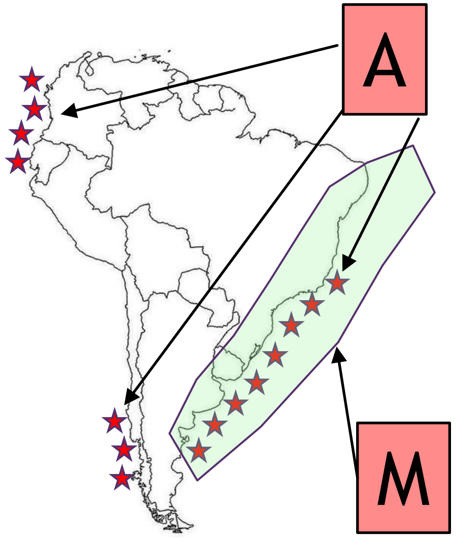
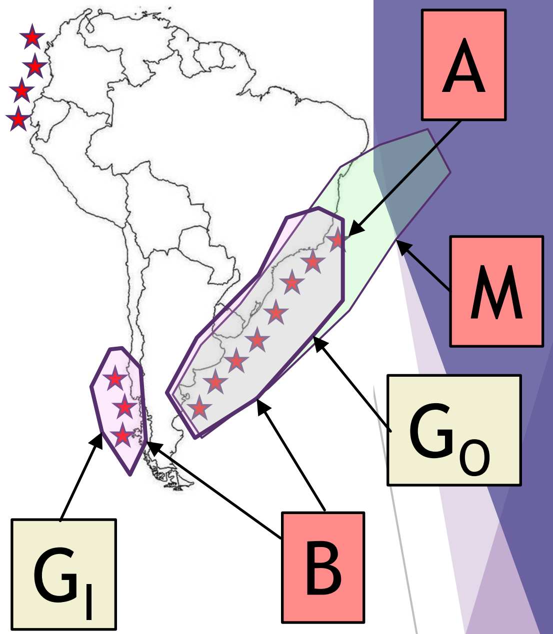
ALGORITHMS

Flávia F. Petean

CONICET
Universidad Nacional de San Martín
INTECH

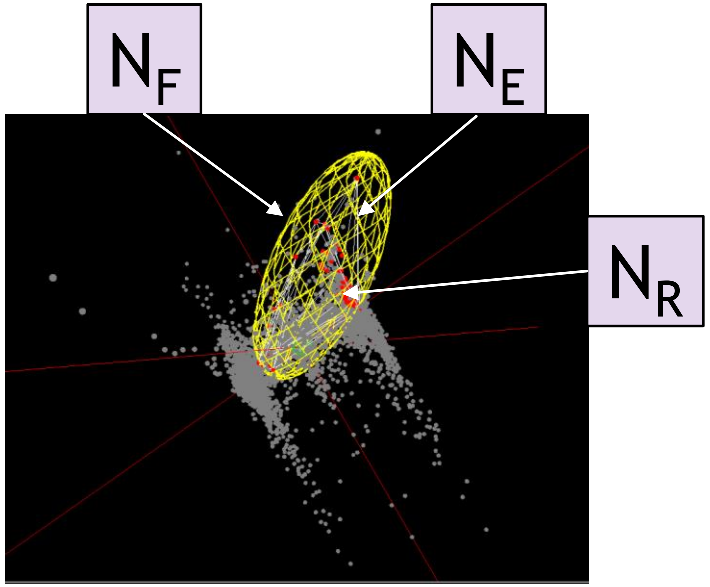
G



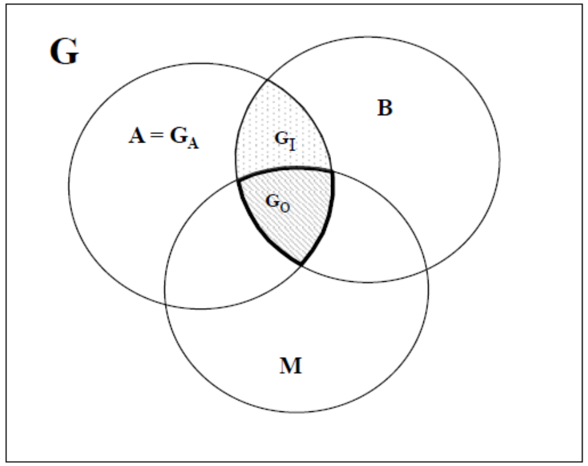


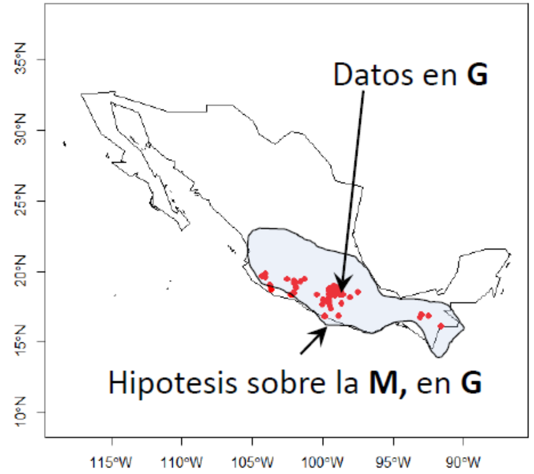
Remembering...

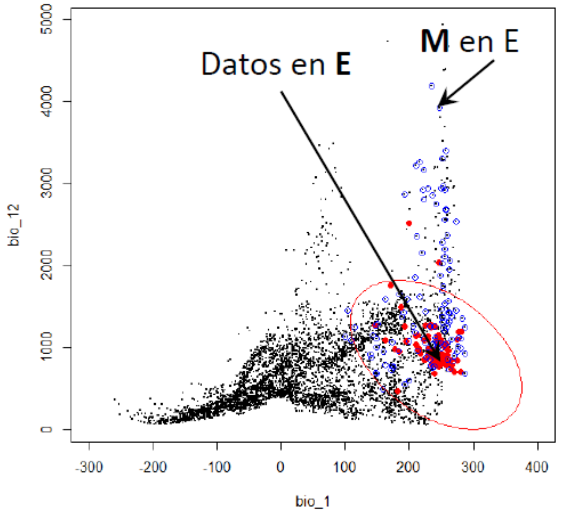
E

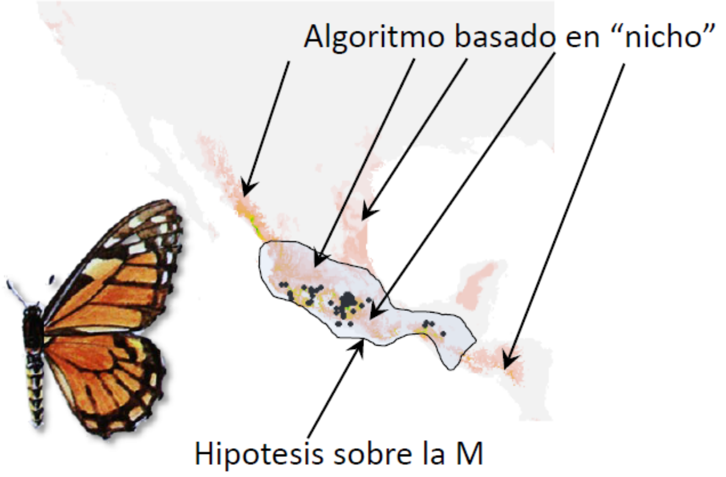


- What we do know:
  - Red dots = observations
- We can hypothesize M = blue area
- Fundamental niche (red ellipse) is unknown unless we have experiments
  - We don't have experiments
- What do algorithms estimate?





