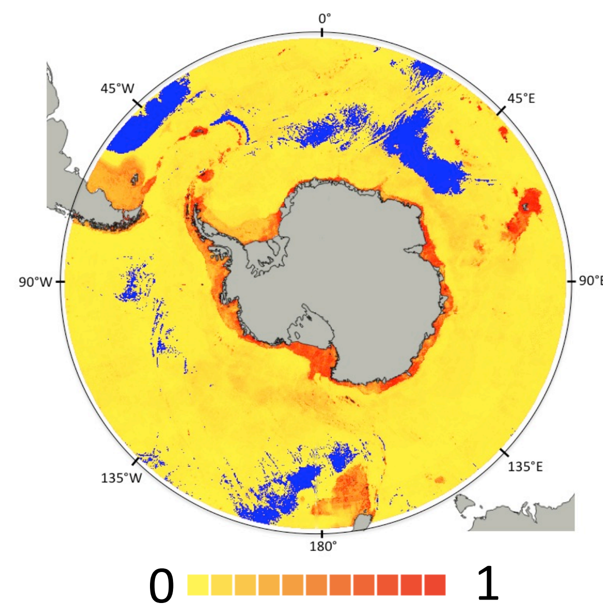


[set of environmental variables]

Presence / absence?	Layer 1 e.g. Depth	Layer 2 e.g. T°	Layer 3 e.g. Salinity
1	-351	0.2	32.4
1	-150	-1.4	32.1
0	-1053	-2	32.8
1	-3042	0.3	31.9
…	…	…	…

SDM

[Predicted distribution]



CALIBRATION: Occurrence dataset

SDM can be run with

- Abundance data (some algorithms)
- Presence- absence data
- Presence-only data

RK: Occurrence and environmental variables selection is the most difficult task for running SDMs !

Generate absence data

Generate absence data

- Experts dires
- Absences surveys (trawls)

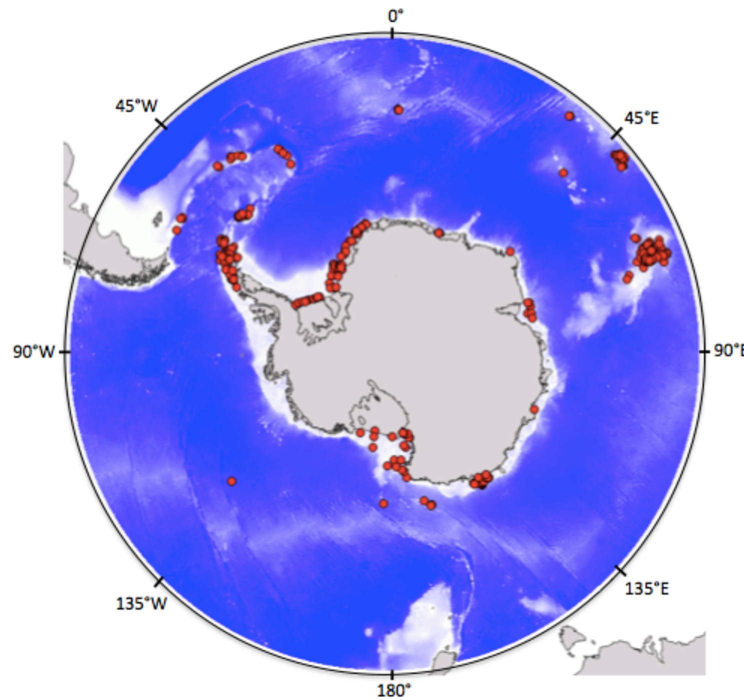
In broad-scale areas
-> difficult to rely on absence records
-> above all if historical compilation of several datasets

In the case of presence-only data, it is necessary to define the environment around which they are located

→ Sampling of background data in the area to calibrate the model

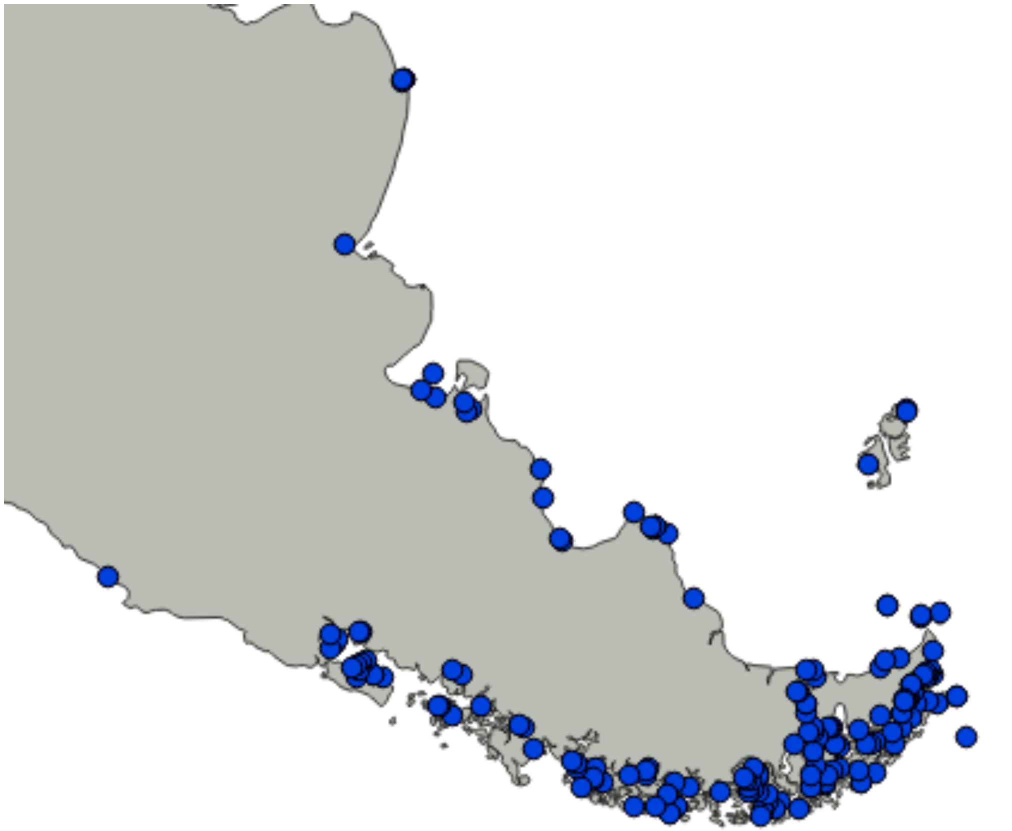
In broad scale areas, difficult to rely on absence data

Presence-only/background SDMs are less reliable and powerful than presence-absence models (Brotons et al. 2004, Wisz & Guisan 2009)

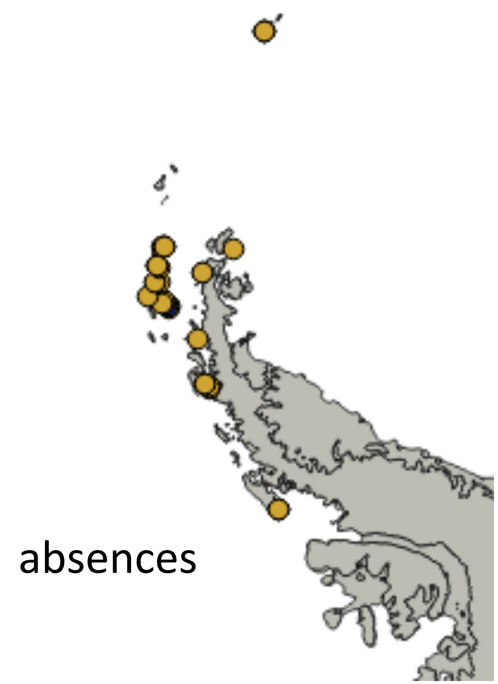


Occurrences of a sea star species in the SO

Presence records *Halicarcinus platanus*



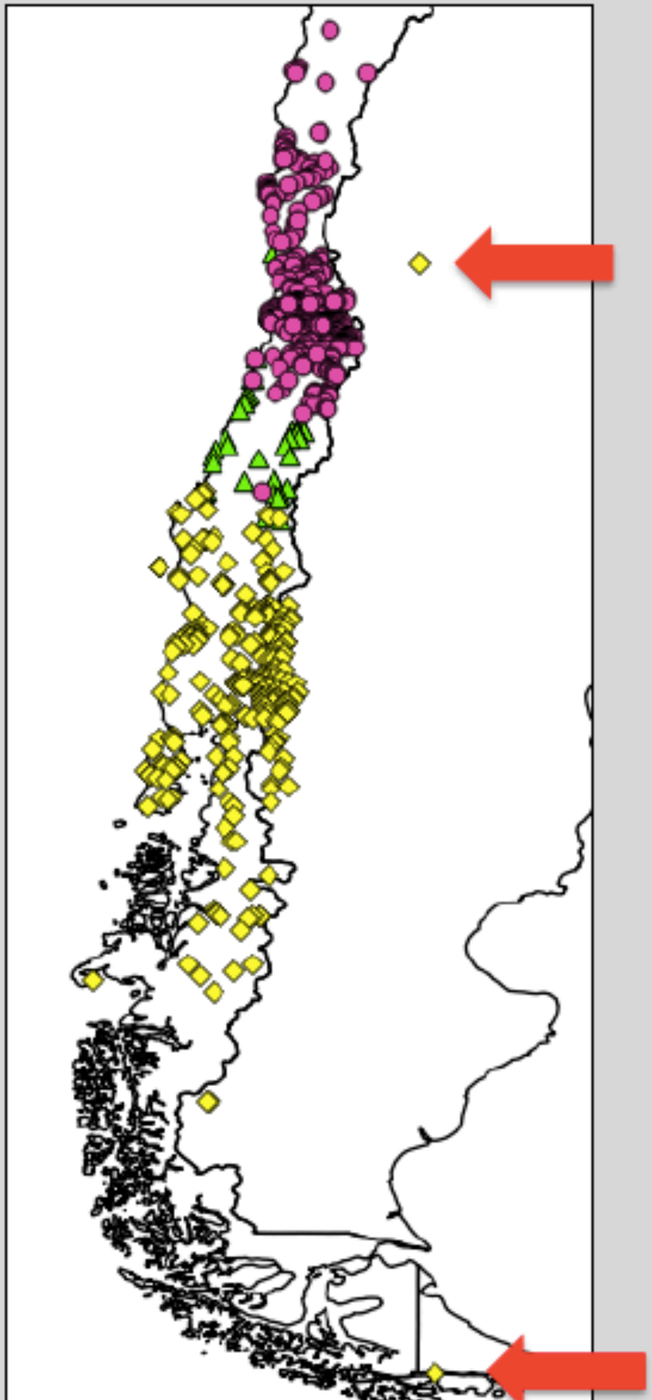
Presence records *Halicarcinus platanus*

