Species Distribution models (Theory and examples)

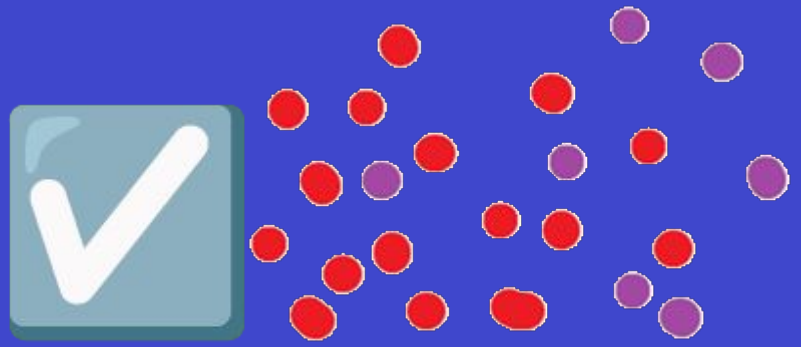
Alliance Bioversity International & CIAT

Before using the gap analysis methodologies (Summary)

- Definition of a model
- Niche concept
- Species distribution models
- Inputs curation
- Examples using R

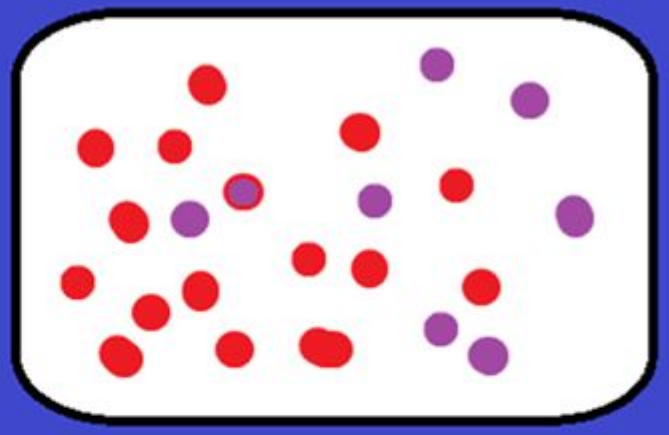
The workflow for a gap analysis

A



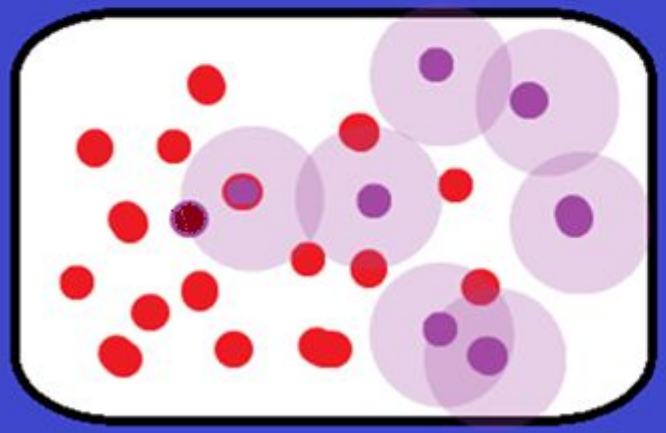
Species occurrences

B



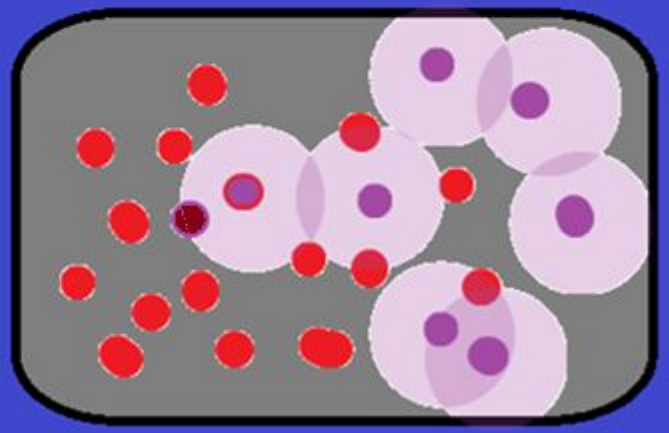
Realized niche

C



Sampled germplasm

D



Where to collect

Germplasm

Other coordinates

Collected germplasm area

Potential area to be collected

CROP TRUST

Alianza Bioversity & CIAT

CGIAR

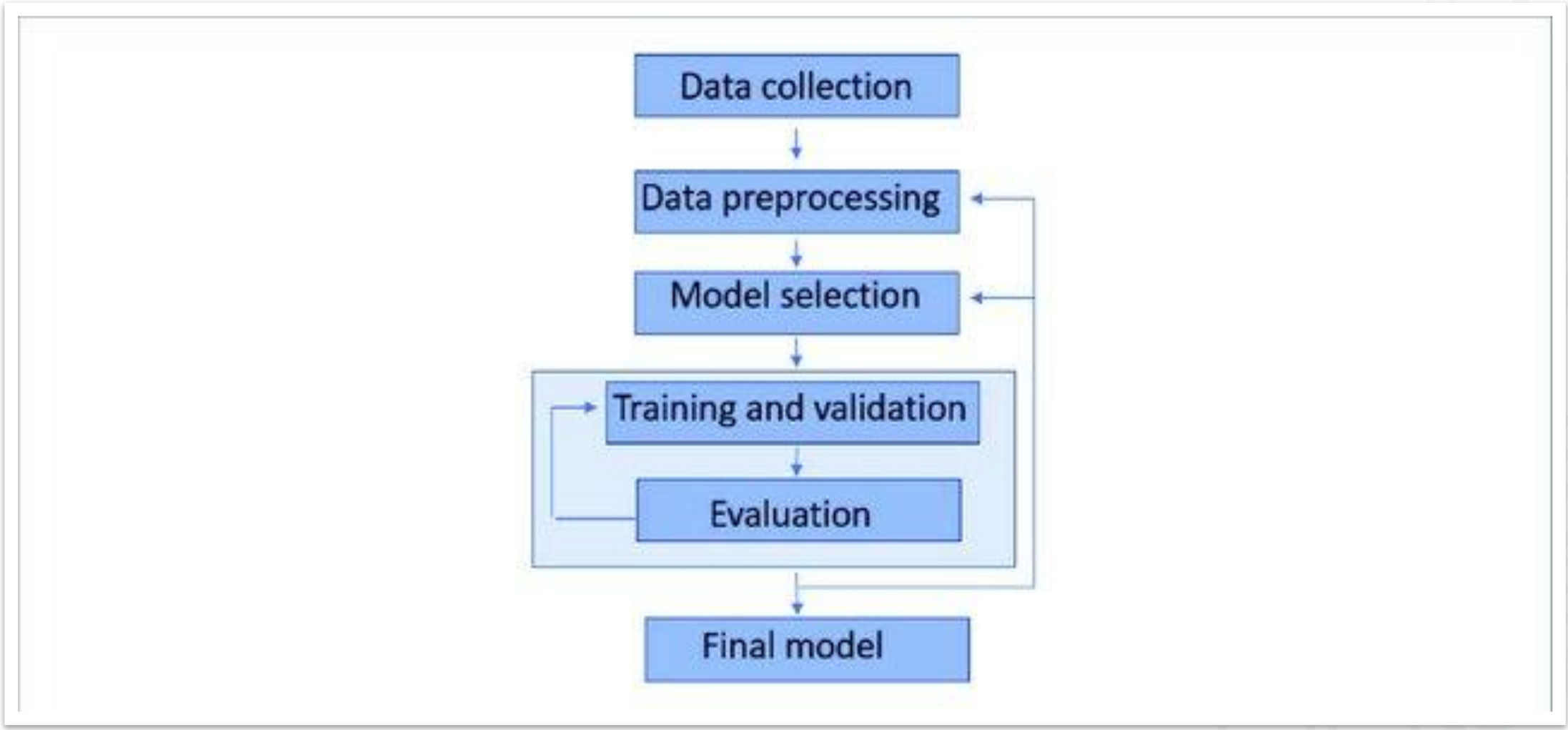
What is a model?

What are Models? Relationships to Theory

A theory is an *abstraction* of some phenomena, usually '*real*' but sometimes imagined in a form that makes the *simplification* or abstraction clear. A model is a simplification of *reality* which takes the theoretical abstractions and puts it into a form that we can manipulate. Simulation is often used to characterise this process of implementation.

In everything we do, we theorise, and more and more frequently we build models to demonstrate theory.

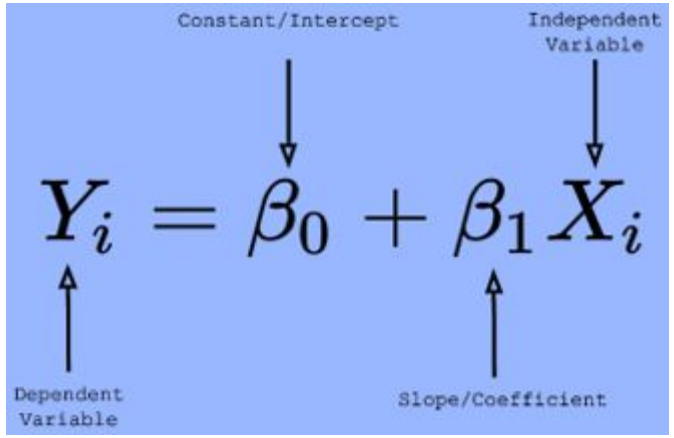
How a model workflow looks like?

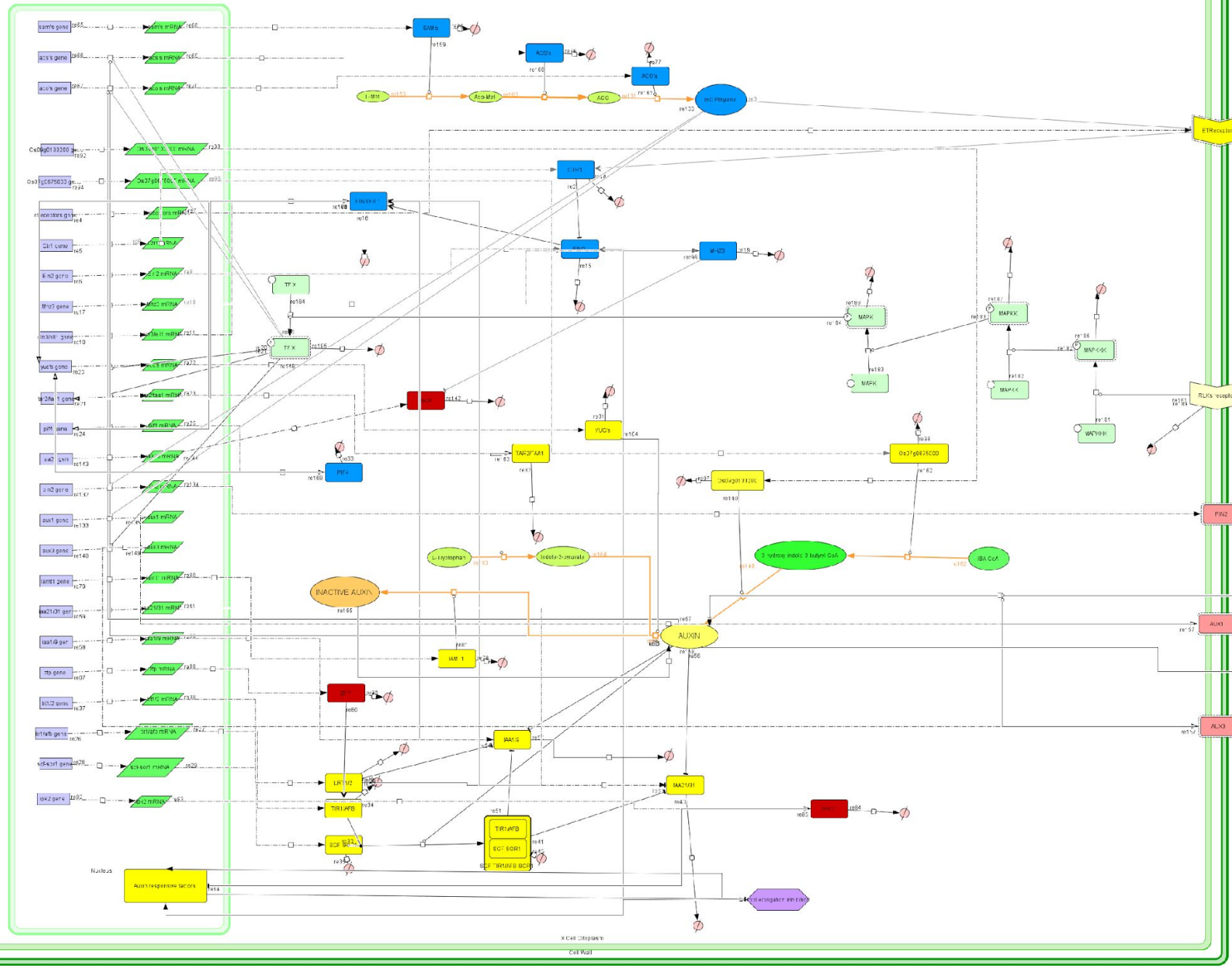


https://iopscience.iop.org/article/10.1088/2633-1357/

Models can be simple or complex

simple





omplex

127 molecules

117 reactions

- Ethylene proteins
- Aluminum
- MAPK
- IBA
- Auxin proteins
- Genes
- Transcripts
- Auxin receptors
- Phytohormone related metabolites

How can obtain the geographical distribution of a species using environmental data?

What is a niche?

American Museum of Natural History
Center for Biodiversity and Conservation
Network of Conservation Educators & Practitioners

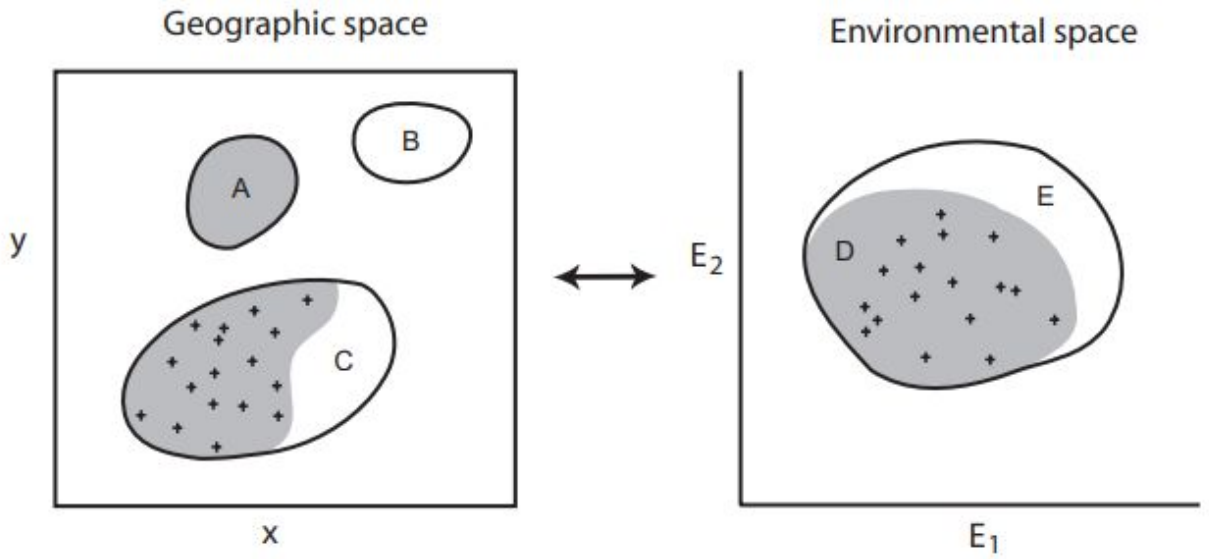
Species' Distribution Modeling for Conservation Educators and Practitioners

Author(s): Richard G. Pearson

Source: *Lessons in Conservation*, Vol. 3, pp. 54-89

Niche and species distribution are not synonyms!

If the environmental conditions encapsulated within the fundamental niche are plotted in geographical space, then we have the potential distribution



Hutchinson's concept of niche: hypervolume of the multidimensional space defined by a set of environmental variables, within which a species can maintain a viable population

\+ Known species occurrence record

Occupied distributional area, $\mathbf{G}_O$ (left panel)/
Occupied niche space, $\mathbf{E}_O$ (right panel)

Abiotically suitable area, $\mathbf{G}_A$ (left panel)/
Scenopoetic existing fundamental niche, $\mathbf{E}_A$ (right panel)

What is a niche?

Biodiversity and Climate Change: Integrating Evolutionary and Ecological Responses of Species and Communities

Sébastien Lavergne,[1] Nicolas Mouquet,[2] Wilfried Thuiller,[1] and Ophélie Ronce[2]

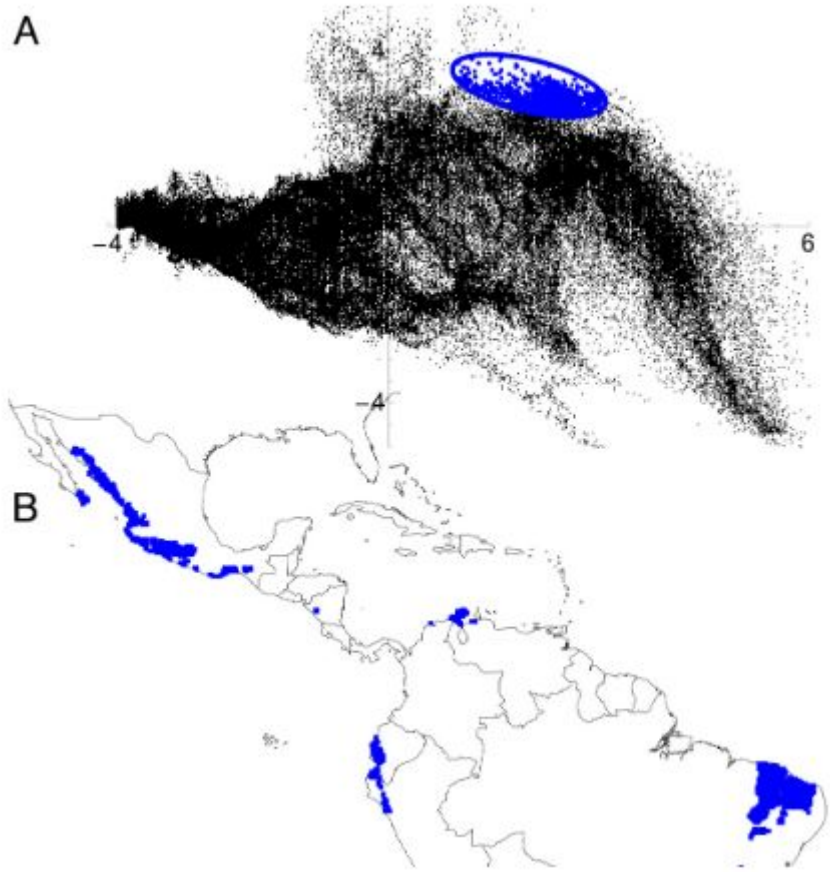
NAS

Niches and distributional areas: Concepts, methods, and assumptions

Jorge Soberón[a,1] and Miguel Nakamura[b]

[a]Biodiversity Institute, University of Kansas, Dyche Hall, 1345 Jayhawk Boulevard, Lawrence, KS 66045; and [b]Centro de Investigación en Matemáticas, A. C. Jalisco s/n, Col. Valenciana, Guanajuato, 36240, México

Edited by Elizabeth A. Hadly, Stanford University, Stanford, CA, and accepted by the Editorial Board August 28, 2009 (received for review March 31, 2009)

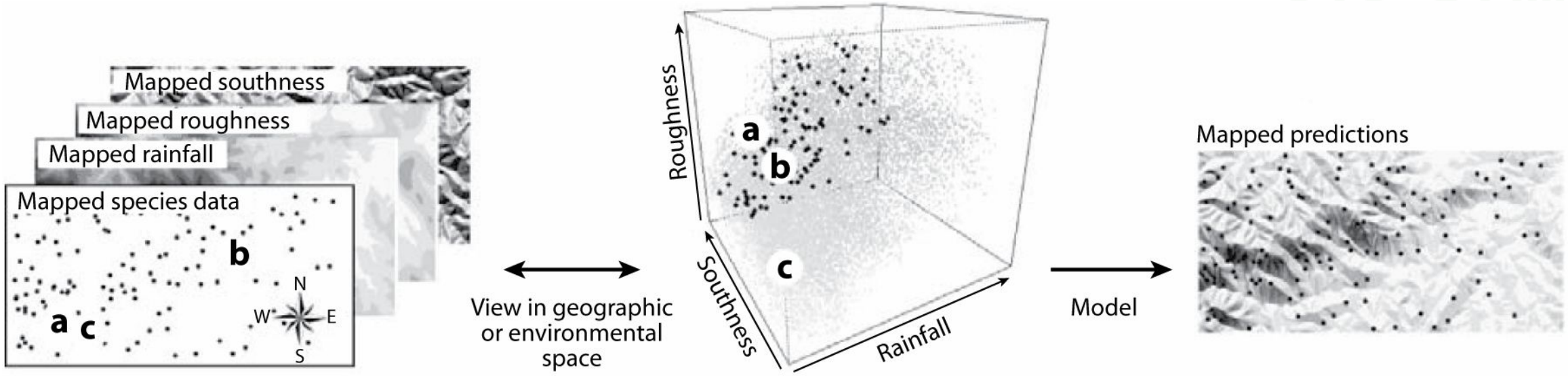


Hutchinson's concept of niche: hypervolume of the multidimensional space defined by a set of environmental variables, within which a species can maintain a viable population

Habitat suitability models: statistical models that predict species' potential distributions by combining known occurrence records with environmental data (so-called niche-based models)

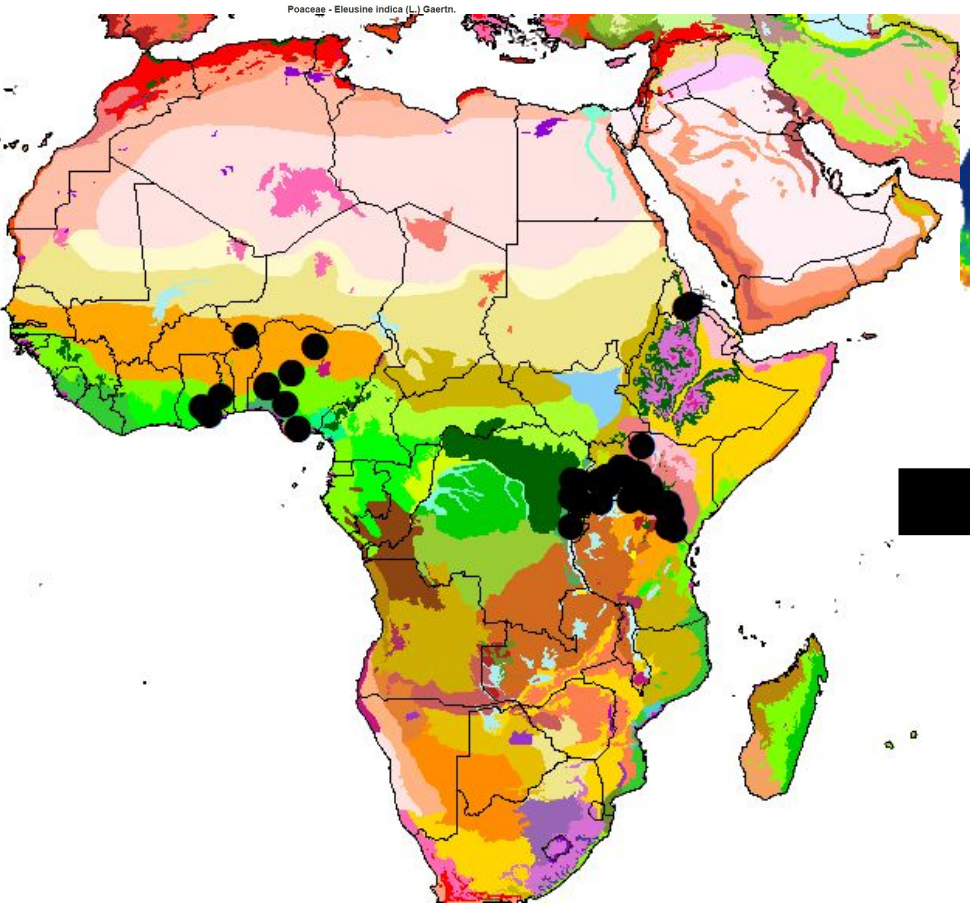
Adaptive potential: the capacity of species or populations to evolve in response to changes in environmental conditions whether biotic or abiotic