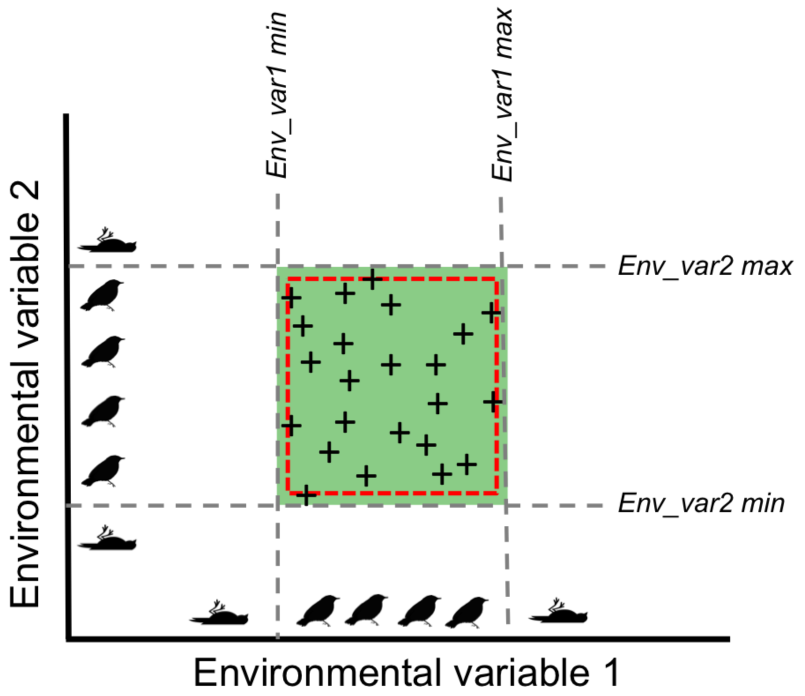


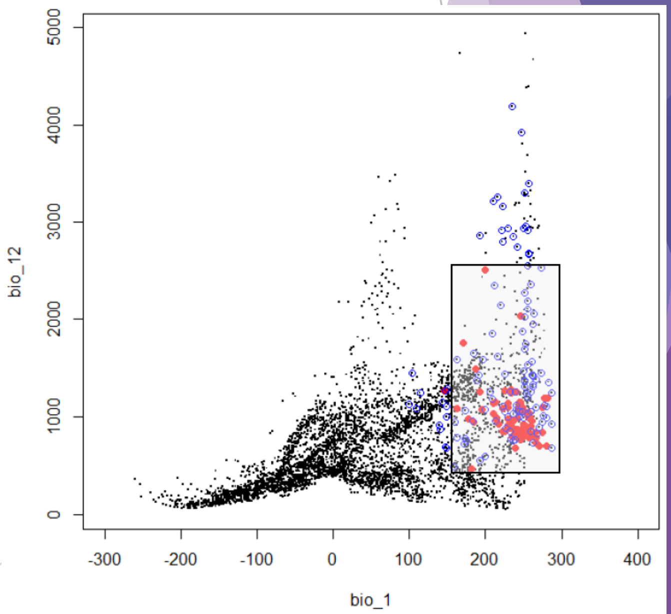


METHODS

Envelope Methods

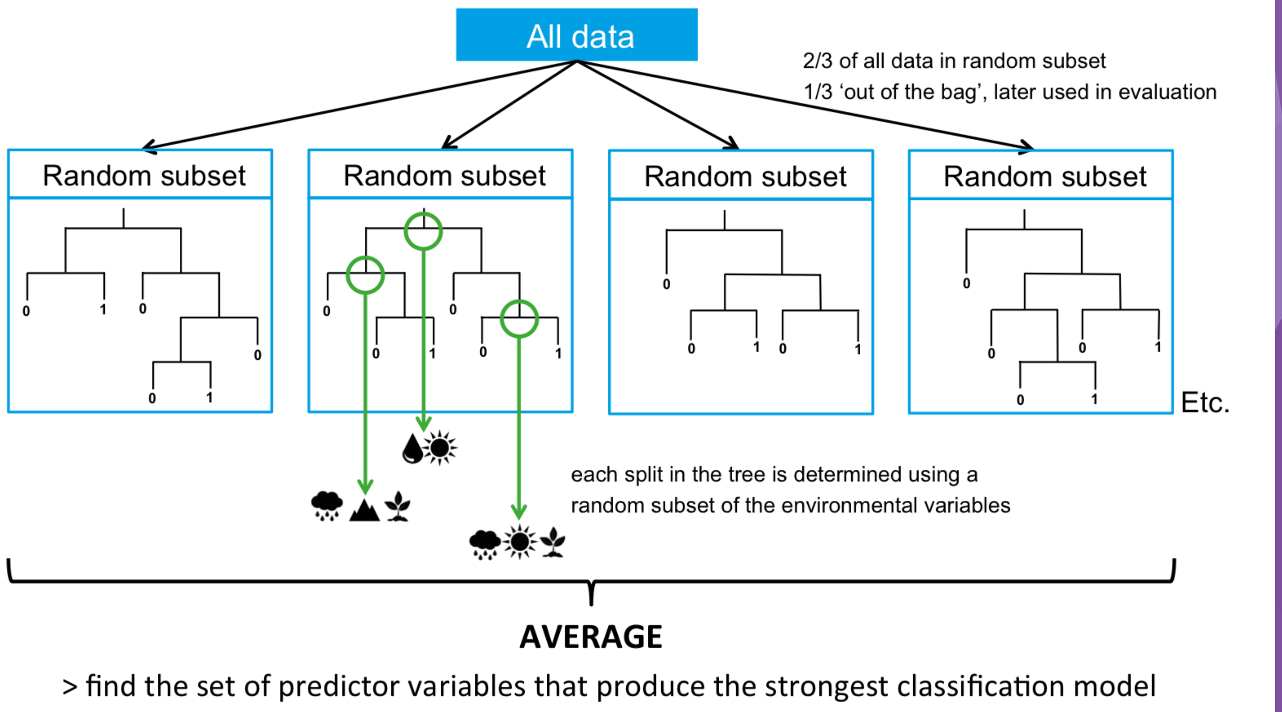
- Bioclim
- Provides ranking of predictor variables
- No absence data required
- Does not provide a probability
- It generates a characterization in E to be projected in G by Hutchinson's duality
- Projection is a region in G where the values of E are in the Bioclim "Box"





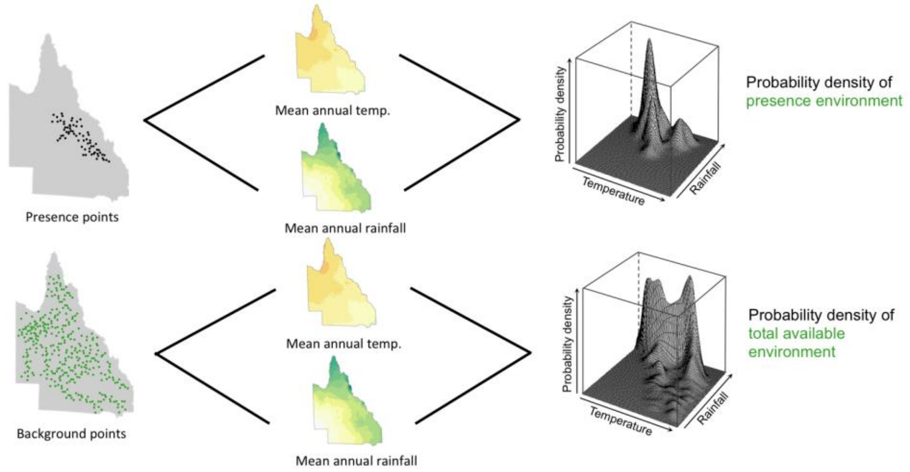
Random Forest Method

- Use of several predictor variables with estimates of importance of each one
- Accuracy even with missing data
- High overfit
- Requires absence data