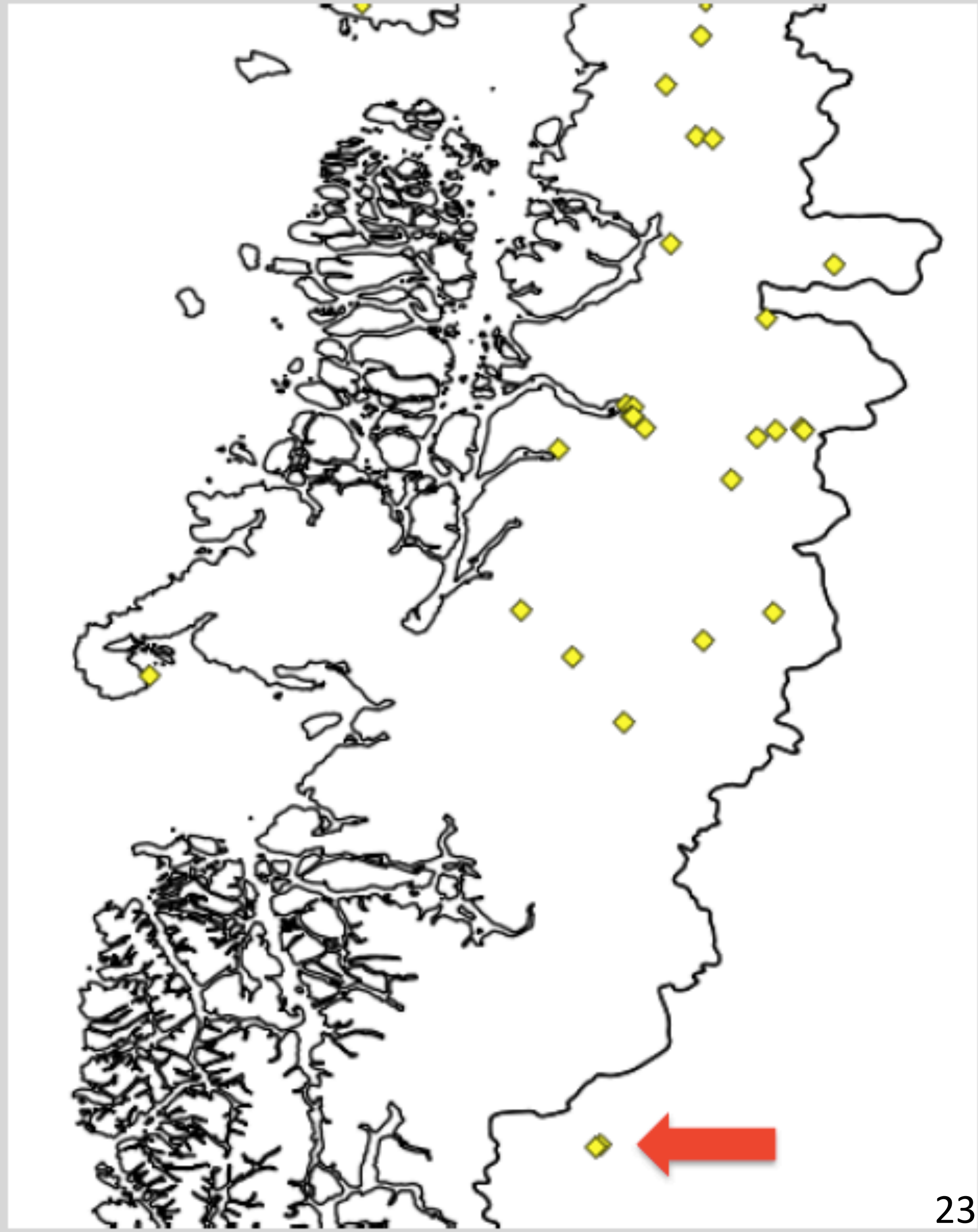


CRUCIAL TO EXPLORE YOUR DATASET

- Plot it, study each occurrence -> reliable or not ?
- Georeferencing errors ?

➔ Essential because it is responsible for strong bias in your SDM (you wrongly calibrate the initial conditions of your model, which conditions your species tolerates…)





PRACTICE !

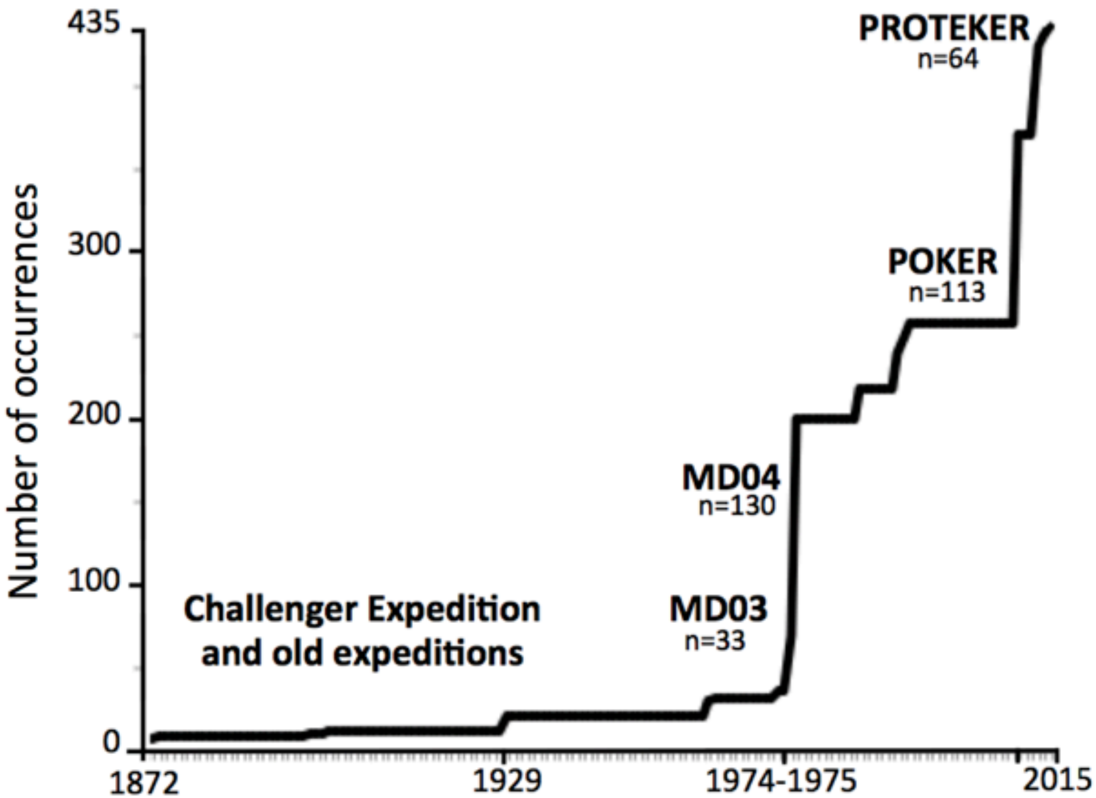
- Plot the occurrence records on the bathymetry layer

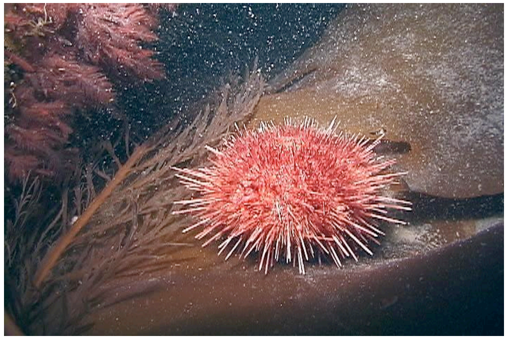
- In the provided example, do you have presence-absence data or presence-only data ? Where is it defined in the code?