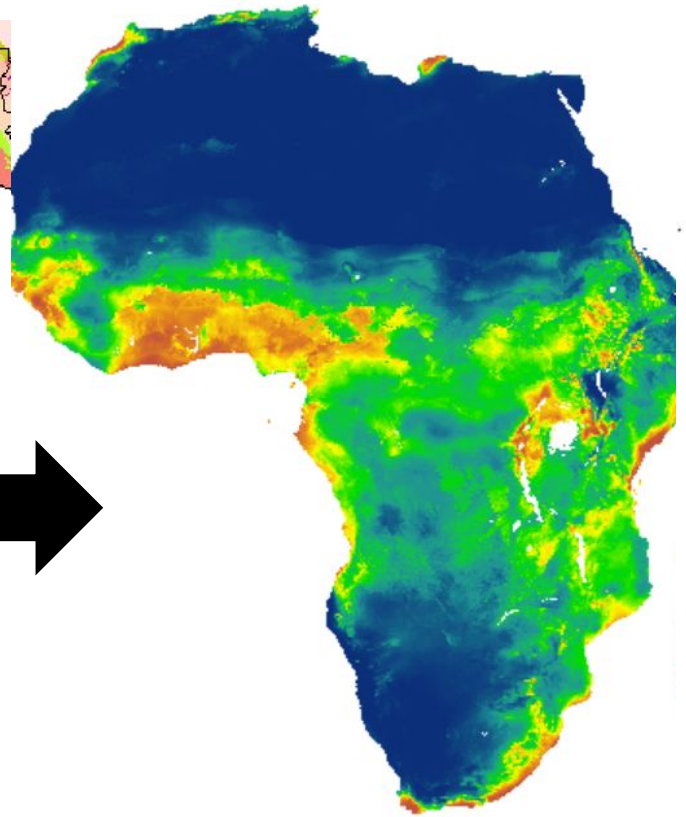
The species distribution model workflow:

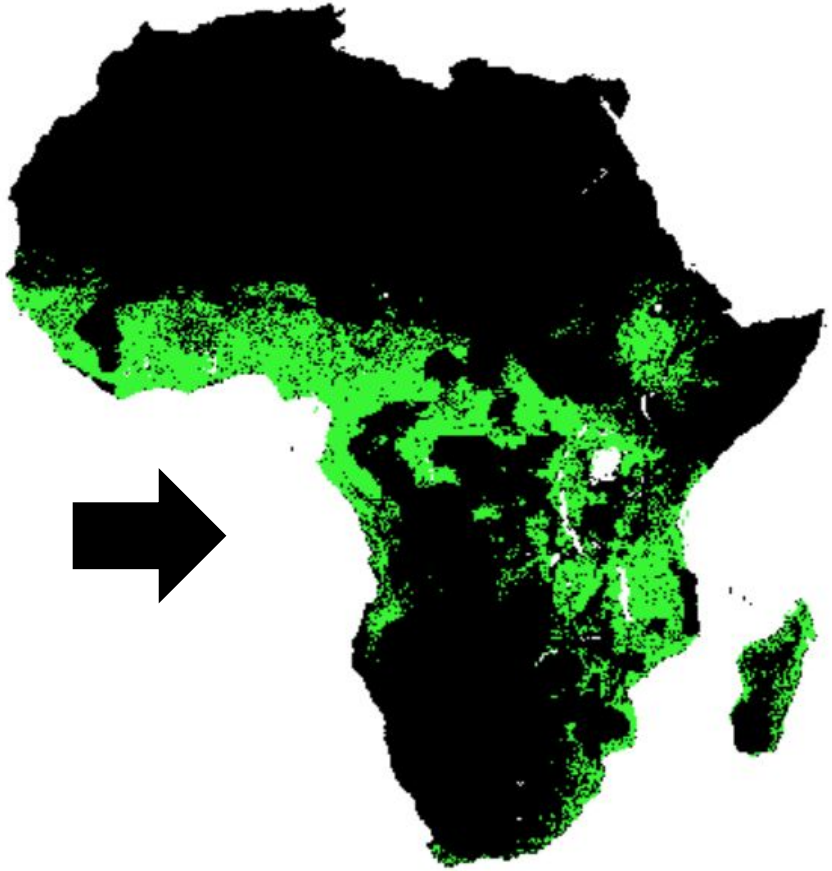


Elith J, Leathwick JR. 2009.
Annu. Rev. Ecol. Evol. Syst. 40:677–97

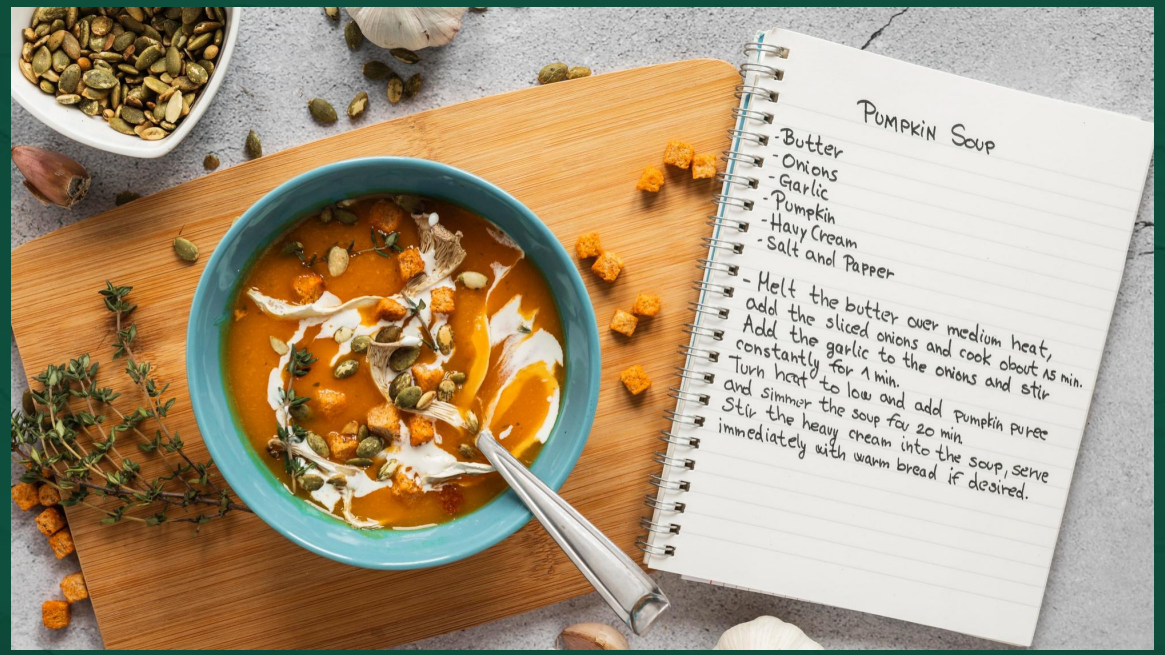
An example using *E. indica*







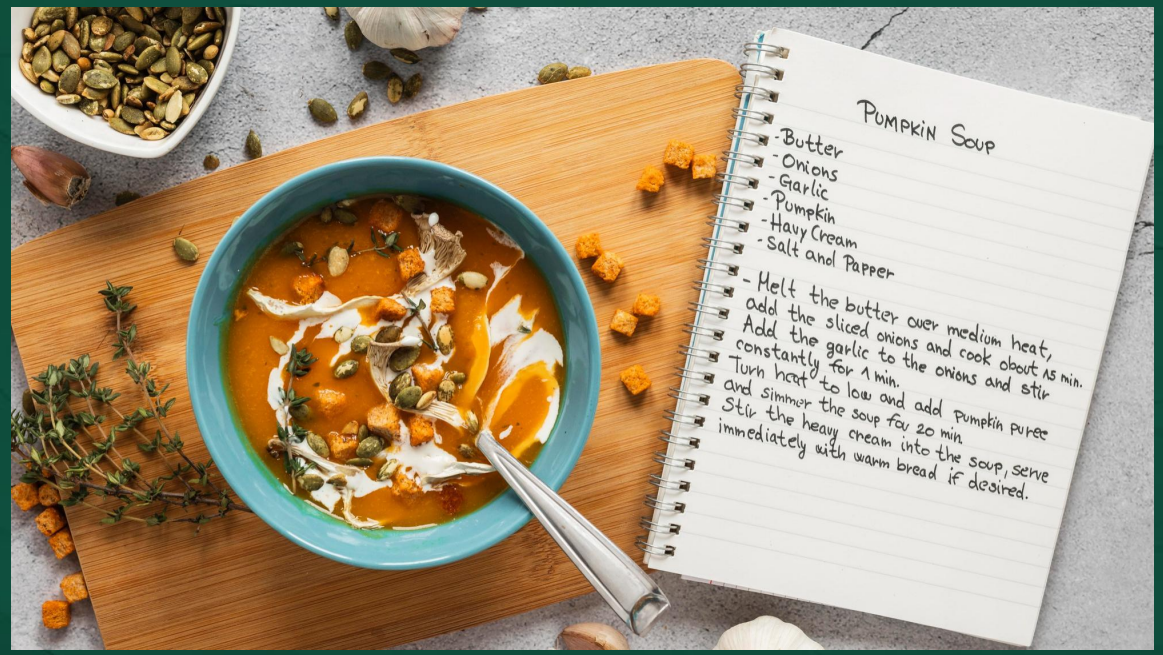
What could be a recipe to create an SDM?



https://www.freepik.com/free-photo/top-view-food-ingredients-with-vegetable-soup-bowl-notebook_12253324.htm#fromView=keyword&page=1&position=0&uuid=d2696821-6a07-481e-8172-69b3b091065c&query=Food+recipe

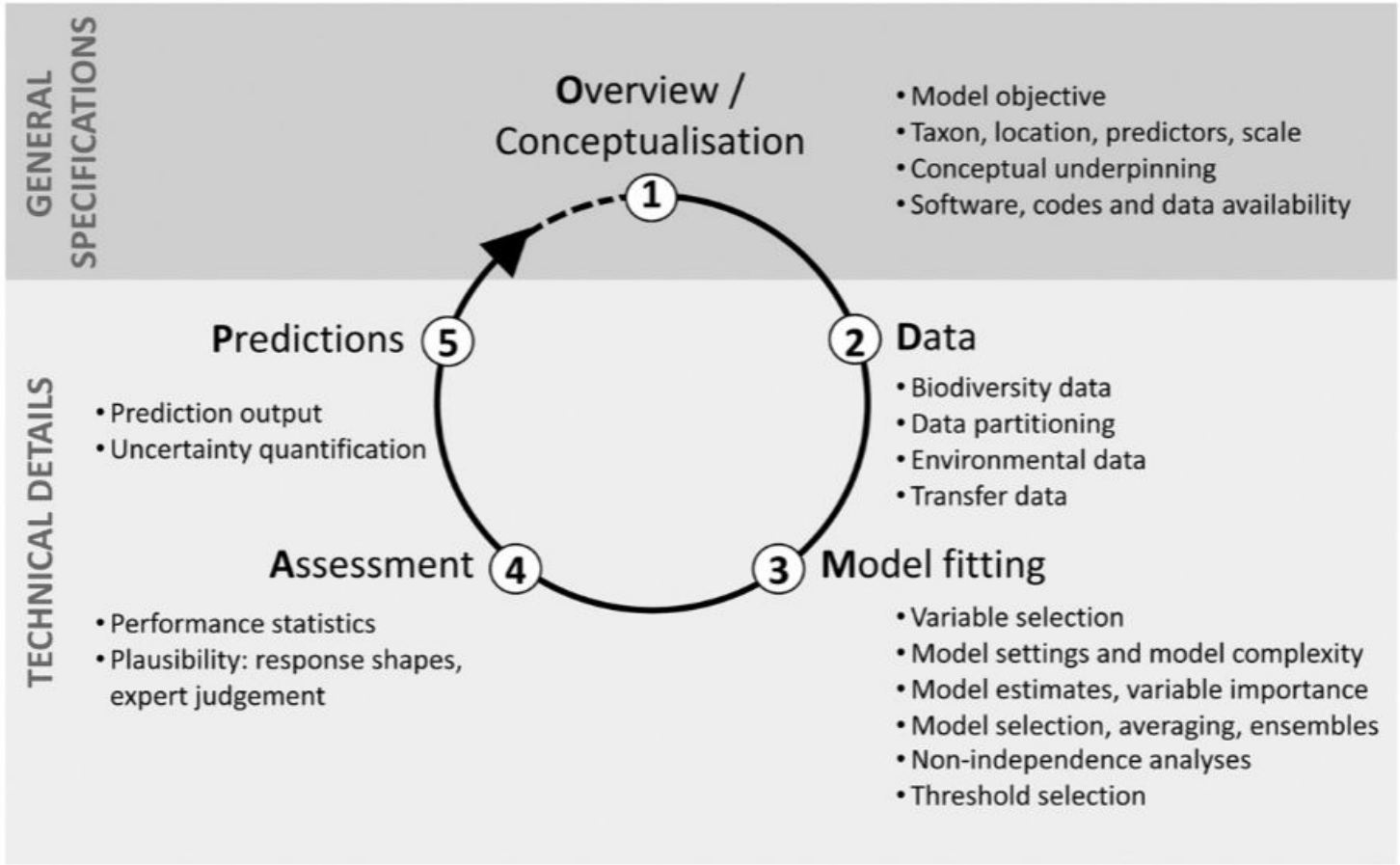
ODMAP!

- Overview
- Data
- Model fitting
- Assessment
- Predictions



https://www.freepik.com/free-photo/top-view-food-ingredients-with-vegetable-soup-bowl-notebook_12253324.htm#fromView=keyword&page=1&position=0&uuid=d2696821-6a07-481e-8172-69b3b091065c&query=Food+recipe

The recipe for a species distribution model (ODMAP)



Zurell et al., 2020

Preprocessing

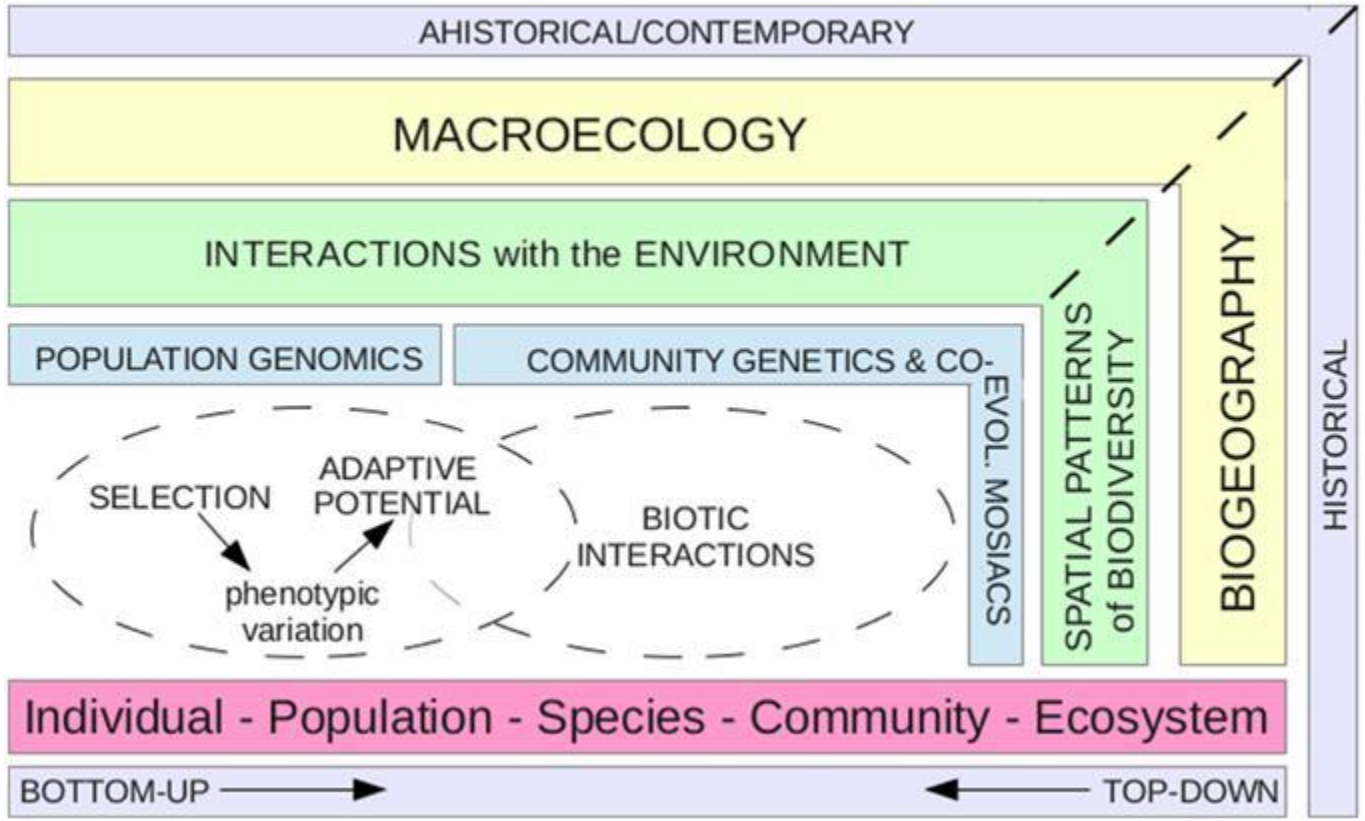
Zurell et al., 2020

Table 1. The five main ODMAP sections and list of ODMAP elements. The full ODMAP v1.0 checklist is available in Supplementary material Table A1.

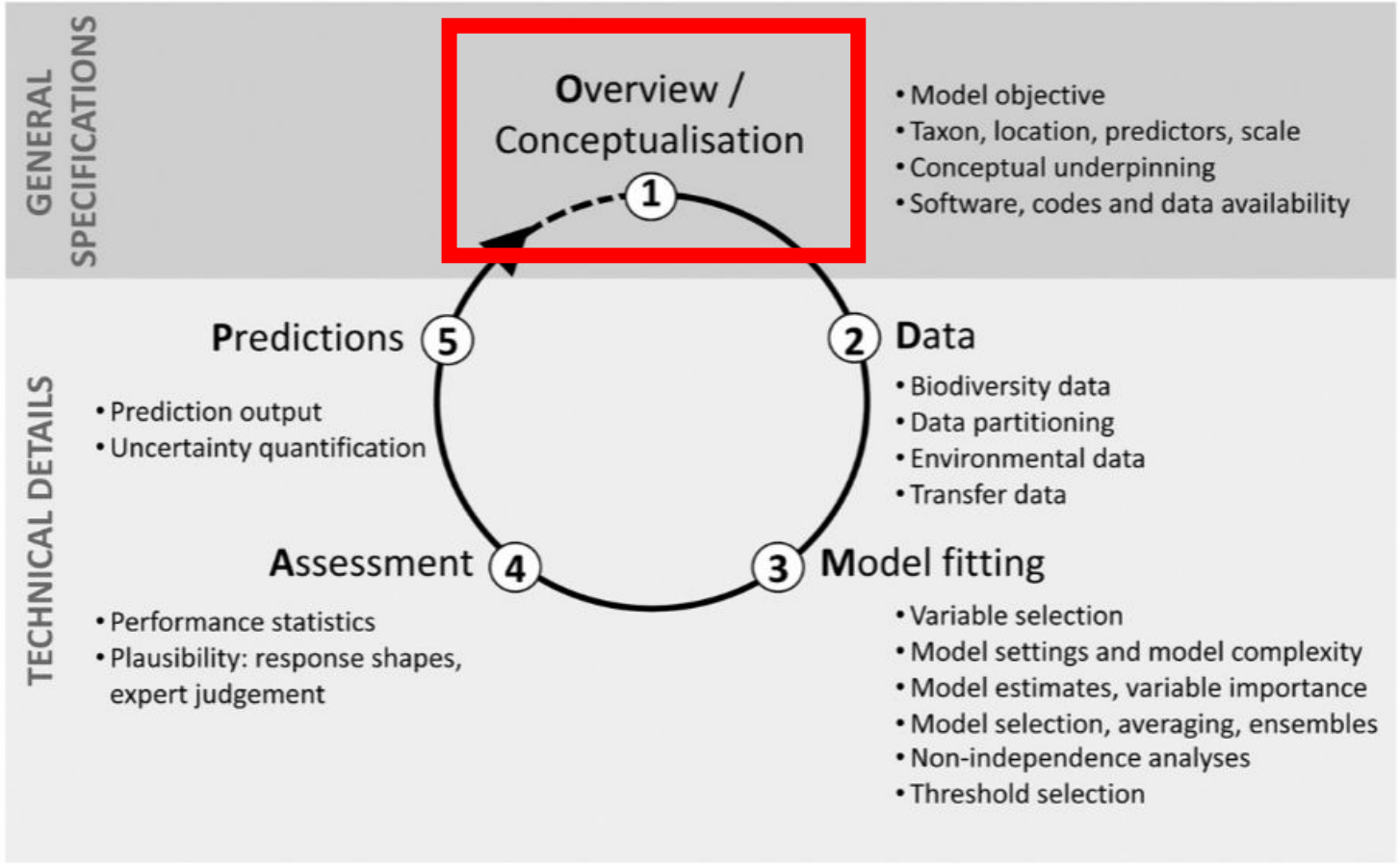
ODMAP section	ODMAP subsection	ODMAP elements
Overview	Authorship	Authors, contact email, title, doi
	Model objective/model purpose	SDM objective/purpose (inference, mapping, transfer), main target output
	Taxon	Focal taxon
	Location	Location of study area
	Scale of analysis	Spatial extent (lon/lat),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

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

Frontiers | Integrating a Population Genomics Focus into Biogeographic and Macroecological Research



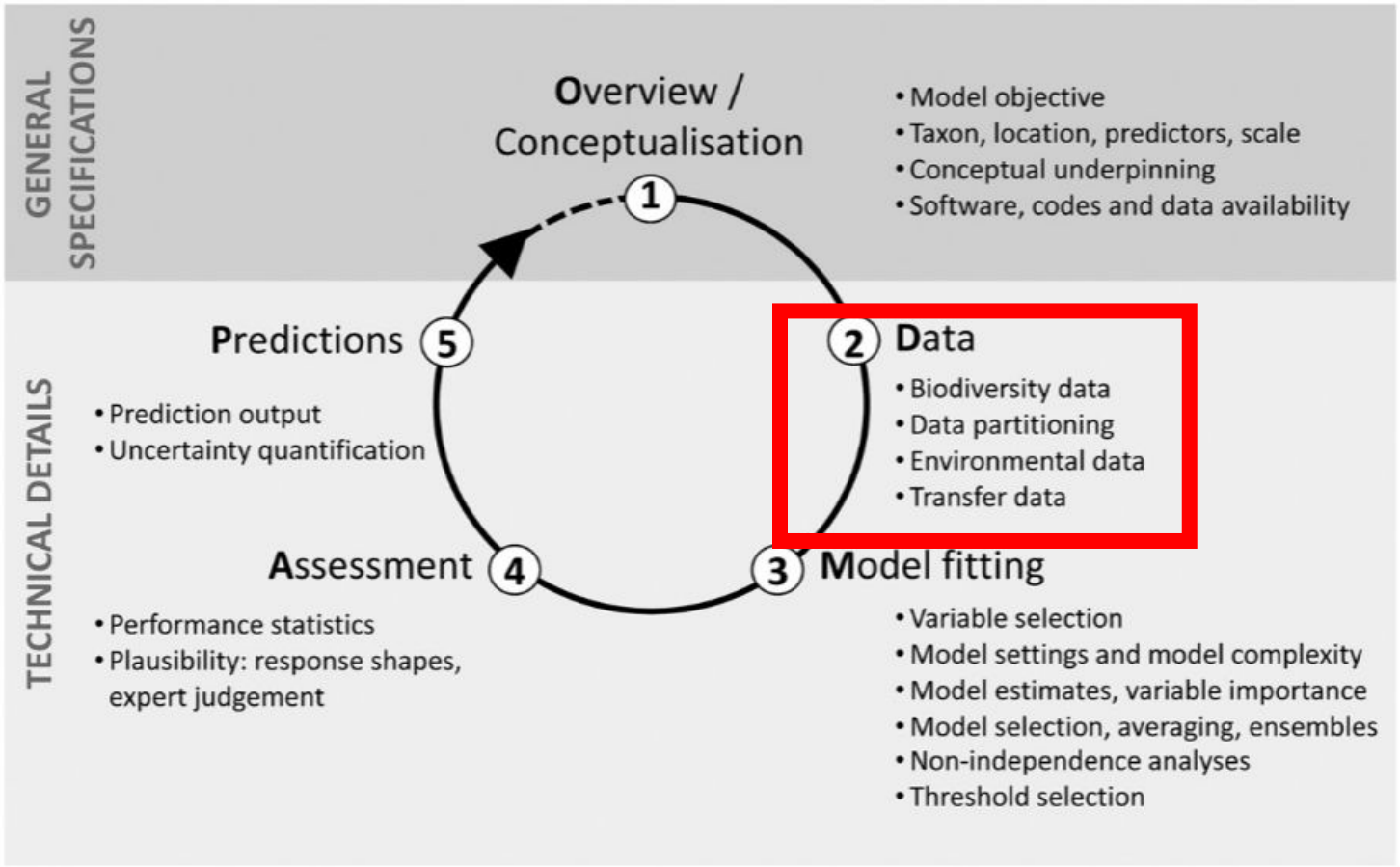
Overview: What is the purpose of make a SDM for gap analysis?



Exercise:

- What is the purpose of a SDM for a gap analysis?
- What geographical area information should I use?
- What geographical extension should I use?

Data (preprocessing)



Basics of data curation

Activity	What respond?
-Taxonomic data	-The species that I am analyzing have other scientific names or it is even a valid name?
-Geographical coordinates	-How accurate is the geographical information found for my species?
-Environmental information	-What environmental information can be used to model the species distribution of the species?

Taxonomic data

wfo-0000888897

Pennisetum glaucum (L.) R.Br.

Prodr. Fl. Nov. Holland. : 195 (1810)

This name is a synonym of Cenchrus americanus (L.) Morrone by Poaceae.

