Species distribution models (SDM) – from a computer perspective

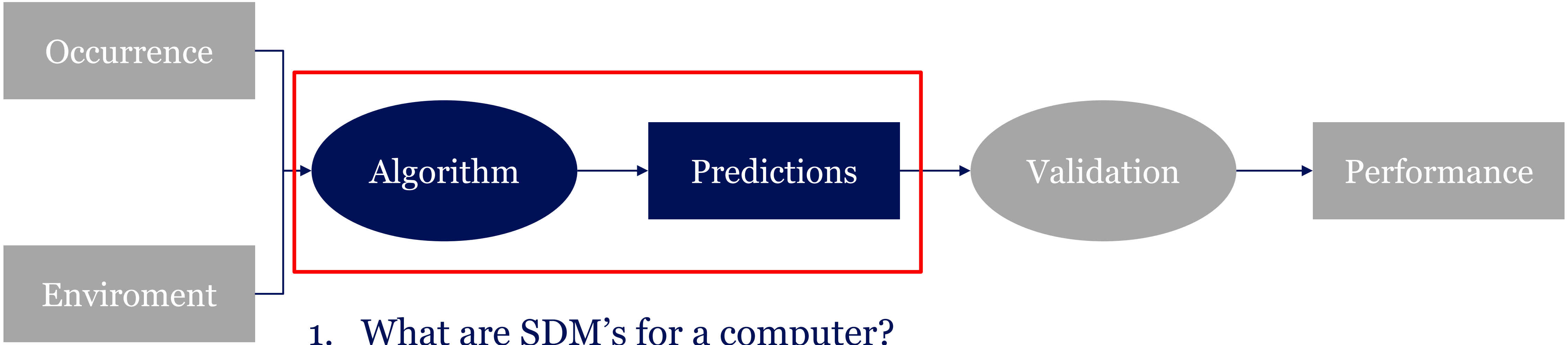
Nuno César de Sá[1],

[1] Institute of Environmental Sciences (CML)

Systematics & Biodiversity, 2020

Data Science Research Programme

Remember where we are:



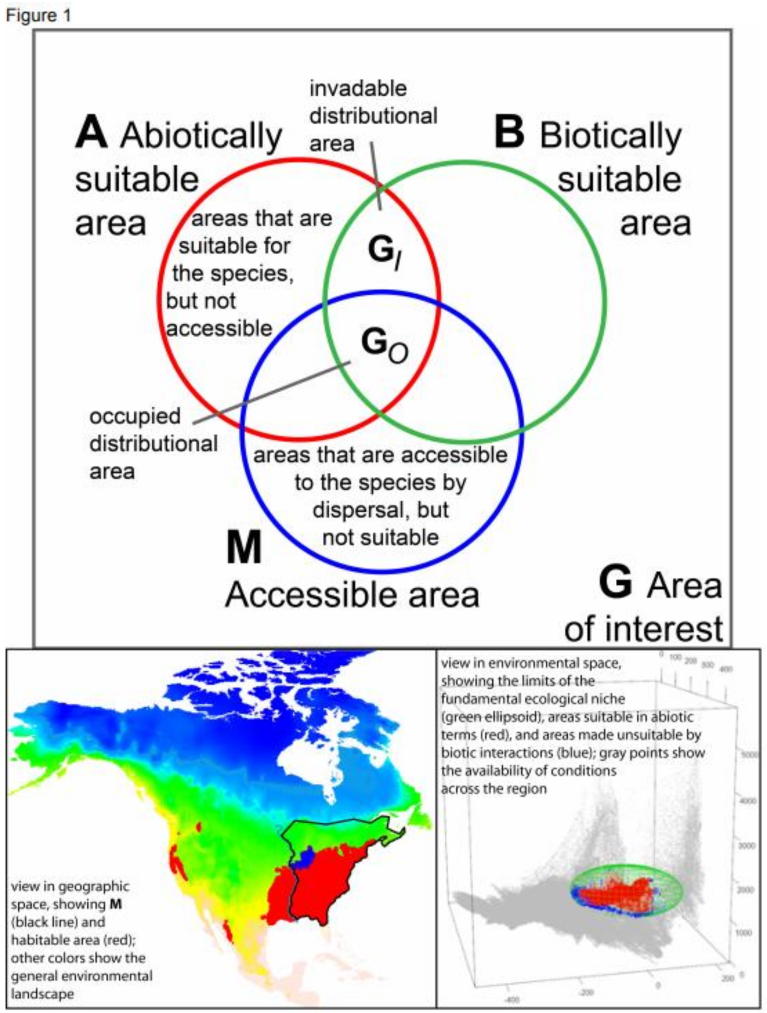
1. What are SDM's for a computer?
2. How does it understand the niche?
3. How does it predict?

SDM – from a computer perspective

- Defining Ecological Niches: Hutchinson
  - n-dimensional **Environmental** space (“hyperspace“)
  - **Sóberon & Peterson**: Abiotic, Biotic, Movement
  - Geographic space (where they are) vs Niche space (why they are there)

- From a more formal point of view:

$$\boldsymbol{EN}(species) = \boldsymbol{f}(Abiotic, Biotic, Movement)$$



From: https://doi.org/10.1515/eje-2015-0014

SDM – from a computer perspective

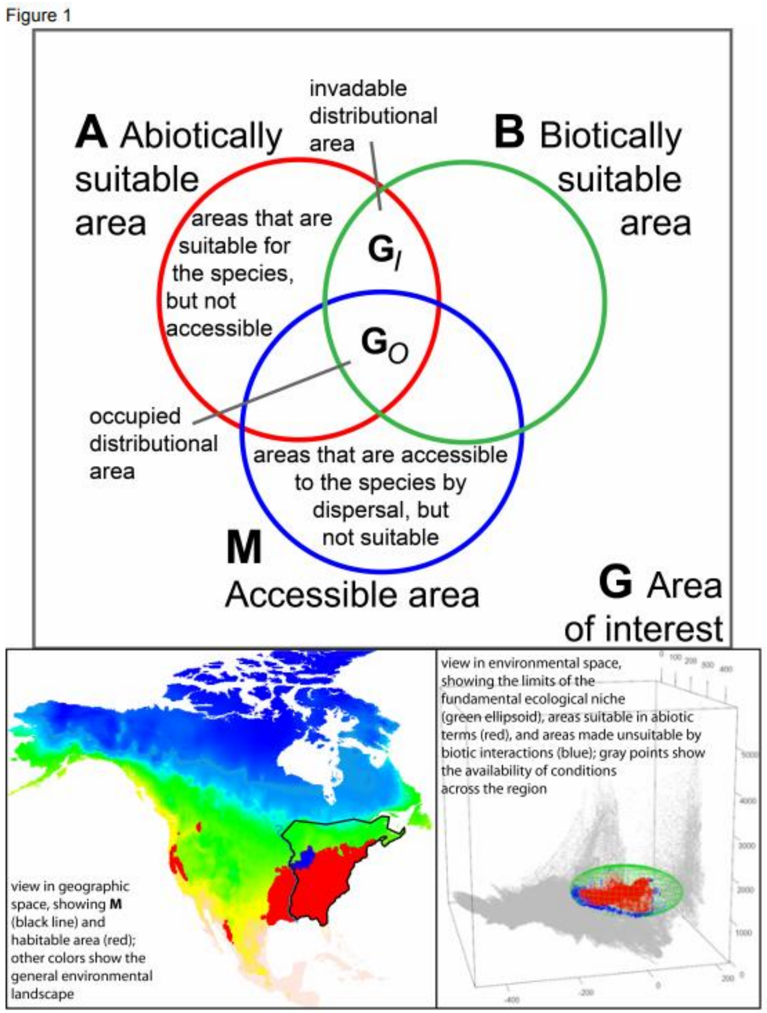
- Defining Ecological Niches: Hutchinson
  - n-dimensional **Environmental** space (“hyperspace“)
  - **Sóberon & Peterson**: Abiotic, Biotic, Movement
  - Geographic space (where they are) vs Niche space (why they are there)

- From a more formal point of view:

$$\boldsymbol{EN}(species) = \boldsymbol{f}(Abiotic, Biotic, Movement)$$

- But.. There is no “Ecological niche value” (EN)
  - And also, no “f(A,B,M)”

- Two **competing approaches**:
  - Probabilistic models “vs” Mechanistic models

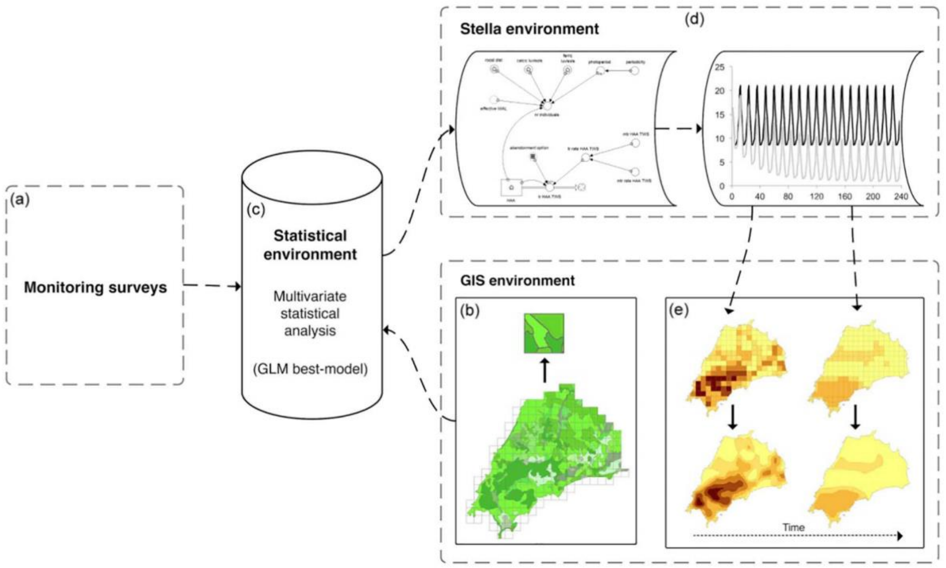


From: https://doi.org/10.1515/eje-2015-0014

SDM – from a computer perspective

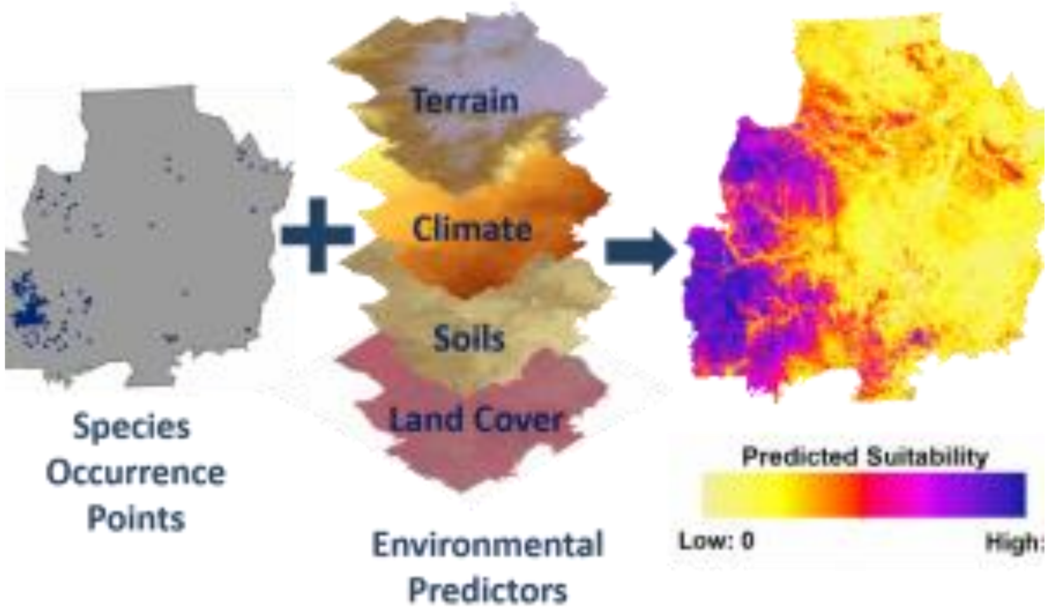
- Two "*competing*" approaches: Focus will be **on** ***Probabilistic models***