SPATIAL AGGREGATION IN OCCURRENCE DATASETS

SPATIAL AGGREGATION IN OCCURRENCE DATASETS

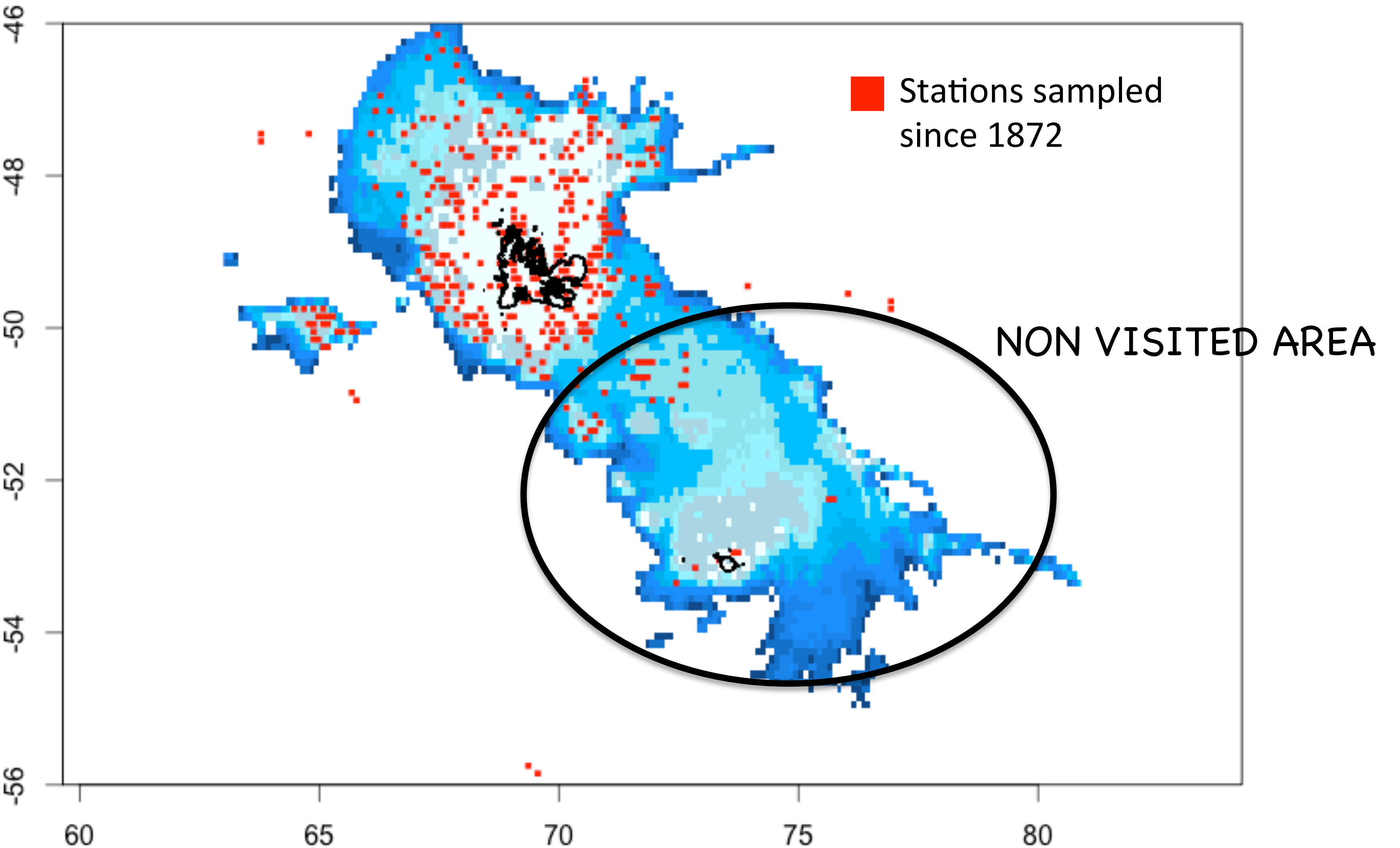
Historical collection
Compilation of datasets





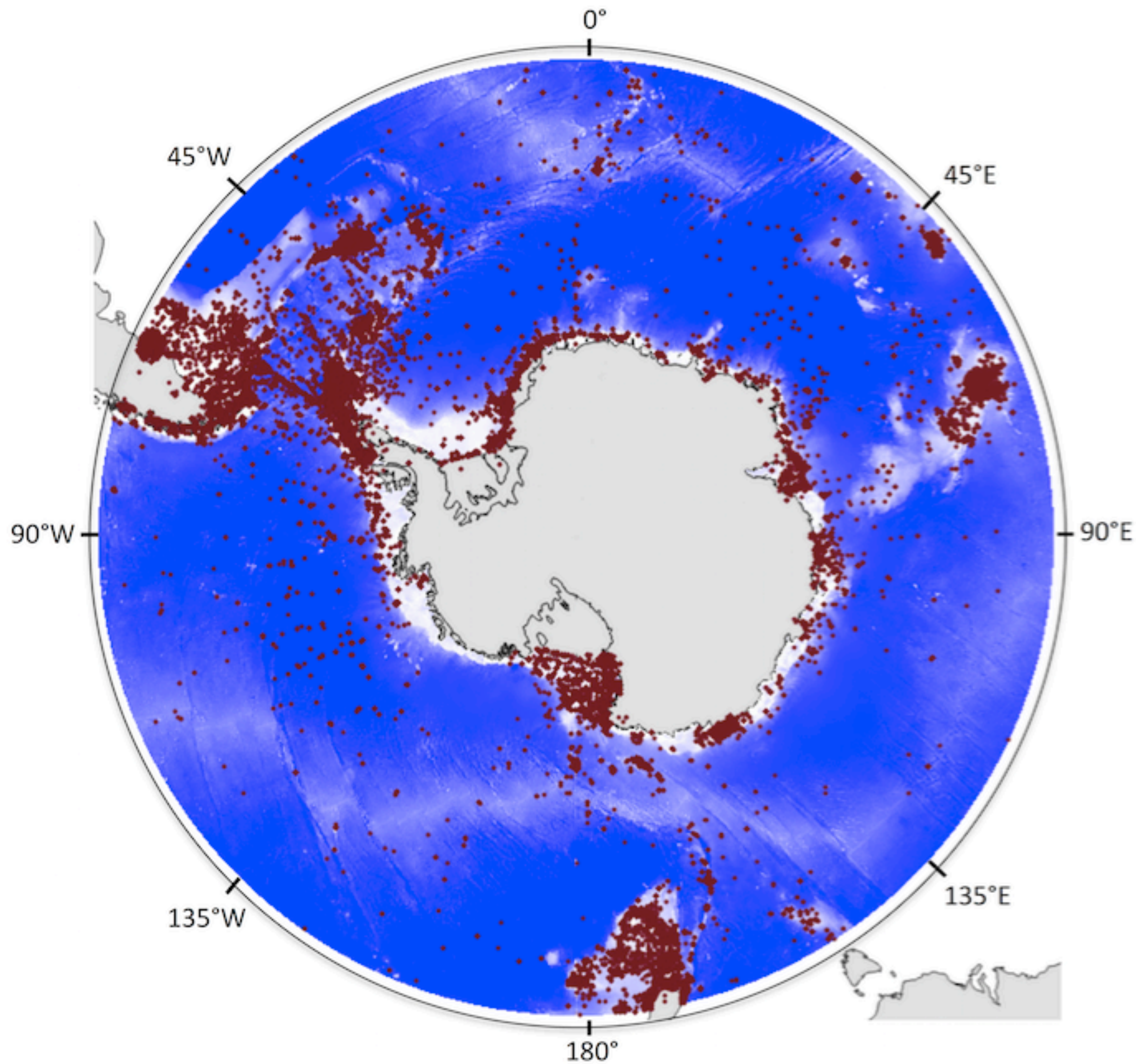
Sea urchins in Kerguelen
(Guillaumot et al. 2018)

SPATIAL AGGREGATION IN OCCURRENCE DATASETS



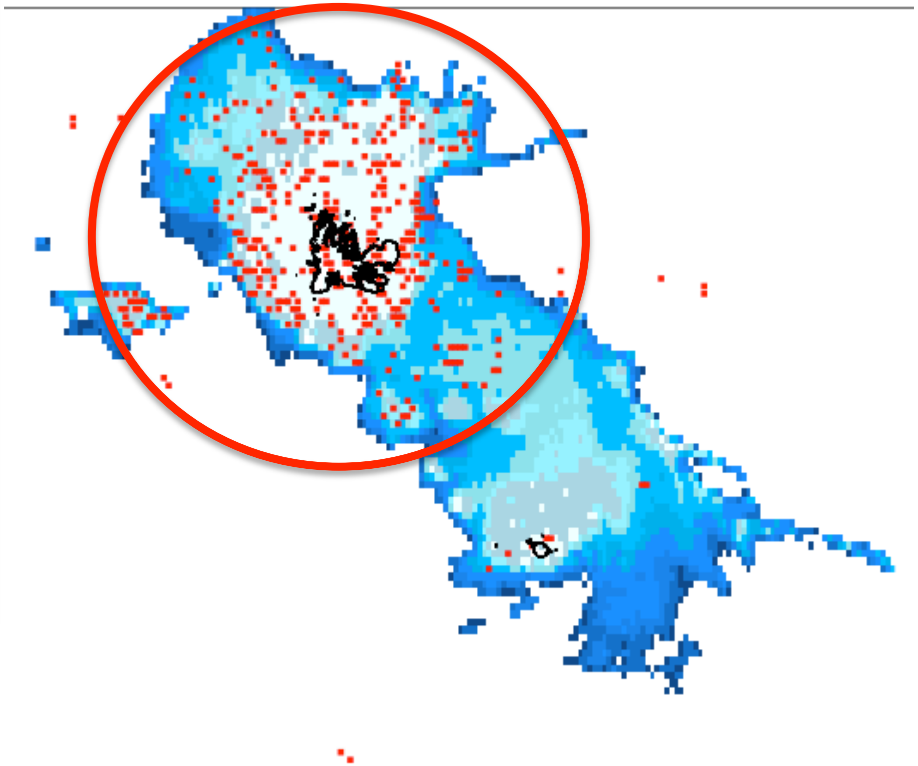
SPATIAL AGGREGATION IN OCCURRENCE DATASETS