Machine learning methods

- MaxEnt
- It gives each pixel a value that is a probability
- The sum of all output values is 1
- Regularization protocol that restricts overadjustment
- Good predictive performance
- Unlike other methods, it provides environmental suitability, not probability of occurrence
- Projection indicates that the points in G are similar to those of observation

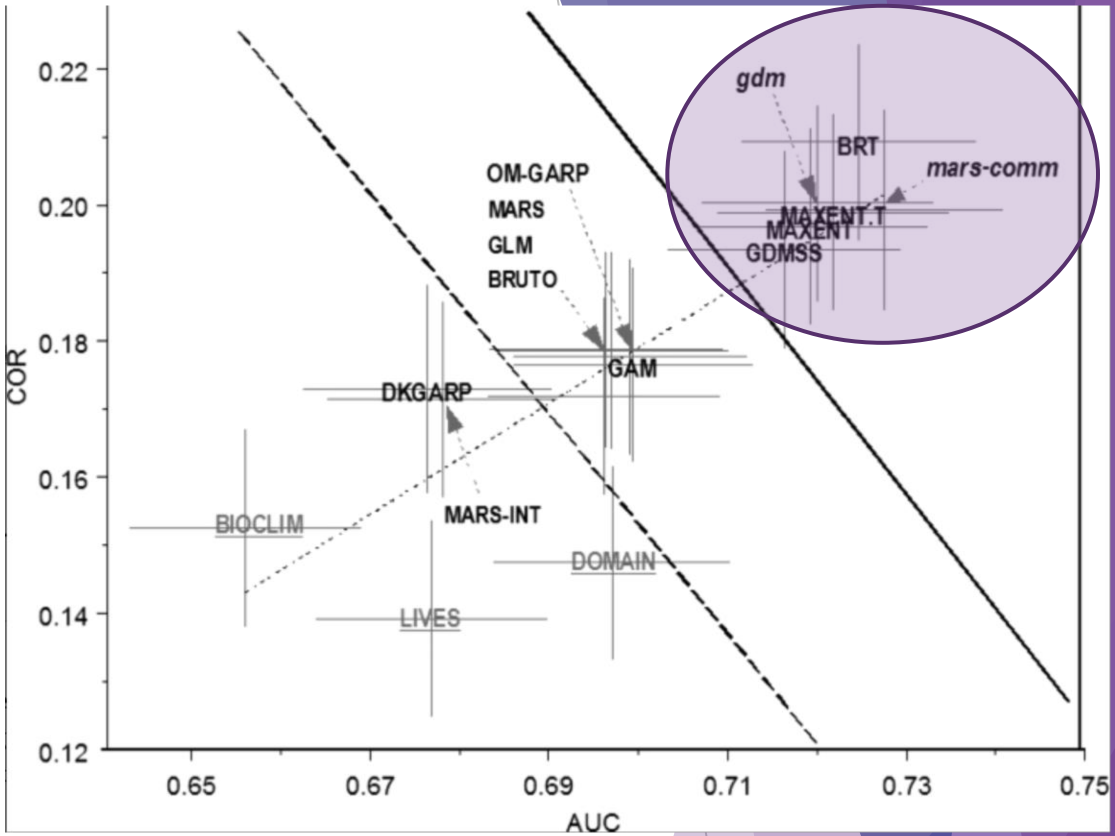


Adapted from Elith et al. (2011) *A statistical explanation of MaxEnt for ecologists*. Diversity and Distributions, 17, 43-57.

Table 4. Modelling methods implemented.

Method	Class of model, and explanation	Data[1]	Software	Std errors?[2]	Contact person
BIOCLIM	envelope model	p	DIVA-GIS	no	CG, RH
BRT	boosted decision trees	pa	R, gbm package	no	JE
BRUTO	regression, a fast implementation of a gam	pa	R and Splus, mda package	yes	JE
DK-GARP	rule sets from genetic algorithms; desktop version	pa	DesktopGarp	no	ATP
DOMAIN	multivariate distance	p	DIVA-GIS	no	CG, RH
GAM	regression: generalised additive model	pa	S-Plus, GRASP add-on	yes	AG,AL,JE
GDM	generalised dissimilarity modelling; uses community data	pacomm	Specialized program not general released; uses Arcview and Splus	no	SF
GDM-SS	generalised dissimilarity modelling; implementation for single species	pa	as for GDM	no	SF
GLM	regression; generalised linear model	pa	S-Plus, GRASP add-on	yes	AG,AL,JE
LIVES	multivariate distance	p	Specialized program not general released	no	JLi
MARS	regression; multivariate adaptive regression splines	pa	R, mda package plus new code to handle binomial responses	yes	JE, FH
MARS-COMM	as for MARS, but implemented with community data	pacomm	as for MARS	yes	JE
MARS-INT	as or MARS; interactions allowed	pa	as for MARS	yes	JE
MAXENT	maximum entropy	pa	Maxent	no	SP
MAXENT-T	maximum entropy with threshold features	pa	Maxent	no	SP
OM-GARP	rule sets derived with genetic algorithms; open modeller version	pa	new version of GARP not yet available	no	ATP

[1] p = only presence data used; pa = presence and some form of absence required – e.g. a background sample; comm = community data contribute to model fitting.
[2] any method can have an uncertainty estimate derived from bootstrapping the modelling; these data refer to estimates that are available as a statistical part of the method.



ECOGRAPHY 29: 129–151, 2006

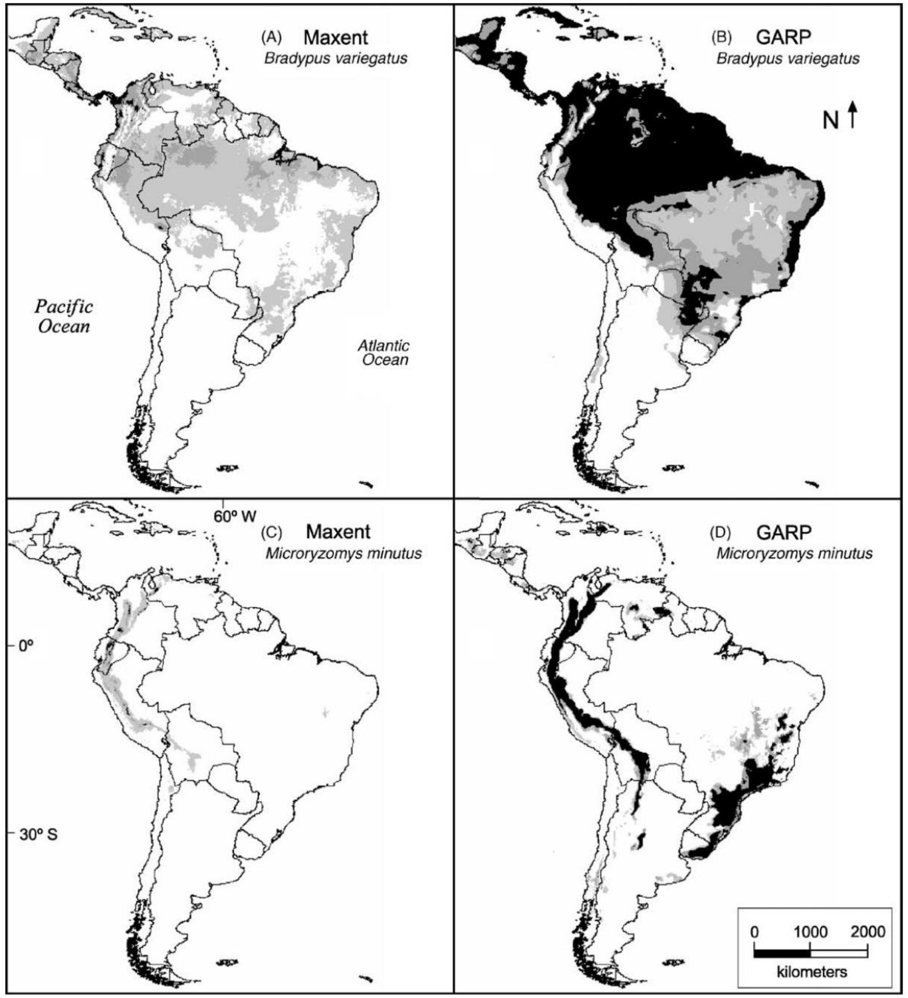
Novel methods improve prediction of species' distributions from occurrence data