Mechanistic models



From: https://doi.org/10.1016/j.biocon.2017.04.013

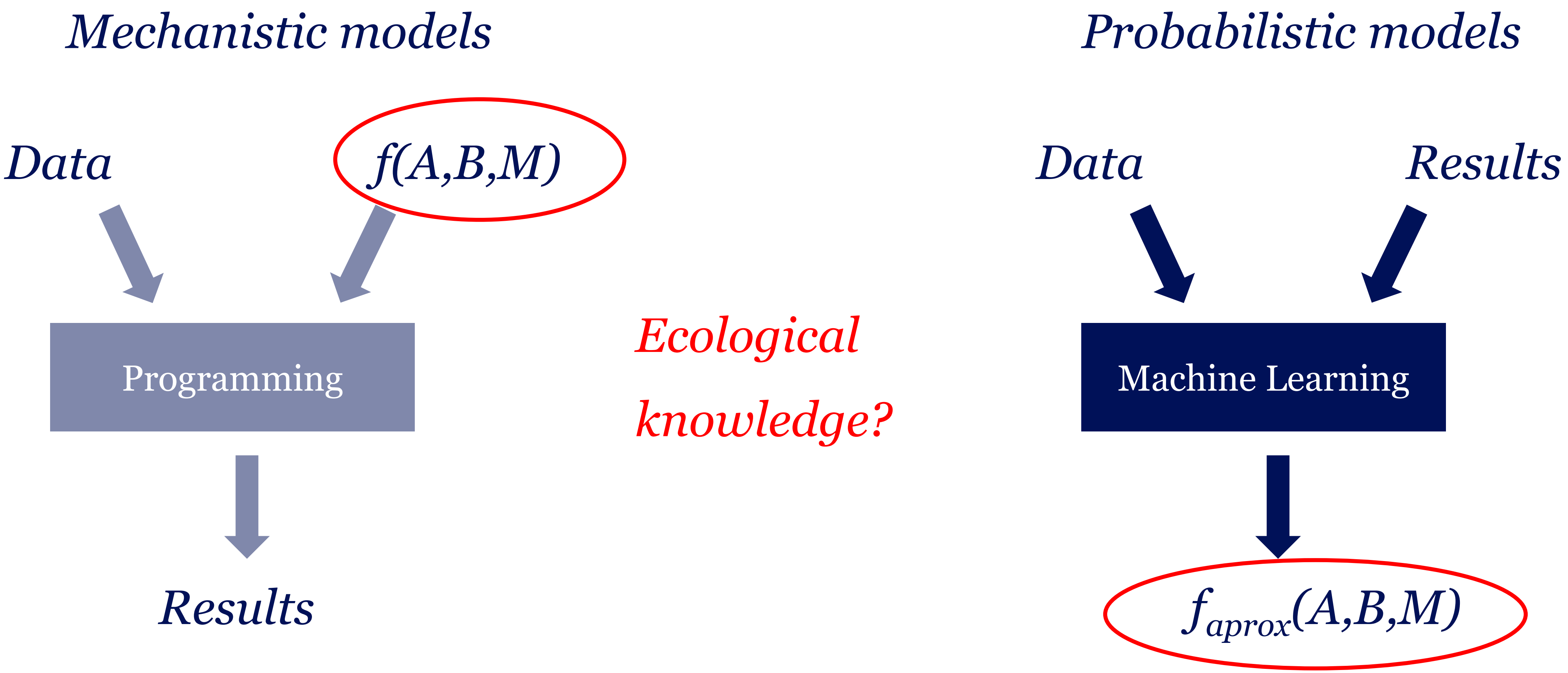
Probabilistic models



From: https://www.natureserve.org/conservation-tools/habitat-suitability-modeling

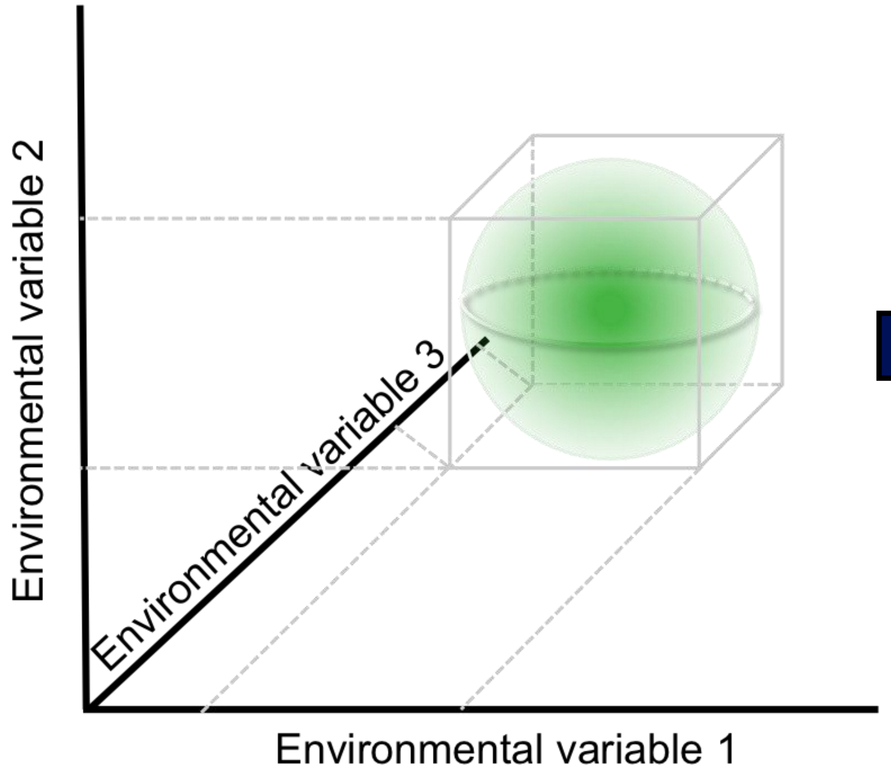
SDM – from a computer perspective

- Two "*competing*" approaches: Focus will be **on** ***Probabilistic models***



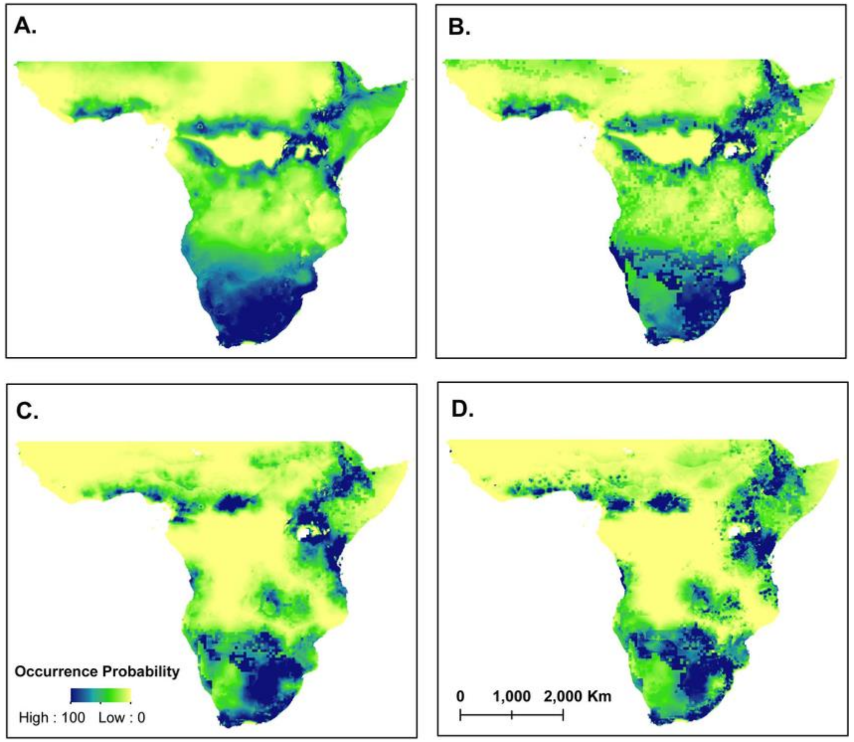
SDM – from a computer perspective

Environmental space – EN



$f(A, B, M)$

Geographic space - EN projected by f(A,B,M)



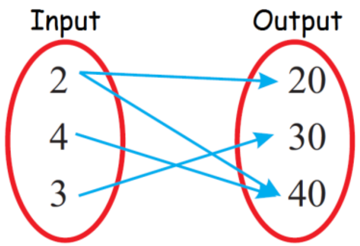
From: https://doi.org/10.4404/hystrix-27.1-11678

If the EN was a simple n-dimentional surface… It would be easy….

SDM – from a computer perspective

- Machine learning:
  - At the core: learning a mapping function
  - Mapping function: “arbitrary” function Translates/transforms input values from one domain to another

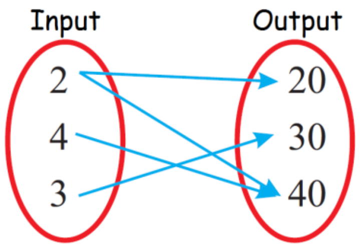
F(input) = Output



SDM – from a computer perspective

- Machine learning:
  - At the core: learning a mapping function
  - Mapping function: “arbitrary” function Translates/transforms input values from one domain to another

F(input) = Output

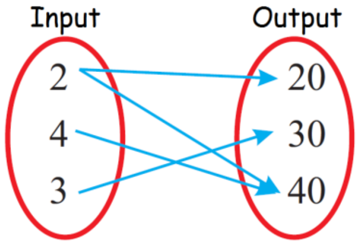


- AKA: Ecological niche models; Habitat/Suitability modelling; Correlative models, Range mapping; etc
  - **For ML people**: Supervised learning
  - Commonly maps: N-dimensions onto 1D probability
  - PS: More common problem in supervised learning is to map N:M dimensions

SDM – from a computer perspective

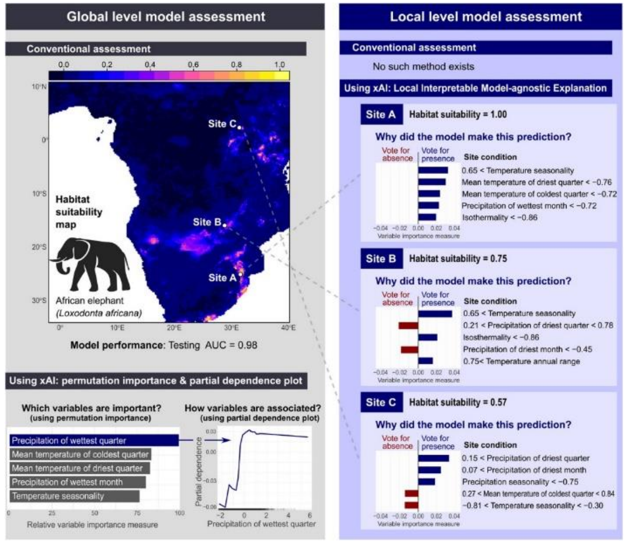
- Machine learning:
  - At the core: learning a mapping function
  - Mapping function: "arbitrary" function Translates/transforms input values from one domain to another

F(input) = Output



- AKA: Ecological niche models; Habitat/Suitability modelling; Correlative models, Range mapping; etc
  - **For ML people**: Supervised learning
  - Commonly maps: N-dimensions onto 1D probability
  - PS: More common problem in supervised learning is to map N:M dimensions

Explainable AI is arriving! (PS: R code available!)

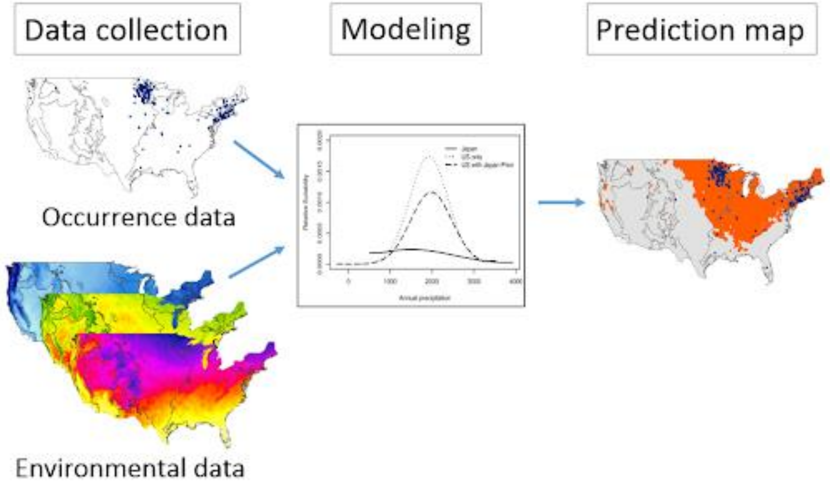


From: https://doi.org/10.1111/ecog.05360

- Risks: (Timnit Gebru)
  - Bad data -> bad model
  - Biases in data -> biases in the model
  - "black-box" (or maybe not… Increasingly debatable!)

SDM – from a computer perspective

- Having occurrence data and environmental data
  - we want to find the P(Species) being present given the environment

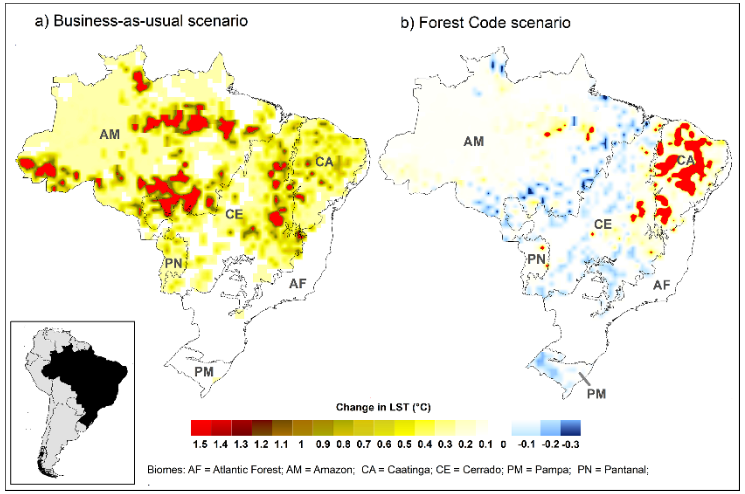


- Formalizing from a Bayesian perspective:

$$\boldsymbol{P}(Species|Env) = \frac{\boldsymbol{P}(Env|Species) \cdot \boldsymbol{P}(Species)}{\boldsymbol{P}(Env)}$$

  - Implies that: P(Species), P(Env) are independent

But, are they?

