

Transdisciplinary Nature Conservation Summer School

Species Modeling for Conservation

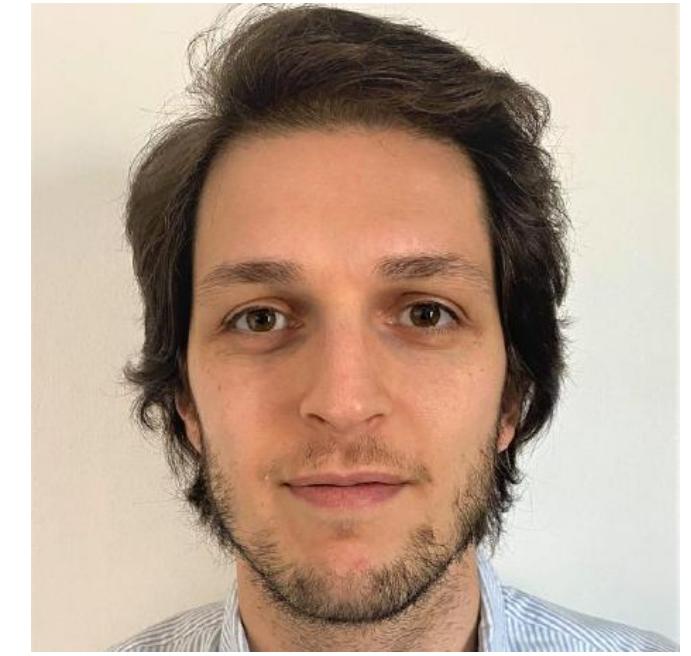
University of Lausanne

Lausanne, Sept. 15th 2022

Federico Riva, PhD

Federico Riva

- BSc and MSc at UniTo
- PhD at UAlberta
- PDF at UAlberta, Carleton U, UNIL
- Landscape ecologist and conservation scientist
- IUCN Butterfly and Moth SSC and Young Professional group