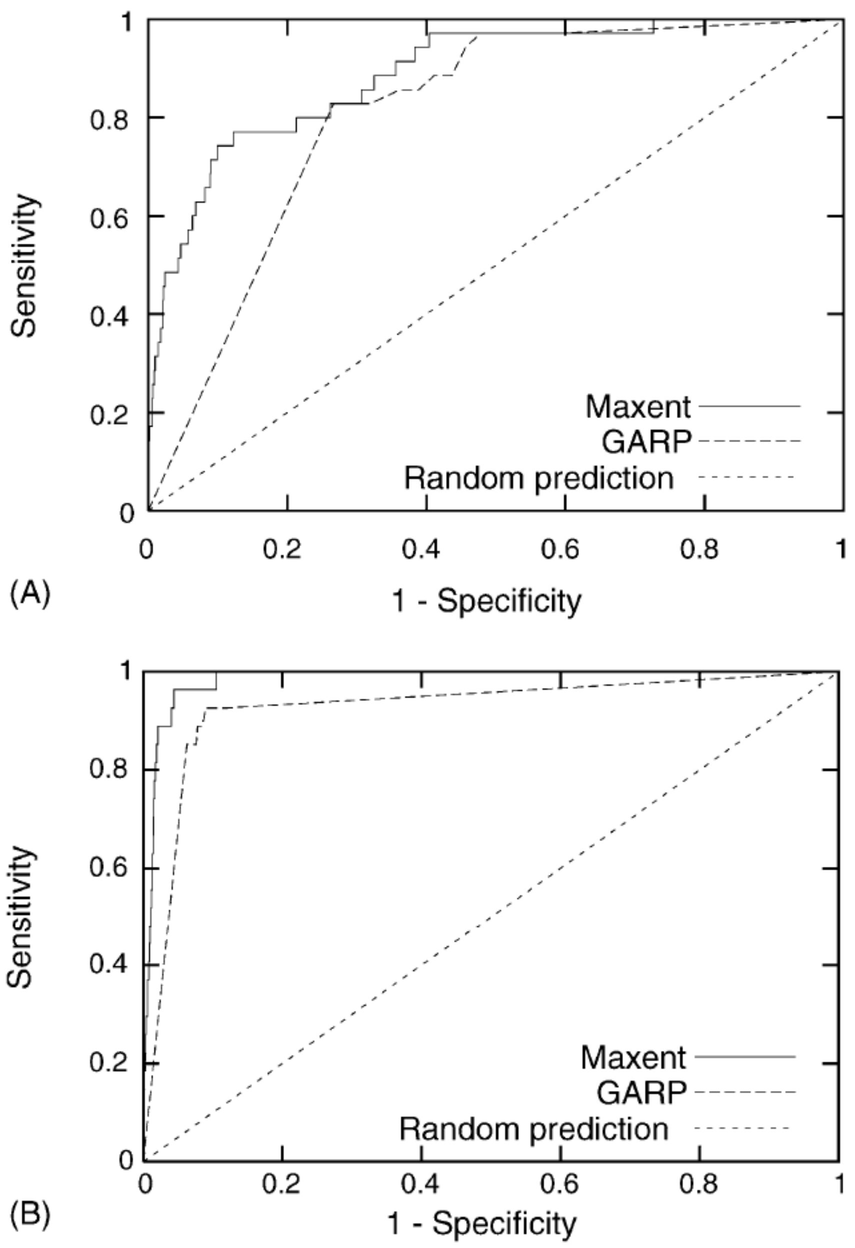
Jane Elith*, Catherine H. Graham*, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, Jin Li, Lucia G. Lohmann, Bette A. Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob McC. Overton, A. Townsend Peterson, Steven J. Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E. Schapire, Jorge Soberón, Stephen Williams, Mary S. Wisz and Niklaus E. Zimmermann

Available online at www.sciencedirect.com
SCIENCE DIRECT
Ecological Modelling 190 (2006) 231–259
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Maximum entropy modeling of species geographic distributions

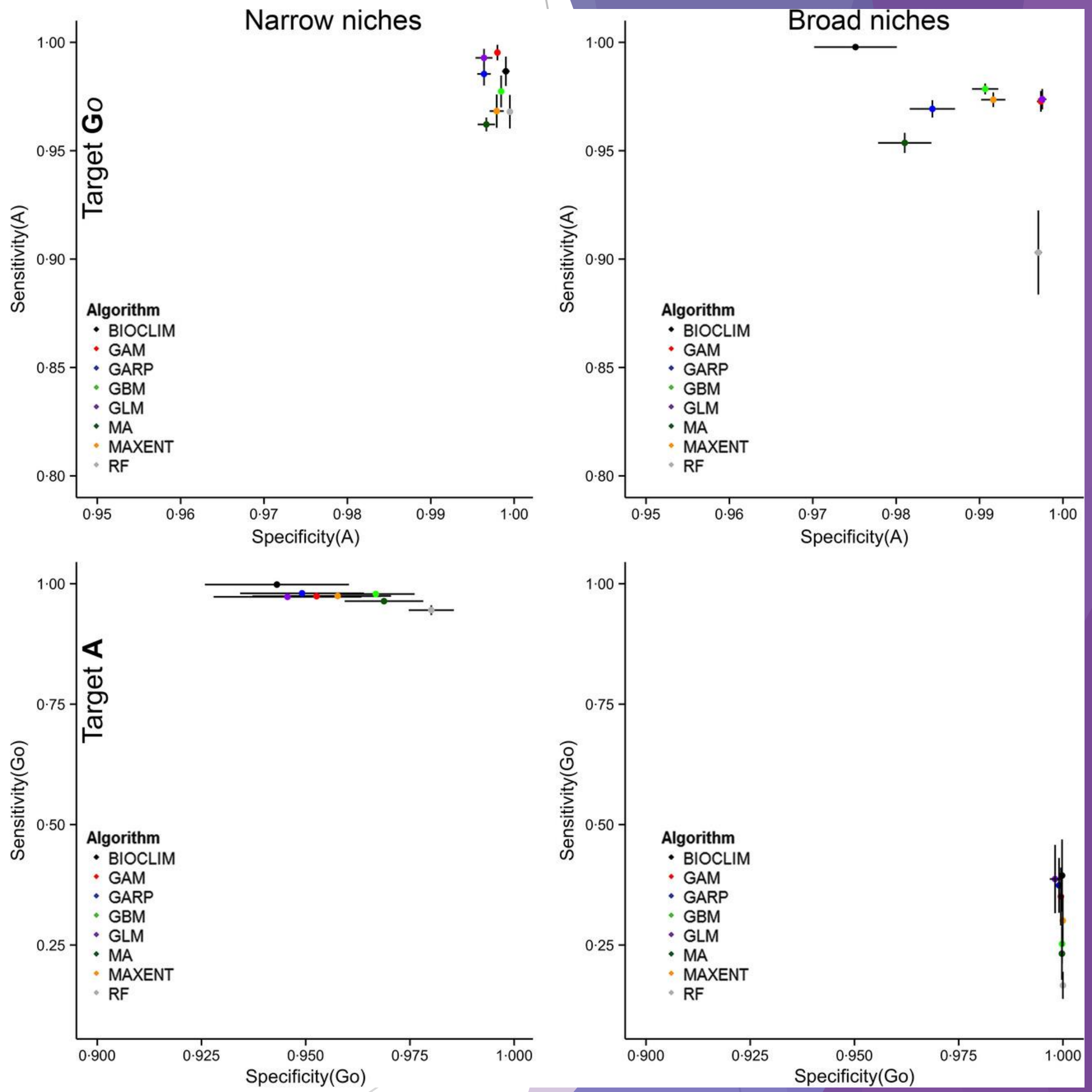
Steven J. Phillips[a,*], Robert P. Anderson[b,c], Robert E. Schapire[d]