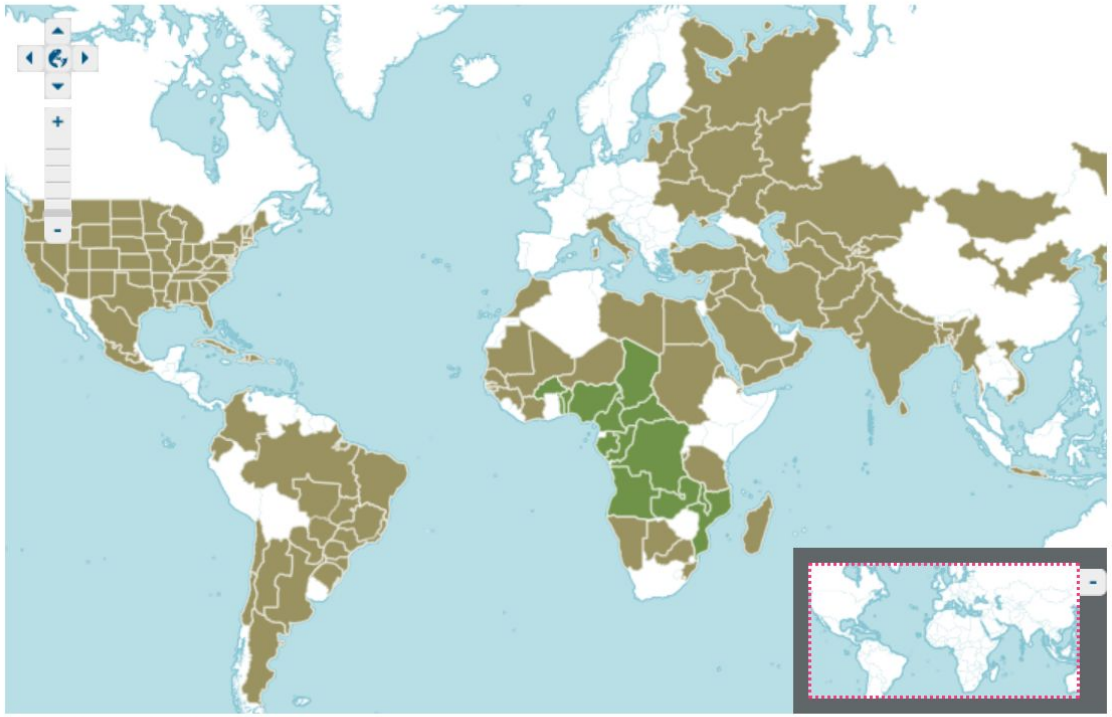
The record derives from **WCSP** (data supplied on 2024-06-04) which reports it as a synonym of Cenchrus americanus (L.) Morrone (record 432725)

wfo-0000917339

Cenchrus americanus (L.) Morrone

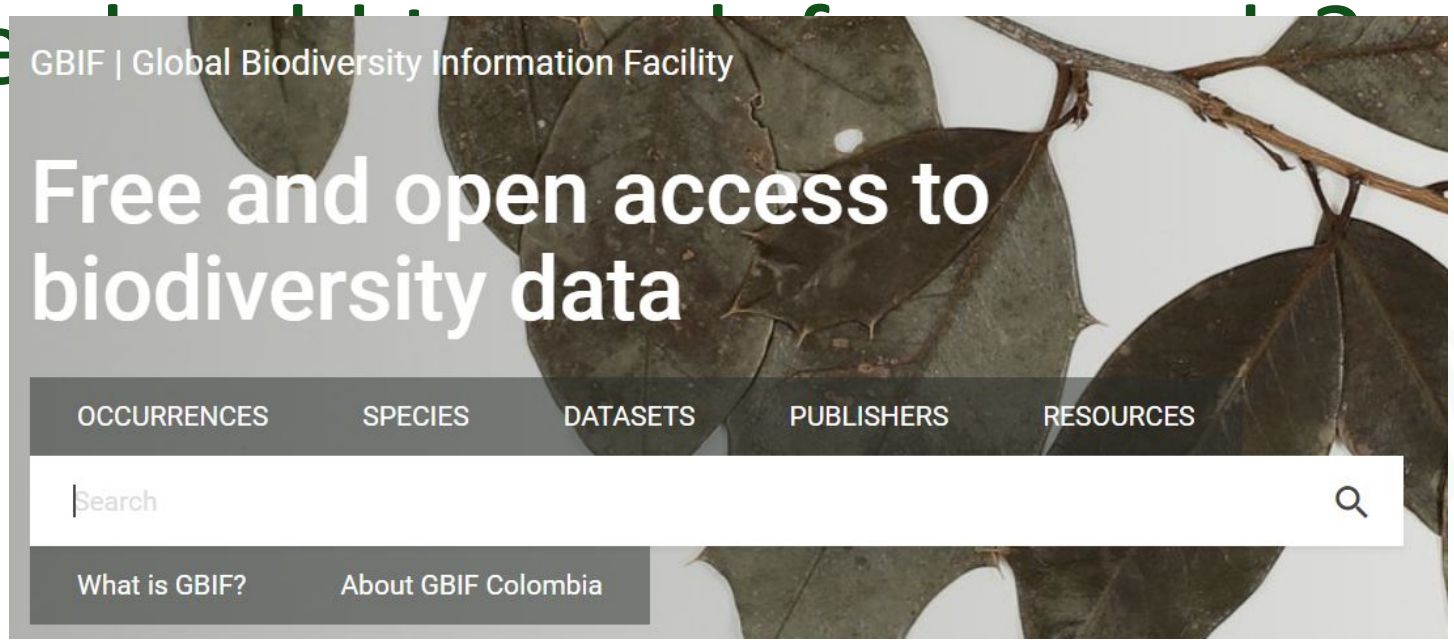
Ann. Bot. (Oxford) 106: 127 (2010)

Distribution Map



Native distribution
Introduced distribution

Where



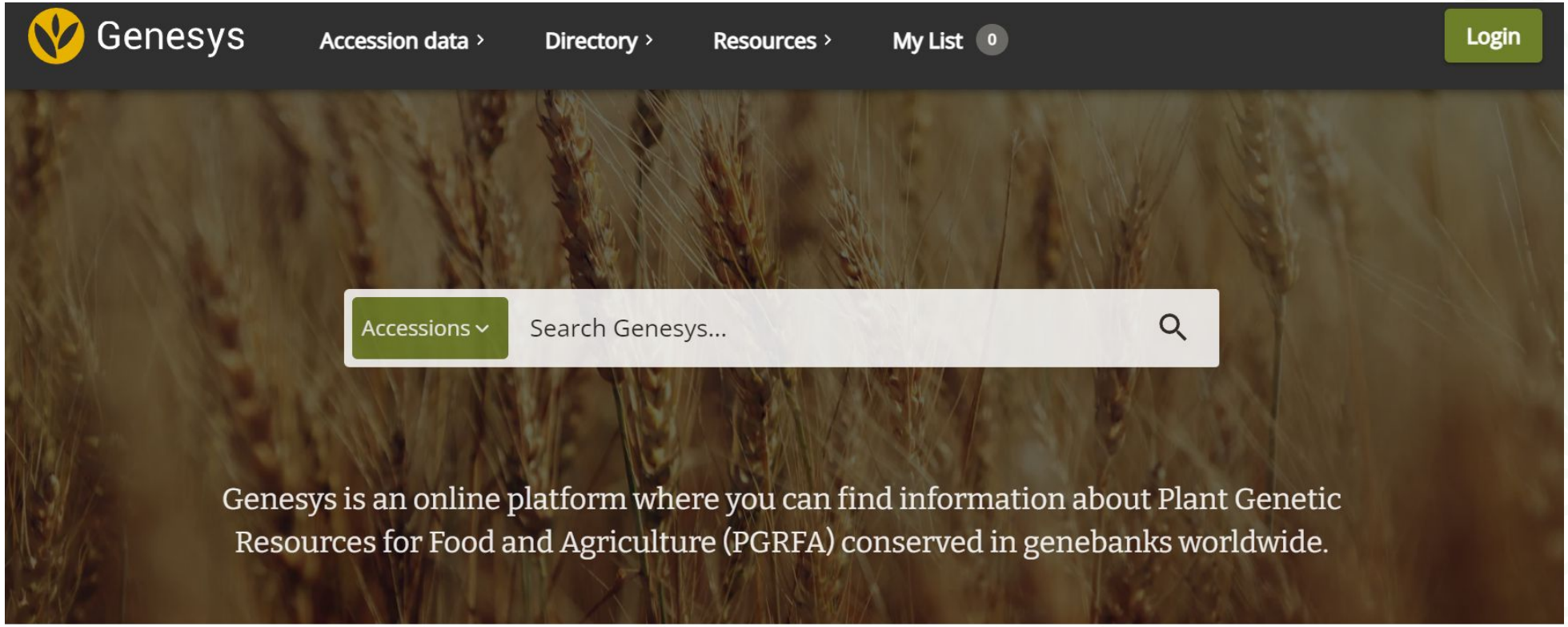
3,037,714,219
Occurrence records

110,870
Datasets

2,352
Publishing institutions

11,962
Peer-reviewed papers using data

GBIF: Dirty information (Needs curation!)



4,413,391
Browse accession records

422
Explore subsets

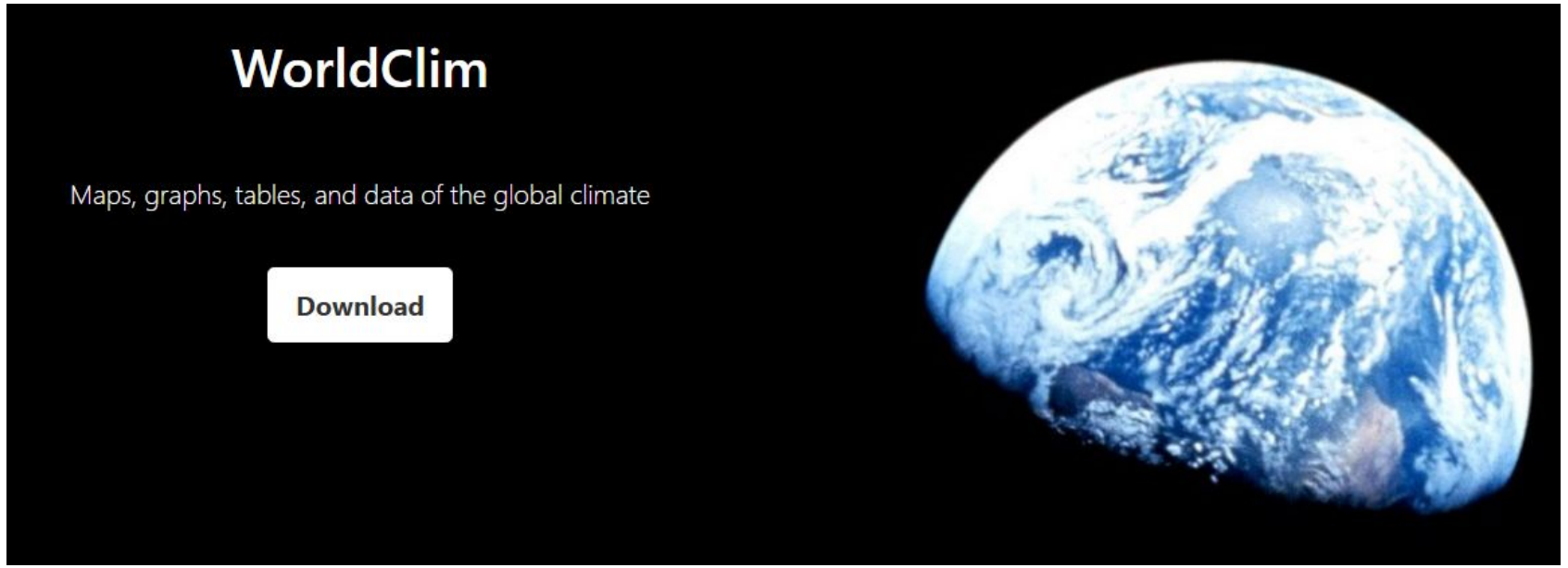
527
Explore C&E Datasets

CROP TRUST
Alianza Bioversity & CIAT
CGIAR

WIEWS - World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture

Plant genetic resources for food and agriculture (PGRFA) are essential to sustainable agriculture and food security. They are the raw materials to meet the current and future needs of crop improvement and adaptation programmes. It is therefore very important to conserve and sustainably use them.

WIEWS is the information system used by FAO for the preparation of periodic, country-driven global assessments of the status of conservation and use of PGRFA. WIEWS also monitors, on the basis of country reports, the implementation of the Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture, adopted in 2011. National Focal Points, appointed by Governments, may provide relevant information through a dedicated Reporting tool.



In R:

```r
#install.packages("geodata")
require(geodata)

#2.5 is 5 km
geodata::worldclim_global(var = "bio",
                          res = 2.5,
                          path = "D:/CGIAR/DS4Climate Action LAC-Bolder Africa - Bolder Africa - Gap analysis/data/runs/input_data/generic_rasters")
```

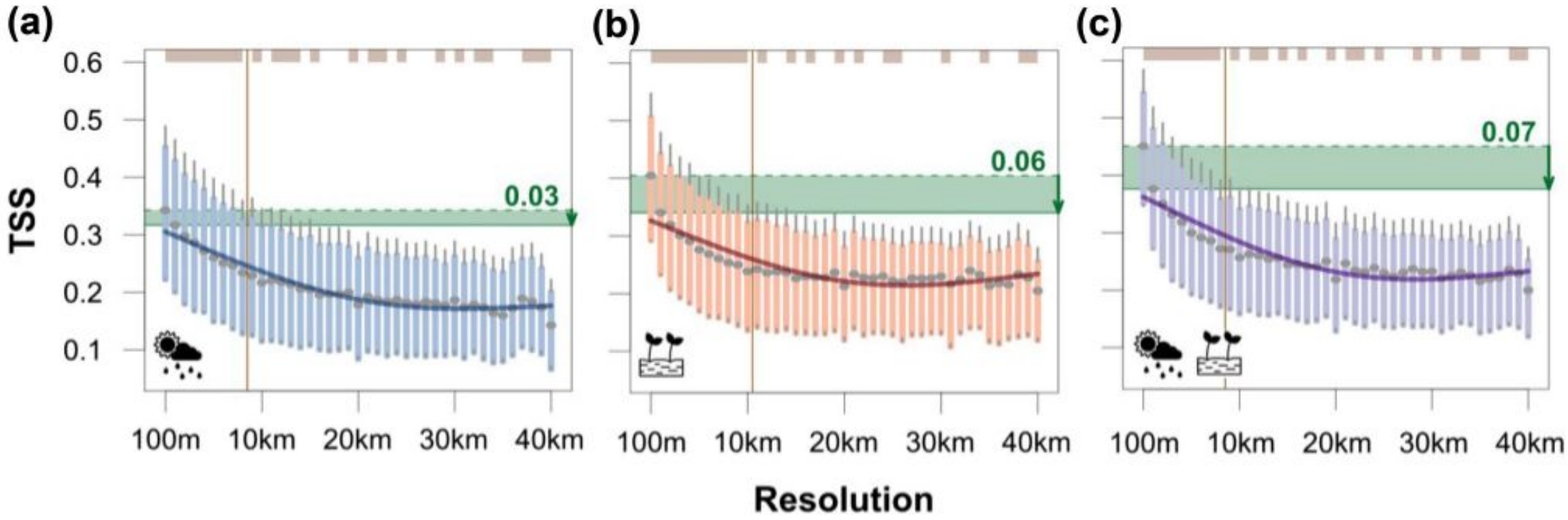
Why the geographical scale matters?

ECOGRAPHY

Research

Resolution in species distribution models shapes spatial patterns of plant multifaceted diversity

Yohann Chauvier, Patrice Descombes, Maya Guéguen, Louise Boulangeat, Wilfried Thuiller and Niklaus E. Zimmermann



What model should I use?

- MaxEnt or maxnet, XGBoost among the prefer ones
- Use an ensembles of models?

Ecological Monographs, 92(1), 2022, e01486

Predictive performance of presence-only species distribution models: a benchmark study with reproducible code

Roozbeh Valavi [1,3], Gurutzeta Guillera-Arroita [2], José J. Lahoz-Monfort [2], and Jane Elith [2]

[1]*School of Biosciences, University of Melbourne, Parkville, Victoria 3010 Australia*
[2]*School of Ecosystem and Forest Sciences, University of Melbourne, Parkville, Victoria 3010 Australia*

Method	Description	R package
GAM	generalized additive model	*mgcv*
GLM	generalized linear model	*stats::glm* and *gam::step.Gam*
Lasso	regularized regression (L1 regularization)	*glmnet*
Ridge regression	regularized regression (L2 regularization)	*glmnet*
MARS	multivariate adaptive regression spline	*earth*
MaxEnt	maximum entropy	*dismo::maxent* (needs maxent.jar)
MaxNet	maximum entropy new implementation	*maxnet*
BRT/GBM	boosted regression trees	*dismo::gbm.step* (relies on the *gbm* package)
cforest	unbiased conditional inference forest	*party::cforest*
RF	random forest	*randomForest*
XGBoost	extreme gradient boosting	*xgboost*
biomod	ensemble framework with up to 10 different models	*biomod2*
SVM	support vector machine	*e1071*

Some terms:

- Presence-only data: Data obtained only from geographical coordinates
- Presence: localities where a species was found
- Background: Points that represents the extent of the study (represent available environment)
- Hyperparameter: It is a configuration variable that controls the learning process of a machine learning model and must be set before training begins
- Feature: real-valued functions based on environmental data
- Entropy: A measure of dispersedness
- Probability density functions: Describe the relative likelihood of random variables over their range; can be univariate or multivariate

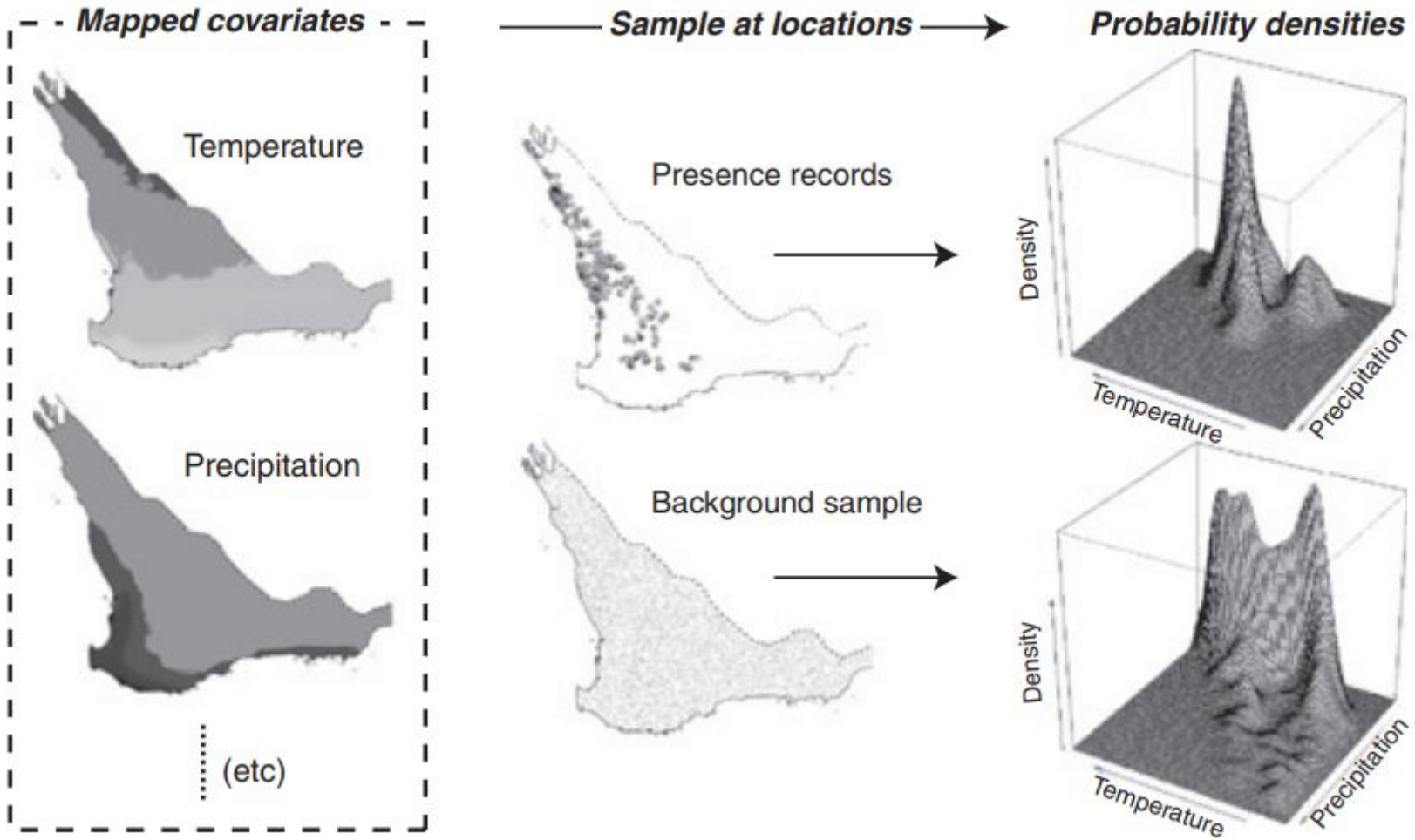
How can obtain the geographical distribution of a species using environmental data?

Diversity and Distributions, (Diversity Distrib.) (2011) 17, 43–57

BIODIVERSITY RESEARCH

A statistical explanation of MaxEnt for ecologists

Jane Elith[1]*, Steven J. Phillips[2], Trevor Hastie[3], Miroslav Dudík[4], Yung En Chee[1] and Colin J. Yates[5]



Features

Feature type	Interpretation	Constraint	Shape
Linear	Continuous variable	The *mean* of each environmental variable at an unknown location should be close to the mean of that variable in known occurrence locations.	
Quadratic	Square of the variable	The *variance* of each environmental variable at an unknown location should be close to the variance of that variable in known occurrence locations.	
Product	Pairs of continuous variables – allows for interactions	The *co-variance* of two environmental variables at an unknown location should be close to the co-variance of those variables in known occurrence locations.	
Threshold	Conversion into binary response based on a threshold	The proportion of predicted occurrences with values above the threshold (binary response = 1) should be close to the proportion of known occurrences.	
Hinge	As threshold type, but response after the threshold (knot) is linear	The mean above the knot of each environmental variable at an unknown location should be close to the mean above the knot of that variable in known occurrence locations.	
Categorical	Categorical variable	The proportion of predicted occurrences in each category should be close to the proportion of observed occurrences in each category.	