All visited pixels for benthic sampling

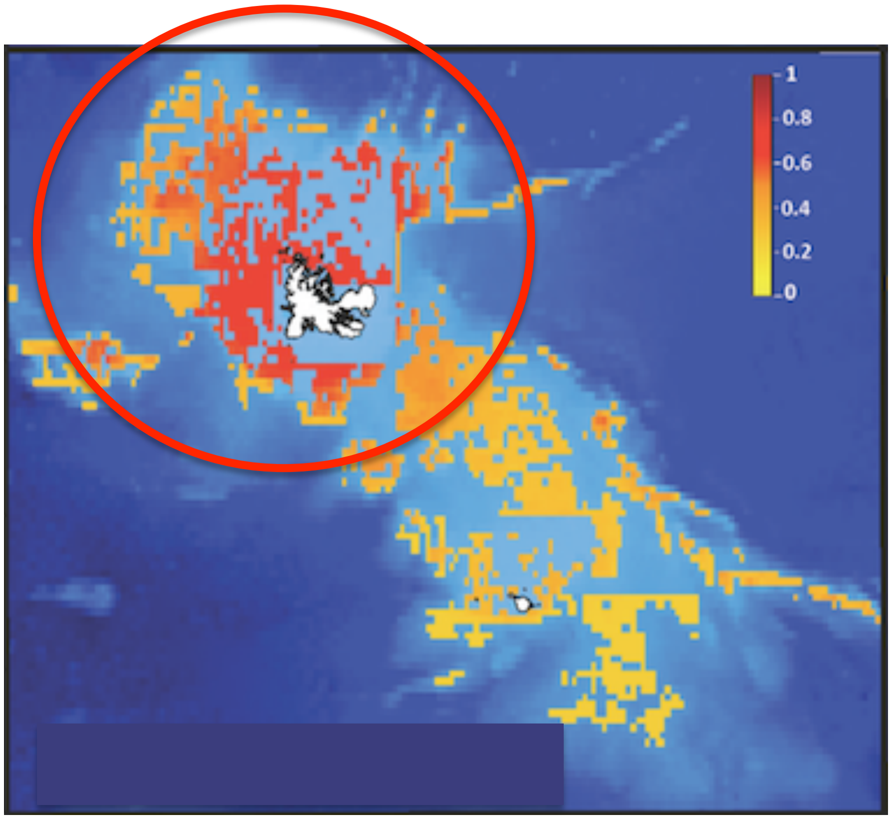


Guillaumot et al. (2019)

SPATIAL AGGREGATION IN OCCURRENCE DATASETS



Aggregated occurrence data



SDM predictions

Guillaumot et al. (2018)

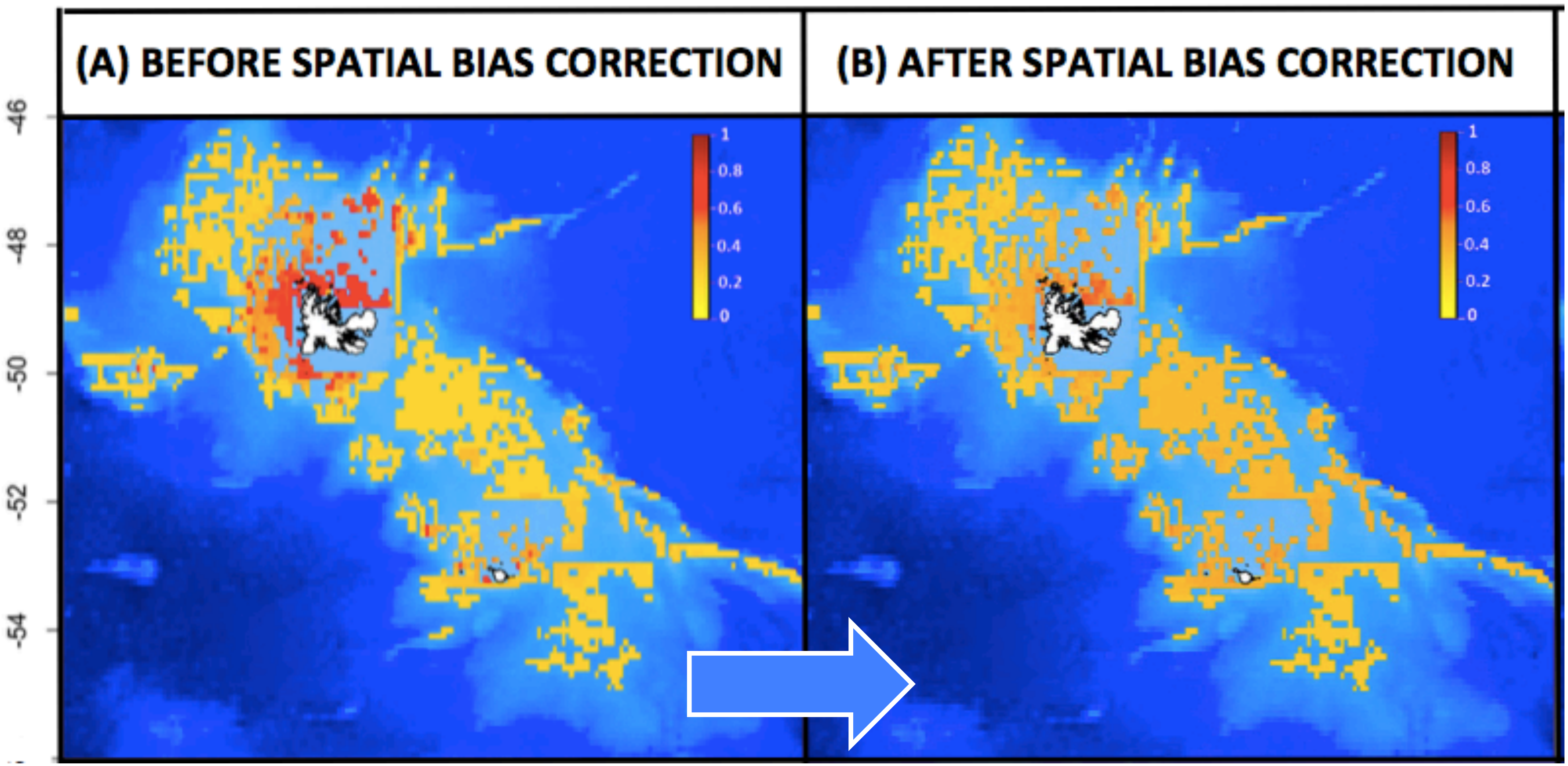
SPATIAL AGGREGATION IN OCCURRENCE DATASETS

SPATIAL AGGREGATION CAN BE MEASURED WITH

- Moran I index
- Variogram

-> both study the relationship between the value (predictions, variance in the result and the distance between points/pixels)

SPATIAL AGGREGATION IN OCCURRENCE DATASETS



APPLY CORRECTIONS !

Guillaumot et al. (2018)

CORRECTION FOR SPATIAL BIAS

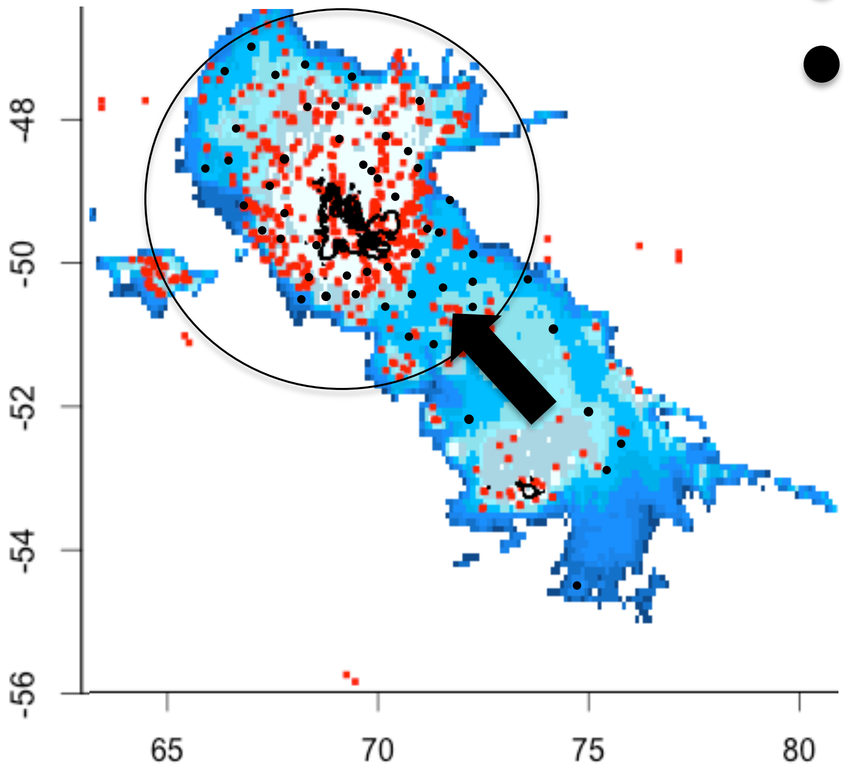
(1) Filter and sample just one occurrence per pixel
('pseudo-replication', Boria et al. 2014)