Maxent

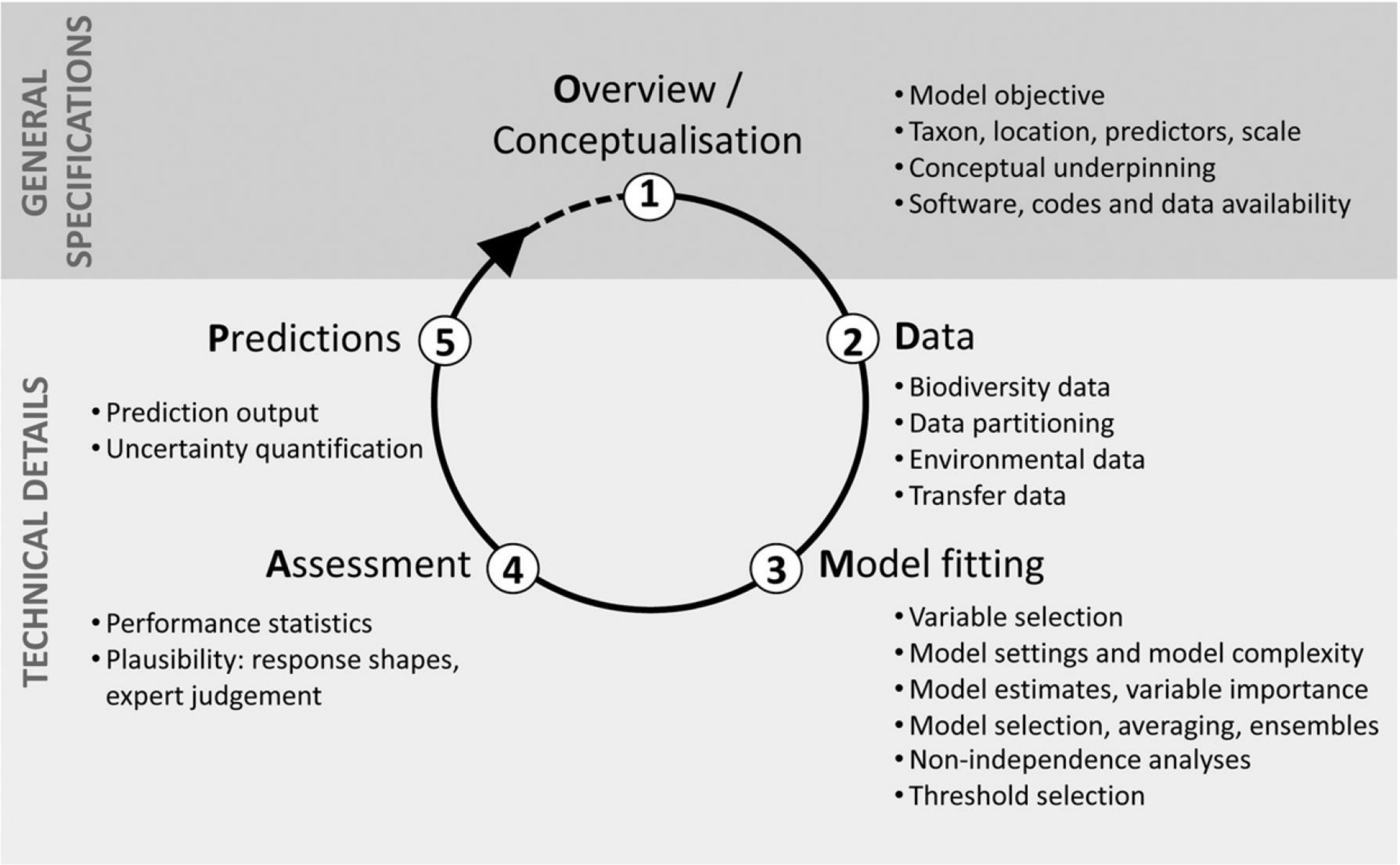
- New Approach to a Complex Problem in Distribution Ecology
- Chosen algorithm
- Good performance in evaluations
- But no algorithm is going to be the best in every situation



Qiao et al., 2015

Describe data and modeling choices

- ODMAP



Zurell et al., 2020

ODMAP section		ODMAP subsection	ODMAP elements
Overview	Obligatory	Authorship	Authors, contact email, title, doi
	Obligatory	Model objective/model purpose	SDM objective/purpose (inference, mapping, transfer), main target output
	Obligatory	Taxon	Focal taxon
	Obligatory	Location	Location of study area
	Obligatory	Scale of analysis	Spatial extent (lon/lat), spatial resolution, temporal extent/time period, temporal resolution, type of extent boundary (e.g. rectangular, natural, political)
	Obligatory	Biodiversity data overview	Observation type, response/data type
	Obligatory	Type of predictors	Climatic, topographic, edaphic, habitat, etc.
	Obligatory	Conceptual model/hypotheses	Hypotheses about biodiversity-environment relationships
	Obligatory	Assumptions	State critical model assumptions (cf. Table 2)
	Obligatory	SDM algorithms	Model algorithms, justification of model complexity, is model averaging/ ensemble modelling used?
	Obligatory	Model workflow	Brief description of modelling steps
	Obligatory	Software, codes and data	Specify software, availability of codes, availability of data
Data	Obligatory	Biodiversity data	Taxon names, taxonomic reference system, ecological level, biodiversity data sources, sampling design, sample size per taxon, country/region mask, details on scaling, data cleaning/filtering, absence data collection, pseudo-absence and background data, potential errors and biases in data
	Objective: mapping/interpolation; Objective: forecast/transfer	Data partitioning	Selection of training data (for model fitting), validation data and test (truly independent) data
	Optional/context dependent	Predictor variables	State predictor variables used, data sources, spatial resolution and extent of raw data, map projection, temporal resolution and extent of raw data, data processing and scaling, measurement errors and bias, dimension reduction
	Objective: forecast/transfer	Transfer data for projection	Data sources, spatial resolution and extent, temporal resolution and extent, models and scenarios used, data processing and scaling, quantification of novel environments
Model	Optional/context dependent	Variable pre-selection	Details on pre-selection of variables
	Obligatory	Multicollinearity	Methods for identifying and dealing with multicollinearity
	Obligatory	Model settings/model complexity	Models settings for all selected algorithms and for extrapolation beyond sample range
	Obligatory	Model estimates	Model coefficients, variable importance
	Optional/context dependent	Model selection/model averaging/ensembles	Model selection strategy, method for model averaging, ensemble method
	Obligatory	Non-independence correction/analyses	Spatial autocorrelation in residuals, temporal autocorrelation in residuals, nested data
	Objective: mapping/interpolation; Objective: forecast/transfer	Threshold selection	Details on threshold selection
Assessment	Obligatory	Performance statistics	Performance statistics estimated on training data, on validation data and on test (truly independent) data
	Obligatory	Plausibility check	Response plots; expert judgements (e.g. map display)
Prediction	Objective: mapping/interpolation; Objective: forecast/transfer	Prediction output	Prediction unit; post-processing steps
	Objective: mapping/interpolation; Objective: forecast/transfer	Uncertainty quantification	Uncertainty through algorithms, input data, parameters, scenarios; visualisation/treatment of novel environments