Ecography 36: 001–012, 2013
doi: 10.1111/j.1600-0587.2013.07872.x

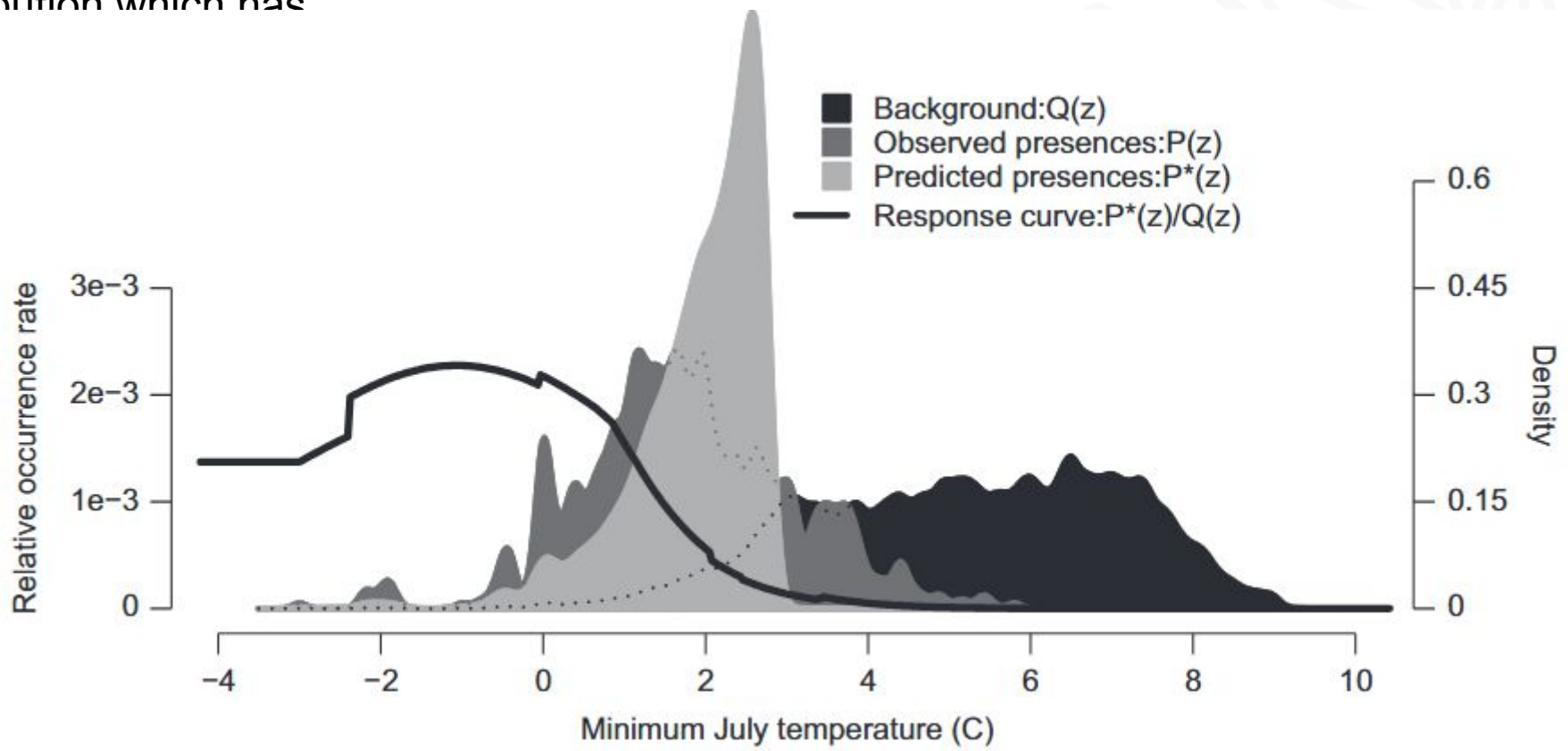
Subject Editor: Niklaus E. Zimmermann. Accepted 27 March 2013

Forum

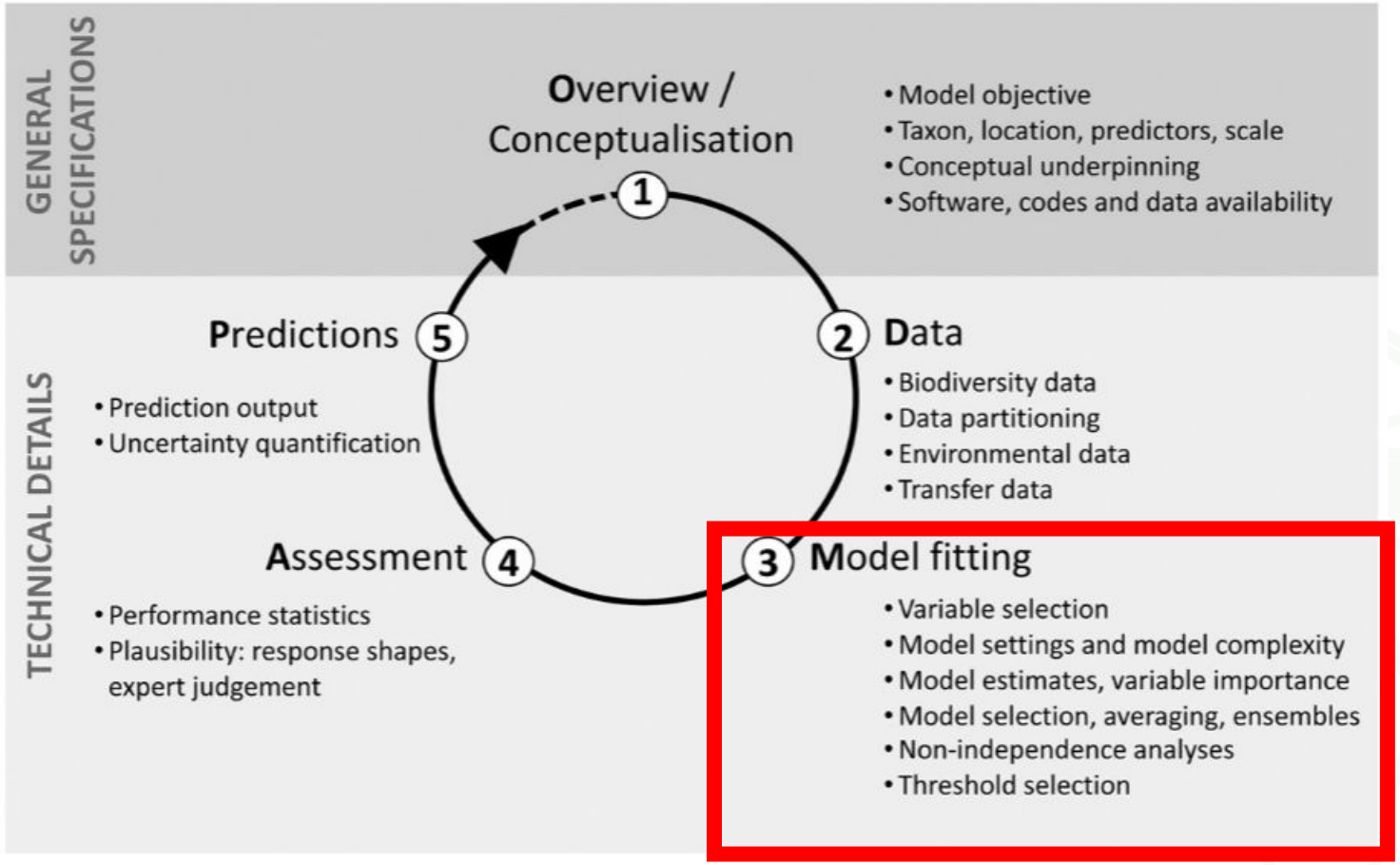
A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter

Cory Merow, Matthew J. Smith and John A. Silander, Jr

Estimates the distribution (geographic range) of a species by finding the distribution which has maximum entropy
Minimizing entropy respe

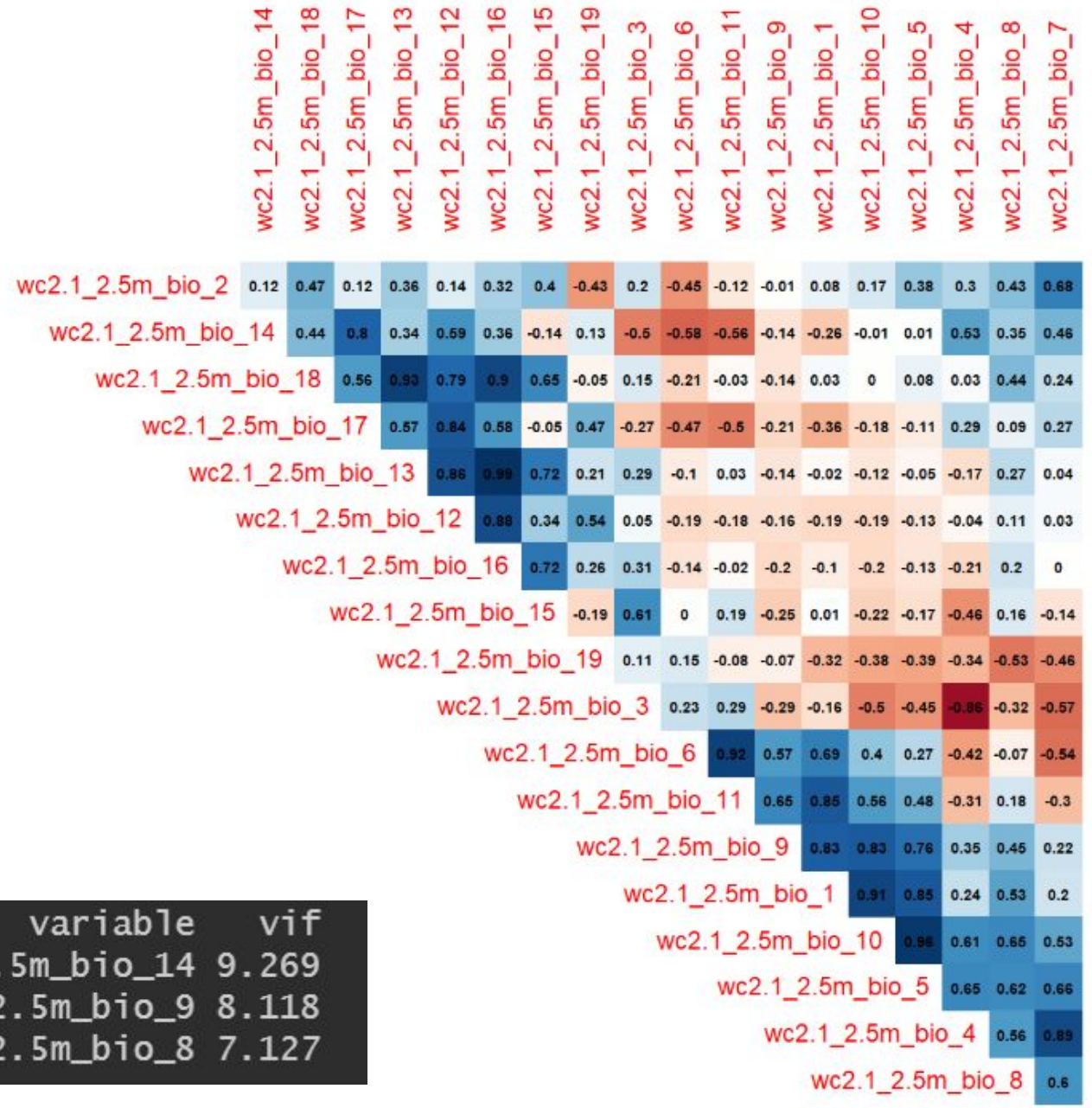


Model fitting



Select variables

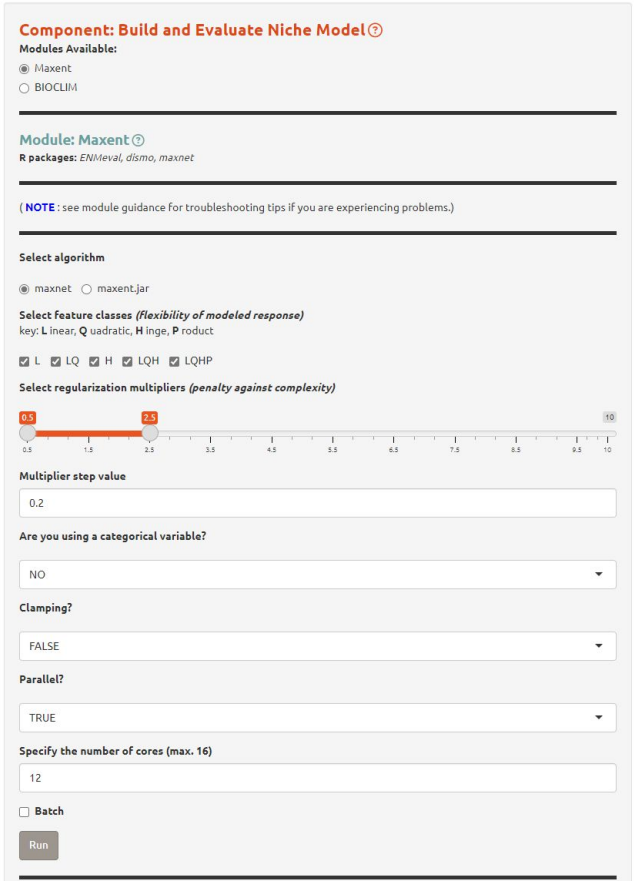
- Avoid Multicollinearity
- Ways to reduce Multicollinearity
  - Correlation (e.g. <0.7)
  - Variance Inflation Factor (VIF <10
  - Principal component Analysis (PC



	variable	vif
17	wc2.1_2.5m_bio_14	9.269
18	wc2.1_2.5m_bio_9	8.118
19	wc2.1_2.5m_bio_8	7.127

Calibration

- Reduce overfitting!



Methods in Ecology and Evolution

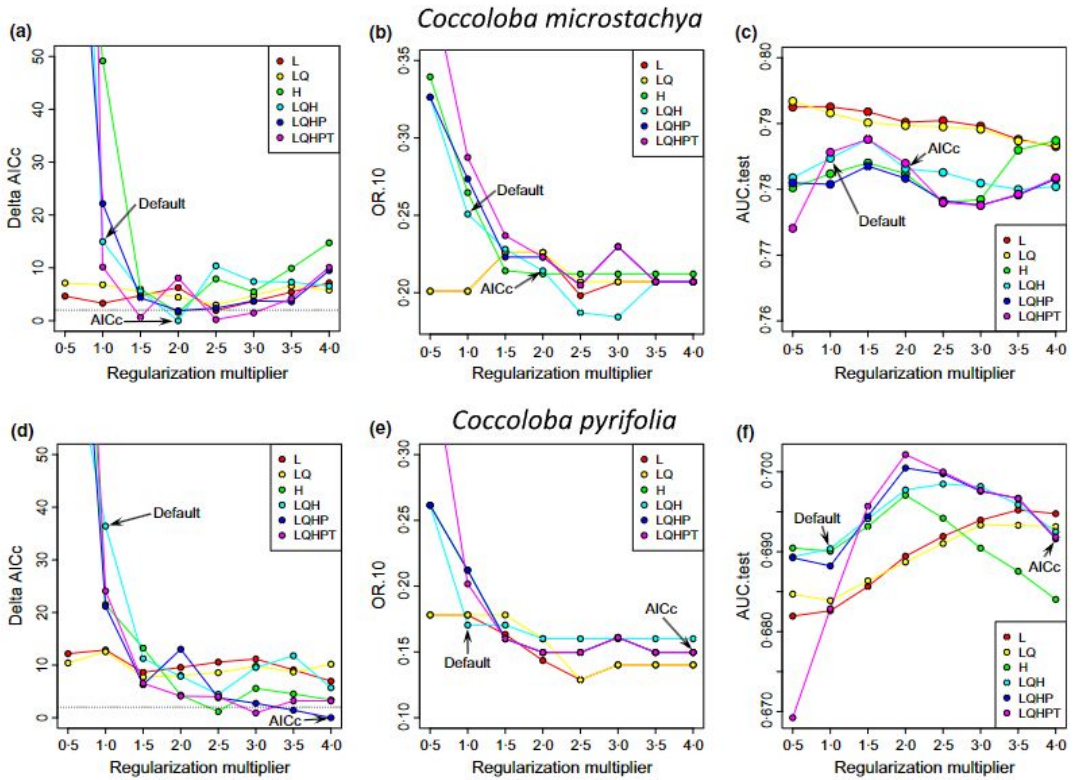
British Ecological Society

Methods in Ecology and Evolution 2014, **5**, 1198–1205 doi: 10.1111/2041-210X.12261

APPLICATION

ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models

Robert Muscarella[1]*, Peter J. Galante[2], Mariano Soley-Guardia[2,3], Robert A. Boria[2], Jamie M. Kass[2,3], María Uriarte[1] and Robert P. Anderson[2,3,4]

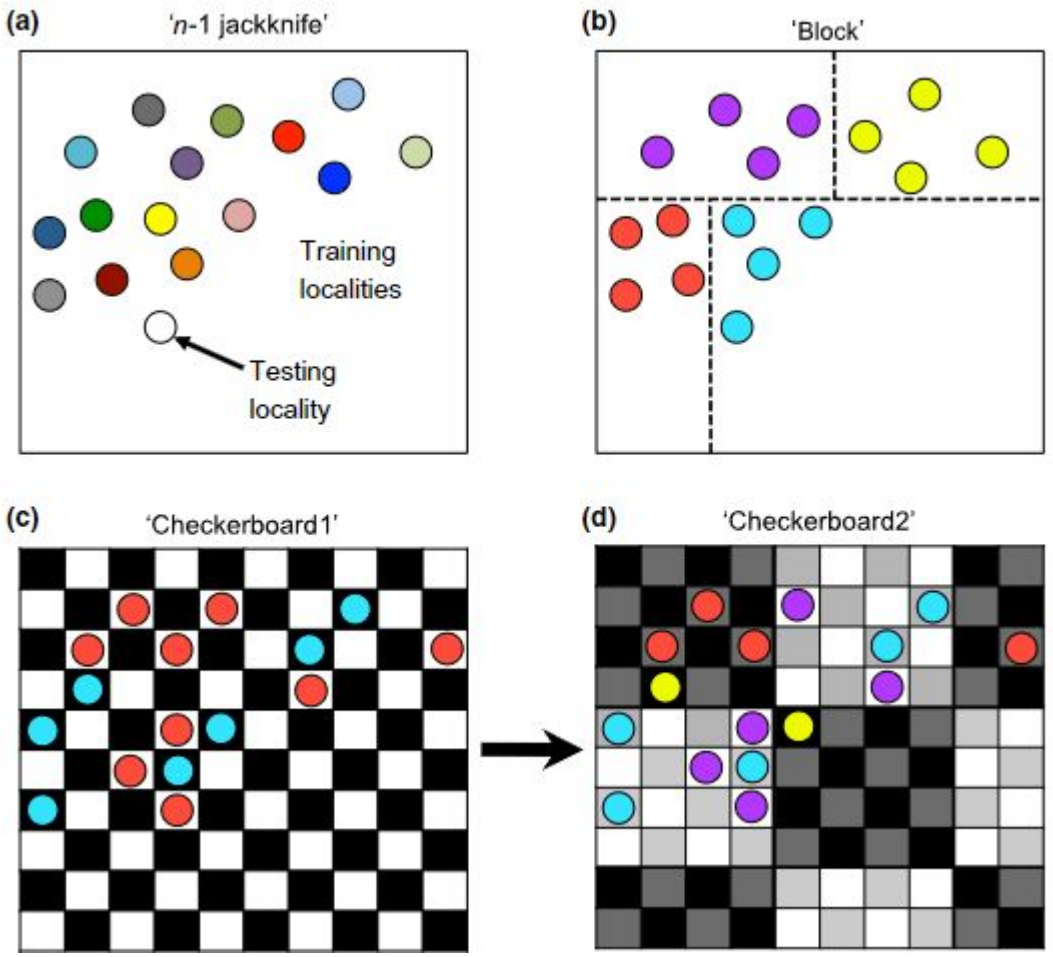


CROP TRUST

Alianza Bioversity & CIAT

CGIAR

Spatial independent evaluation

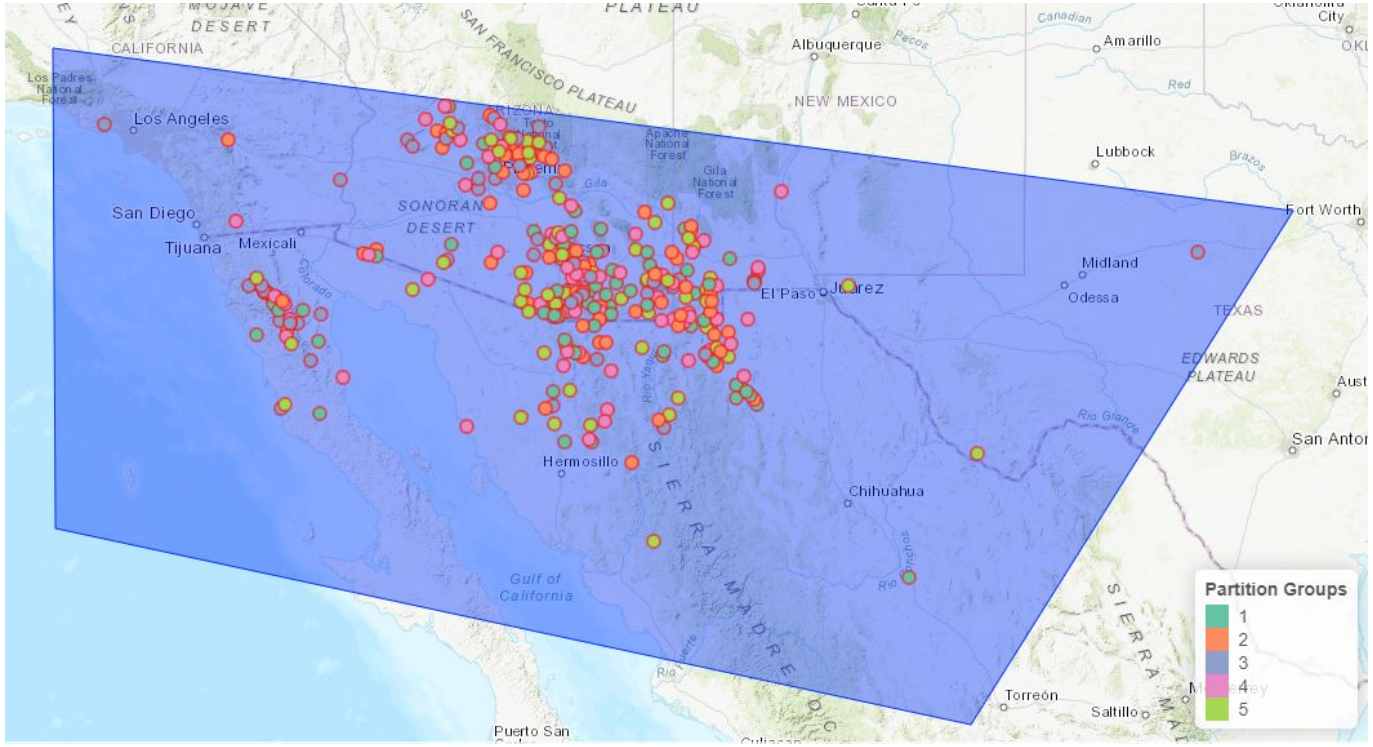


Methods in Ecology and Evolution
British Ecological Society
Methods in Ecology and Evolution 2014, **5**, 1198–1205
doi: 10.1111/2041-210X.12261

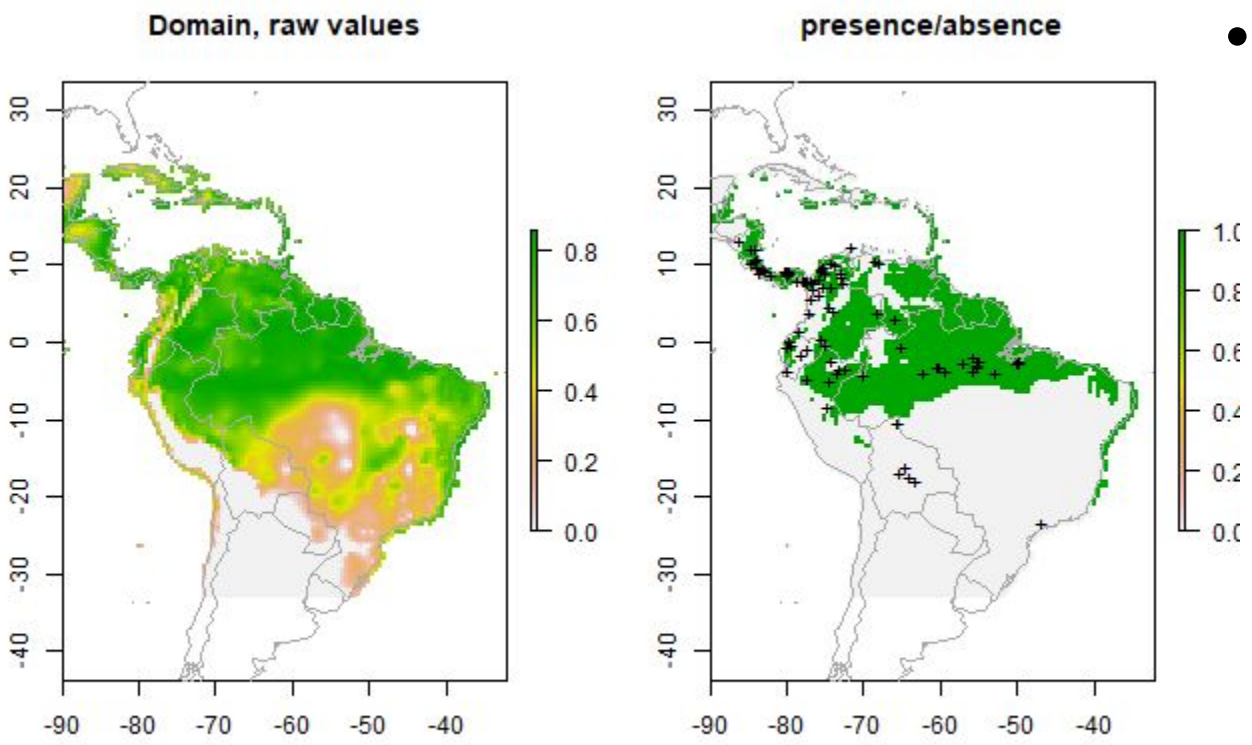
APPLICATION

ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models

Robert Muscarella[1]*, Peter J. Galante[2], Mariano Soley-Guardia[2,3], Robert A. Boria[2], Jamie M. Kass[2,3], María Uriarte[1] and Robert P. Anderson[2,3,4]



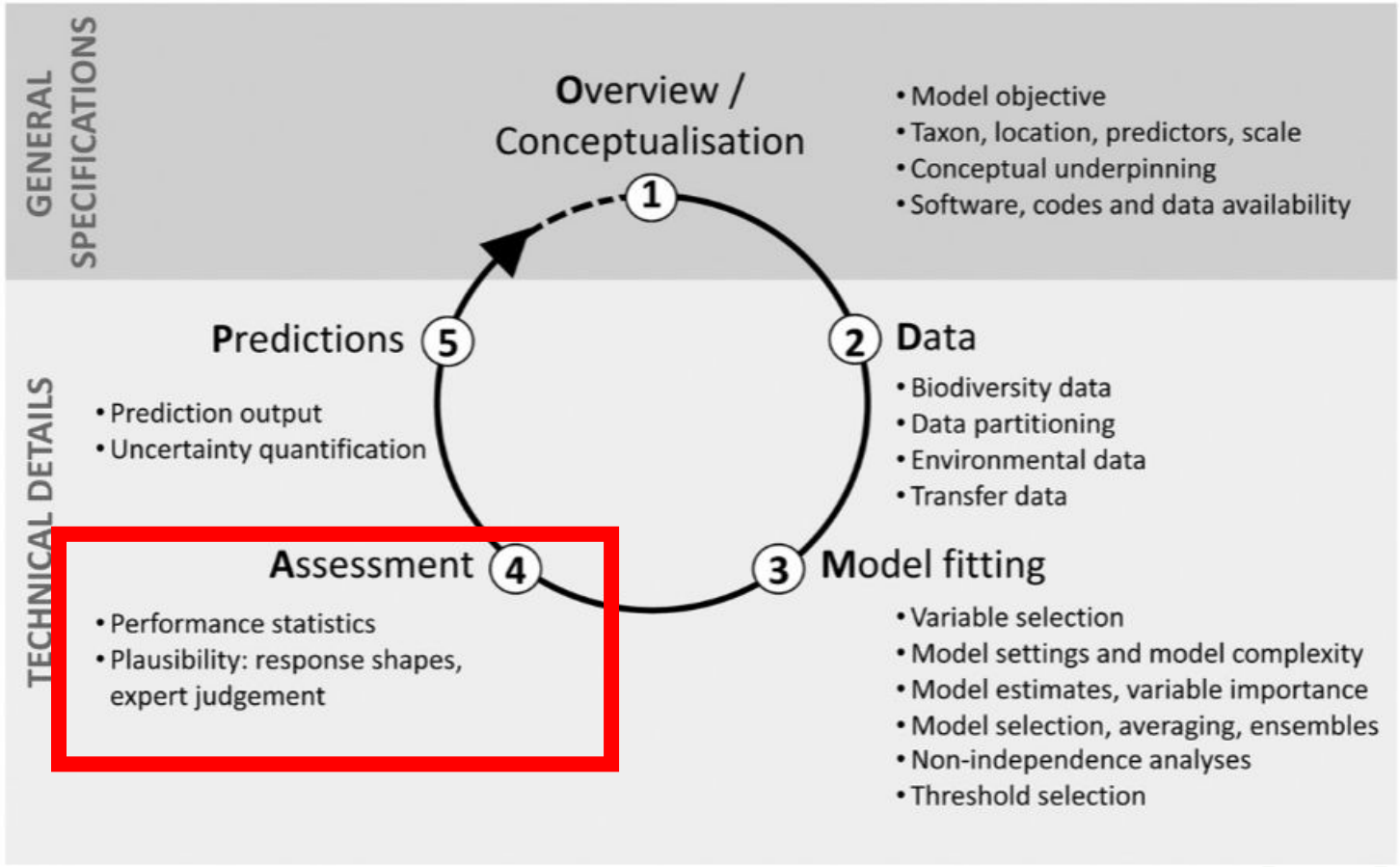
What cut-off value should I use to say that a species is present or absent?