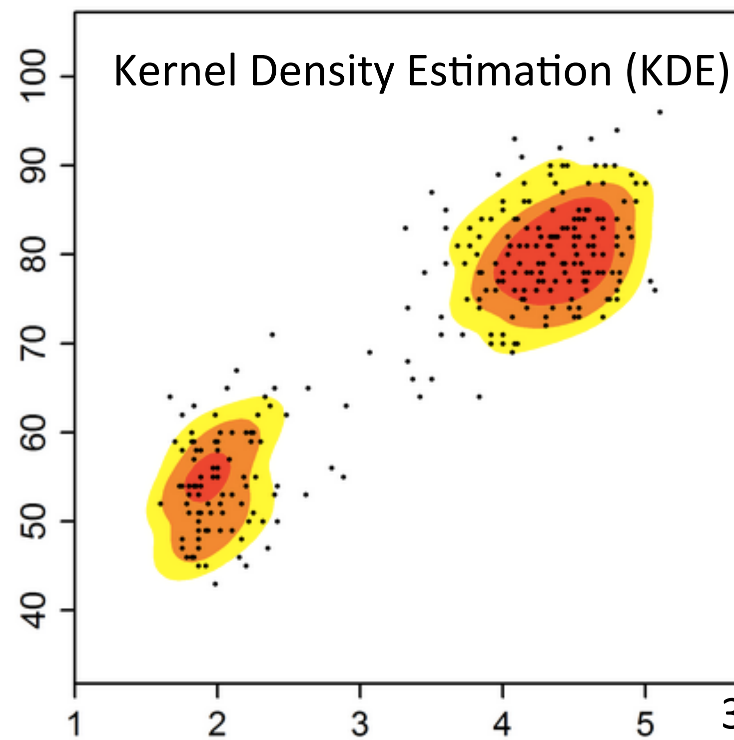
CORRECTION FOR SPATIAL BIAS

(2) Target-background approach: sample background data following the spatial pattern (Phillips et al. 2009)

Presence-only records

Background records





CORRECTION FOR SPATIAL BIAS

(2) Target-background approach: sample background data following the spatial pattern (Phillips et al. 2009)

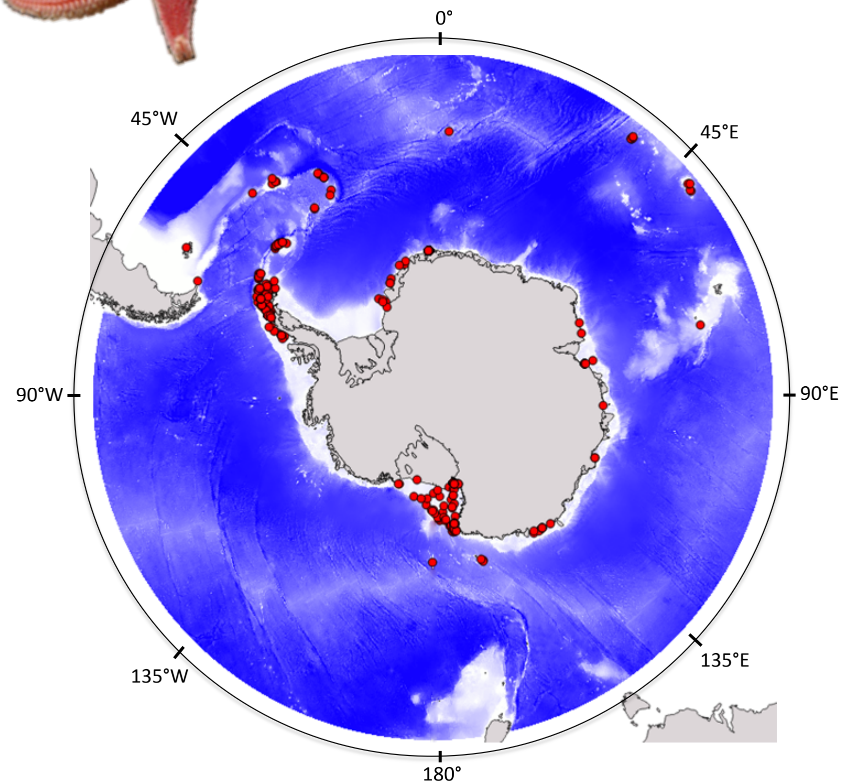
You can also :
- Generate disks around the presences and sample the background data inside these disks

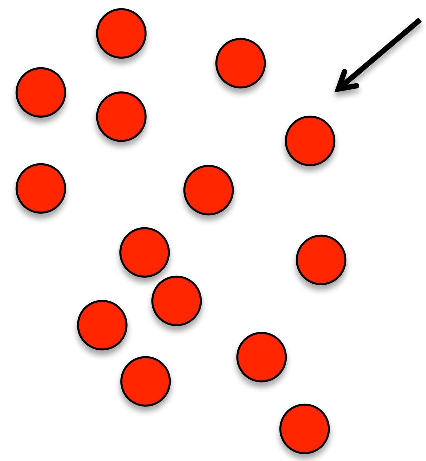
- Sample background data in areas where an associated species is present