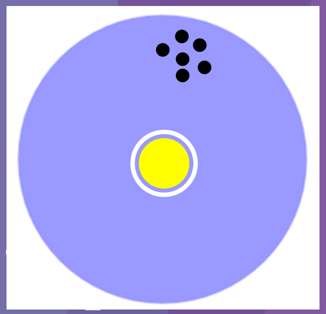
Obligatory; Objective: mapping/interpolation; Objective: forecast/transfer; Optional/context dependent.

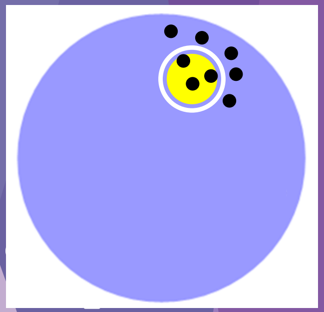
Zurell et al., 2020

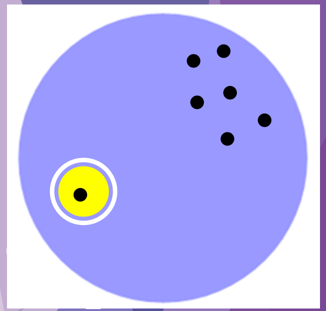
UNCERTAINTY

What is uncertainty?

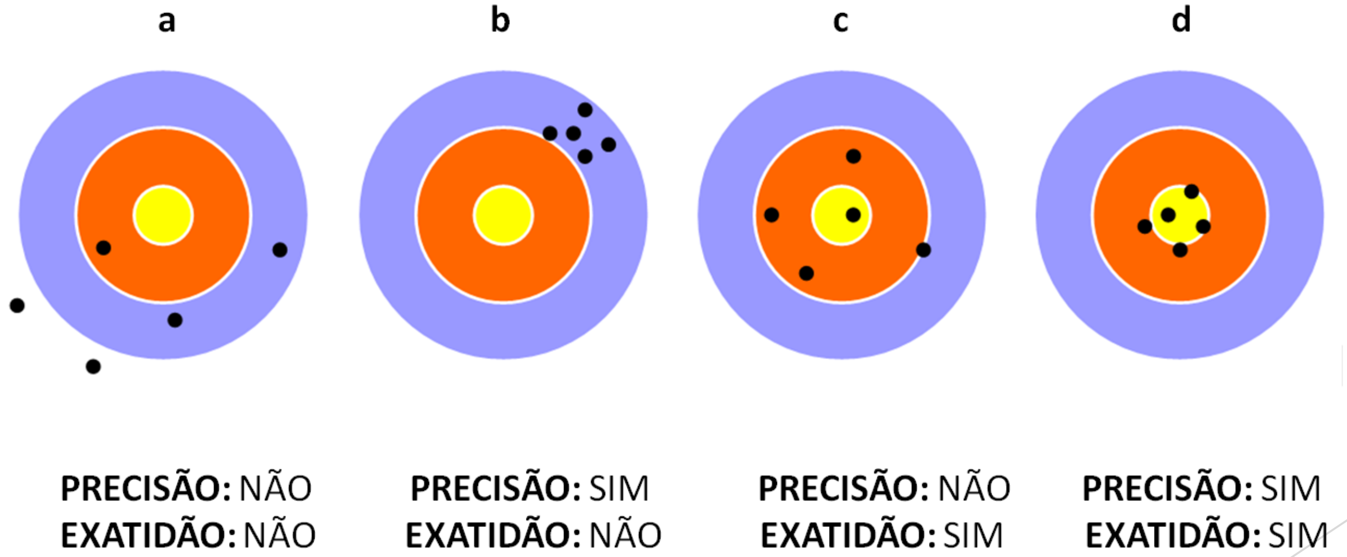
- What if we don't know the reality to compare with the models?
  - You can't measure the variation
    - Uncertainty!







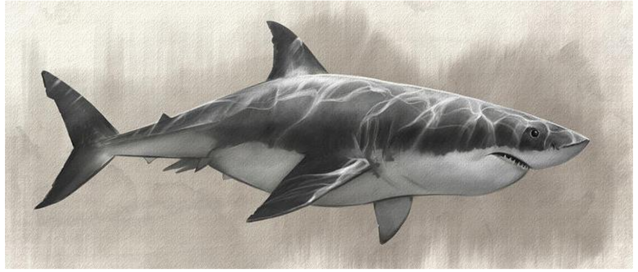
- Lack of knowledge of how well a model represents reality
- It is associated with, but not the same thing as, the error or variation of the models

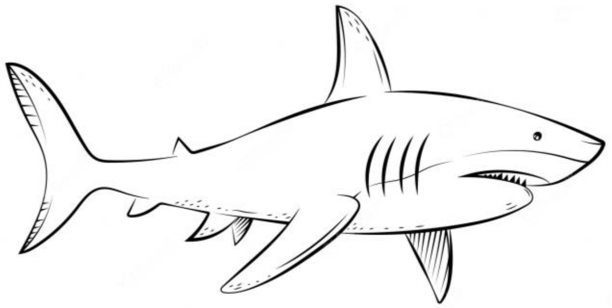


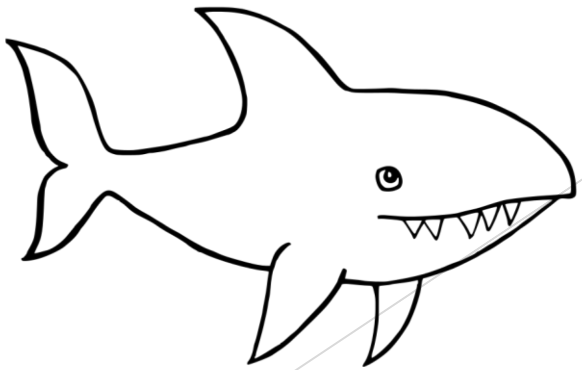
Uncertainty in **a** and **c** >>> uncertainty in **b** and **d**

Sources of uncertainty

- What points of occurrence?
- Accuracy of the points of occurrence?
- What modeling parameters?
- Which algorithms?
- What environmental data?