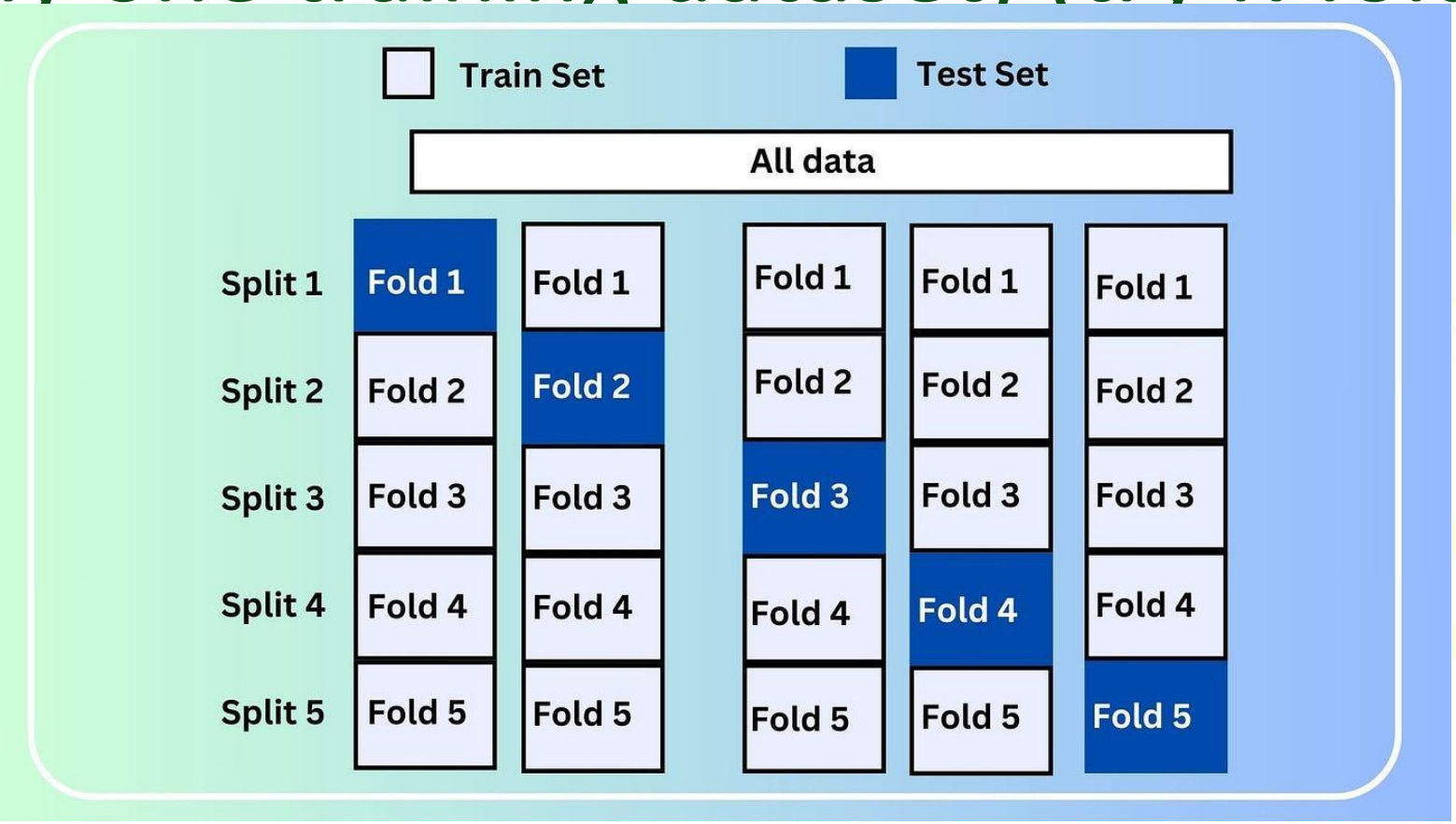
- Threshold: Designating areas above a suitability level as within species range!
  - Maximum sum of specificity and sensitivity
  - Minimum training presence
  - 10th percentile training presence

https://babichmorrowc.github.io/post/2019-04-12-sdm-threshold/

Assessment


GENERAL SPECIFICATIONS
Overview / Conceptualisation
1
• Model objective
• Taxon, location, predictors, scale
• Conceptual underpinning
• Software, codes and data availability
TECHNICAL DETAILS
Predictions
5
• Prediction output
• Uncertainty quantification
2
Data
• Biodiversity data
• Data partitioning
• Environmental data
• Transfer data
Assessment
4
• Performance statistics
• Plausibility: response shapes, expert judgement
3
Model fitting
• Variable selection
• Model settings and model complexity
• Model estimates, variable importance
• Model selection, averaging, ensembles
• Non-independence analyses
• Threshold selection

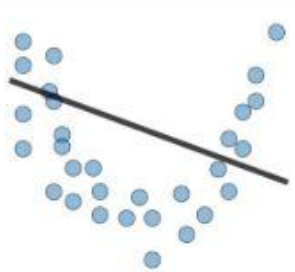
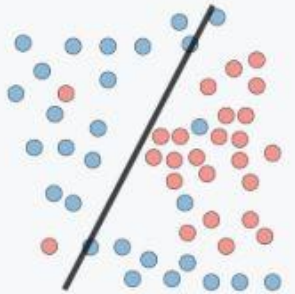
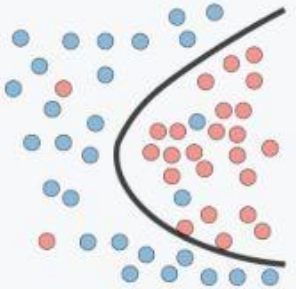
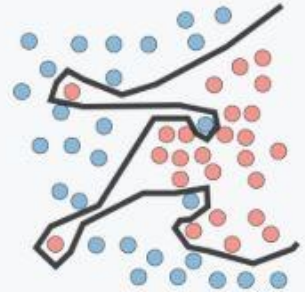
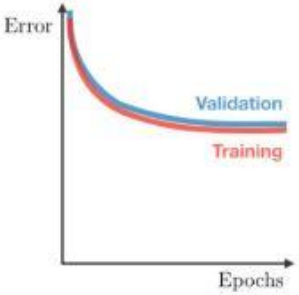
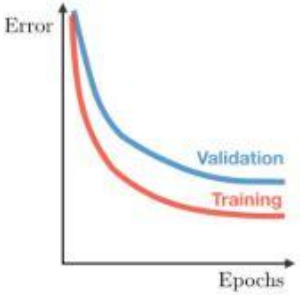
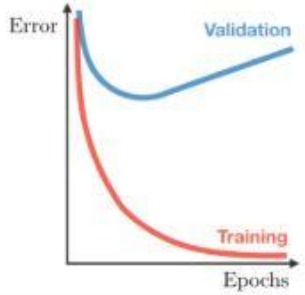
How can obtain good results? (Do not use only one training dataset) (try K-fold)



https://medium.com/@bididudy/the-essential-guide-to-k-fold-cross-validation-in-machine-learning-2bcb58c50578

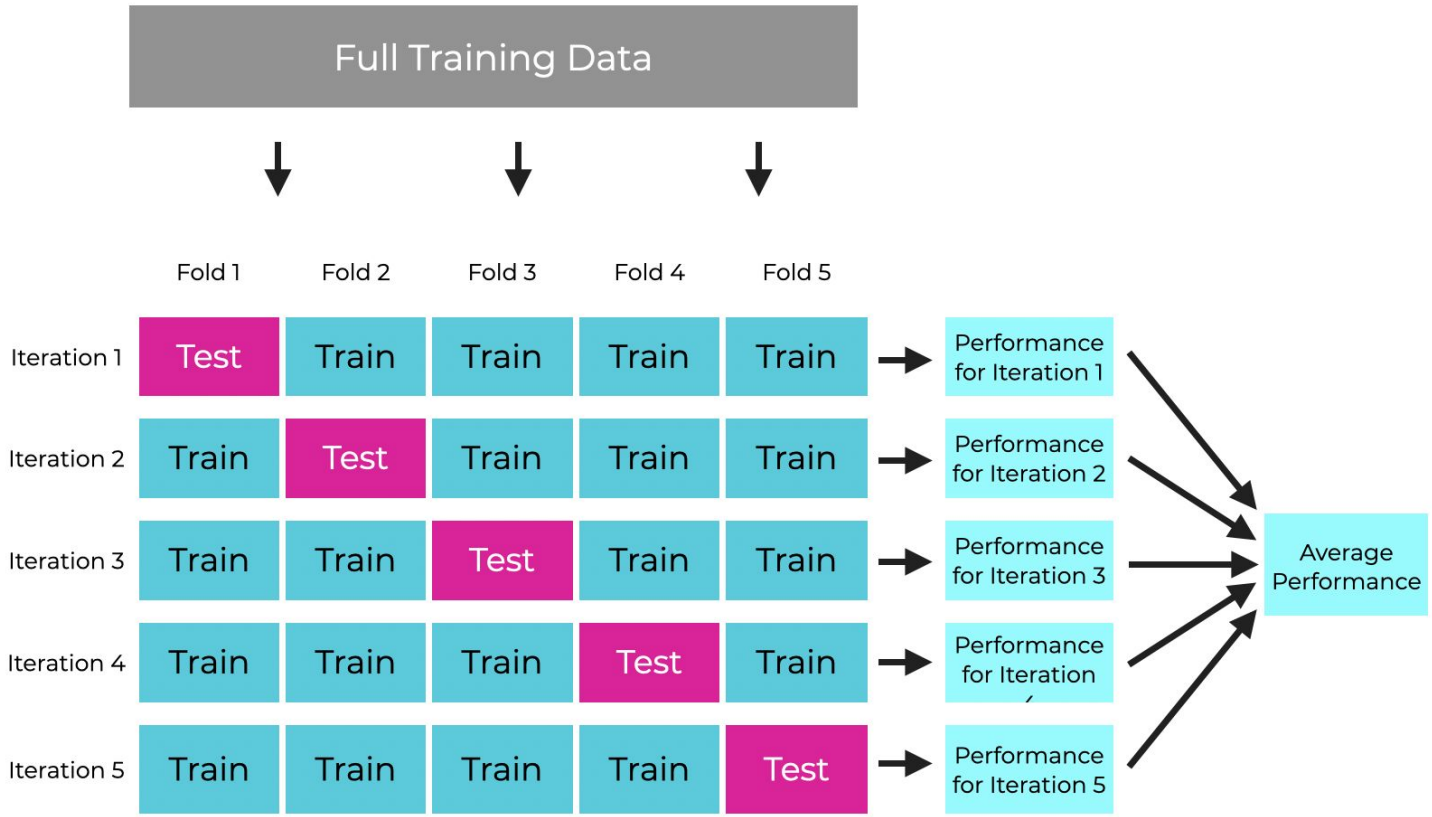
There are no silver bullet

- When too good is not good?

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

https://www.kaggle.com/discussions/getting-started/166897

Crossvalidation (Evaluation)



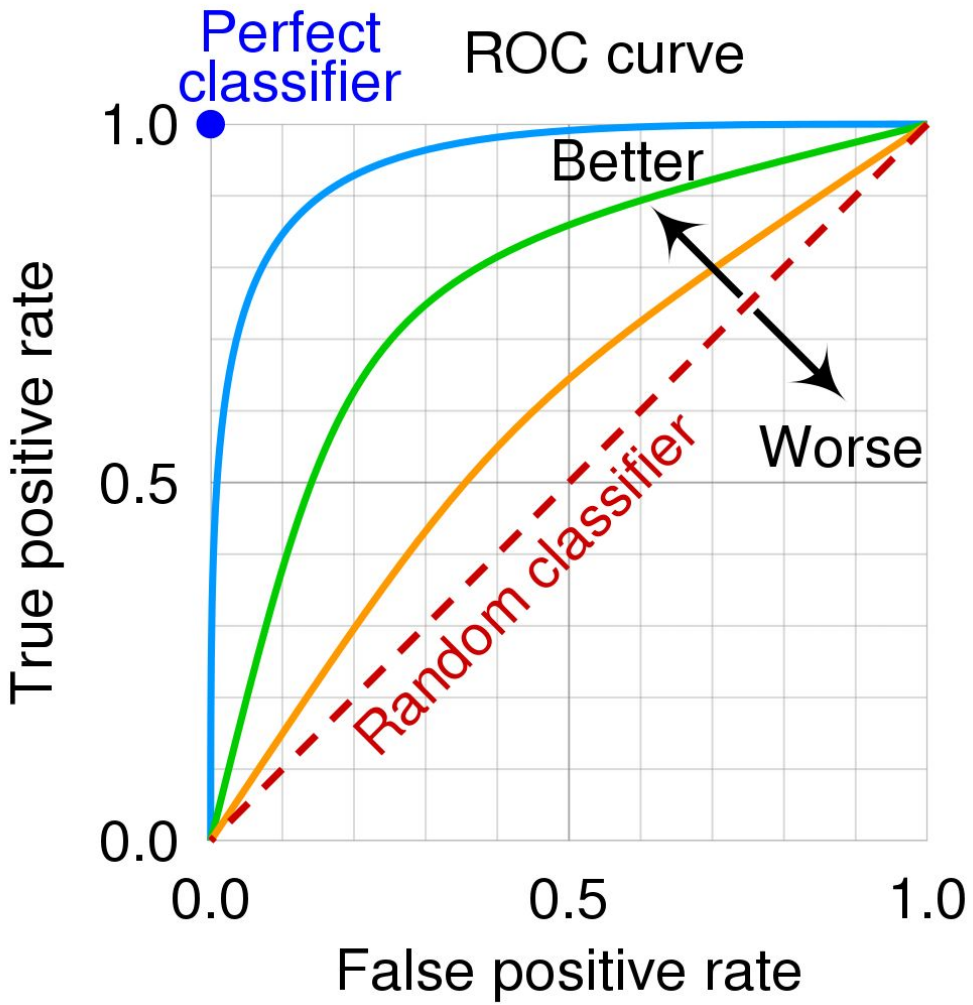
https://www.sharpsightlabs.com/blog/cross-validation-explained/

Evaluation (Confusion matrix)

Actual Class \ Predicted Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

https://encord.com/glossary/confusion-matrix/

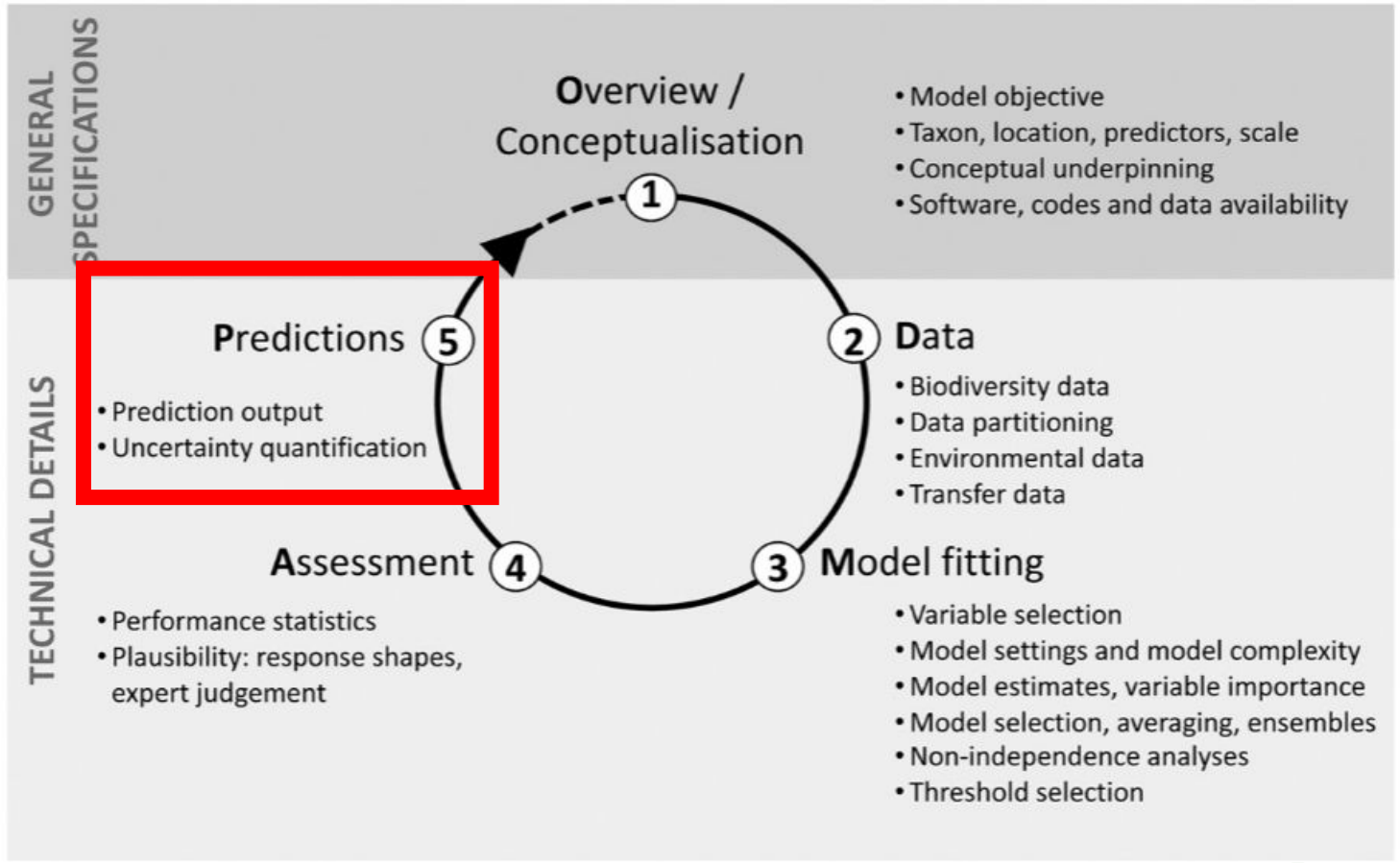
AUC-ROC (receiver operating characteristic curve)



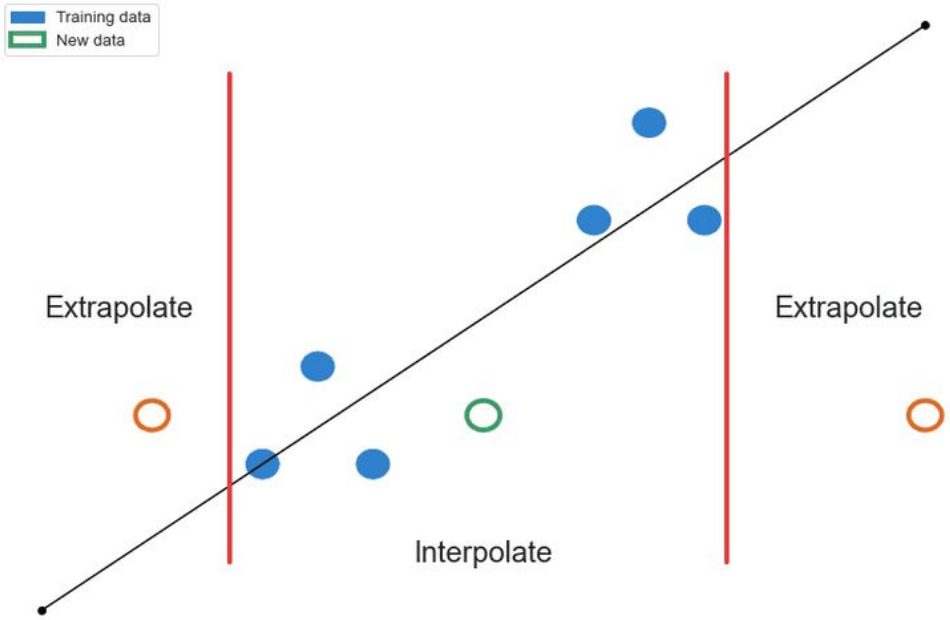
- AUC = 1: Perfect prediction
- AUC > 0.9: Excellent
- AUC 0.7-0.9: Good
- AUC 0.5-0.7: Poor
- AUC = 0.5: Random prediction

https://community.deeplearning.ai/t/roc-curve-and-auc-area-under-the-roc-curve/224140/2

Predictions



The multivariate environmental similarity



Methods in Ecology and Evolution

Methods in Ecology and Evolution 2010, **1**, 330–342 doi: 10.1111/j.2041-210X.2010.00036.x

The art of modelling range-shifting species

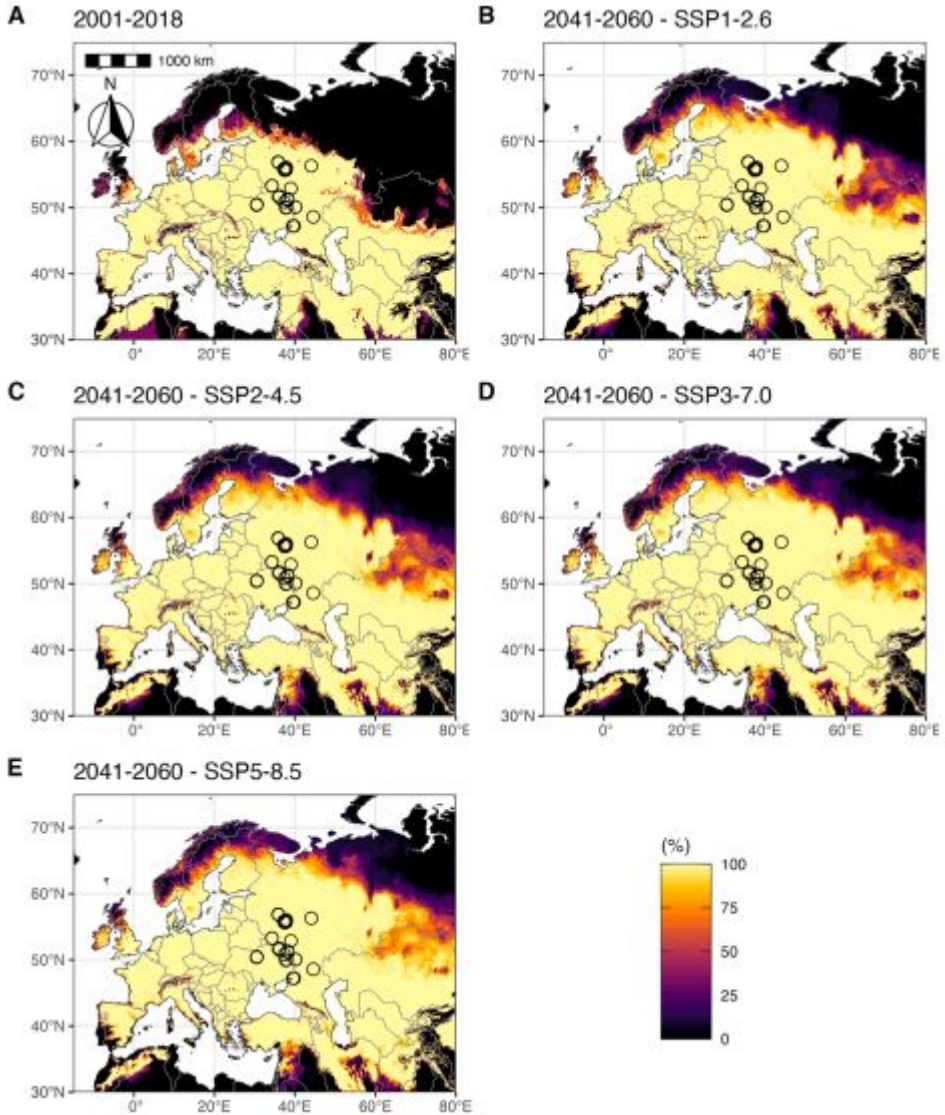
Jane Elith[1]*, Michael Kearney[2] and Steven Phillips[3]

[1]School of Botany, The University of Melbourne, Parkville 3010, Australia; [2]Department of Zoology, The University of Melbourne, Parkville 3010, Australia and [3]AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, USA

- Negative MESS: Model extrapolation
- Positive MESS: values within environmental ranges of presence records

The MES of a point P is calculated as follows:

1. Let min_i be the minimum value of variable V_i over the reference point set, and similarly for max_i.
2. Let p_i be the value of variable V_i at point *P*.
3. Let f_i be the percent of reference points whose value of variable V_i is smaller than p_i.
4. Then the similarity of *P* with respect to variable V_i is:
   - $(p_i - min_i) / (max_i - min_i) * 100$ if $f_i = 0$
   - $2 * f_i$ if $0 < f_i \leq 50$
   - $2 * (100 - f_i)$ if $50 \leq f_i < 100$
   - $(max_i - p_i) / (max_i - min_i) * 100$ if $f_i = 100$
5. Finally, the multivariate similarity of *P* is the minimum of its similarity with respect to each variable.