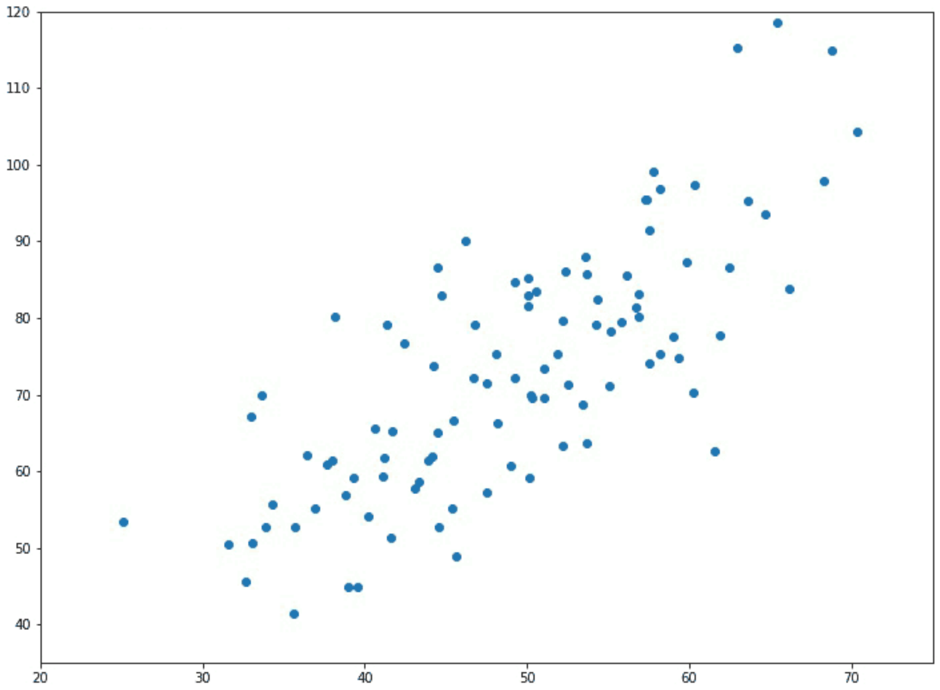


From: https://doi.org/10.1371/journal.pone.0213368

- Objective: learn a function that maps environmental variables onto a 1D probability of occurrence

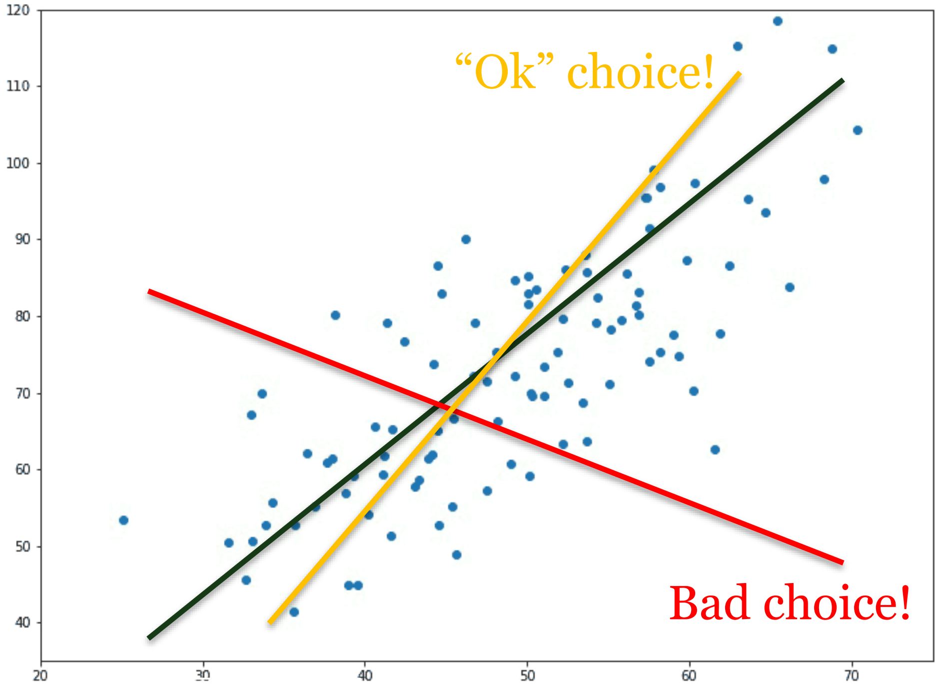
SDM – from a computer perspective

- ***How do machines actually, learn?***
- Consider you are asked to find the best linear model that fits a set of data
  - And you skipped the linear algebra class
  - But you know that: Y =mX + C is a linear equation
  - And m is the "slope" and C is the "bias"



SDM – from a computer perspective

- ***How do machines actually, learn?***
- Consider you are asked to find the best linear model that fits a set of data
  - <u>And you skipped the linear algebra class</u>
  - But you know that: Y =mX + C is a linear equation
  - And m is the “slope” and C is the “bias”
- A smart solution would be to draw a line that “more or less fits” the data
  - Solve the system for m and C: $\begin{cases} Y_x = mX + C \\ Y_0 = C \end{cases}$



Good choice!

You’ll only sucessfully “cheat” if you <u>choose a good enough linear model</u>.. AKA trial and error

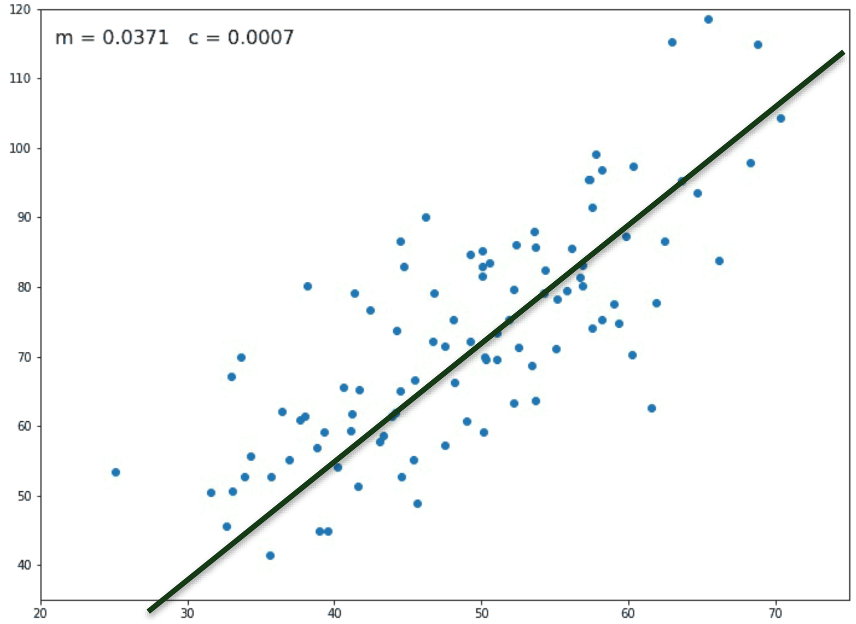
SDM – from a computer perspective

- Machine's learn by "trial & error".
  - *Computers excel at quickly trying everything*

- They find the best parameters so that:
  - Some measurement of error is minimized
  - E.g: || Y(parameters) – E(Y) || ≈ O
  - This is often called the "objective/cost/loss" :

$$Cost(m, C) = \left|Y_{(m,C)} - E(Y)\right|$$

- Finding the best set of parameters <u>becomes an optimization problem</u>
  - As in, finding the optimal solution to a problem



Good choice!

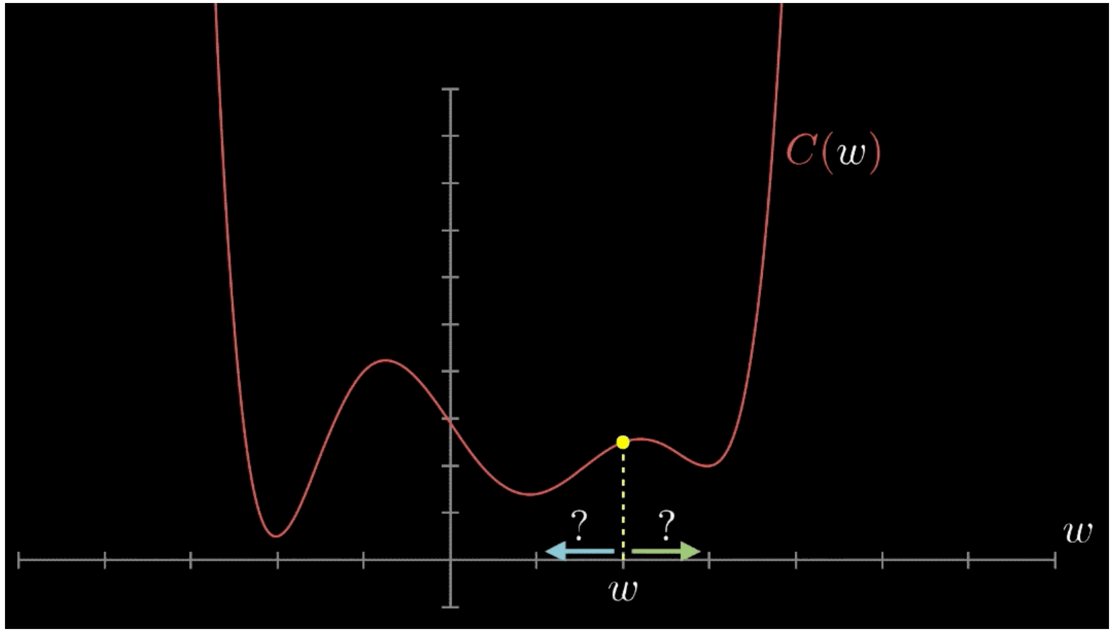
The model "unit"

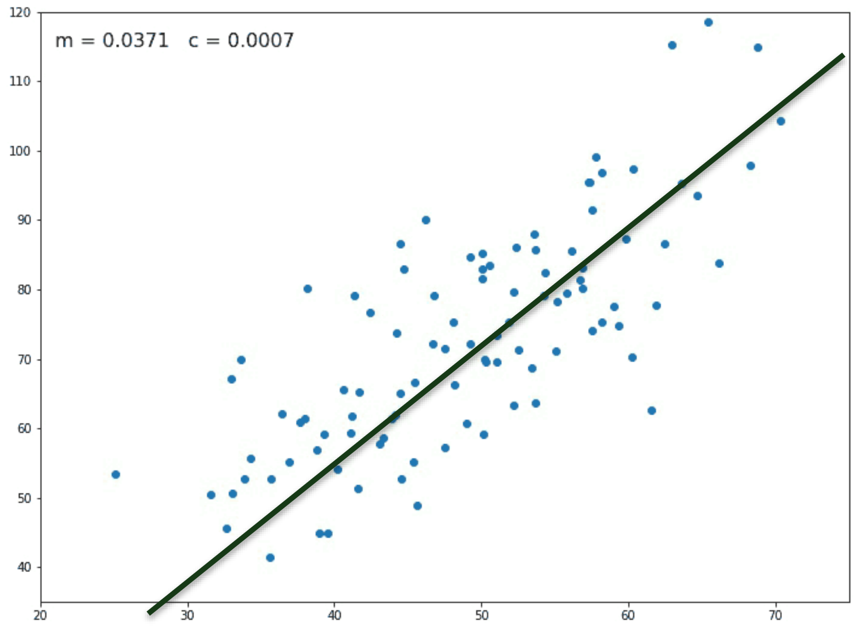
$$\boxed{Y = m\,X + C}$$

Parameters

SDM – from a computer perspective

3blue1brown: ["But what is a neural network?" - Youtube playlist]





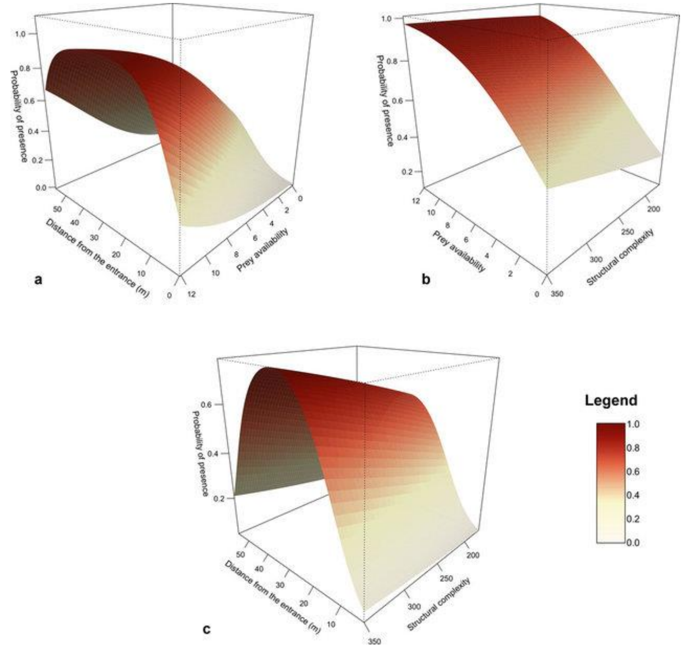
- If we would just try different random sets of parameters we might never find the best solution
- So computers generally <u>start random but then in each iteration focus on the sets that have the best results</u>

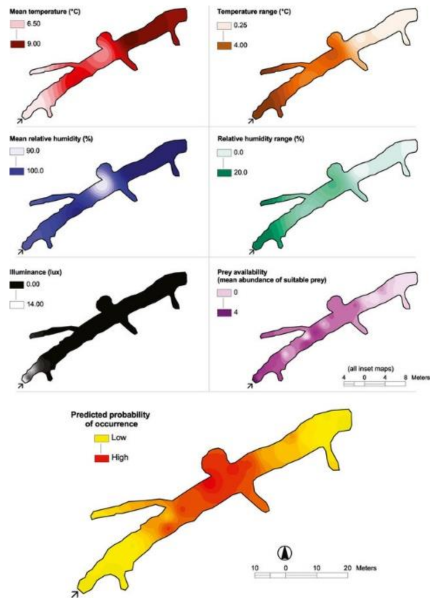
SDM – from a computer perspective

- What we hope in the end is:

Learn a function: P(Species|Env): Geographical space

$$f(A, B, M) = P(species)$$





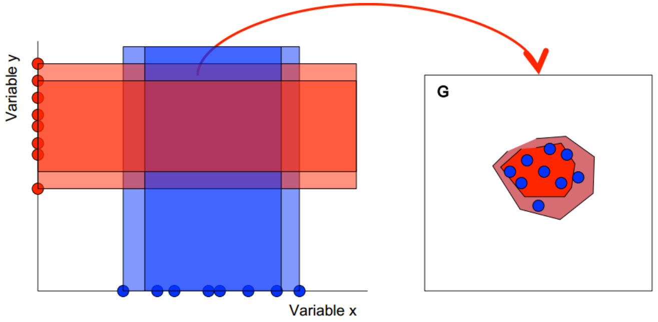
From: https://doi.org/10.1111/ivb.12113

Any questions? 30s

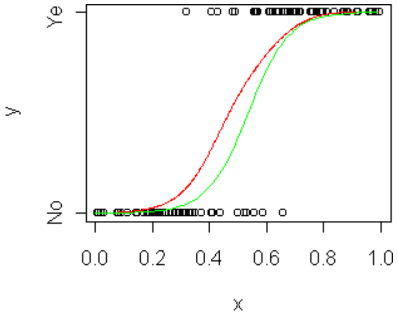
Universiteit Leiden
The Netherlands

SDM – from a computer perspective

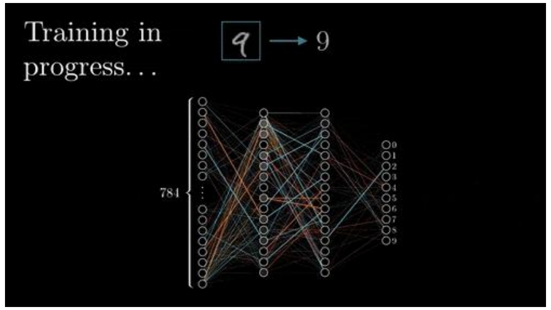
- Bioclimatic envelope
  - Find best “rule” (aka threshods) that groups the data into similar sets
  - AKA: minimizing the impurity



- Generalized Linear model (Logistic)
  - Finds the best parameters in a linear model structure that minimizes the ||Y – Y_expected||

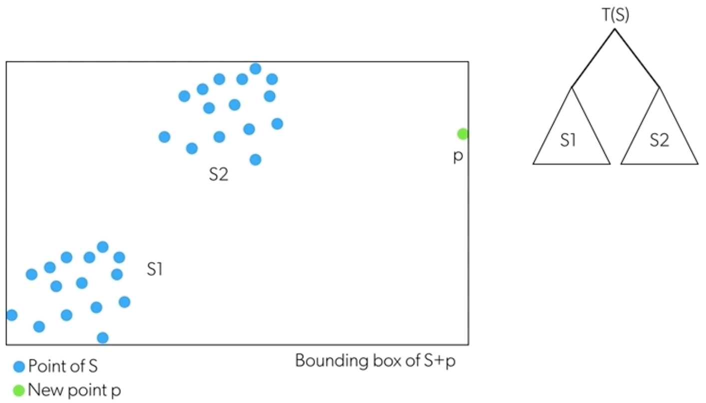


- Artificial neural networks
  - Sets of directed input/output linear models compressed into non-linear functions
  - Minimizes the difference between the ||Y – Y_expected||

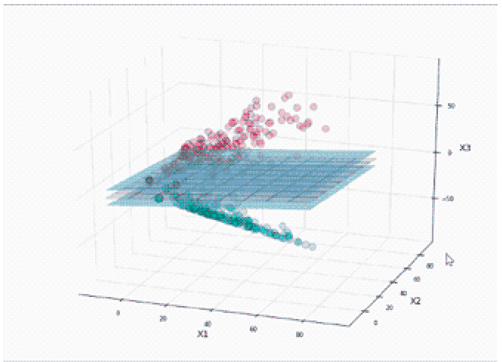


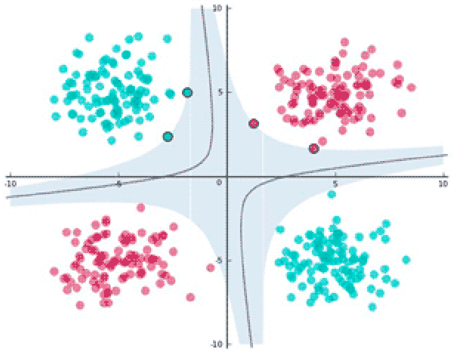
SDM – from a computer perspective

- Random forests and Decision tree's
  - Creates a hierarchy of rules that bins similar groups into the same **by minimizing their impurity**
  - Final output is a weighted sum of regression models on each bin



- Support vector machine
  - Projects data into another space
  - Find the best n-dimensional plane that separates the data on the new space AKA maximizes separability





- There are dozens/hundreds of algorithms, but our focus will in MAXENT

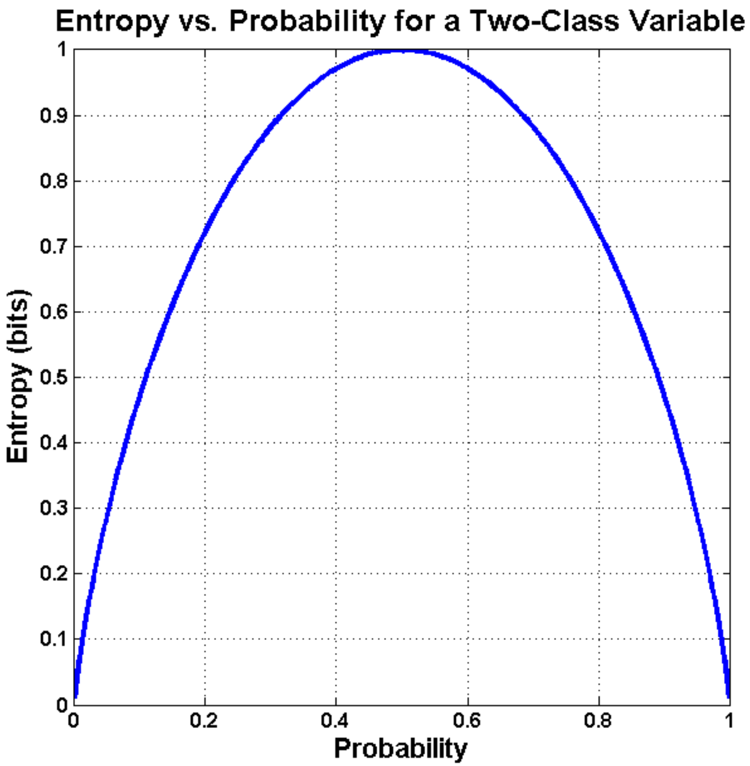
SDM – from a computer perspective

- MAXENT
  - Maximum entropy principle
  - AKA the best model is the one that is constrained by the data but has the highest uncertainty
  - **One of the benefits of MAXENT SDM is that they only needs presence points**

- Entropy can be defined as measure of disorder or uncertainty

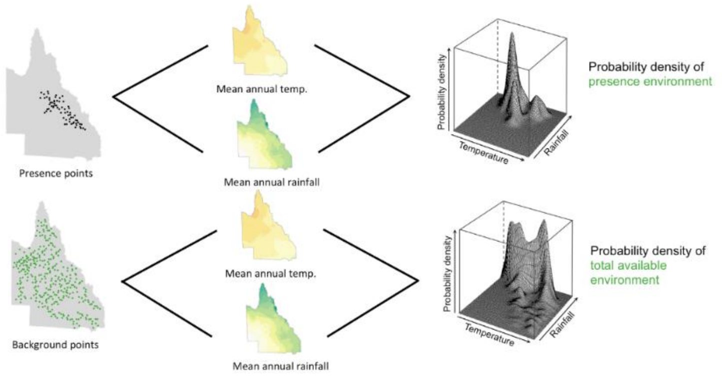
- If a coin toss is unbiased, each side has 50% probability
  - The entropy is maximized –> you are the most uncertain about the outcome
  - If the coin has a bias -> less entropy

- Applying the **Maximum entropy principle**:
  - For any possible probability distribution, choose the most uncertain possible
  - OR: Maximize a measure of uncertainty



SDM – from a computer perspective

- The most uncertain case for a species distribution:
  - A P(Species|Env) where it species is equally likely to occur in every location
  - In MAXENT: Background distribution

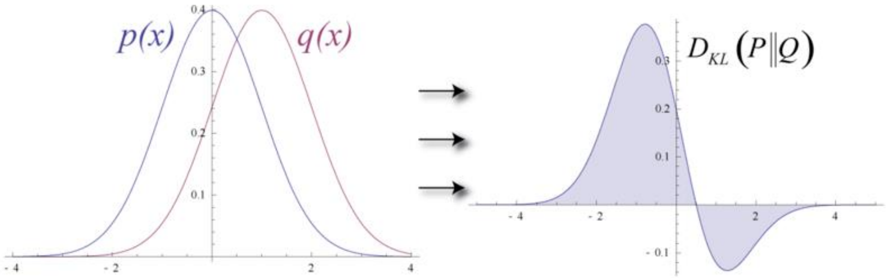


From: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1472-4642.2010.00725.x

- Imposes the structure of a Gibbs distribution to the environment:

$$P(X = x) = \frac{1}{Z(\beta)} \exp(-\beta E(x)).$$

- MAXENT in machine learning:
  - Minimizes the difference between the observed distribution (presences) and the random background distribution (generated pseudo-absences)
  - Uses the relative entropy formula to measure their difference



SDM – from a computer perspective

- How does the <u>model maths </u>look like?
  - Deeply discussed in Phillips, [2004]
  - RE is the relative entropy operator symbol
  - π-tilde – is the observed probability distribution
  - q_λ- is the background distribution parametrized by the Gibbs measure (also)
  - $\beta_i|\lambda_i|$-> are regulurization parameters to curb over fitting.

- Notice, it changes only the parameters of the background distribution (q_λ) – and not the presence distribution

- In SDM terms, MAXENT aims to:
  - Find the most uncertain distribution that is constrained by the observed presence points
  - Os, in other words: Starts with the most random distribution possible and finds the most random distribution possible that still predicts the presences

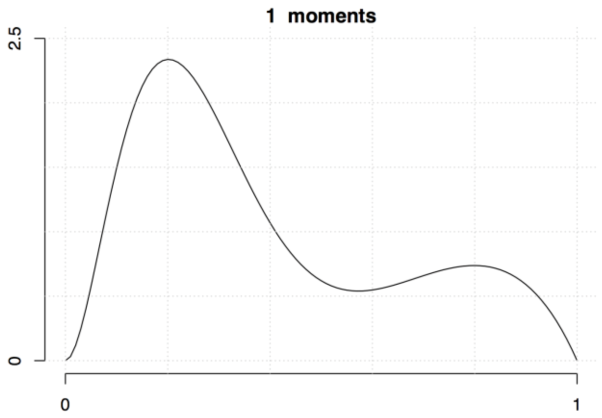
$$\mathrm{RE}(\tilde{\pi} \parallel \boxed{q_{\boldsymbol{\lambda}}}) + \boxed{\textstyle\sum_j \beta_j|\lambda_j|}$$

Regularization parameters

Feature weights

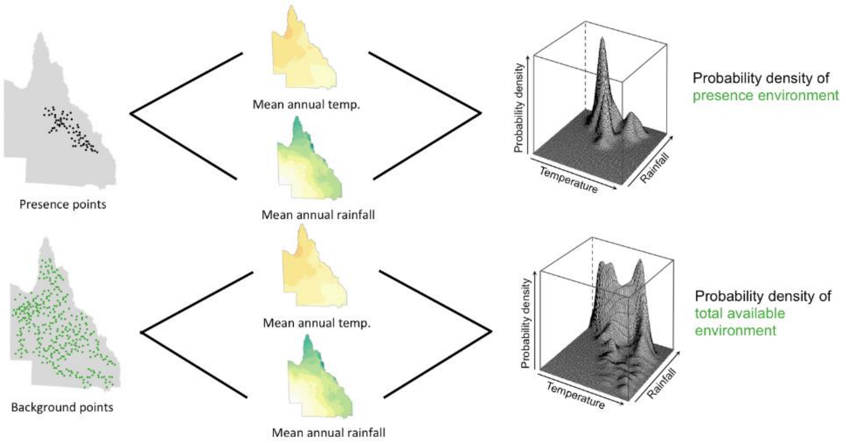
$$q_\lambda(x) = \frac{\exp\left(\sum_{j=1}^{n} \boxed{\lambda_j} f_j(x)\right)}{\boxed{Z_\lambda}}$$

Normalization constant



SDM – from a computer perspective

Mine both probability distributions from the data and learn the parameters

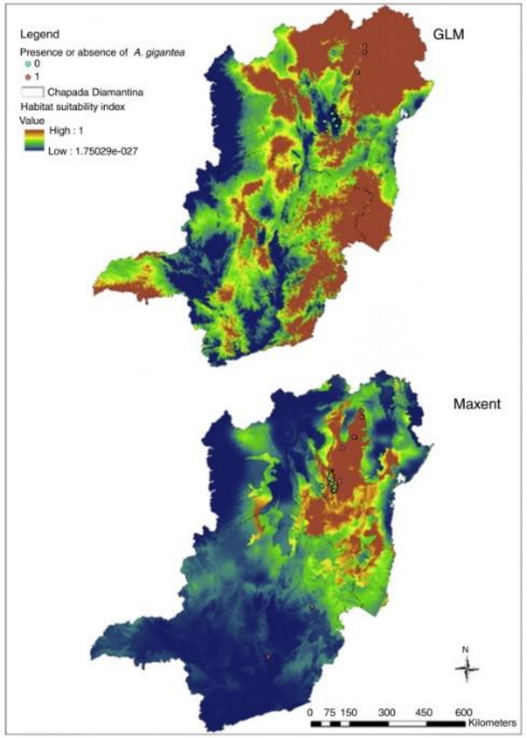


From: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1472-4642.2010.00725.x

Apply the bayes theorem!