More options in Phillips et al. (2009) and in the biomod2 R package

CONSEQUENCES OF DATA AGGREGATION ON MODEL VALIDATION

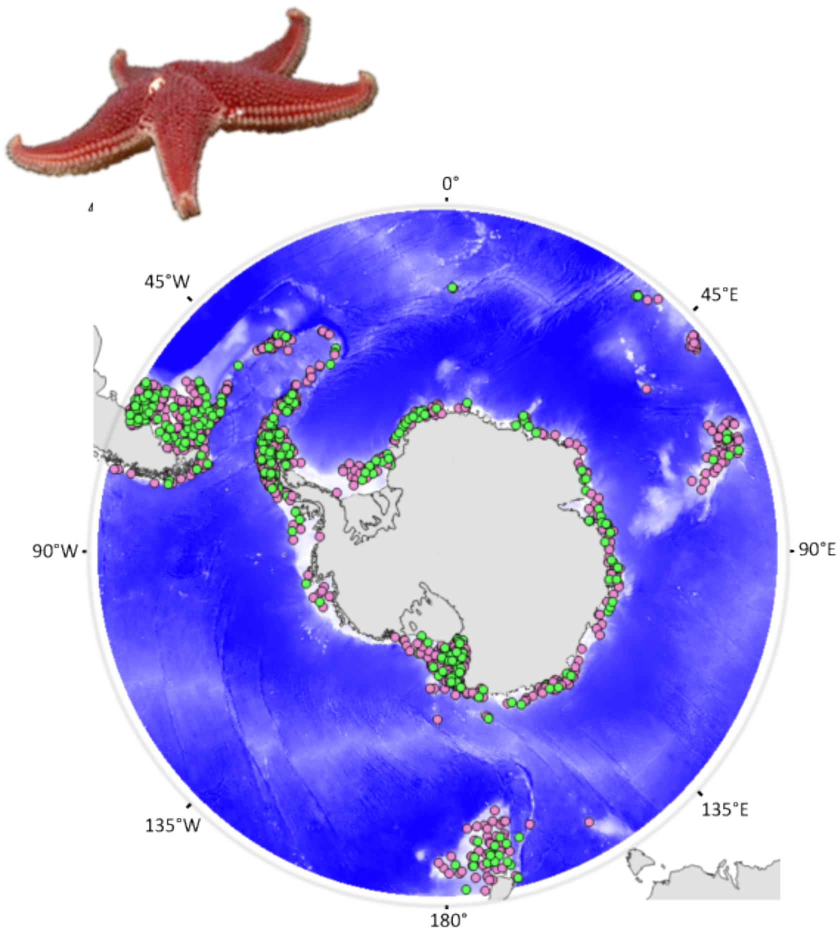
CONSEQUENCES OF DATA AGGREGATION ON MODEL VALIDATION

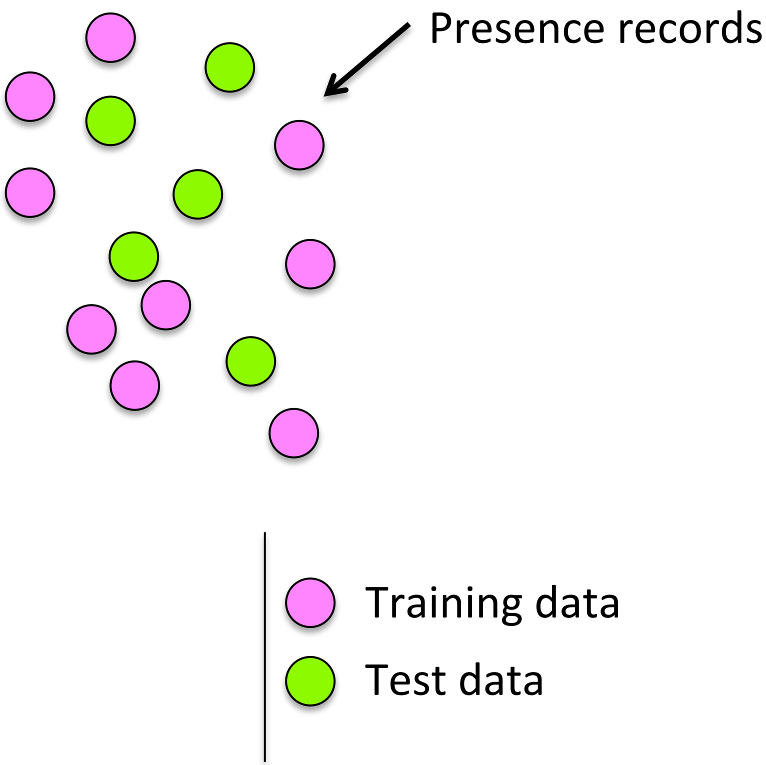




Guillaumot *et al. (2019)*

CONSEQUENCES OF DATA AGGREGATION ON MODEL VALIDATION



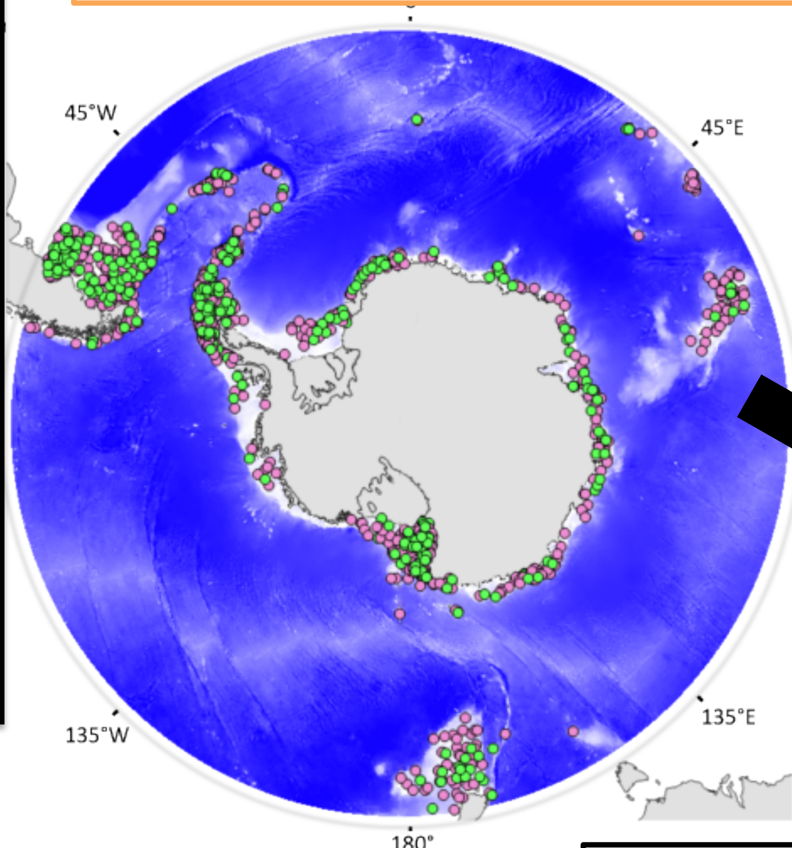


Guillaumot *et al. (2019)*

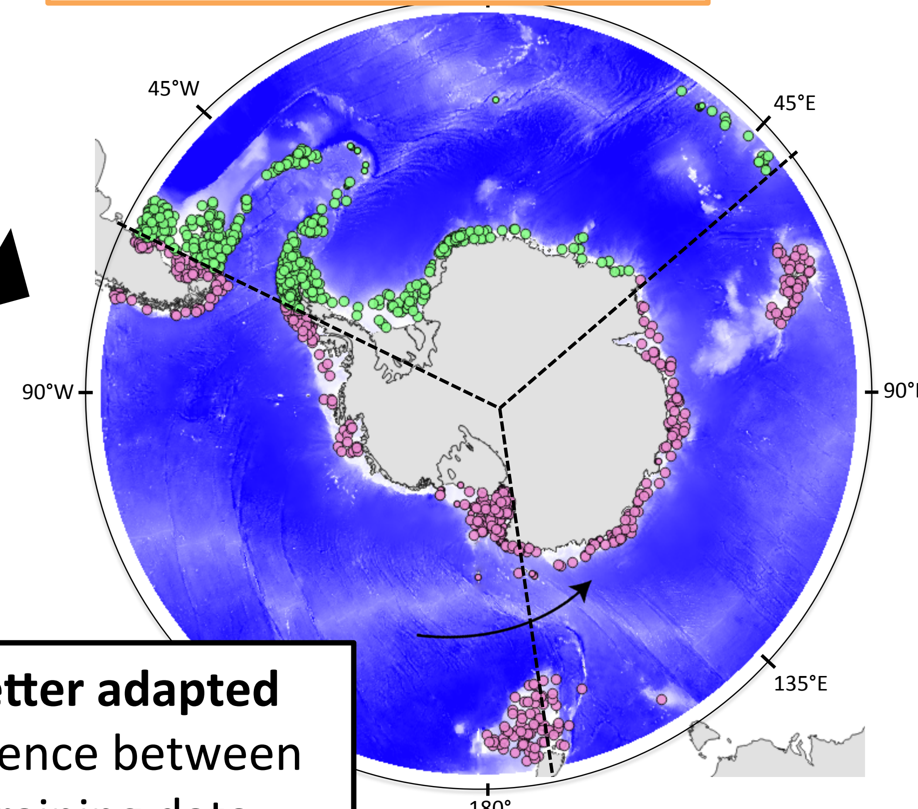
Training data

Test data

Standard cross-validation



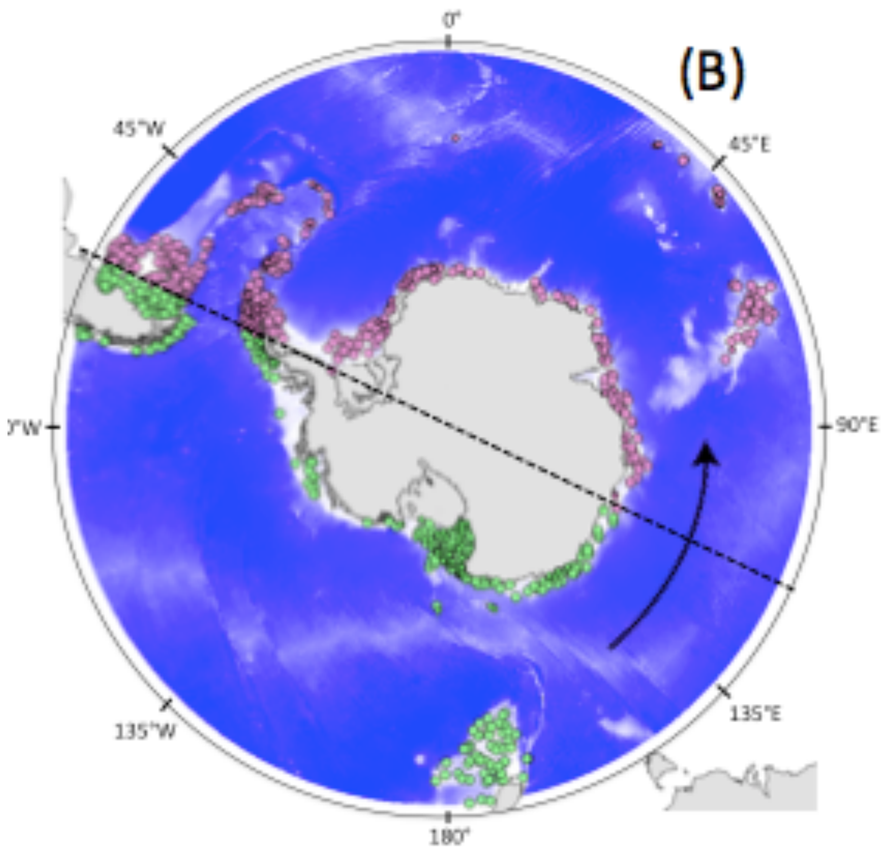
Spatial cross-validation

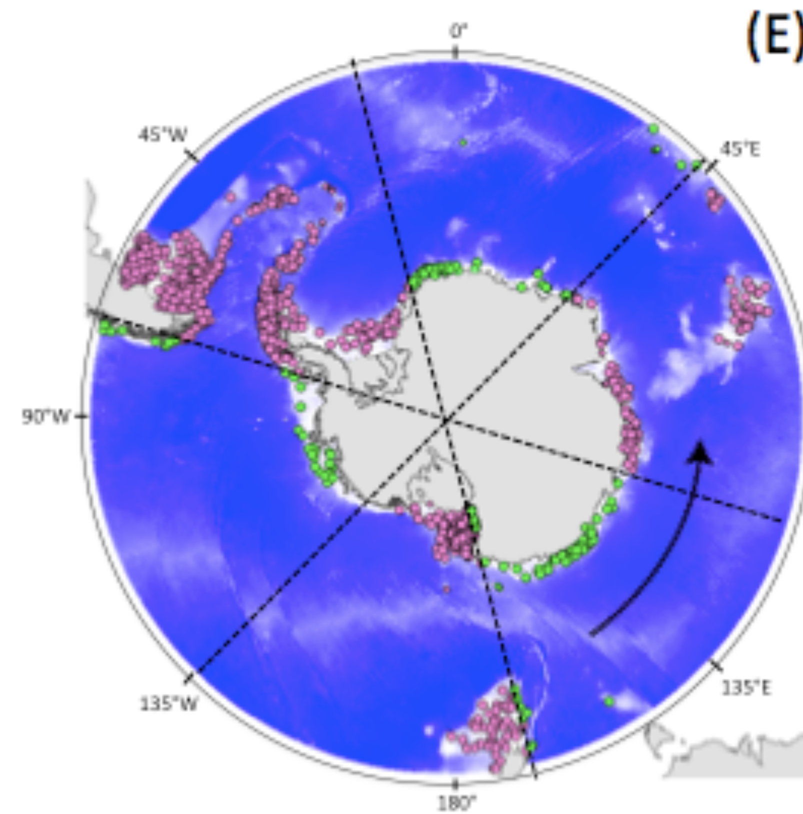


Method better adapted
-> independence between
test and training data

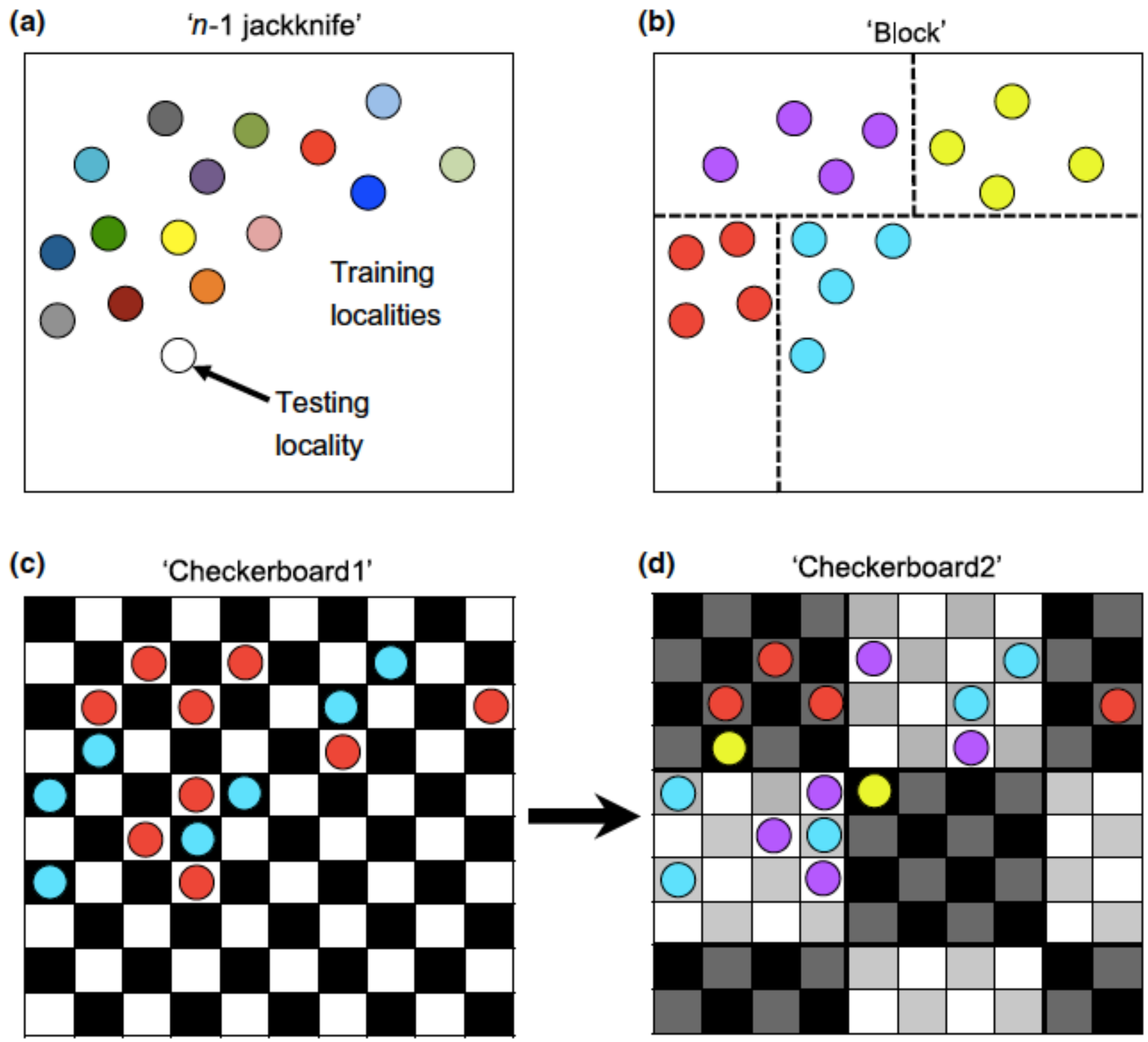
Guillaumot *et al. (2019)*

More cross-validation designs & comparisons in Guillaumot et al. (2019)