Trees, Forests and People 16 (2024) 100559

Contents lists available at ScienceDirect

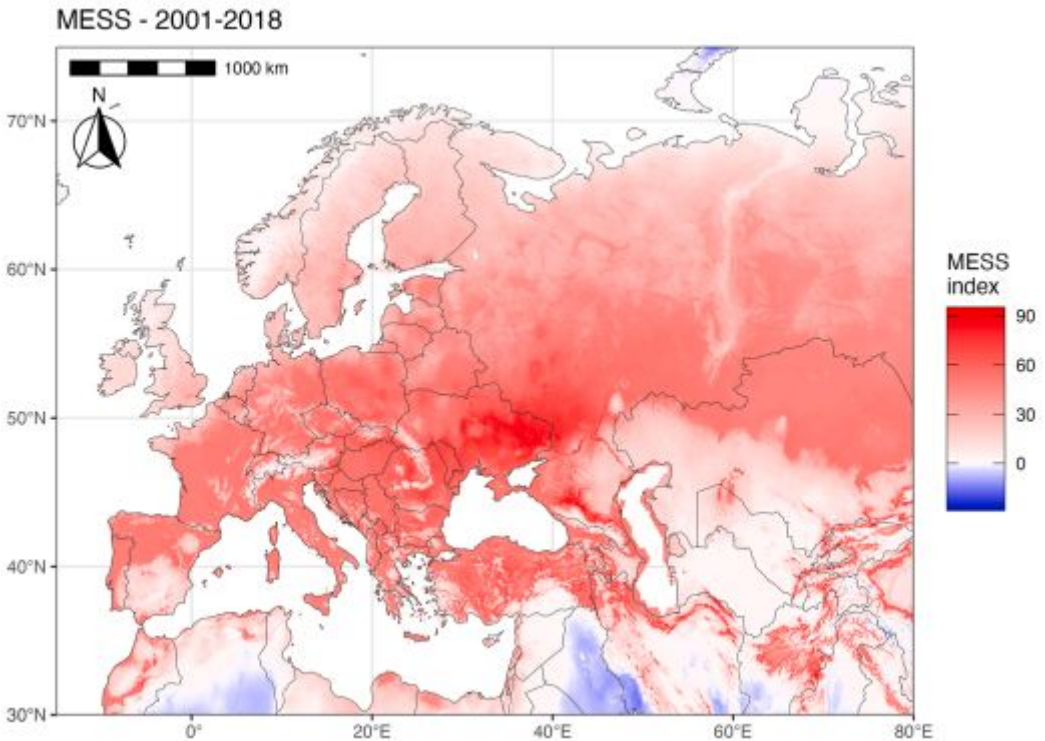
Trees, Forests and People

journal homepage: www.sciencedirect.com/journal/trees-forests-and-people

Modelling the potential range of *Agrilus planipennis* in Europe according to current and future climate conditions

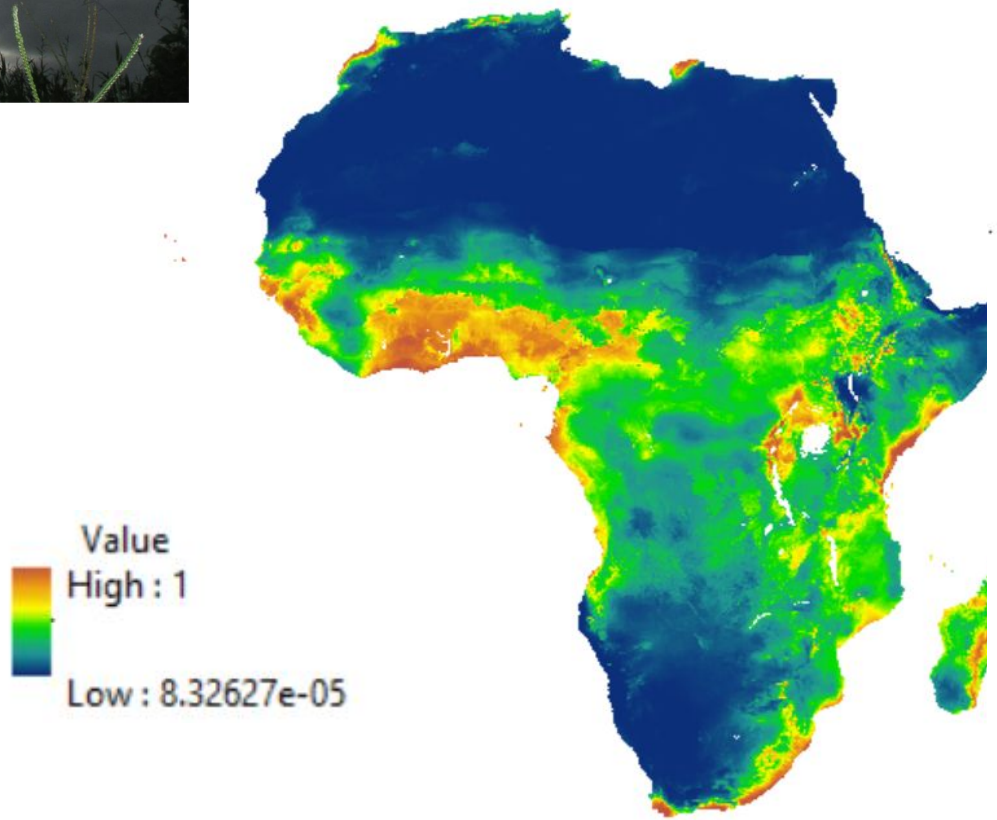
Jean-Pierre Rossi [a,*], Raphaëlle Mouttet [b], Pascal Rousse [b], Jean-Claude Streito [a]

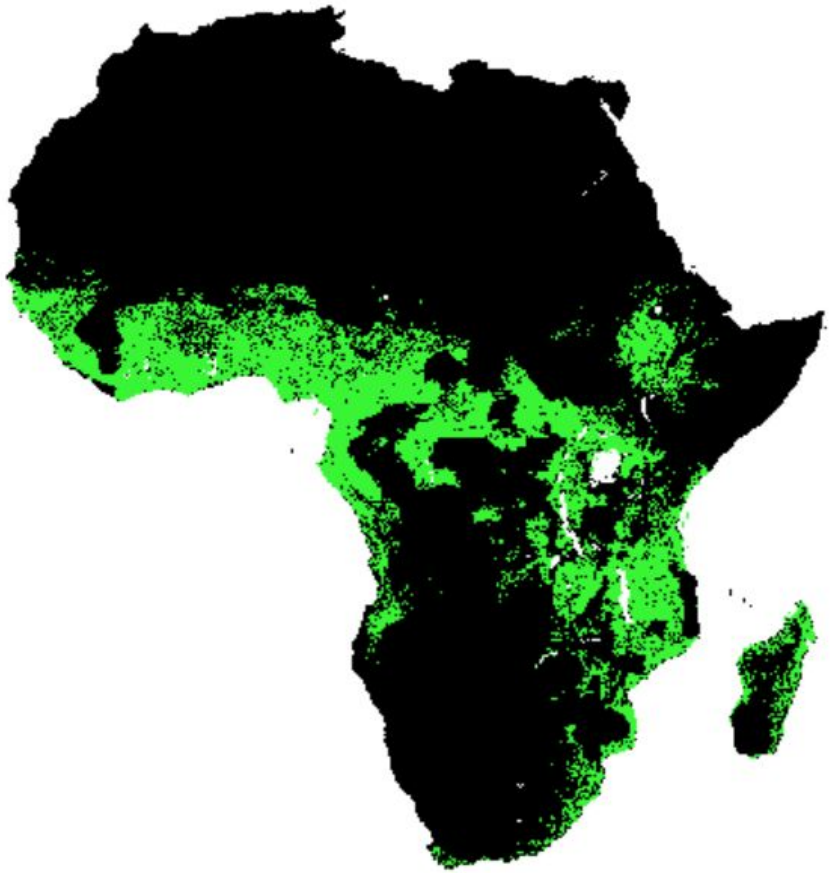
[a] CBGP (Centre de Biologie pour la Gestion des Populations), INRAE, CIRAD, IRD, Institut Agro, 755 Avenue du Campus Agropolis, CS 30016, 34988, Montferrier-sur-Lez, France
[b] ANSES, Plant Health Laboratory, Entomology and Botany Unit, 755 Avenue du Campus Agropolis, CS 30016, F-34988 Montferrier-sur-Lez, France



Poaceae - Eleusine indica (L.) Gaertn.
© Cyril CRUSSON / CIRAD

AUC	presences	absences	sensitivity	specificity	TSS	Threshold (max TSS)
0.850	1255	12550	0.852	0.664	0.517	0.357





Exercise: Use Bactris gasipaes data to get a SDM!

```
library(wallace)
run_wallace()
```

WORKFLOW

Wallace (v2.2.0) currently includes ten components, or steps of a possible workflow. Each component includes two or more modules, which are possible analyses for that step.

Components:

1. *Obtain Occurrence Data*

- Query Present Database
- User-specified Occurrences

2. *Obtain Environmental Data*

- WorldClim
- EcoClimate
- User-specified Environmental Data

3. *Process Occurrence Data*

- Select Occurrences on Map
- Remove Occurrences by ID
- Spatial Thin

4. *Process Environmental Data*

- Select Study Region by Extent
- Draw Study Region
- User-specified Study Region

5. *Characterize Environmental Space*

- Environmental Ordination
- Occurrence Density Grid
- Niche Overlap

6. *Partition Occurrence Data*

What is *Wallace*?

Welcome to *Wallace*, a flexible application for reproducible ecological modeling, built for community expansion. The current version of *Wallace* (v2.2.0) steps the user through a full niche/distribution modeling analysis, from data acquisition to visualizing results.

The application is written in `R` with the web app development package `shiny`. Please find the stable version of *Wallace* on CRAN, and the development version on Github. We also maintain a *Wallace* website that has some basic info, links, and will be updated with tutorial materials in the near future.

Wallace is designed to facilitate spatial biodiversity research, and currently concentrates on modeling species niches and distributions using occurrence datasets and environmental predictor variables. These models provide an estimate of the species' response to environmental conditions, and can be used to generate maps that indicate suitable areas for the species (i.e. its potential geographic distribution; Guisan & Thuiller 2005; Elith & Leathwick 2009; Franklin 2010a; Peterson et al. 2011). This research area has grown tremendously over the past two decades, with applications to pressing environmental issues such as conservation biology (Franklin 2010b), invasive species (Ficetola et al. 2007), zoonotic diseases (González et al. 2010), and climate-change impacts (Kearney et al. 2010).

Also, for more detail, please see our initial publication in *Methods in Ecology and Evolution* and our follow-up in *Ecography*.

Kass J. M., Vilela B., Aiello-Lammens M. E., Muscarella R., Merow C., Anderson R. P. (2018). *Wallace*: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolultion*, 9(4): 1151-1156. DOI: 10.1111/2041-210X.12945

Kass, J.M., Pinilla-Buitrago, G.E, Paz, A., Johnson, B.A., Grisales-Betancur, V., Meenan, S.I., Attali, D., Broennimann, O., Galante, P.J., Maitner, B.S., Owens, H.L., Varela, S., Aiello-Lammens, M.E., Merow, C., Blair, M.E., Anderson R.P. (2022). *wallace* 2: a shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. *Ecography*, 2023(3): e06547. DOI: 10.1111/ecog.06547.

Who is *Wallace* for?

We engineered *Wallace* to be used by a broad audience that includes graduate students, ecologists, conservation practitioners, natural resource managers, educators, and programmers. Anyone, regardless of programming ability, can use *Wallace* to perform an analysis, learn about the methods, and share the results. Additionally, those who want to disseminate a technique can author a module for *Wallace*.

Component: Obtain Occurrence Data

Modules Available:

- ○ Query Database (Present)
- ◉ User-specified

Module: User-specified Occurrences

R packages:

Upload Occurrence CSV

Browse... No file selected

☐ Do you want to define delimiter-separated and decimal values?

Load Occurrences

Module Developers: *Jamie M. Kass, Gonzalo E. Pinilla-Buitrago, Robert P. Anderson*

Module: User-specified Occurrences

R packages:

Upload Occurrence CSV

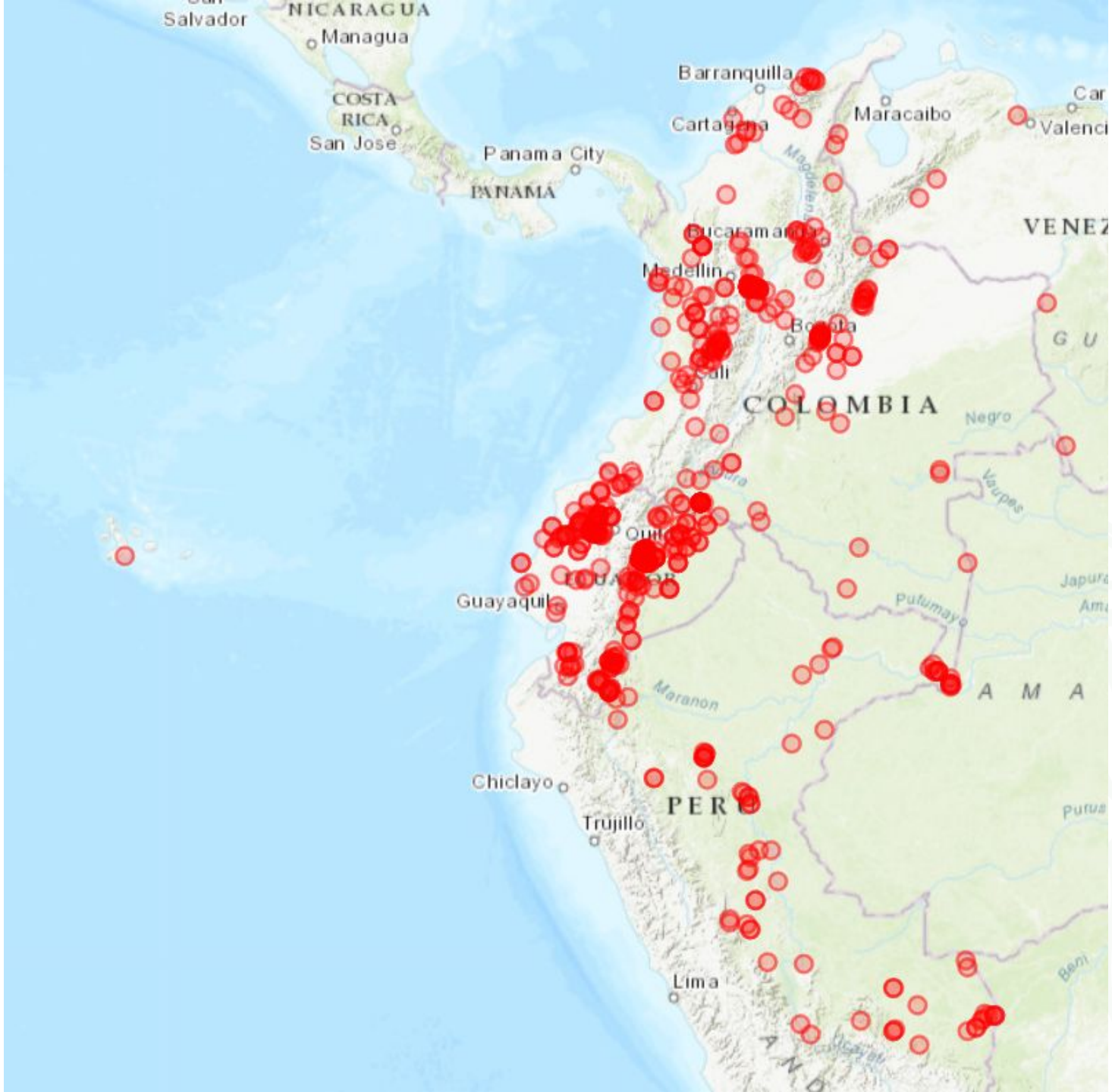
Browse... occurrences.csv

Upload complete

☐ Do you want to define delimiter-separated and decimal values?

Load Occurrences

Module Developers: *Jamie M. Kass, Gonzalo E. Pinilla-Buitrago, Robert P. Anderson*


Salvador
NICARAGUA
Managua
COSTA
RICA
San Jose
Panama City
PANAMA
Barranquilla
Cartagena
Maracaibo
Valenci
Bucaramanga
Medellin
Bogota
Cali
COLOMBIA
VENEZ
G U
Negro
Vaupes
Guayaquil
Putumayo
Japura
Maranon
A M A
Chiclayo
Trujillo
PERU
Lima
Purus
Beni

Component: Obtain Environmental Data

Modules Available:

- (●) WorldClim Bioclims
- ( ) ecoClimate
- ( ) User-specified

Module: WorldClim Bioclims

R packages: *raster, geodata*

Select WorldClim bioclimatic variable resolution

2.5 arcmin

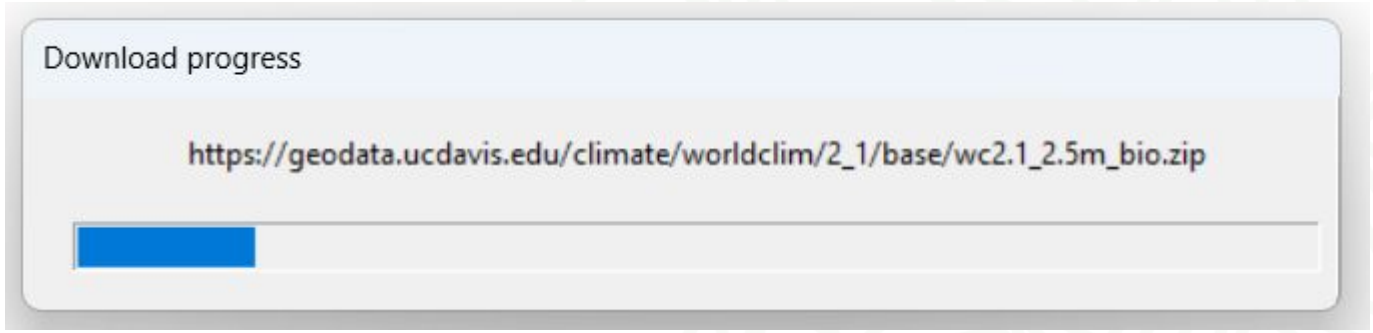
[ ] Save to memory for faster processing and save/load option

Select bioclim variables

bio01, bio02, bio03, bio04, bio05, bio06, bio07, bio08, bio09, bic

[ ] **Batch**

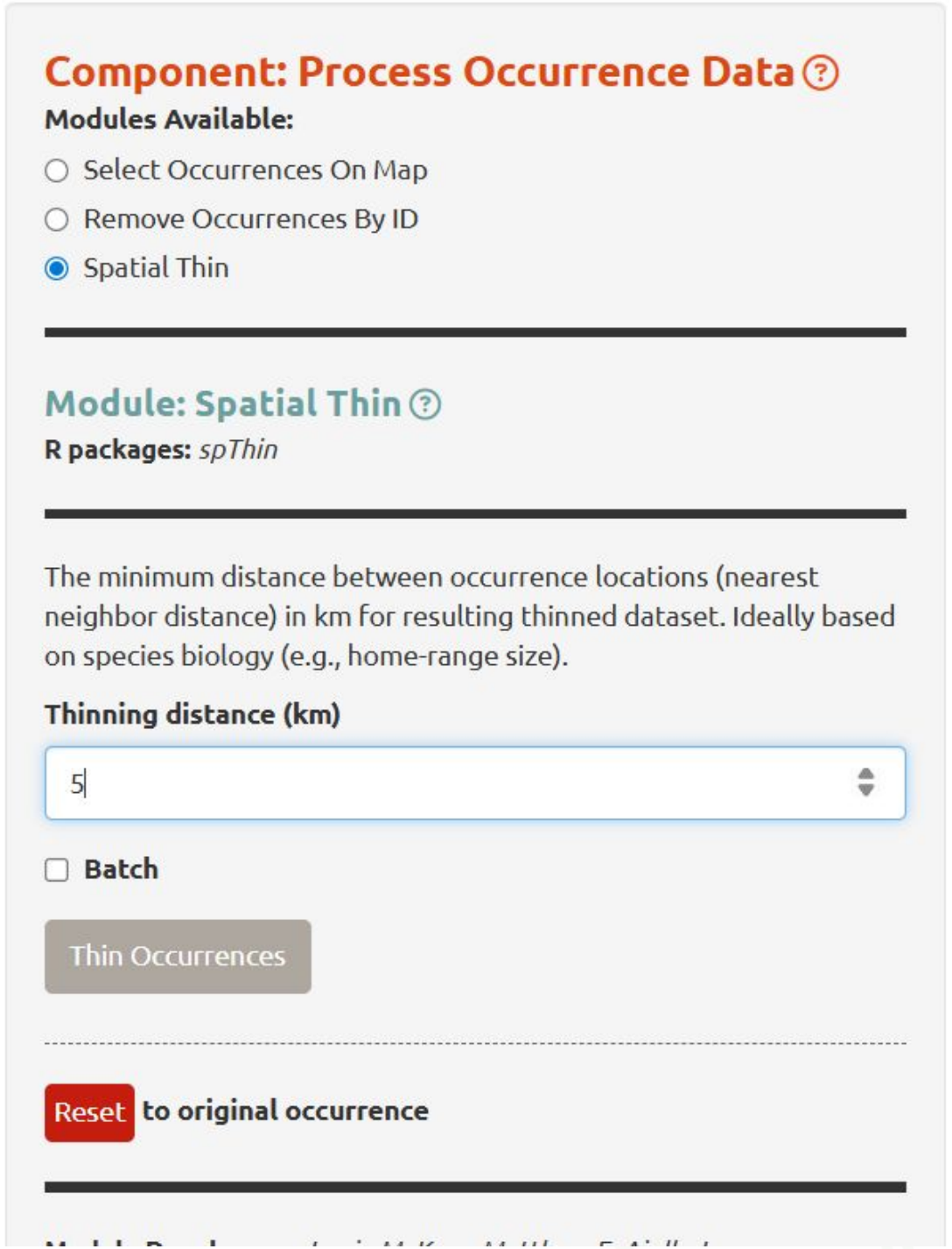
Load Env Data



Map | Occurrences | Results | Component Guidance | Module Guidance | Save

```
class      : RasterStack
dimensions : 4320, 8640, 37324800, 19  (nrow, ncol, ncell, nlayers)
resolution : 0.04166667, 0.04166667  (x, y)
extent     : -180, 180, -90, 90  (xmin, xmax, ymin, ymax)
crs        : +proj=longlat +datum=WGS84 +no_defs
names      :       bio01,       bio02,       bio03,       bio04,       bio05,       bio06,       bio07,       bio08,       bio09,
min values :  -54.759167,    1.000000,    9.063088,    0.000000,  -30.760000,  -72.503998,    1.000000,  -66.328003,  -56.601334,  -38.1
max values :    31.16667,    21.97300,   100.00000,  2377.62402,    48.46000,    26.45000,    72.68000,    37.96800,    37.66133,    38.
```

Reduce spatial bias!