$$P(S|E) = \frac{P(S|E) \cdot P(S)}{P(E)}$$

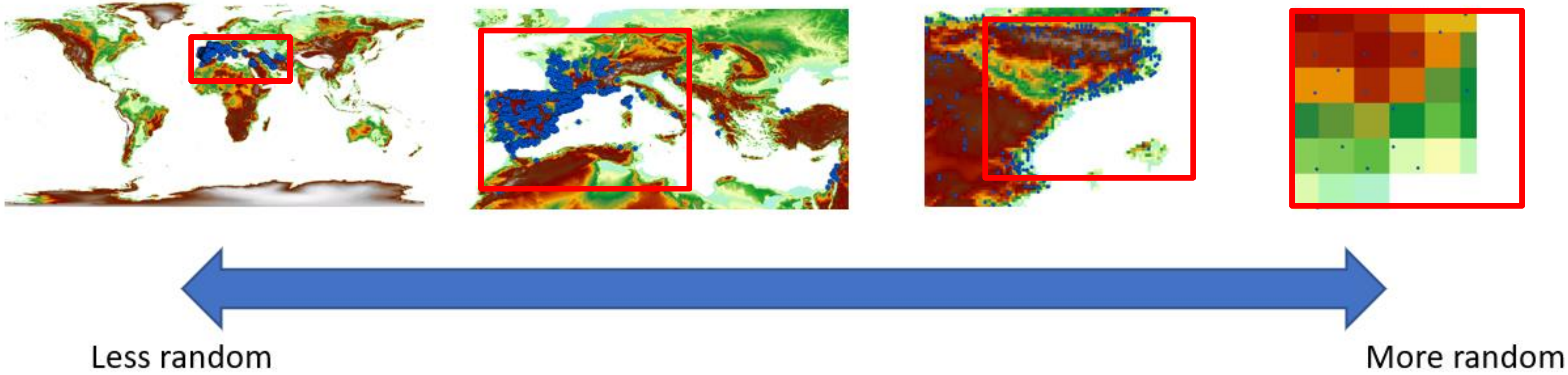
And voilá



From: https://doi.org/10.1016/j.ncon.2015.03.001

SDM – from a computational perspective

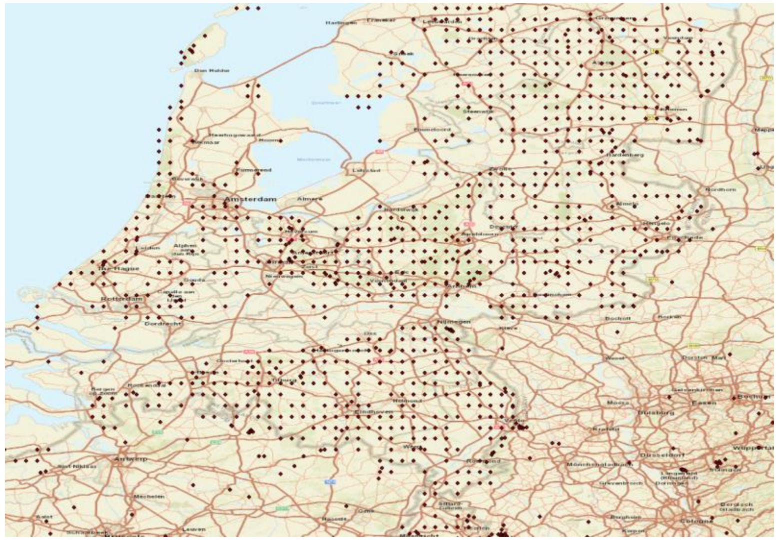
- SDM can be limited by:
  - Environmental space not being representative of the process -> Problem of scale

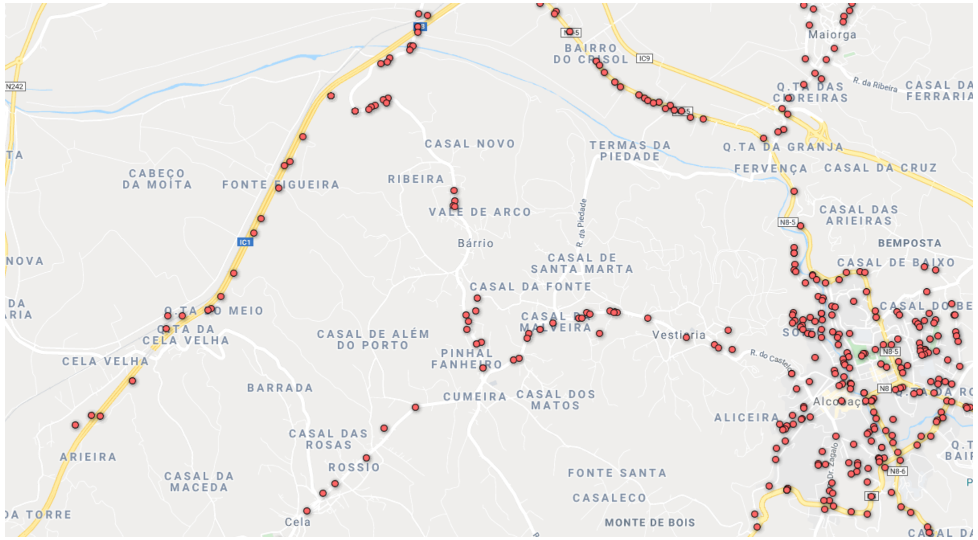


- The processes defining the distribution of the species vary according to the scale that you are using!
  - Machine learning, with bad data in will just produce a bad model -> it will learn how to do random predictions.

SDM – from a computational perspective

- SDM can be limited by:
  - Spatial biases

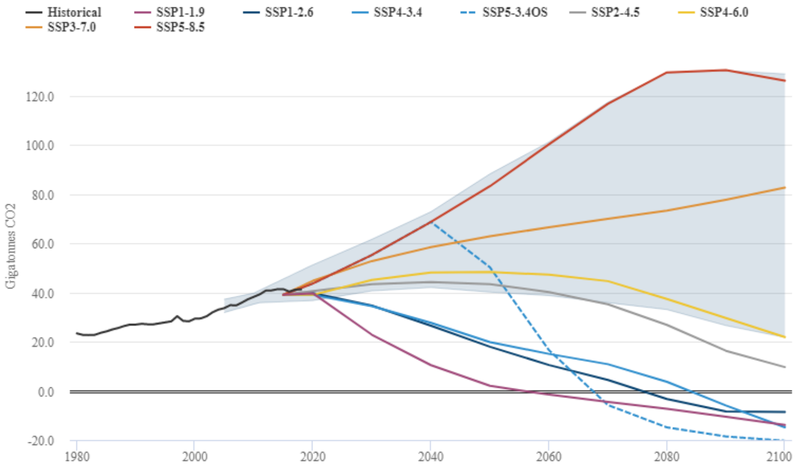




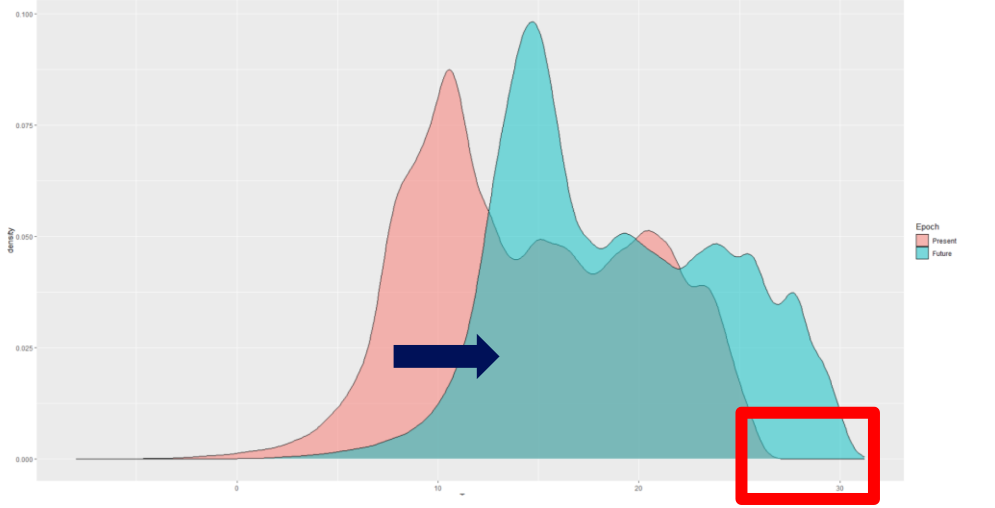
- Different sampling designs are also part of the processes that Machine learning learns
  - Meaning, if your data sampling strategy is biased, your model, will also learn that bias

SDM – from a computational perspective

- SDM can be limited by:
  - Non-analogous conditions in time



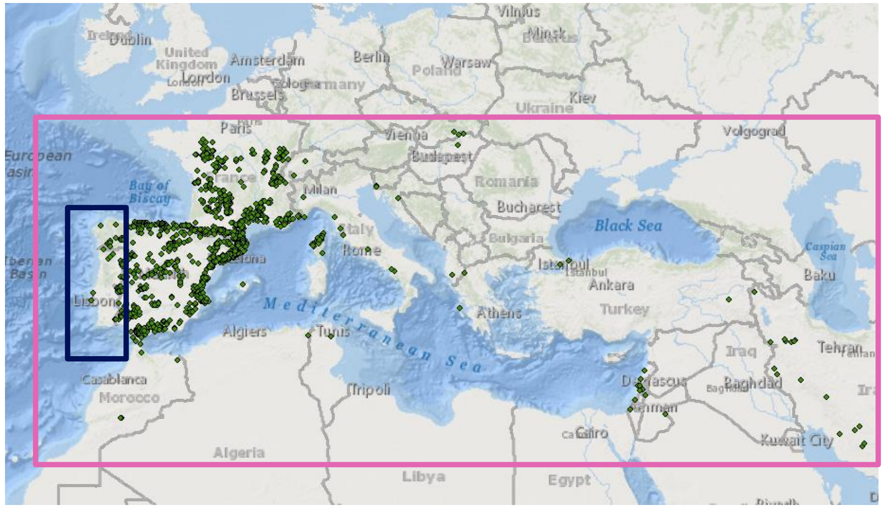
Present vs 50 years in the future (rocky road scenario!)



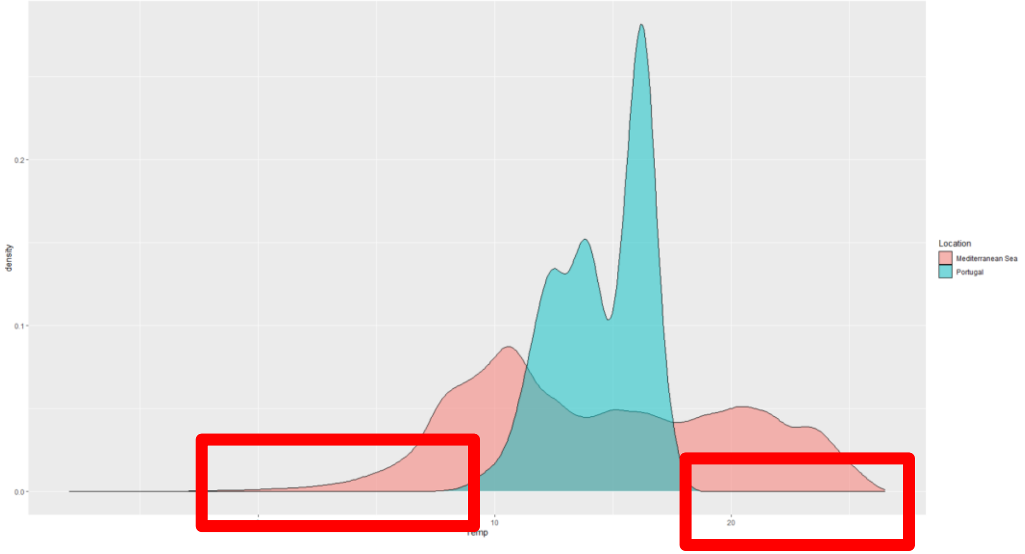
- Above we see a shift both in the distribution (arrow) as well as a range of values outside the training range(red box) for the Mean Annual temperature in the study area
- These non-analogous conditions imply that your model is extrapolating on that range

SDM – from a computational perspective

- SDM can be limited by:
  - Non-analogous conditions in space



Temperature distribution for Portugal and the Mediterranean



- A model trained in Portugal would have to extrapolate if used in the Mediterranean regions.
- Mismatch is an environmental space problem occurs frequently with Invasive Alien Species

SDM – from a computer perspective

- So in summary:
  - All ML models have a set of parameters and structure which is optimized to fit observations
  - Machine learning is able to “understand” the processes but can’t (necessarely) explain them to humans

- The ML behaviour is fully driven by:
  - Model structure
  - Parameter
  - Cost function and optimization method

- At the end of the procedure you have a function (f) that:
  - maps an n-hyperdimensional space onto another n-hyperdimensional space
  - In SDM: **n-Environmental space** to probabilistic space -> which then can be projected to a **Geographical space**

SDM – from a computer perspective

- Error sources:
  - Spatial biases: are a process that can (will) get learned by the algorithm
  - Scale: affects the driving spatial processes defining the distributions -> your model learns what you give it
  - Non-analogous conditions in time & space: The model has to extrolate

- Not all models work the same:
  - MAXENT works “fine” even with reduced sample numbers – maximum entropy principle