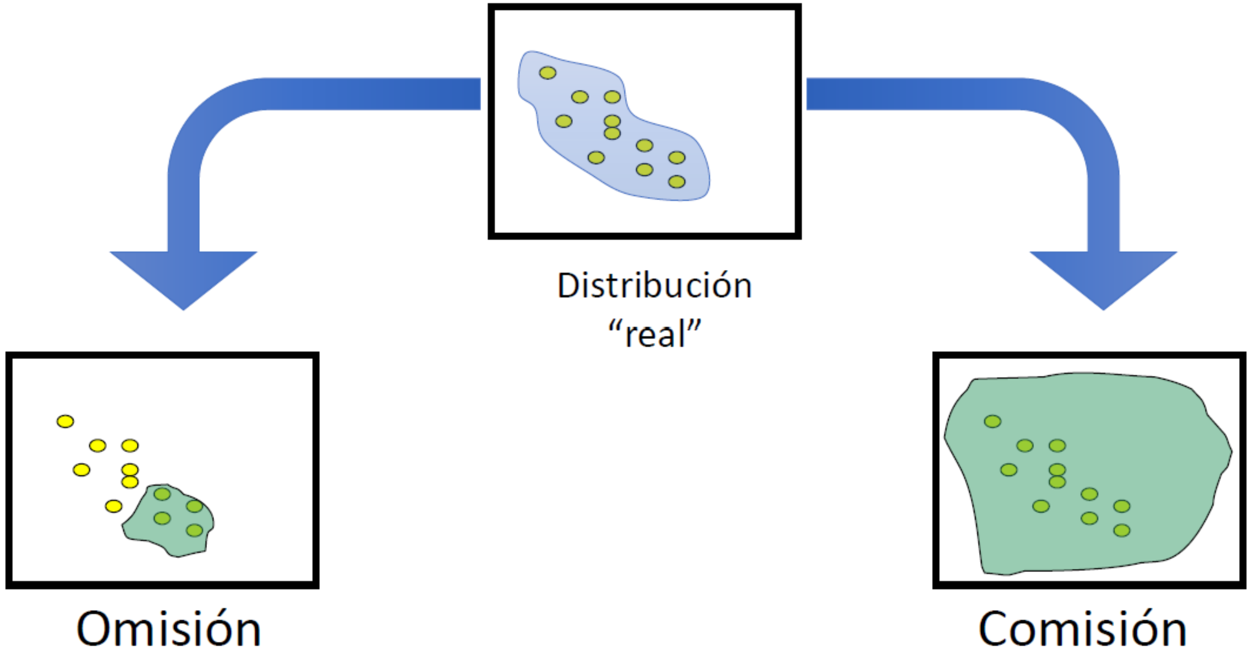






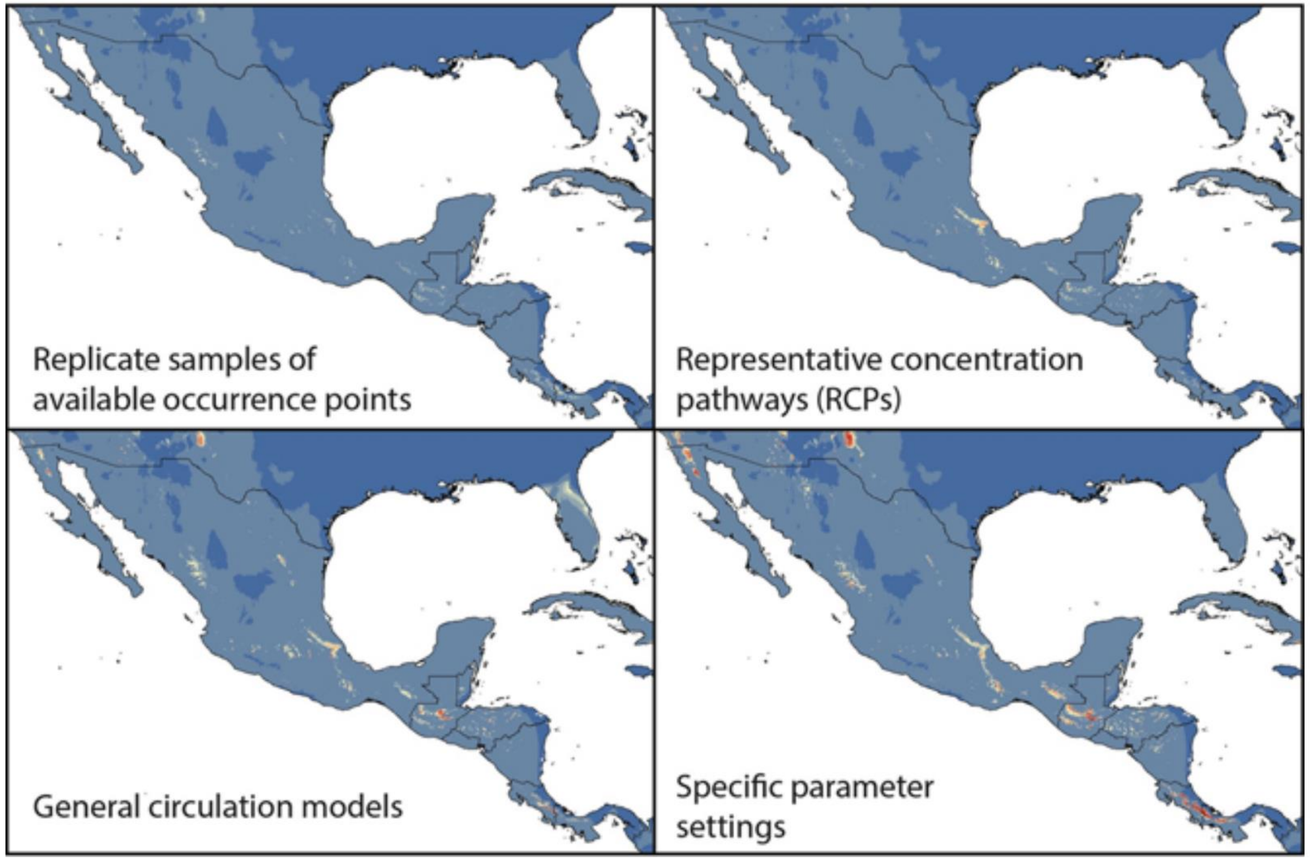
Mistakes in niche models

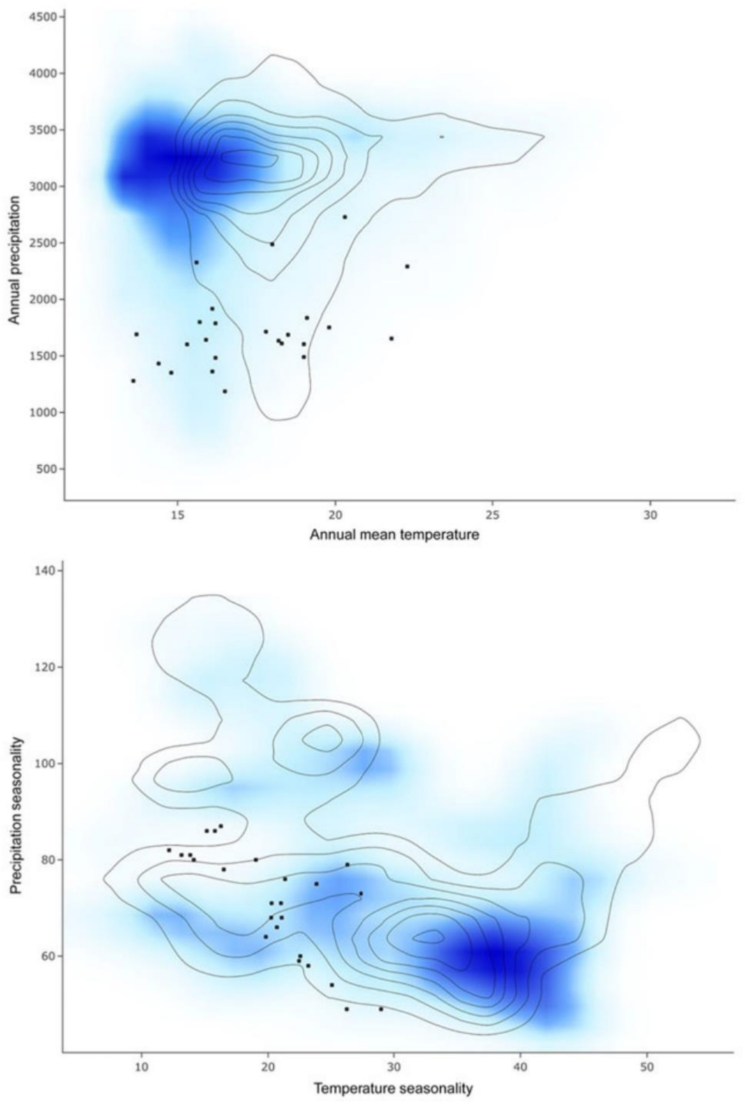
- In the data
  - Taxonomic
  - Geographical
  - Absence?
- In the procedures
- In the biology and ecology of species



Variation origin

- Sampling
- Environmental data
- Extrapolation (time and/or space)
- Parameters

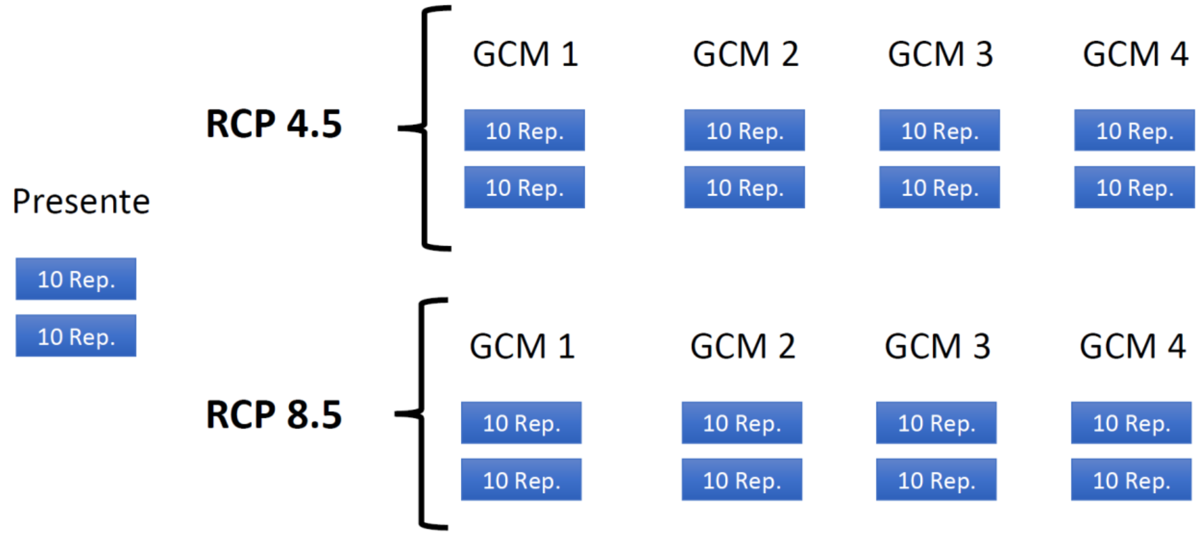




Peterson et al., 2018

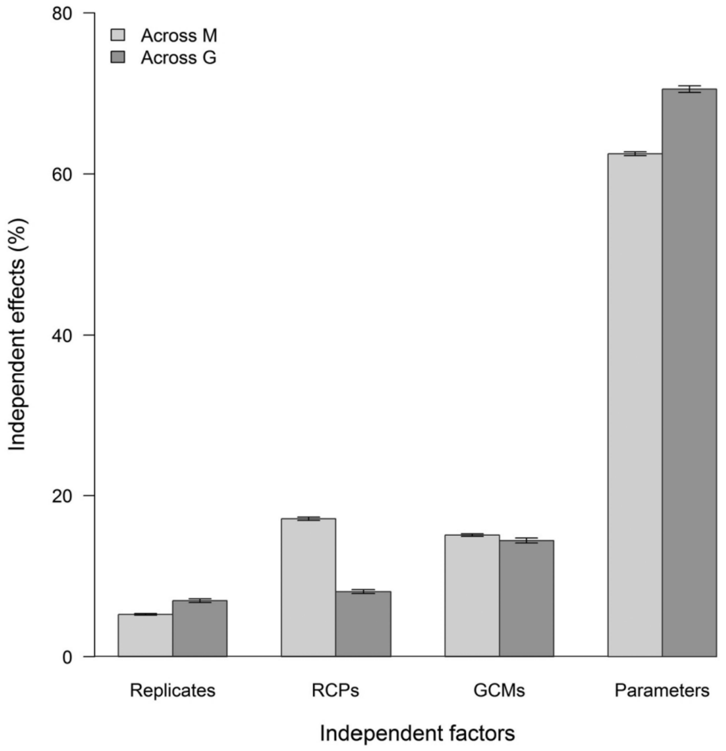
Statistical quantification of variation

1. Data extraction and organization
   - Random sampling points in the study area
   - Point data extraction
   - Group data according to factors

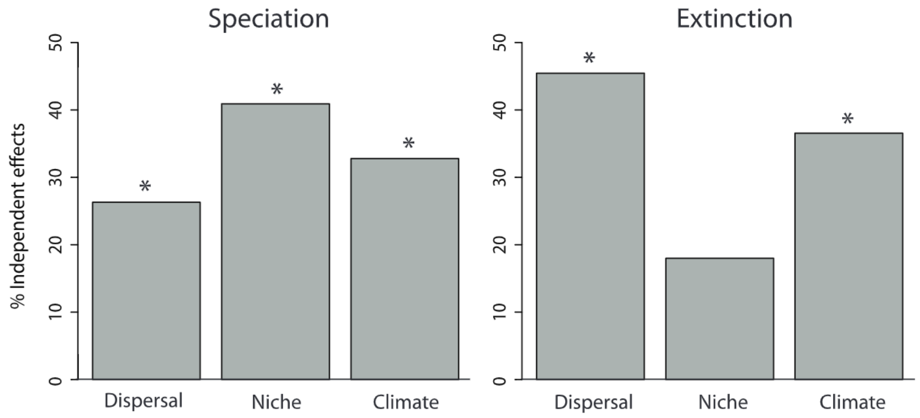


Peterson et al., 2018

Statistical quantification of variation



2. Hierarchical Partition Analysis of Variance
   - Independent and combined effect of each of the factors on the recorded variance
3. Bootstrap
   - Random sampling of data to detect variation in measured effects
4. Statistical significance
   - Comparison of measured effects with a null distribution created by randomizing data across factors



Araújo & Guisan, 2006
Qiao et al., 2016
Peterson et al., 2018

So…

- Errors and variations generate uncertainty in the NMS
- Uncertainty can be reduced by avoiding mistakes, but it cannot be eliminated
- Variation is an important part of models and should be considered
- Representing the variation allows you to reflect levels of uncertainty;
  - it is better to represent it than to assume that a single model is showing reality