Create background

Component: Process Environmental Data

Modules Available:

- (●) Select Study Region
- ( ) Draw Study Region
- ( ) User-specified Study Region

Module: Select Study Region by Extent

R packages: *sp, sf*

Step 1: *Choose Background Extent*

Background Extents:

- ( ) bounding box
- (●) minimum convex polygon
- ( ) point buffers

Study region buffer distance (degree)

1

[ ] **Batch**

Select

Step 2: *Sample Background Points*

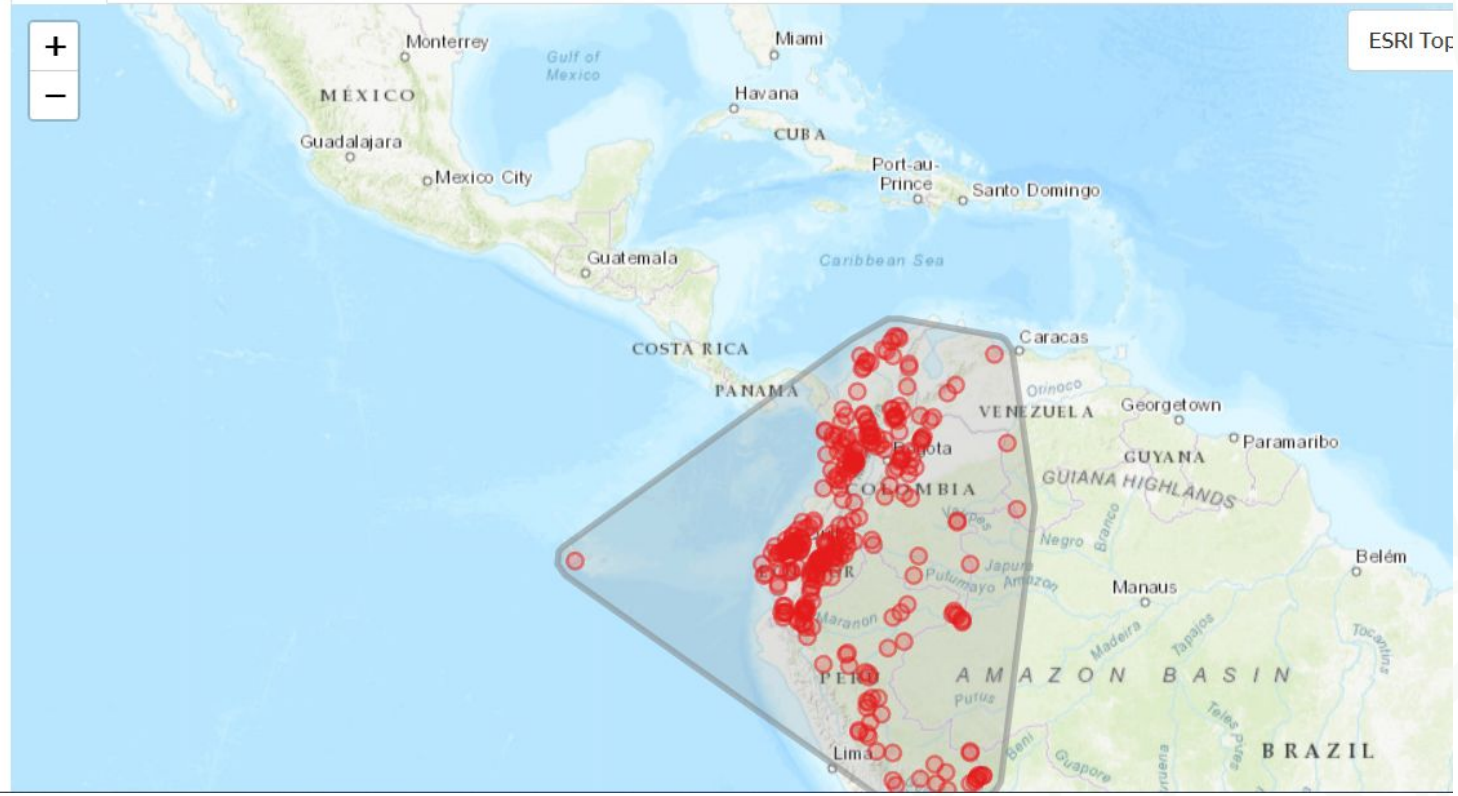
Species menu

Bactris_gasipaes

Log window

bio16, bio17, bio18, bio19 at 2.5 arcmin resolution.
> ***Bactris gasipaes*** | Worldclim variables ready to use.
> ***Bactris gasipaes*** | Total records thinned (5 km) to 301 localities
> ***Bactris gasipaes*** | Study extent: minimum convex polygon. Buffered by 1 degrees.

Map | Occurrences | Results | Component Guidance | Module Guidance | Save



Step 2: *Sample Background Points*

Mask predictor rasters by background extent and sample background points

No. of background points

10000

☐ **Batch**

Sample

Component: Partition Occurrence Data

Modules Available:

- ○ Non-spatial Partition
- ◉ Spatial Partition

Module: Spatial Partition

R packages: *ENMeval*

Options Available:

Block (k = 4)

☐ **Batch**

Partition

Module Developers: *Jamie M. Kass, Bruno Vilela, Bethany A. Johnson, Robert P. Anderson*

ENMeval references

Automated Tuning and Evaluations of Ecological Niche Models

Package Developers: Jamie M. Kass, Robert Muscarella, Peter J. Galante, Corentin Bohl, Gonzalo E. Buitrago-Pinilla, Robert A. Boria, Mariano Soley-Guardia, Robert P. Anderson

CRAN | documentation

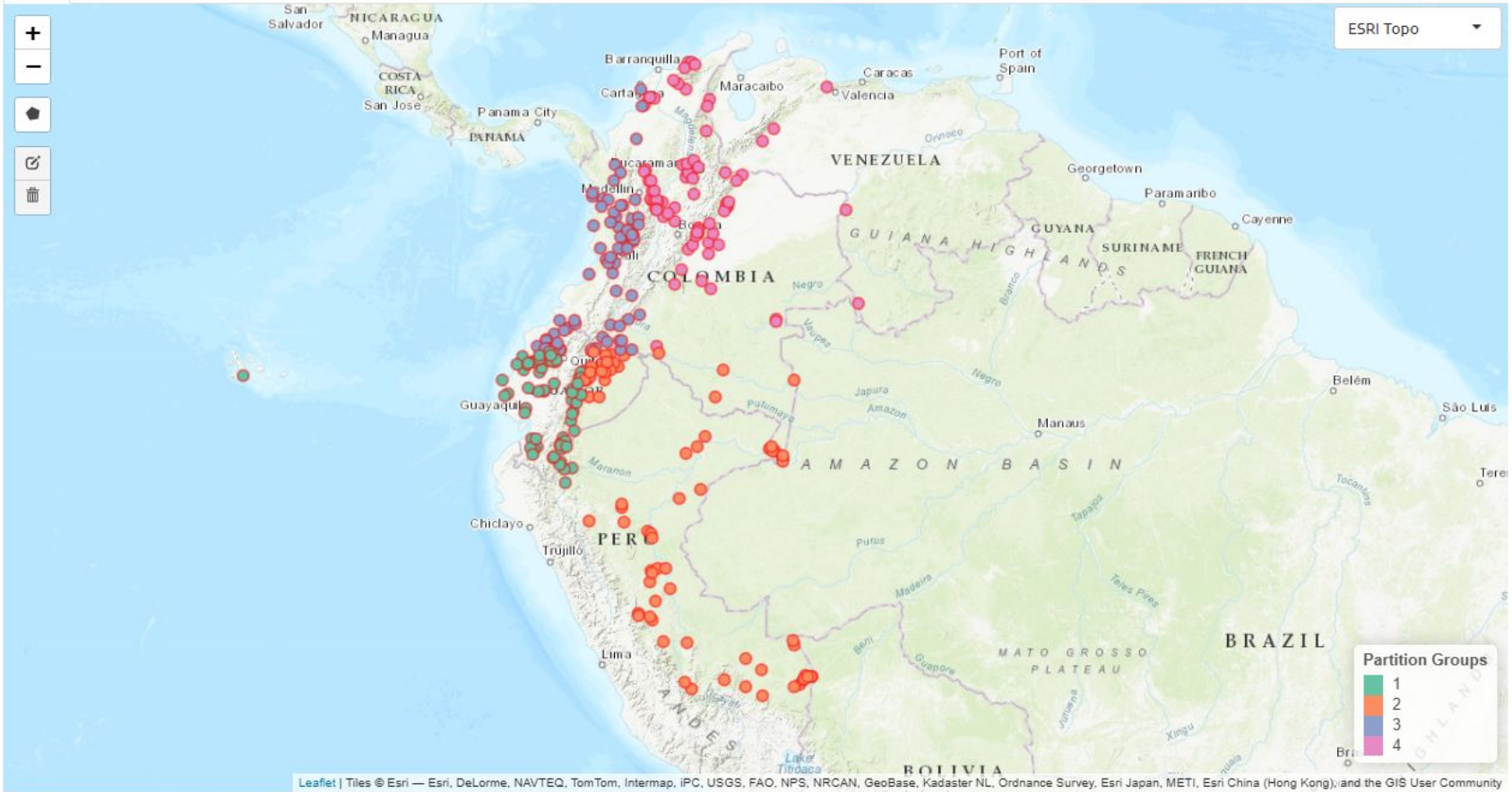
Species menu

Bactris_gasipaes

Log window

> ***Bactris gasipaes*** | Study extent: minimum convex polygon. Buffered by 1 degrees.
> ***Bactris gasipaes*** | Environmental data masked.
> ***Bactris gasipaes*** | 10000 random background points sampled out of 169172 total points.

Map | Occurrences | Results | Component Guidance | Module Guidance | Save



Component: Build and Evaluate Niche Model

Modules Available:

- (•) Maxent
- ( ) BIOCLIM

Module: Maxent

R packages: *ENMeval, dismo, maxnet*

(**NOTE** : see module guidance for troubleshooting tips if you are experiencing problems.)

Select algorithm

(•) maxnet ( ) maxent.jar

Select feature classes *(flexibility of modeled response)*
key: **L** inear, **Q** uadratic, **H** inge, **P** roduct

[x] L [x] LQ [ ] H [x] LQH [x] LQHP

Select regularization multipliers *(penalty against complexity)*

Multiplier step value

Multiplier step value

1

Are you using a categorical variable?

NO

Clamping?

FALSE

Parallel?

TRUE

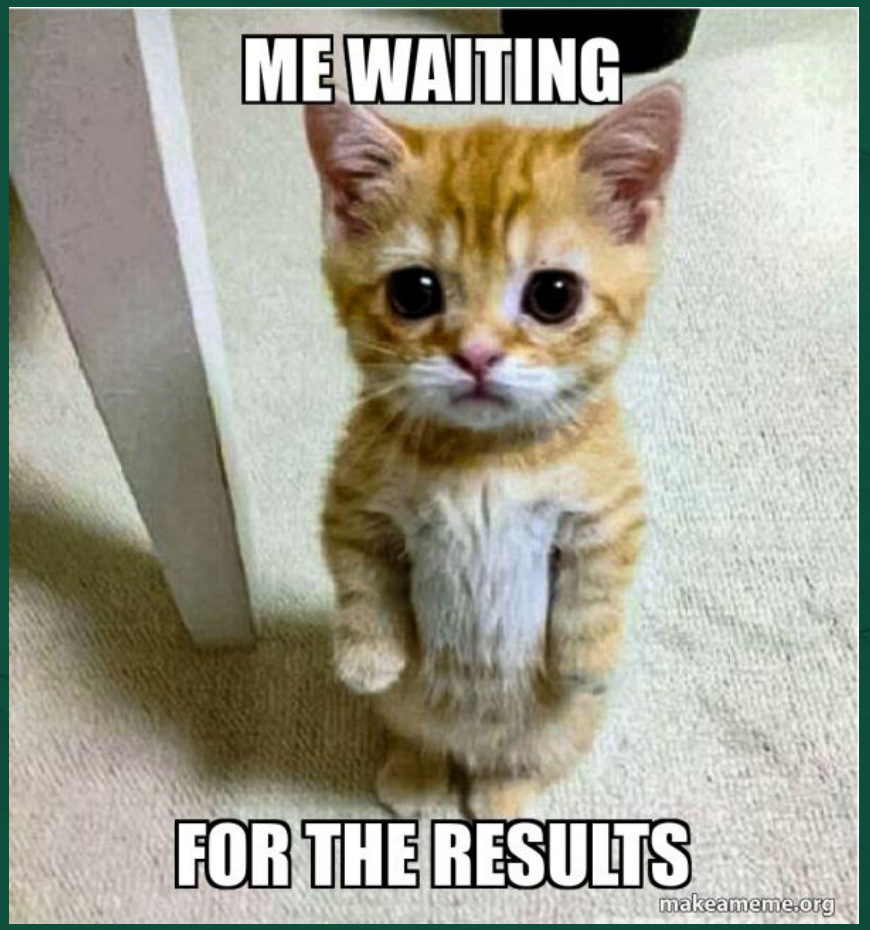
Specify the number of cores (max. 16)

8

[ ] **Batch**

Run

Be patient, ENMEval is calibrating the model!



Which one to use?

Evaluation | Lambdas

Evaluation statistics: full model and partition averages

	fc	rm	tune.args	auc.train	cbi.train	auc.diff.avg	auc.diff.sd	auc.val.avg	auc.val.sd	cbi.val.avg	cbi.val.sd	or.10p.avg	or.10p.sd	or.mtp.avg
1	L	1	fc.L_rm.1	0.744	0.951	0.081	0.08	0.766	0.092	0.871	0.077	0.047	0.039	0.00
2	LQ	1	fc.LQ_rm.1	0.787	0.927	0.138	0.038	0.795	0.121	0.832	0.066	0.113	0.109	0.00
3	LQH	1	fc.LQH_rm.1	0.866	0.983	0.081	0.074	0.813	0.105	0.835	0.111	0.143	0.142	0.00
4	LQHP	1	fc.LQHP_rm.1	0.87	0.981	0.091	0.065	0.808	0.109	0.812	0.119	0.17	0.141	0.0
5	L	2	fc.L_rm.2	0.743	0.951	0.095	0.073	0.762	0.098	0.851	0.073	0.057	0.062	0.00
6	LQ	2	fc.LQ_rm.2	0.779	0.928	0.133	0.039	0.792	0.117	0.832	0.082	0.09	0.084	0.00
7	LQH	2	fc.LQH_rm.2	0.839	0.988	0.088	0.07	0.823	0.113	0.831	0.089	0.07	0.107	0.00
8	LQHP	2	fc.LQHP_rm.2	0.842	0.984	0.096	0.05	0.817	0.115	0.762	0.16	0.063	0.092	0.00
9	L	3	fc.L_rm.3	0.741	0.945	0.099	0.072	0.763	0.1	0.847	0.093	0.063	0.062	0.00
10	LQ	3	fc.LQ_rm.3	0.774	0.931	0.129	0.046	0.79	0.114	0.784	0.116	0.07	0.073	0.00

Time to interpret outcomes

Component: Visualize Model Results

Modules Available:

- (•) Map Prediction
- ( ) Maxent Evaluation Plots
- ( ) Response Curves
- ( ) BIOCLIM Envelope Plots

Module: Map Prediction

R packages: *dismo*

Set threshold

No threshold

Prediction output

(•) cloglog ( ) logistic ( ) raw

Plot

Module Developers: *Jamie M. Kass, Robert Muscarella, Bruno Vilela, Gonzalo E. Pinilla-Buitrago, Bethany A. Johnson, Robert P. Anderson*

dismo references

Species Distribution Modeling

Package Developers: Robert J. Hijmans, Steven Phillips, John Leathwick, Jane Elith

CRAN | documentation

Species menu

Bactris_gasipaes

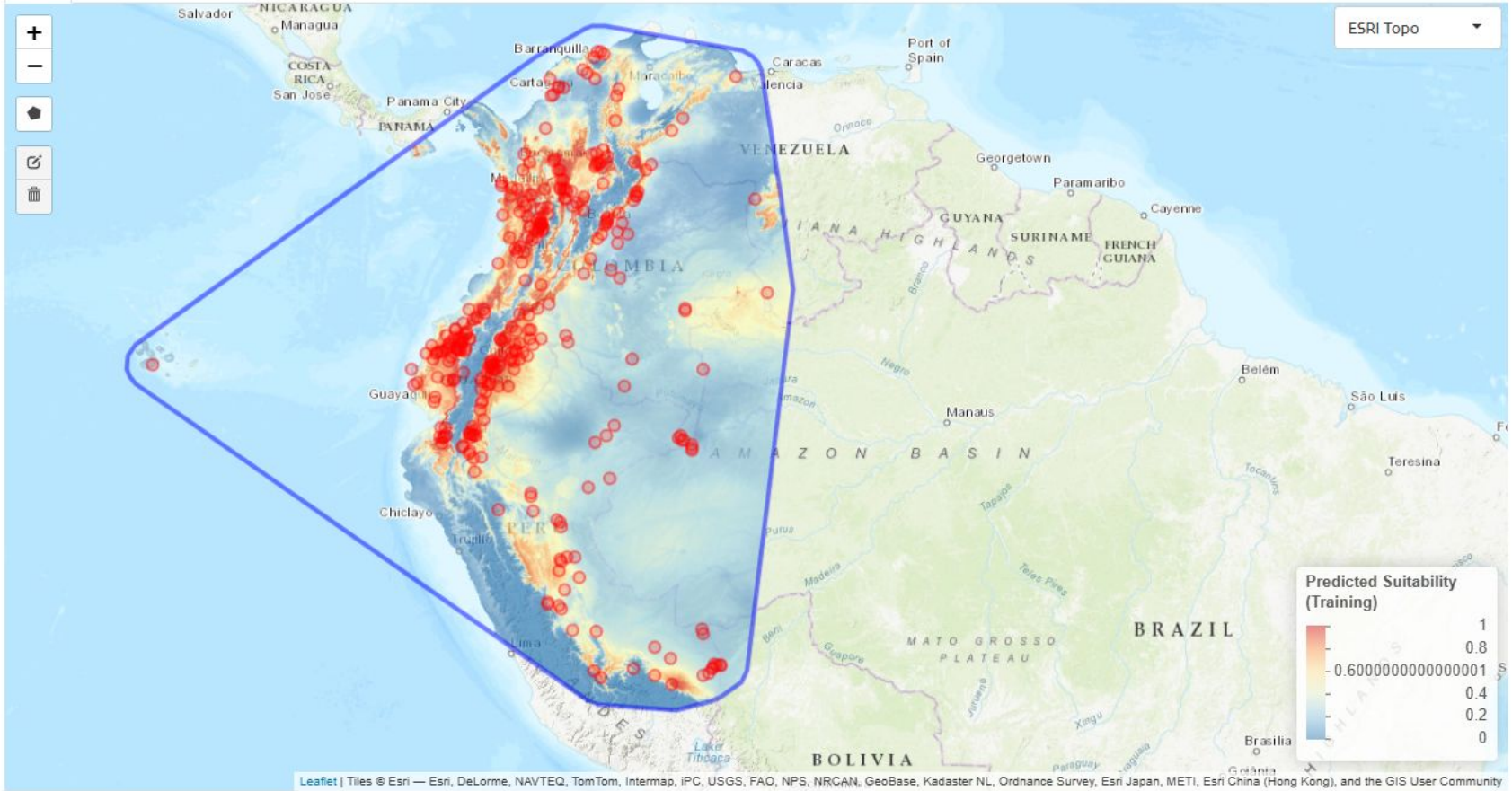
Select model:

fc.LQH_rm.2

Log window

> ***Bactris gasipaes*** | maxnet model prediction plotted.
> ***Bactris gasipaes*** | p10 threshold selected for maxnet cloglog (2.30e-01).
> ***Bactris gasipaes*** | maxnet model prediction plotted.
> ***Bactris gasipaes*** | maxnet cloglog model prediction plotted.

Map | Occurrences | Results | Component Guidance | Module Guidance | Save



Thresholding for gap analysis!

Component: Visualize Model Results

Modules Available:

- (●) Map Prediction
- ( ) Maxent Evaluation Plots
- ( ) Response Curves
- ( ) BIOCLIM Envelope Plots

Module: Map Prediction

R packages: *dismo*

Set threshold

10 Percentile Training Presence

Plot

Module Developers: *Jamie M. Kass, Robert Muscarella, Bruno Vilela, Gonzalo E. Pinilla-Buitrago, Bethany A. Johnson, Robert P. Anderson*

dismo references

Species Distribution Modeling

Package Developers: Robert J. Hijmans, Steven Phillips, John Leathwick, Jane Elith

CRAN | documentation

Species menu

Bactris_gasipaes

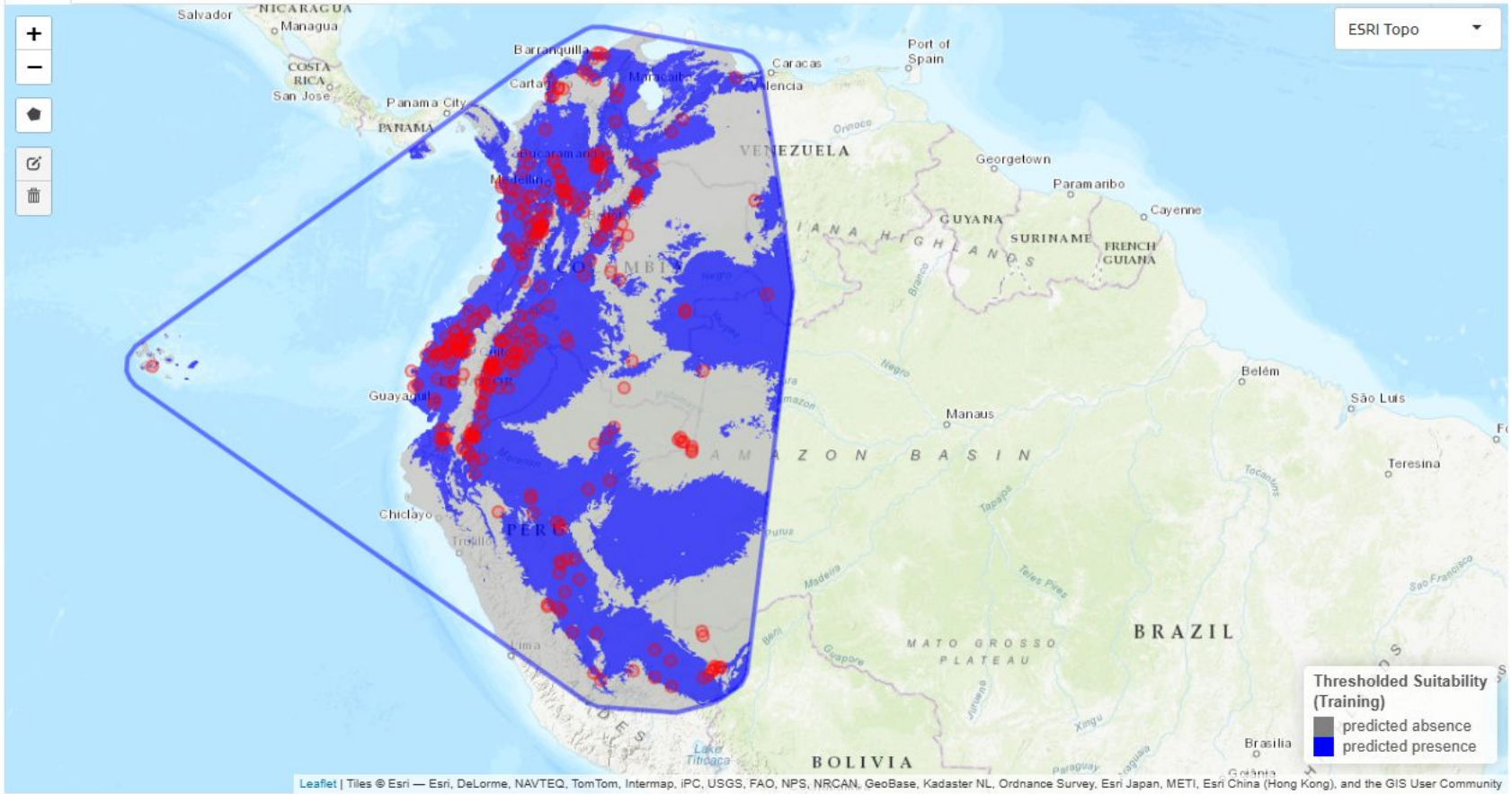
Select model:

fc.LQH_rm.2

Log window

(1.28e-02).
> ***Bactris gasipaes*** | maxnet model prediction plotted.
> ***Bactris gasipaes*** | p10 threshold selected for maxnet cloglog (2.30e-01).
> ***Bactris gasipaes*** | maxnet model prediction plotted.