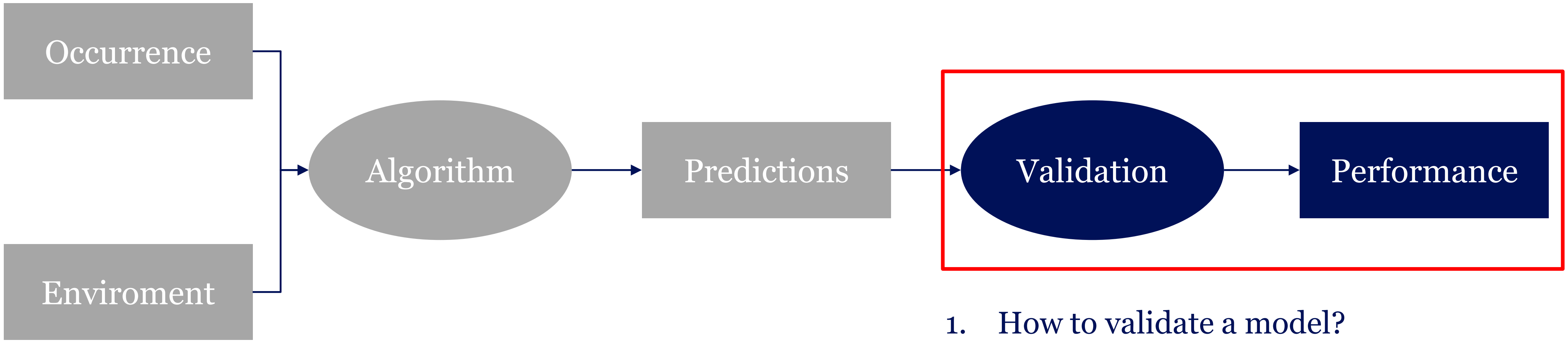
- **Most important for a good SDM model:**
  - Occurrence data reflecting the ecological niche of the species
  - A scale that reflects the ecological processes that you want to investigate <- one of the big problems
  - A model that “maximizes” your objectives: Excellent predictions? Deep Learning

Any questions? 1min

Next: Validating your model

Universiteit Leiden
The Netherlands

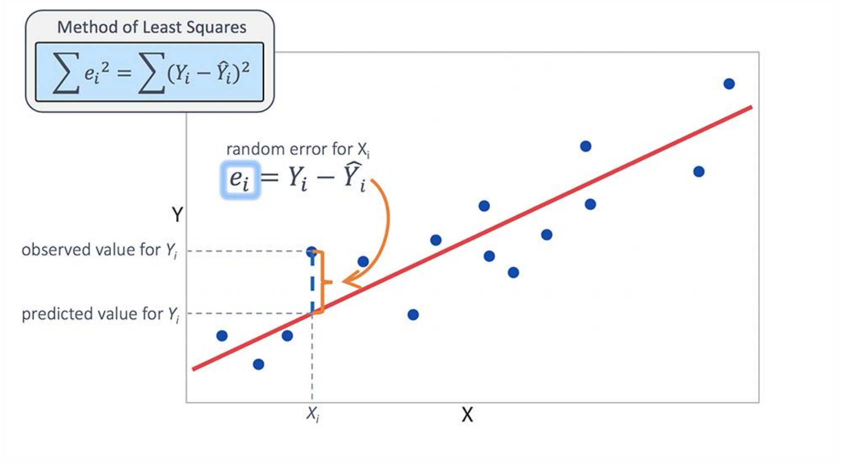
Remember where we are:



1. How to validate a model?
2. How to interpret the relations on the data?
3. What are the limitations of the model?

SDM – Model validation & performance

- In ML there are two main types of validations:
  - Regressions errors: Root mean squared error, Mean absolute error etc
  - Classification errors: Overall acuraccy, K, jaccard etc.



- SDM is a supervised classification exercise
  - Binary: the species is either Present or Absent
  - Even if in some cases, the output is a Probability
  - Threfore, use classification validation metrics

- Classification errors:
  - Mostly based on the concept of Confusion matrix (next)

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$		

SDM – Model validation & performance

- Confusion matrix: (aka error matrix)
  - Shows the agreement between our predictions and the validation data

- The values represent the times our predictions agreed or disagreed with the validation data
  - Type I error -> our model predicted the species where it is absent
  - Type II error -> our model didn't predict the species where it is present

- Remember, our outputs of MAXENT are probabilities [0 to 1] so:
  - We have to select a minimum probability (e.g. 50%) after which the species is present

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Type I error

Type II error

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$		

SDM – Model validation & performance

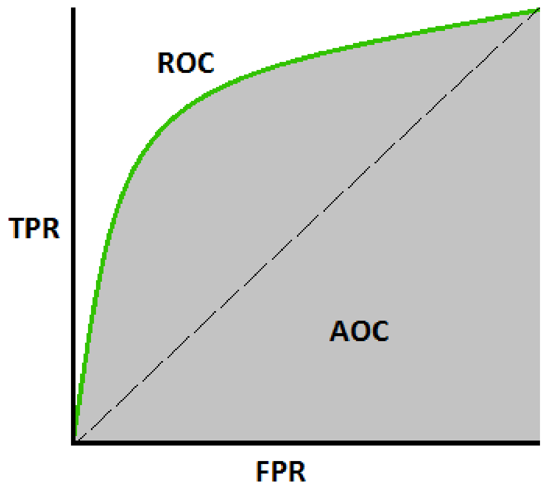
- Problem!
  - Absences are especially difficult to define in the context of Species Distribution Models
  - MAXENT does not need to use Presence and Absence data

- Without Presence/Absence data we cannot use the confusion matrix

- The alternative:
  - An estimate of the models ability to discriminate the species from the enviroment
  - This metric is know as Area under the curve of the Receiver operating characteristic (AUC-ROC)

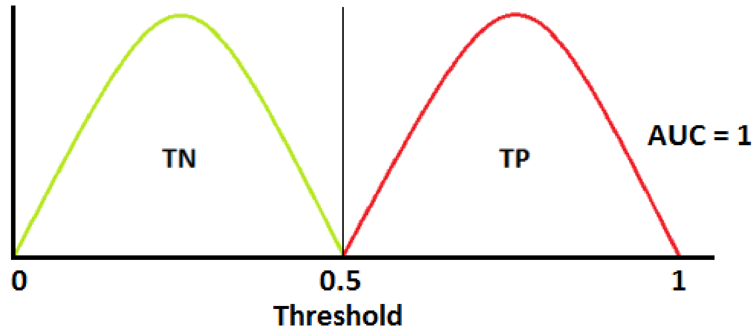
SDM – Model validation & performance

- Receiver operating characteristic curve (ROC)
  - Ilustrates the ability of the binary classifier by varying the probability threshold
  - Notice: both axis vary from 0 to 1



- TPR = Sensitivity
  - Tells something about the ability to predict true positives
- FPR = 1 – Specificity
  - Tells something about the ability to identify true negatives

- Area-under the curve (AUC):
  - Integral of the area under the ROC curve.
  - When the discrimination power is perfect: AUC=1



SDM – Model validation & performance

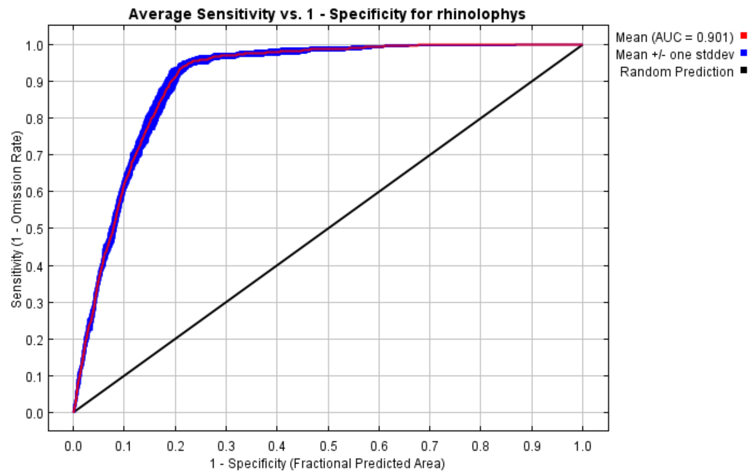
- Problem still remains in MAXENT!
  - No Presences and absences available!

- MAXENT actually uses the Area of the predictions to produce the AUC-ROC:

$$TPR_{maxent} = 1 - A_{FOR} = 1 - \frac{A_{FN}}{A_{FN} + A_{TN}}$$

$$FPR_{maxent} = 1 - A_{PPV} = 1 - \frac{A_{TP}}{A_{TP} + A_{FP}}$$

- Where the A_{ij} represents:
  - The actual geographical area predicted as FN, TN, TP,FP



		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

SDM – Model validation & performance

"*When AUC statistics are applied to presence-only data and pseudo-absences, the maximum achievable AUC value is no longer 1, BUT 1- a/2; where a stands for the true species' distribution, which we typically do not know*" (Phillips, 2006)

Less prevalent

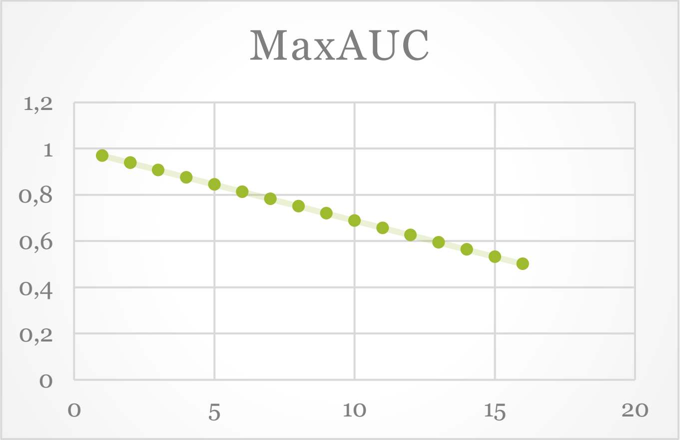
x			
		x	
x			
	x		

$$1-\frac{\frac{4}{16}}{2}=1-0.125=0.875$$

More prevalent

x		x	
	x	x	x
x		x	x
	x	x	

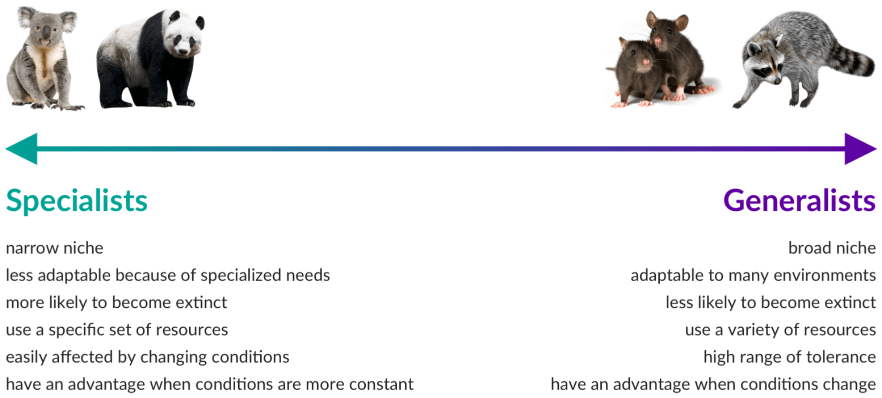
$$1-\frac{\frac{10}{16}}{2}=0.6875$$



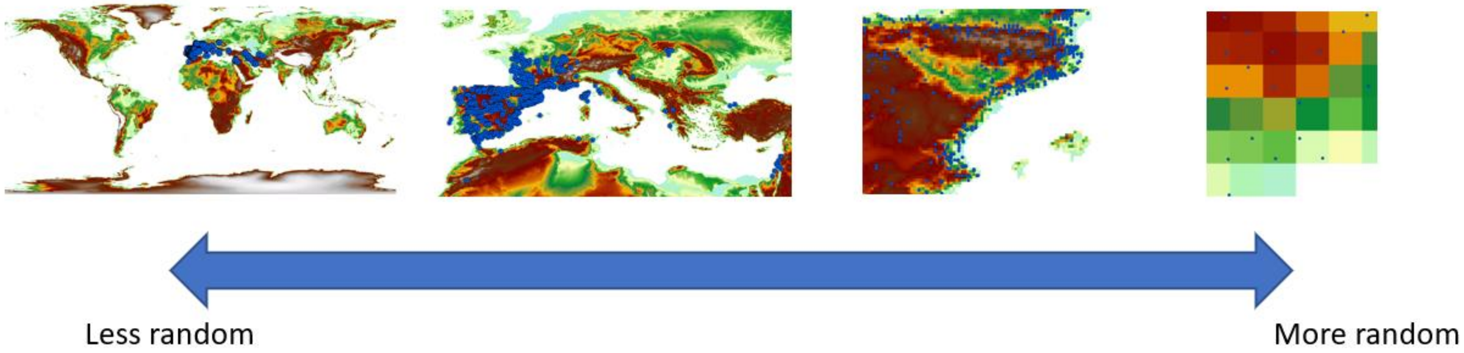
Less prevalent — More prevalent

SDM – Model validation & performance

The implication that the AUC is vulnerable to generalist/specialist species



And also highly sensitive to the Scale:



SDM – Model validation & performance

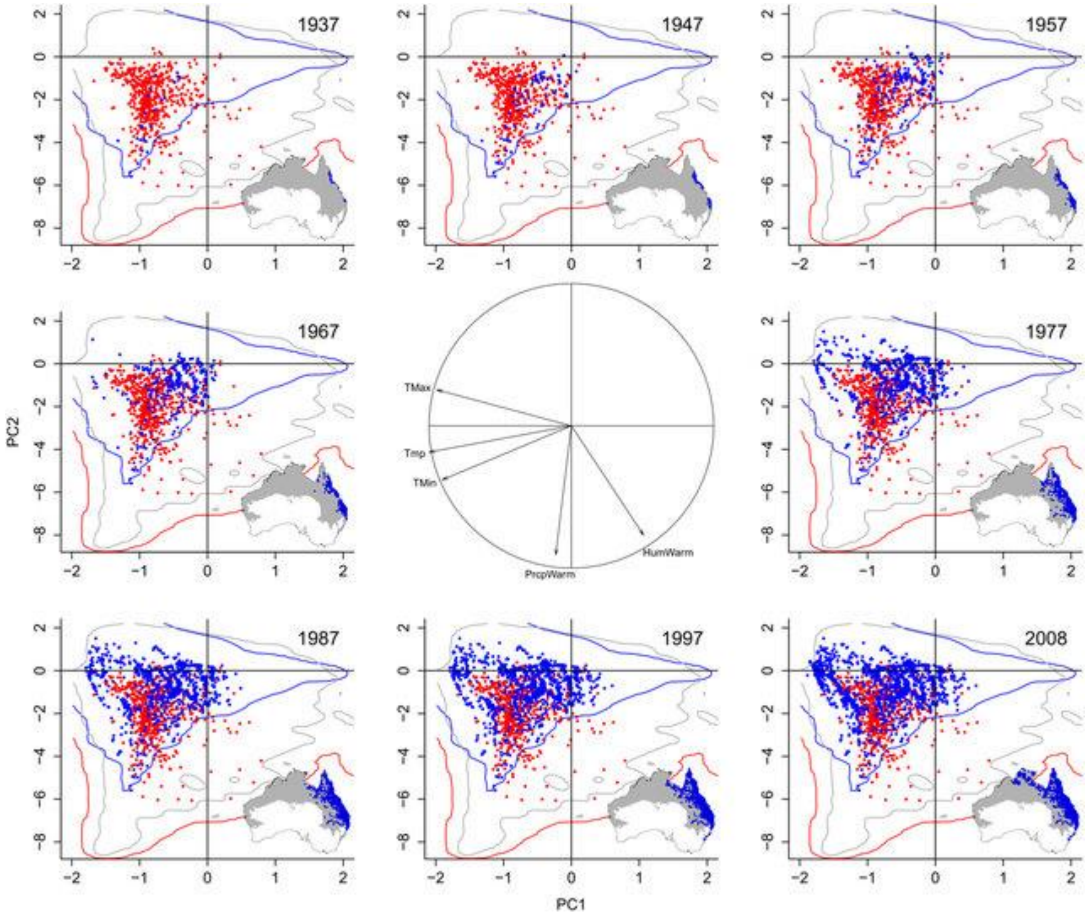
- In summary:
  - SDM can and should be validated using regular classification error metrics
  - MAXENT is a special case where absence data is “not available”
- MAXENT is usually validated using the AUC-ROC curve
  - Software provides it! So don’t worry
- AUC is highly affected by prevalence & scale
  - MAXENT goes around it by considering background data as pseudo-absence
- There are other methods to validate! Of course!
  - E.g. a Leiden research proposed a Null-model approach (Neils Raes, 2007)

Any questions? 30s

Universiteit Leiden
The Netherlands

SDM – Model validation & performance

- Non- analogous conditions
  - In time (e.g. Climate change)
  - In space (e.g. A new geographical region)
- There are other methods but not explored today
- We'll focus on the solutions MAXENT offers
  - Model response curves & clamping
  - Multivariate Environmental Similarity Surfaces (MESS) & Most dissimilar variable (MoD)
  - Variable importance & Jacknife testing

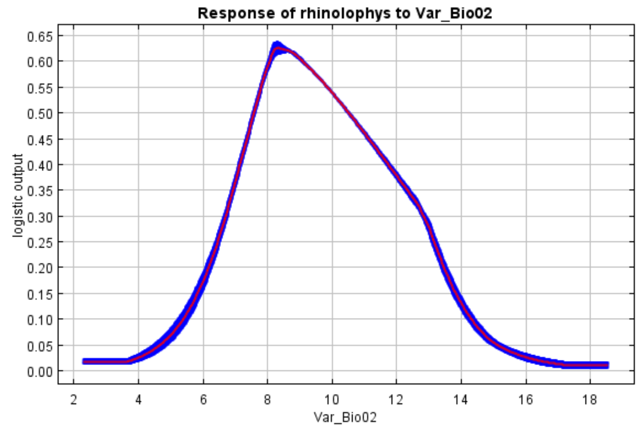


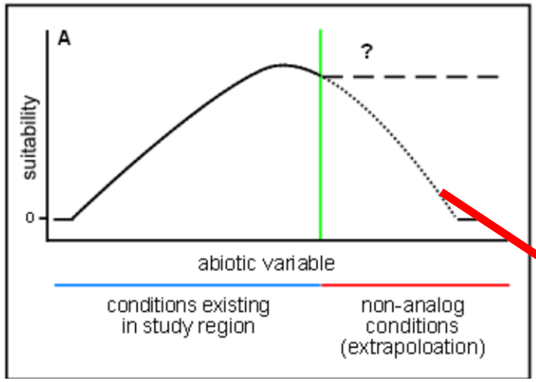
From: https://www.pnas.org/content/111/28/10233

SDM – Model validation & performance

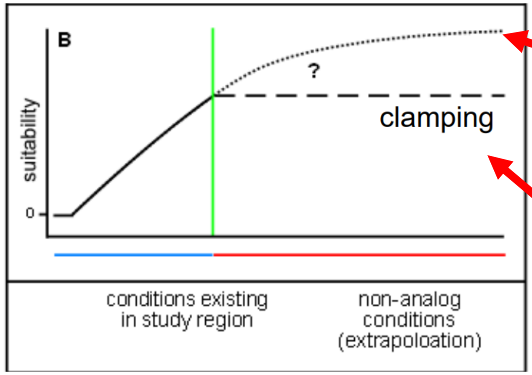
- Model responses
  - MAXENT provides a univariate probability response plot
  - These show howprobability varies according to each enviromental factor
  - You can explore these outputs to know if you're model was trained close to the limits of the species range
- Clamping:
  - Tells MAXENT what to do when the variable is outside the training range

Good example!





"Fade by clamping" option

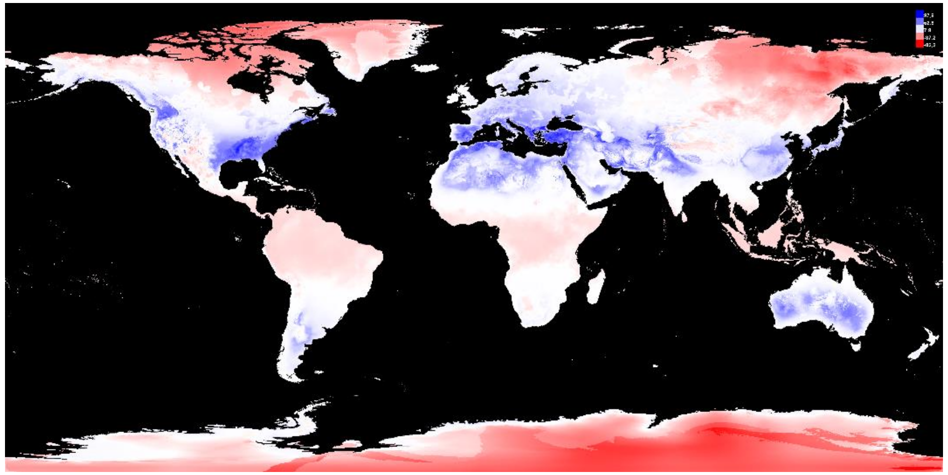


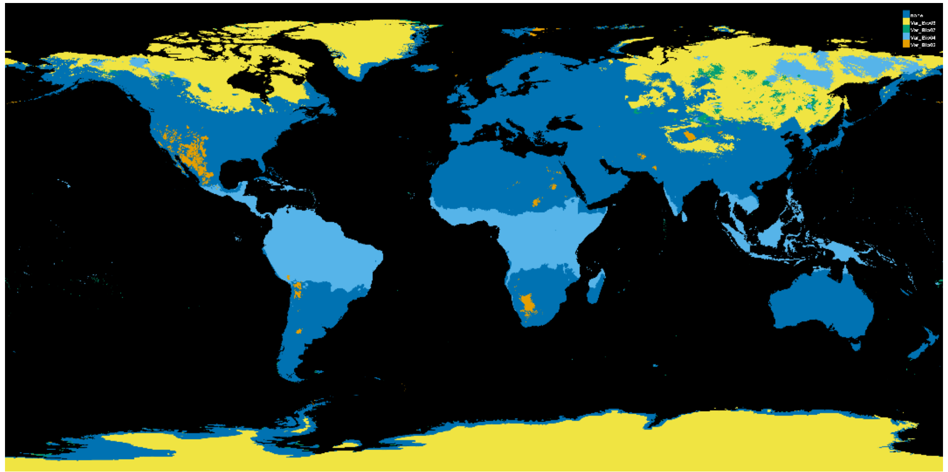
"Extrapolate" option

"Clamping" option

SDM – Model validation & performance

- Algorithms described in: (Elith,2010; Supporting information)
- Multivariate Environmental Similarity Surfaces (MESS)
  - Produces a map estimating the similarity [-100, 100];
  - With 0 being "equal" and -100 or 100 being totally (negatively or positively) dissimilar
- Most Dissimilar Variables (MoD)
  - Maps which of the various environmental had the most dissimilar MESS for that particular location
- These models should be used to identify regions where we should be suspicious of our model performance
  - Notice: Clamping might (will) affect these surfaces!





SDM – Model validation & performance

- Variable importance:
  - These provide an estimate of how significant X variable was for the MAXENT model
  - **Critical for ecologists**: These are the variables "driving" the function that defines the distribution of the species.

$$\mathrm{RE}(\tilde{\pi} \parallel q_{\boldsymbol{\lambda}}) + \underbrace{\boxed{\textstyle\sum_j \beta_j |\lambda_j|}}_{\text{Regularization parameters}}$$

- Percent contribution:
  - Δ changes in AUC gain based on Regularization parameters – "remember the formula"
  - Each scalar is in function of each variable, so this is used to measure is contribution

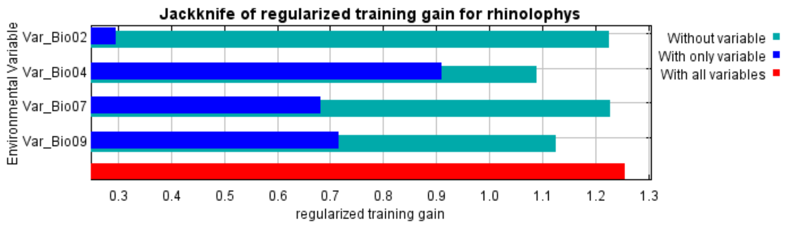
- Permutation importance:
  - Changes in training AUC by excluding/including the given variable
  - And then normalizes, to provide a % of contribution

Variable	Percent contribution	Permutation importance
_Bio04	52.9	50.8
_Bio12	36.3	31.9
_Bio01	6.6	12.9
_Bio15	3.2	2.6
_Bio19	1.1	1.9

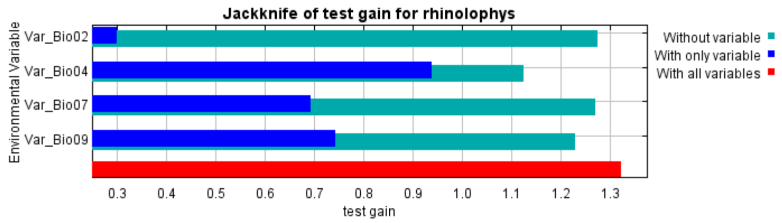
- Take note:
  - % contribution ≈ Permutation importance is a Good sign
  - These estimates are highly affected by autocorrelation between environmental variables

SDM – Model validation & performance

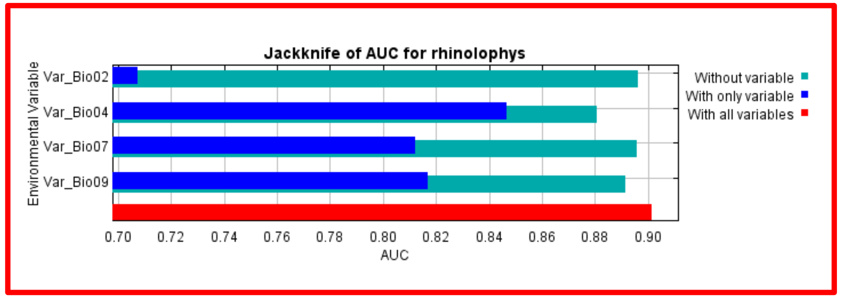
- Jacknife testing
  - Asses the variation in accuraccy (aka, gain or AUC) if a model is ran using only variable X or excluding variable X – "jacknife"
  - The "non-used" variables are set to the mean value of the distribution



- Mostly the same information as variable importance
  - Remember: Ecologists should care for what drives the distribution



- MAXENT provides 3 versions:
  - Based on the training acuracy gain
  - Based on the test accuraccy gain (data that was left out!)
  - Based on the AUC <- This is the most helpful as it relates directly with the AUC-ROC report



SDM – Model validation & performance

- In summary:
  - Non-analogous conditions – how to deal with them?
  - Species with unstable niche (e.g. exotic species) often break the assumptions of Hutchinson niche theory

- MAXENT diagnostic's:
  - Response curves –> Indication of how your model responds to the environmental variable
  - MESS & MoD -> Indication of where and why is your mode extrapolating
  - Variable importance -> indication of which variables are more important for the species distributions
  - Jacknife testing -> Same as above

- Your model performance should be interpreted using all the diagonostics offered:
  - Otherwise.. You might find yourself predicting giraffes in the artic