And generalised to all areas in Muscarella et al. (2014)

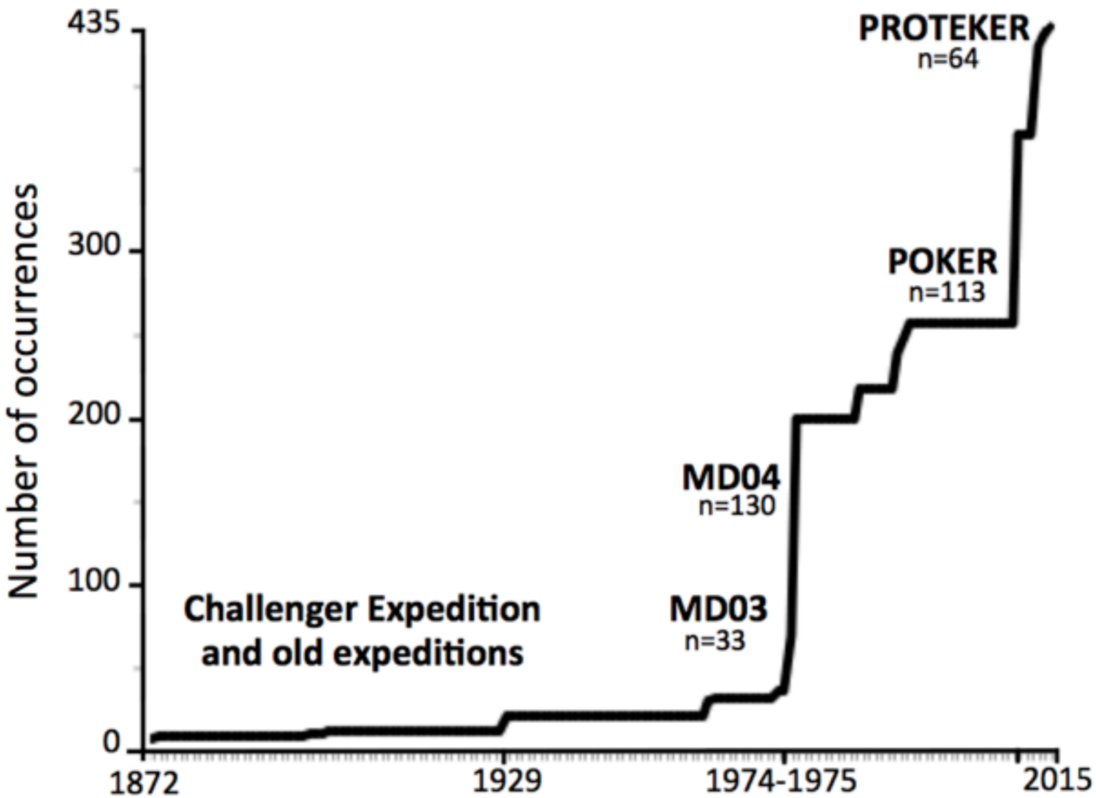


Little outline of this part ! =)

-> occurrence dataset used to calibrate the models
-> introduction of the use of background data
-> datasets spatially aggregated
=> why?
⇒How to measure it ?
⇒Consequences on SDM predictions
⇒Methods to correct it
⇒Consequences on model validation & corrections

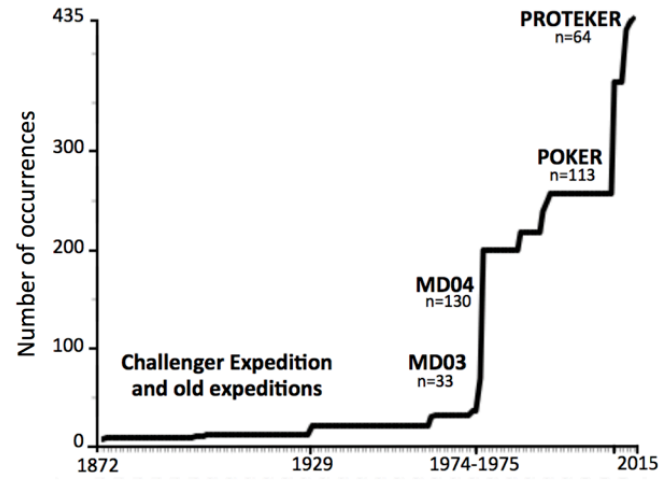
-> temporal biases
-> extrapolation

TEMPORAL BIASES



TEMPORAL BIASES

- **Old & recent datasets mixed together…**

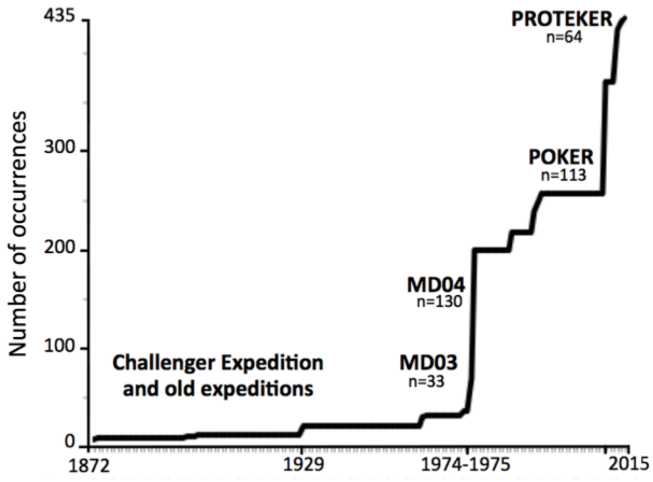


➔ Changes in species preferences to environmental conditions ?
➔ Population migrations ?
➔ Past environmental conditions have changed ? => species niche has changed??