Map | Occurrences | Results | Component Guidance | Module Guidance | Save



Component: Visualize Model Results

Modules Available:

- (•) Map Prediction
- ( ) Maxent Evaluation Plots
- ( ) Response Curves
- ( ) BIOCLIM Envelope Plots

Module: Map Prediction

R packages: *dismo*

Set threshold

No threshold

Prediction output

(•) cloglog ( ) logistic ( ) raw

Plot

Module Developers: *Jamie M. Kass, Robert Muscarella, Bruno Vilela, Gonzalo E. Pinilla-Buitrago, Bethany A. Johnson, Robert P. Anderson*

dismo references

Species Distribution Modeling

Package Developers: Robert J. Hijmans, Steven Phillips, John Leathwick, Jane Elith

CRAN | documentation

Species menu

Bactris_gasipaes

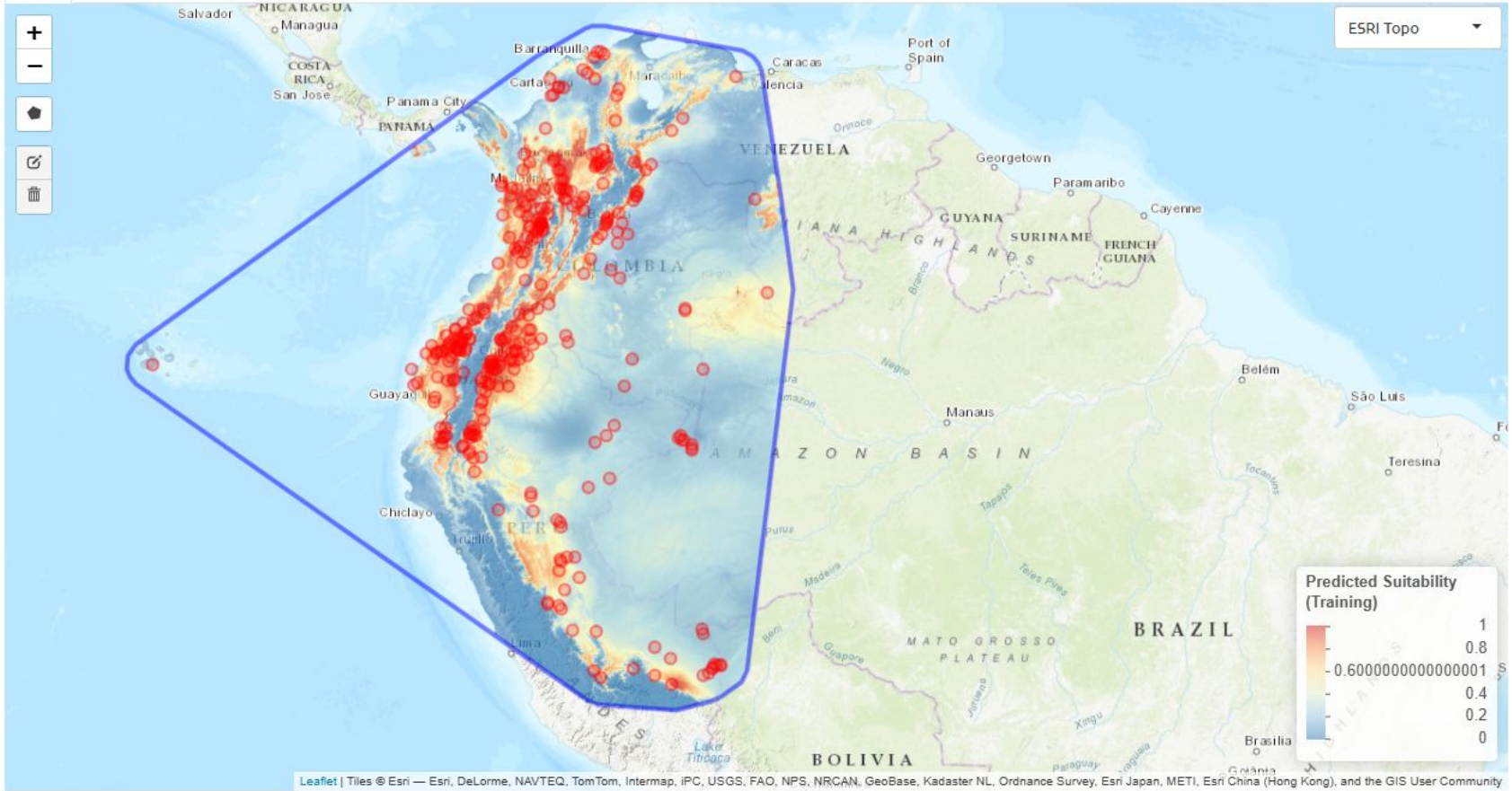
Select model:

fc.LQH_rm.2

Log window

> ***Bactris gasipaes*** | maxnet model prediction plotted.
> ***Bactris gasipaes*** | p10 threshold selected for maxnet cloglog (2.30e-01).
> ***Bactris gasipaes*** | maxnet model prediction plotted.
> ***Bactris gasipaes*** | maxnet cloglog model prediction plotted.

Map | Occurrences | Results | Component Guidance | Module Guidance | Save



I want to Project my model to other countries

Component: Model Transfer

Modules Available:

- (●) Transfer to New Extent
- ( ) Transfer to New Time
- ( ) Transfer to User Environments
- ( ) Calculate Environmental Similarity

Module: Transfer to New Extent

R packages: *dismo*

Step 1: *Choose Study Region*

Select method

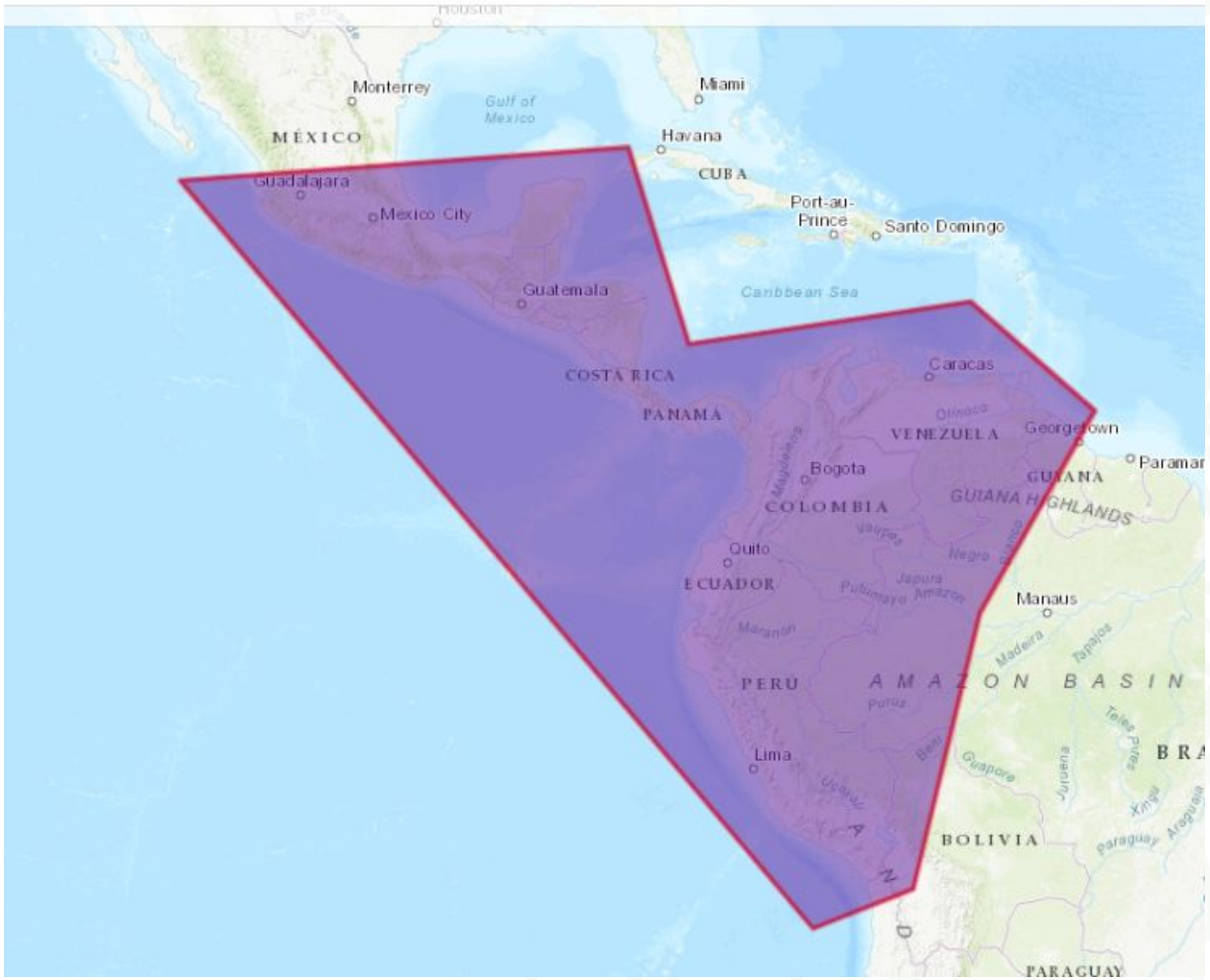
Draw polygon

Draw a polygon and select buffer distance

Study region buffer distance (degree)

0

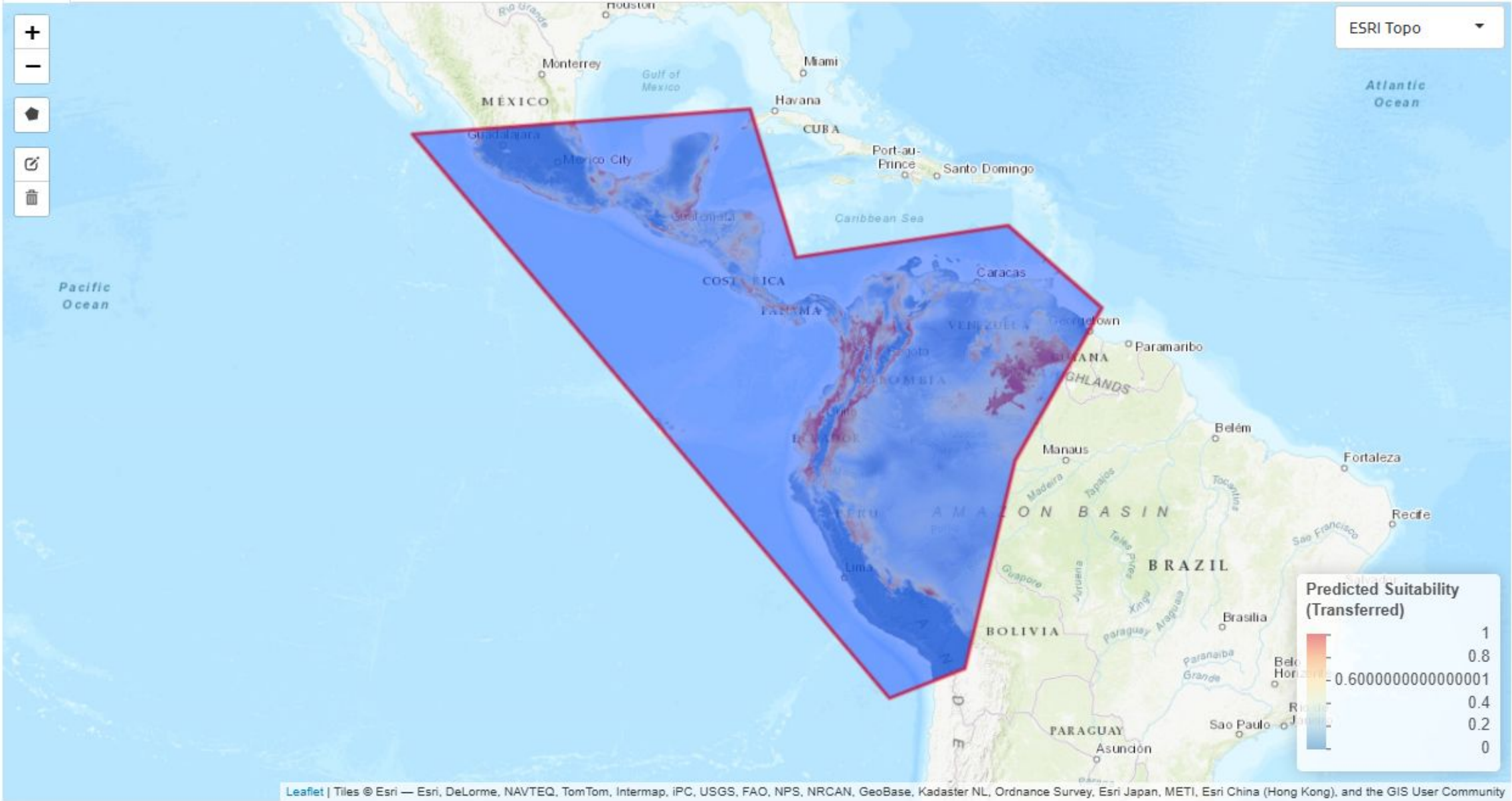
Create



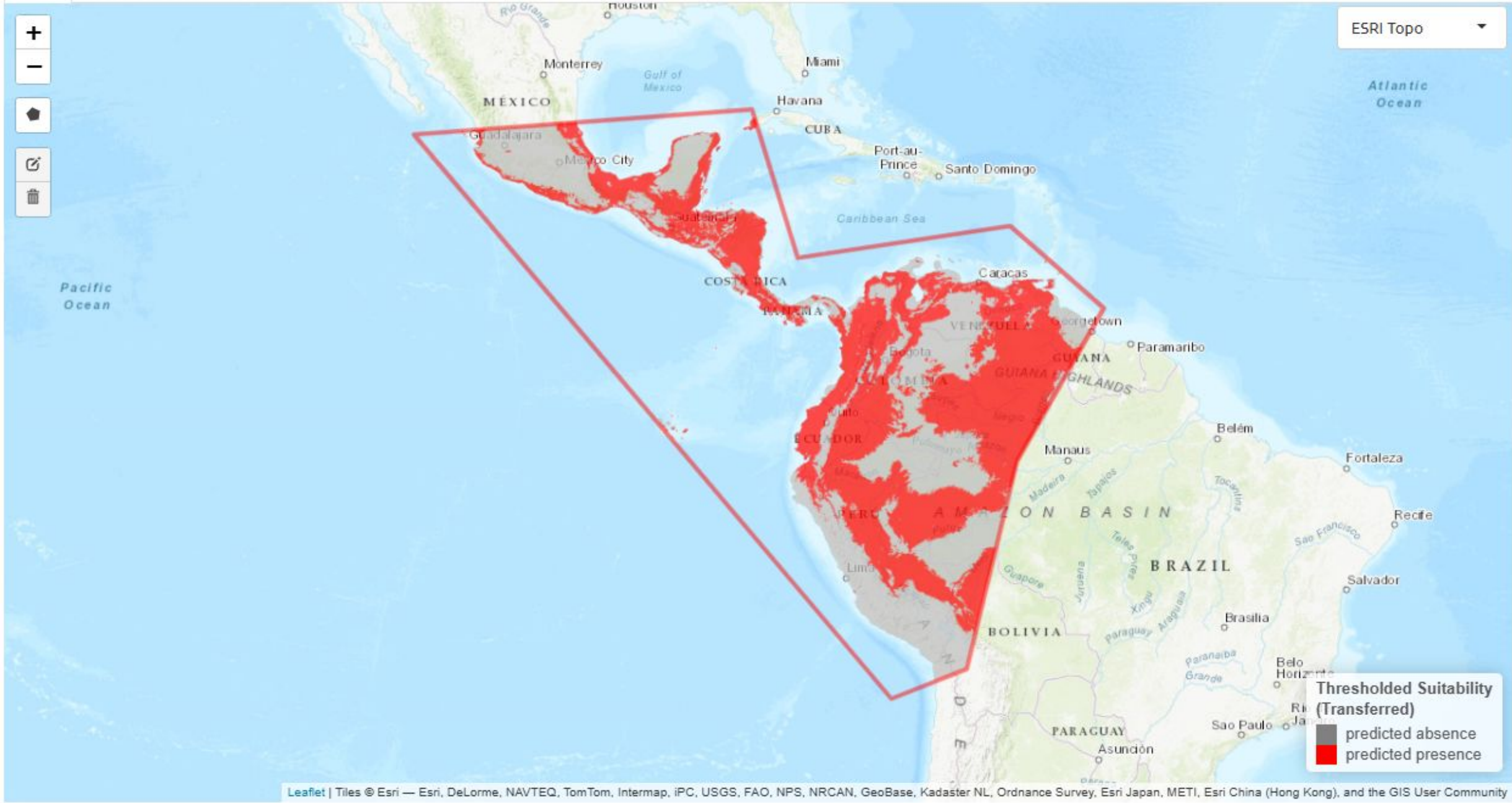
fc.LQH_rm.2

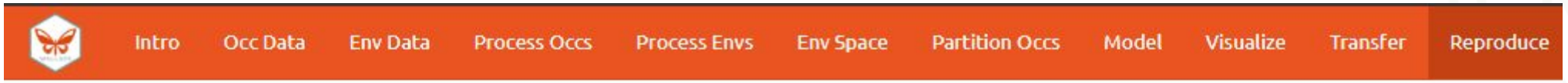
> ***Bactris gasipaes*** | Transfer of model to new area with cloglog output.

Map | Occurrences | Results | Component Guidance | Module Guidance | Save



Map | Occurrences | Results | Component Guidance | Module Guidance | Save





Component: Reproduce

Modules Available:

- (●) Session Code
- ( ) Metadata
- ( ) Reference Packages

Module: Download Session Code

R packages: *rmarkdown, knitr*

Select download file type

Rmd

Download Session Code

Module Developers: *Jamie M. Kass, Gonzalo E. Pinilla-Buitrago, Bruno Vilela, Robert P. Anderson*

Module: *Session Code*

BACKGROUND

Over the decade of the 2010s, scientific practice increasingl
the area of modeling species niches/distributions has advan
community-driven standards (see Fitzpatrick et al. 2021 for
et al. 2019), standardized metadata frameworks (RMMS, Me
reporting (ODMAP, Zurell et al. 2020). These tools facilitate
model, indicating whether it meets minimal standards for ap
leveraging `ENMeval 2.0` and `rangeModelMetadata`, *Wallace*
(which also form the basis of ODMAP reporting) and allows
Wallace promotes documentation and downstream assessm
information that includes sources of input data, methodolo
Module: *Download Session Code*) is a file that can be re-run
and dependent packages). Many intermediate and advanced
Additionally, *Wallace* now provides citations of the particula
Reference Packages).

Via the *Session Code* module, the user can download files th
executable code that can reproduce them). This functionalit
Fitzpatrick et al. 2021).

CROP TRUST · Alianza Bioversity & CIAT · CGIAR

Saving

Note: To save your session code or metadata, use the Reproduce component

Save Session

By saving your session into an RDS file, you can resume working on it at a later time or you can share the file with a collaborator.

⚠ The current session data is large, which means the downloaded file may be large and the download might take a long time.

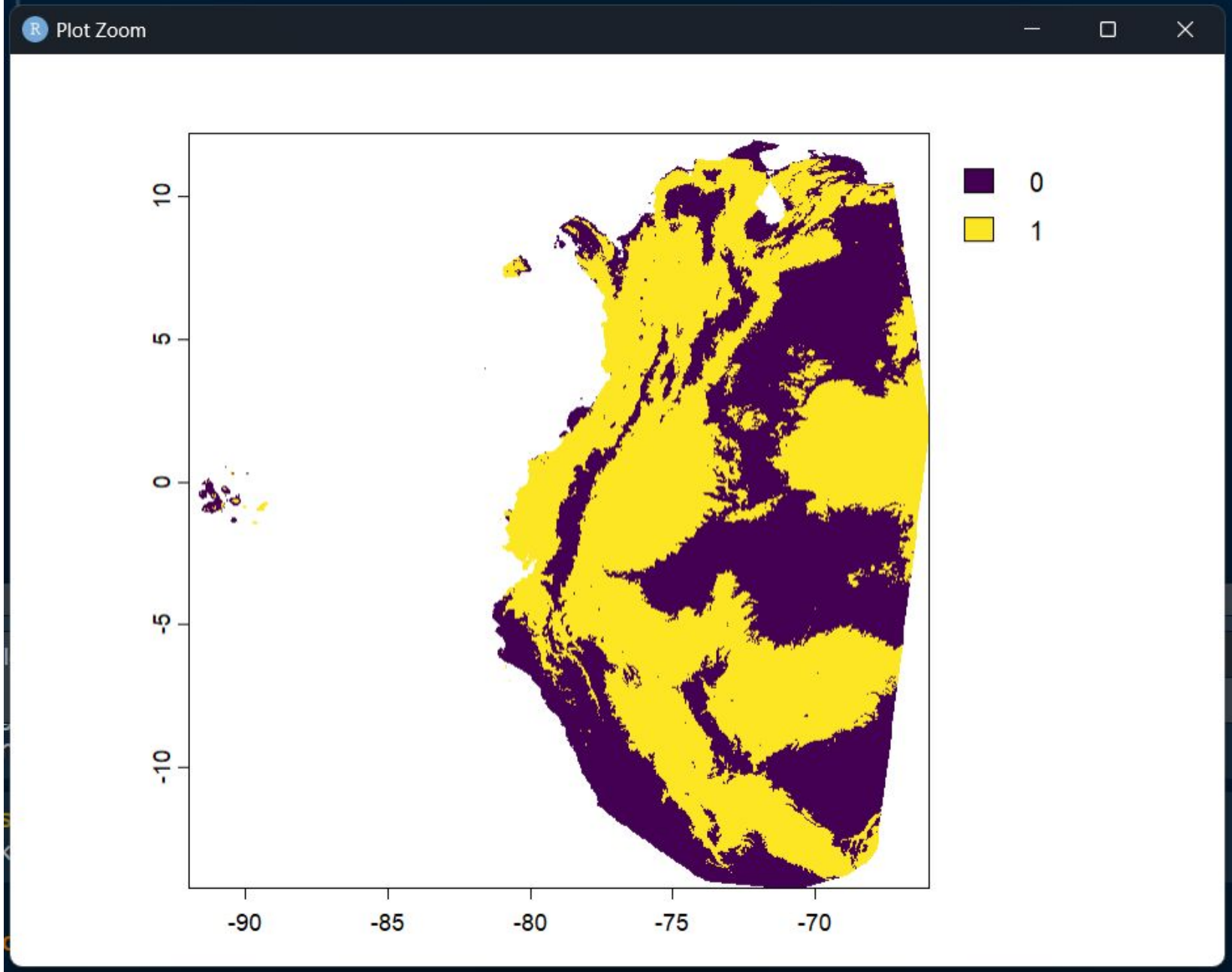
Save Session

Download Data

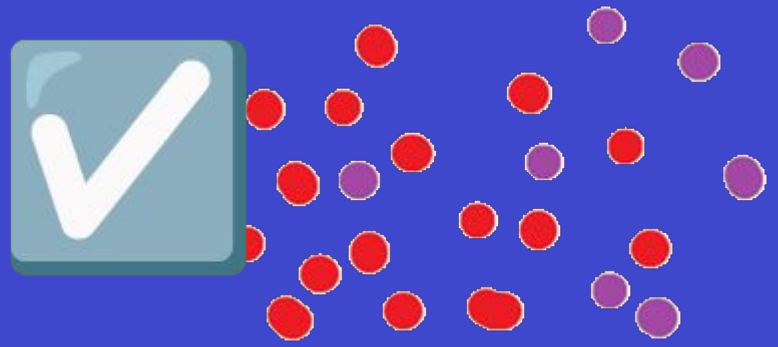
Download data/results from analyses from currently selected module

Download Bioclim plot	PNG file	
Download Maxent plots	ZIP file	
Download Response plots	ZIP file	
Download current prediction (Select download file type**)	GeoTIFF	Prediction file

```r
require(terra)
plot(rast("D:/BOLDER/Bactris_gasipaes_fc.LQH_rm.2_p10.tif"))
```

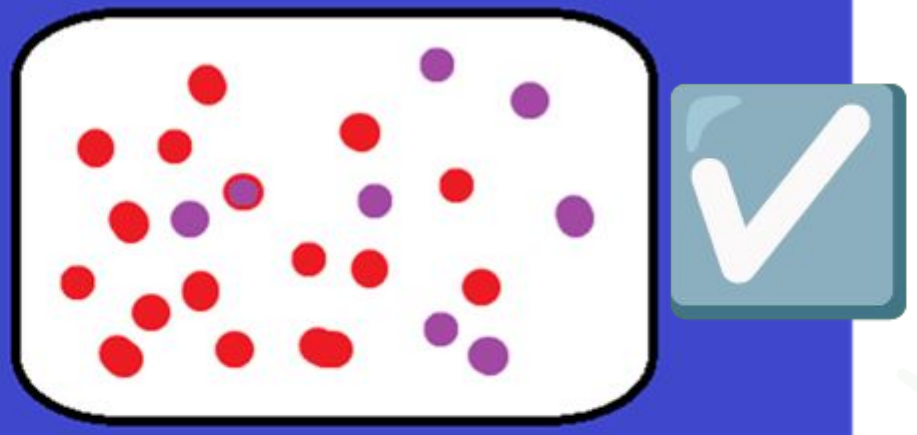


A



Species occurrences

B



Realized niche

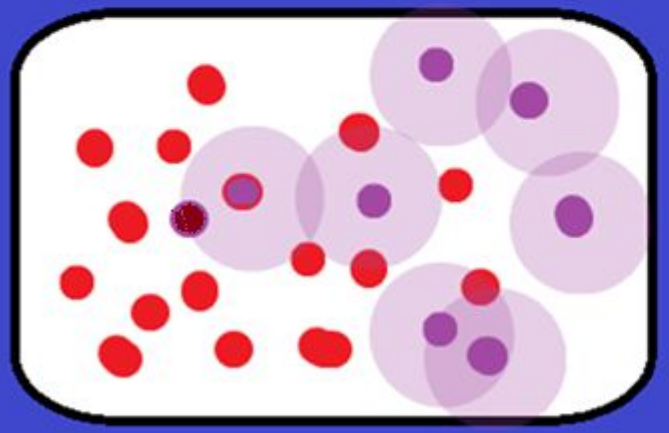
Germplasm

Other coordinates

Collected germplasm area

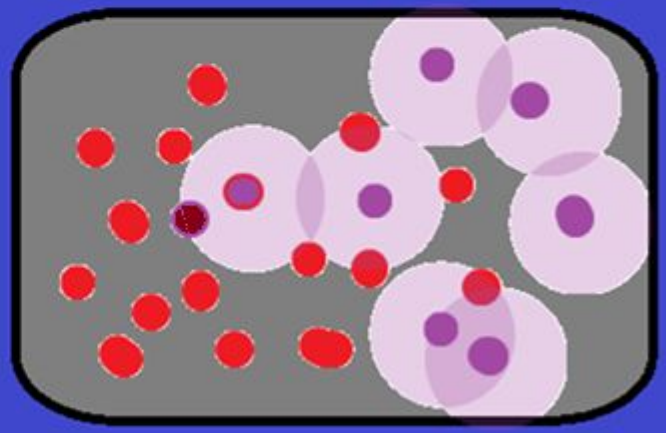
Potential area to be collected

C



Sampled germplasm

D



Where to collect

CROP TRUST

Alianza Bioversity & CIAT

CGIAR

Questions?