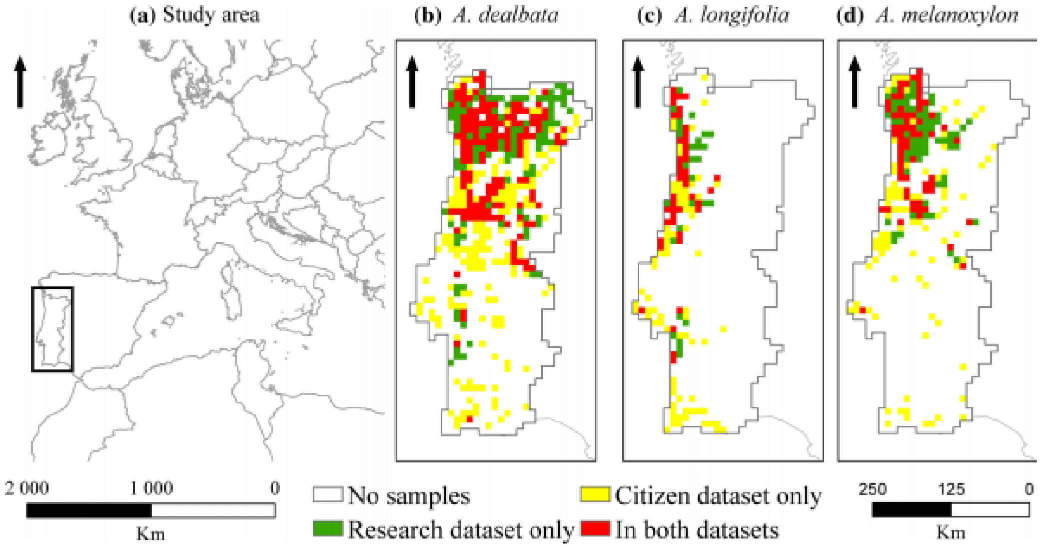
Two case studies:

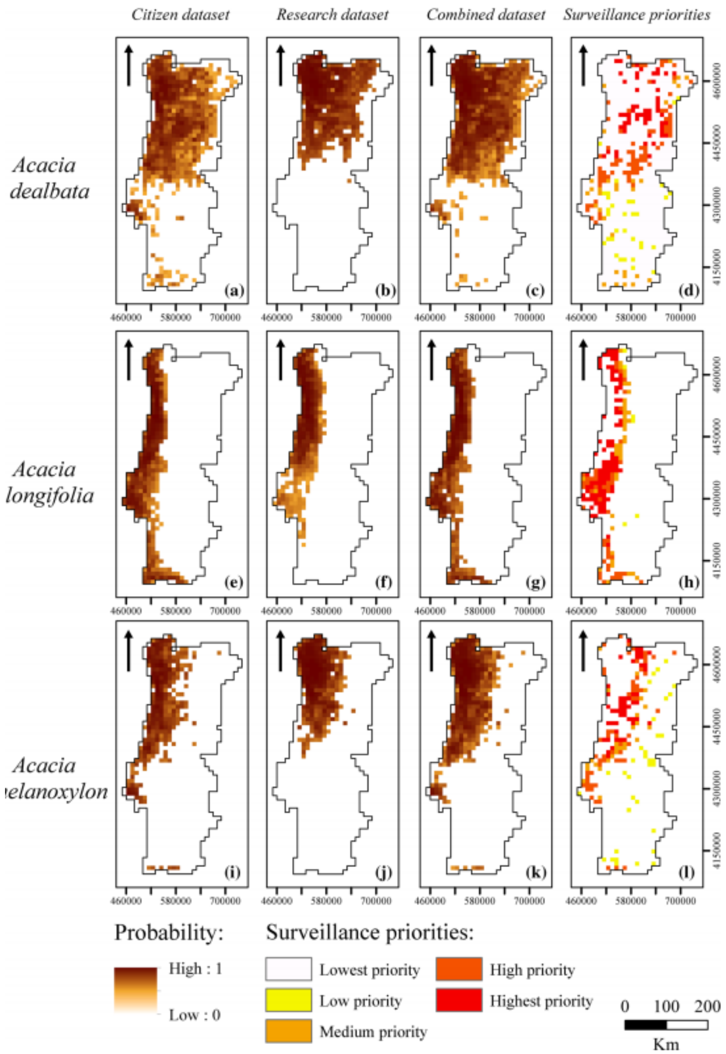
Shameless self promotion!

Universiteit Leiden
The Netherlands

SDM – Model validation & performance

- Using Citizen science to improve SDM (César de Sá, 2019)
  - Compared how data from citizens coupled with scientific surveys improves SDM coverage for invasive alien species
  - Identified areas of highest surveillance priority (aka likelihood of invasion but no data is available:

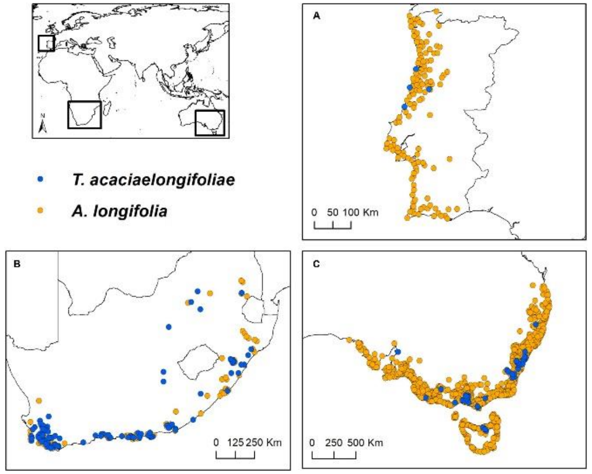


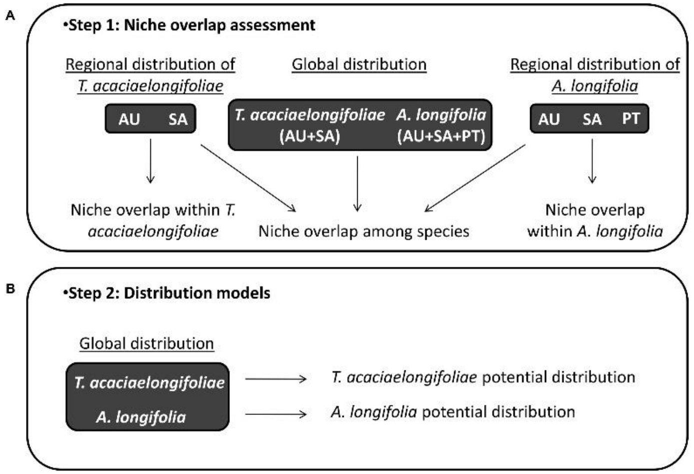


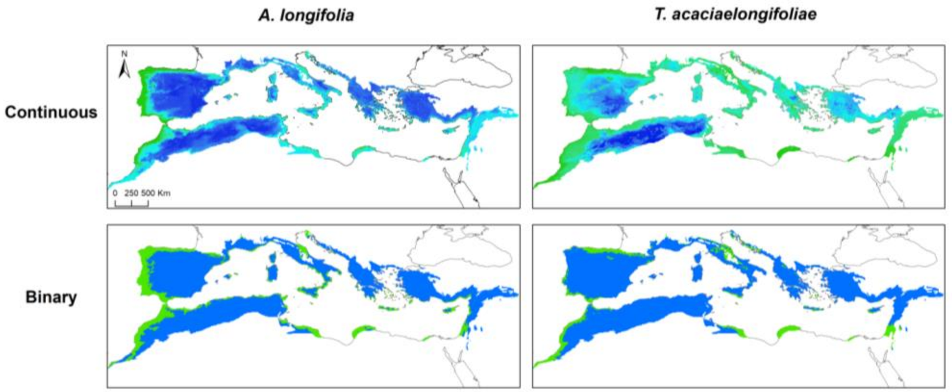
SDM – Model validation & performance

- Predicting the distribution of a biocontrol agent introduced in Portugal (Dinis, 2020)
  - The first biocontrol introduction in continental europe!
  - Measured niche overlap to assess if there is significant change & predicted its distribution

	Pairs	Expansion	Stability	Unfilling
Acacia longifolia	AU-PT	0.084	0.916	0.257
	AU-SA	0.102	0.898	0.211
	SA-PT	0.06	0.94	0.004
Trichilogaster acaciaelongifoliae	AU-SA	0.561	0.439	0.254
Among species	AlSA-TaSA	0.039	0.961	0.053
	AlAU-TaAU	0.001	0.999	0.227
	AlAll-TaAll	0.013	0.987	0.106







Thank you!

Q & A time

In the afteroon, you get to try this on your SPECIES!

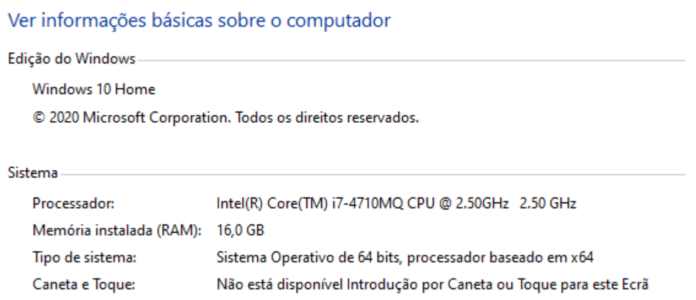
Setting up MAXENT & JAVA

And starting by Downloading the climate data during lunch

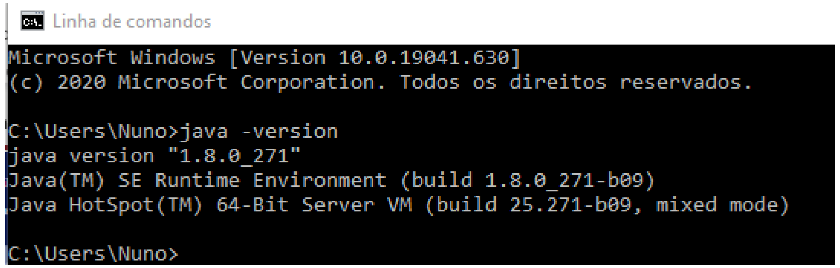
Universiteit Leiden
The Netherlands

Setting up Java

1. For Java:
   1. Go to https://www.java.com/en/download/help/download_options.xml
   2. Follow the steps to download and install the software for your operating system
   3. Unsure about your OS? In windows, go to your system properties:



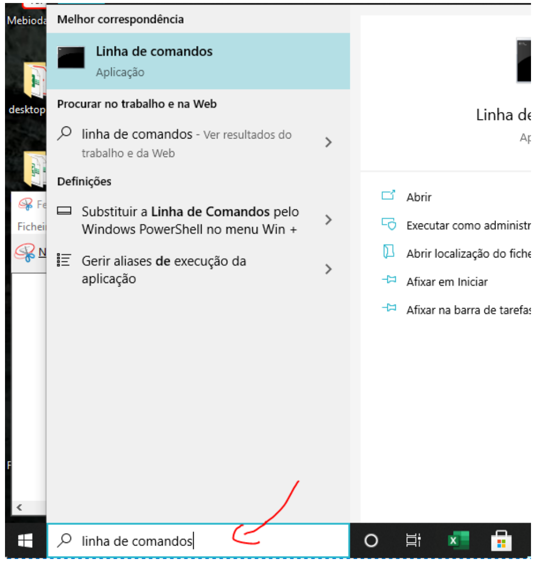
2. Check Java installation:
   1. Go to the search box and write "Command line" (or in your own OS language)
   2. On the DOS box that opens, type: java – version



```
Linha de comandos
Microsoft Windows [Version 10.0.19041.630]
(c) 2020 Microsoft Corporation. Todos os direitos reservados.

C:\Users\Nuno>java -version
java version "1.8.0_271"
Java(TM) SE Runtime Environment (build 1.8.0_271-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.271-b09, mixed mode)

C:\Users\Nuno>
```



   3. If all is well, you will have a similar output

Setting up MAXENT:

1. Go to https://biodiversityinformatics.amnh.org/open_source/maxent/

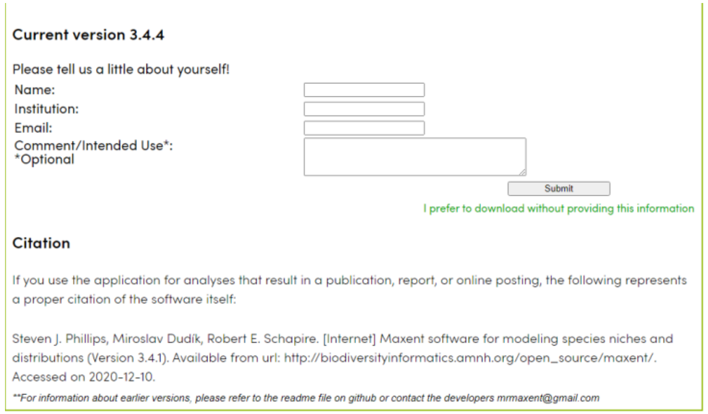
2. Tell them a bit about yourself
    - You can lie, but remember it is a sin

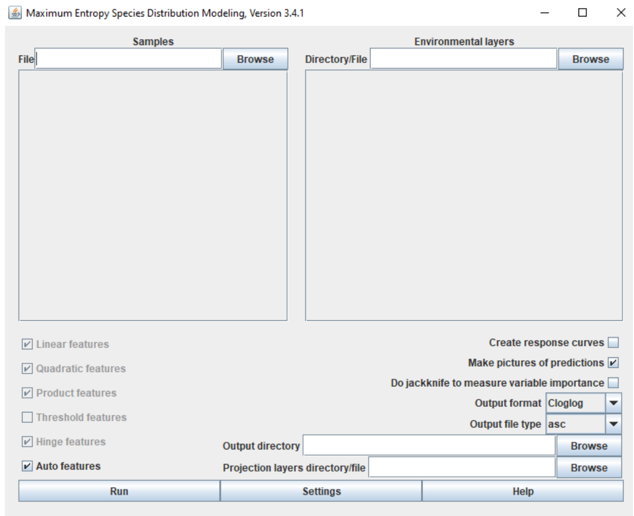
3. Download the software and unzip it to a folder

4. Click on the .Jar file to activate

5. If you see the screen, should be fine

PS: Would be nice to cite them on your report





Downloading the climate data

1. Go to the manual
    - It's already in brightspace!

- Follow the steps to download the climate data detailed there

In summary:

1. Go to https://www.worldclim.org/
2. Download:
    - Historical bioclimatic data (bio 5m) at 5 minutes resolution
    - Future scenario bioclimatic data (bc) data: IPSL-CM6A-LR / SSP370