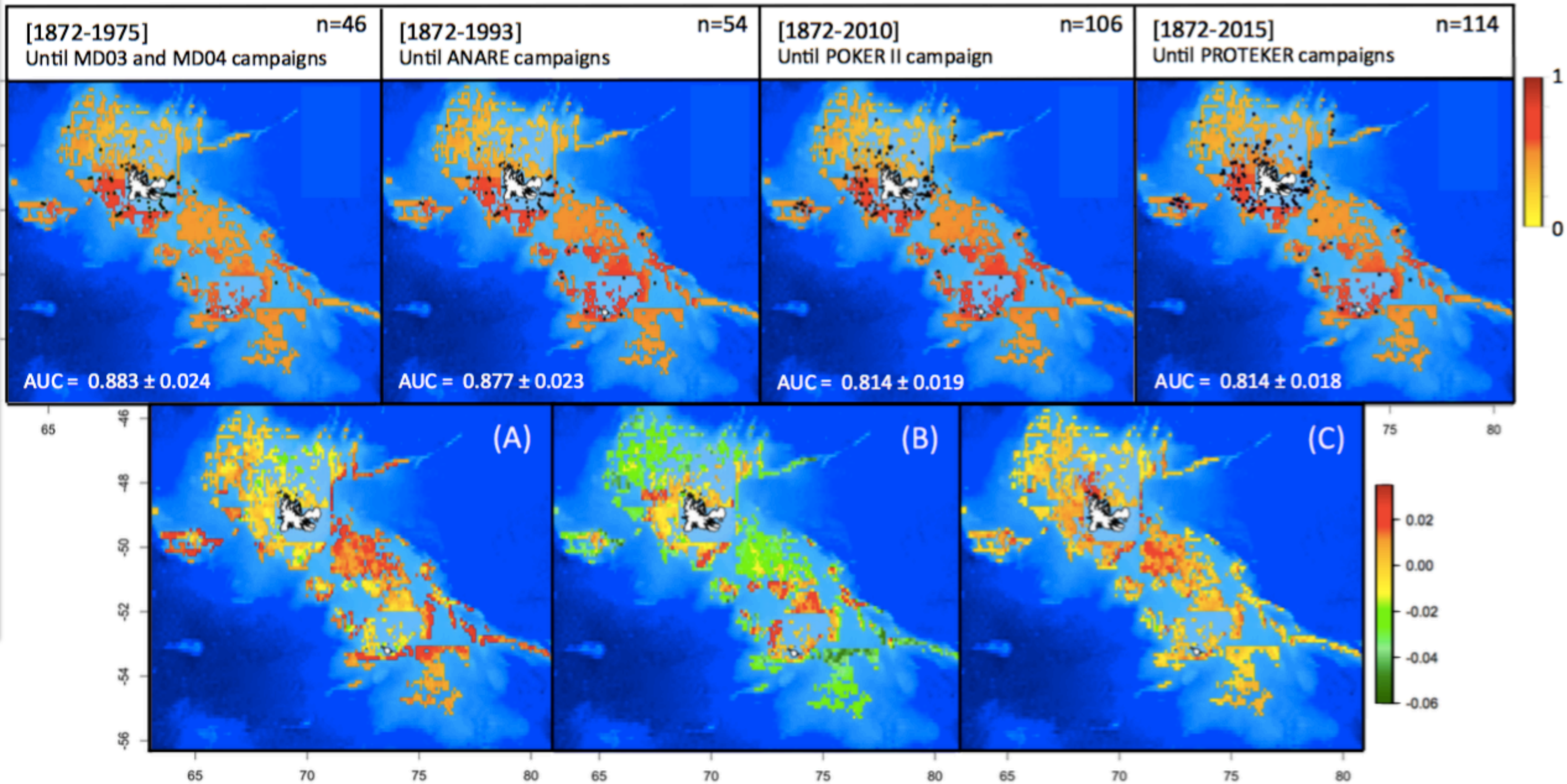
STRONG ASSUMPTIONS…BE CAREFUL WITH INTERPRETATION

TEMPORAL BIASES

- **Old & recent datasets mixed together…**
- **Biases linked to the number of occurrences and addition of new data**



CALIBRATION: Occurrence dataset

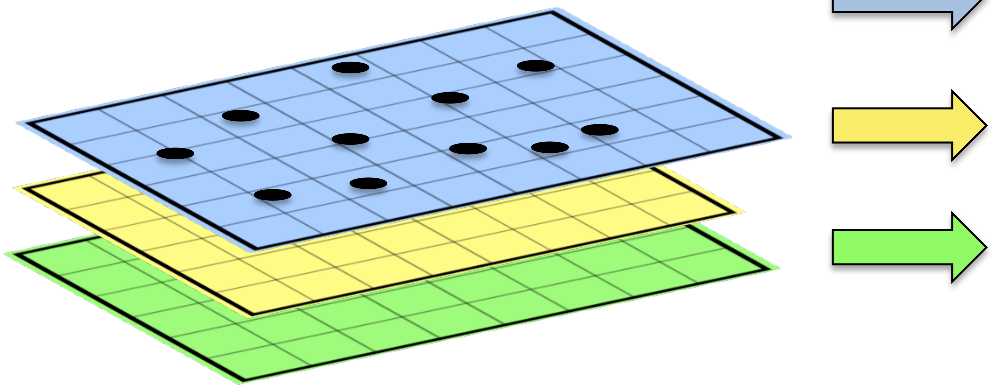


Guillaumot et al. (2018)

EXTRAPOLATION…

EXTRAPOLATION…

Presence records



Descriptor A interval [a1, a2]

Descriptor B interval [b1, b2]

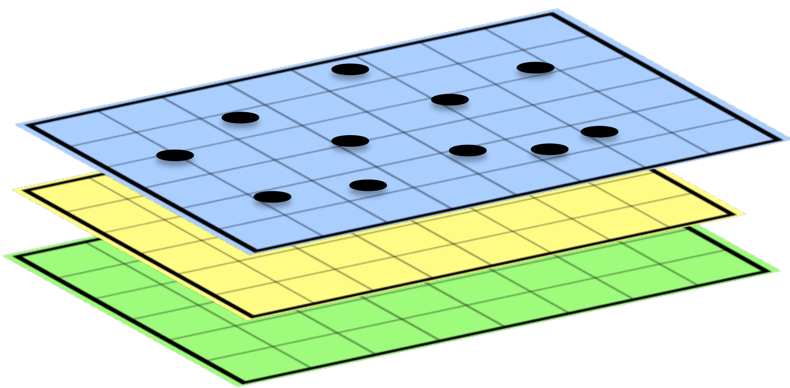
Descriptor C interval [c1, c2]

…

Guillaumot *et al. (in prep.)*

EXTRAPOLATION…

Presence records



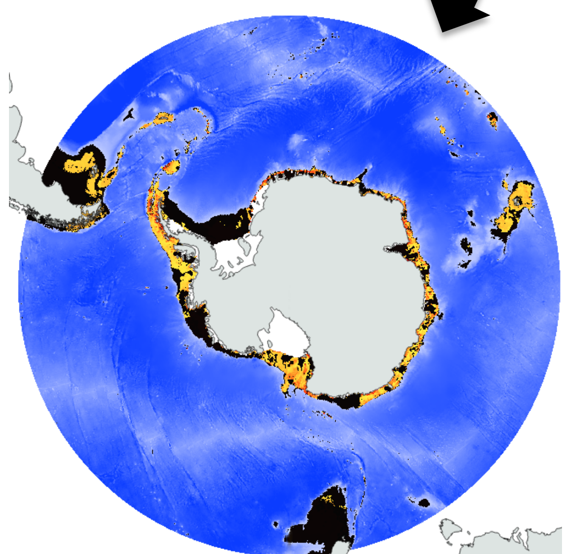
Descriptor A interval [a1, a2]

Descriptor B interval [b1, b2]

Descriptor C interval [c1, c2]

…

MESS: Multivariate Environmental Similarity Surface (Elith et al. 2010)



More than 60% of the area: extrapolation !
➔To take into consideration

Guillaumot *et al. (in prep.)*

Questions ???

EXTRA PRACTICE

Have you spotted in your code where you can change the layer of environmental variables on which you will project your model ? If you want for example to make a future projection ?