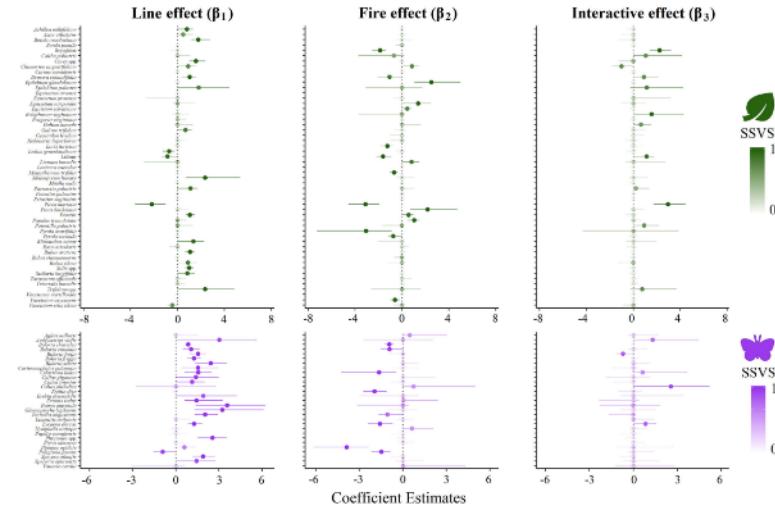


@riva_ecology

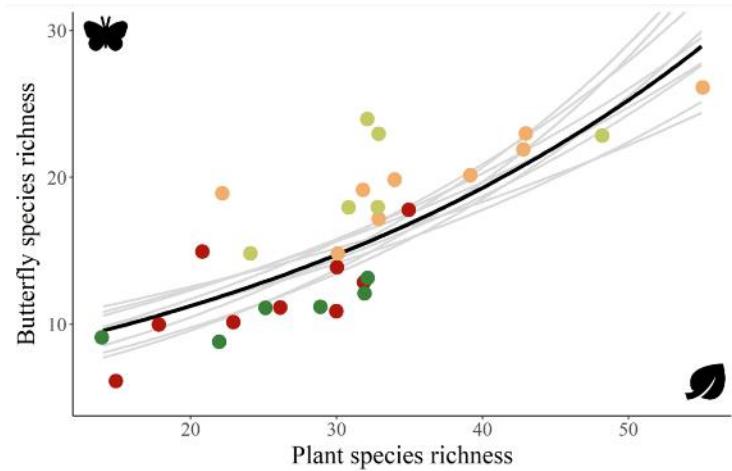


Federico.riva.1@unil.ch

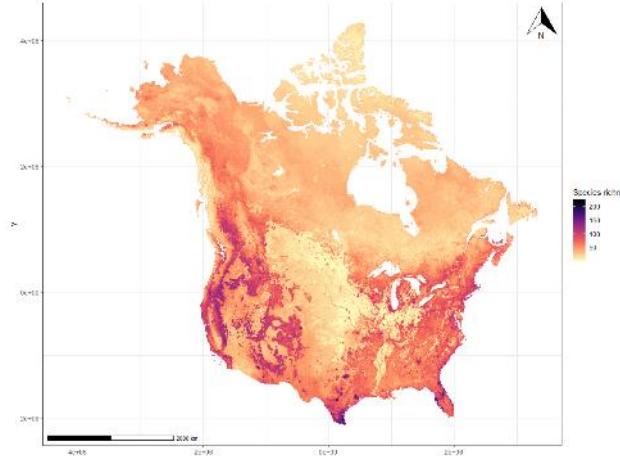
Community ecology



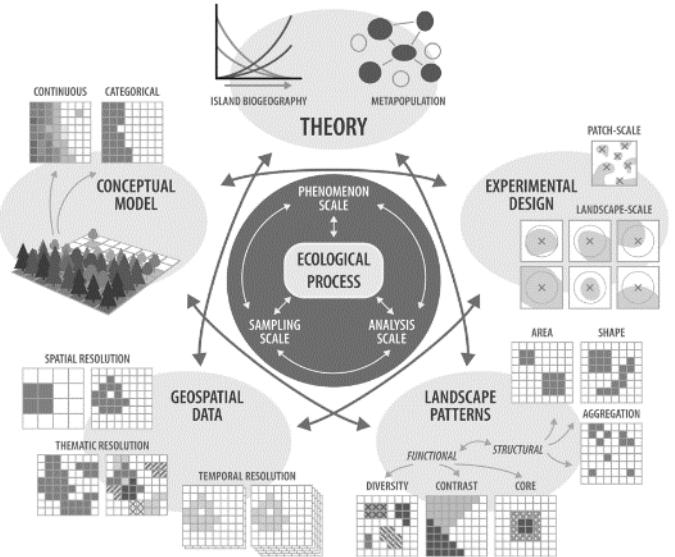
Macroecology



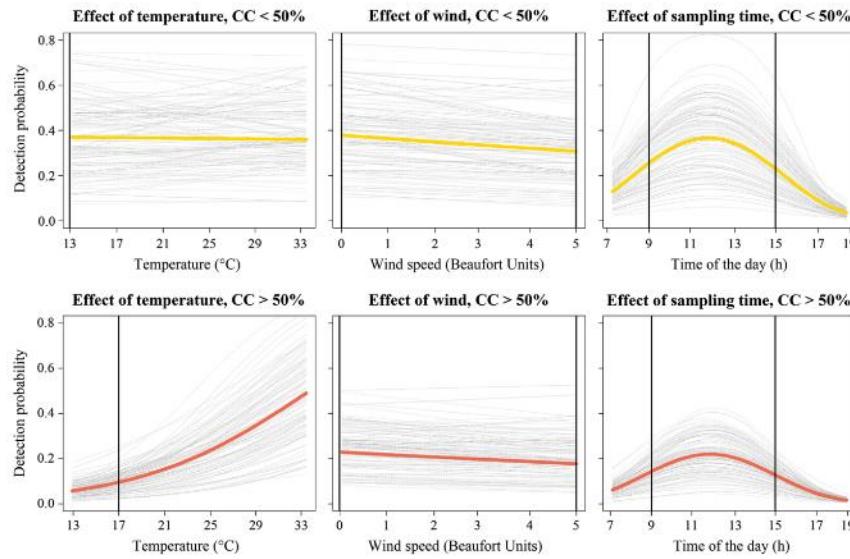
Biogeography



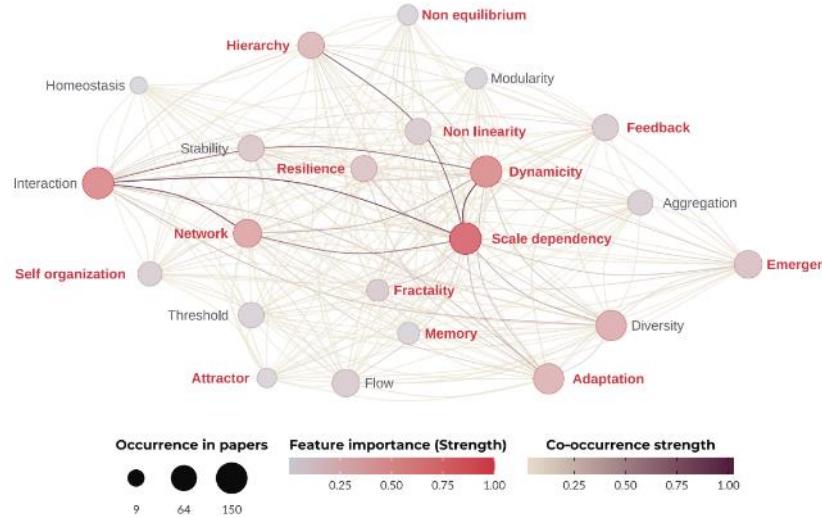
Landscape ecology

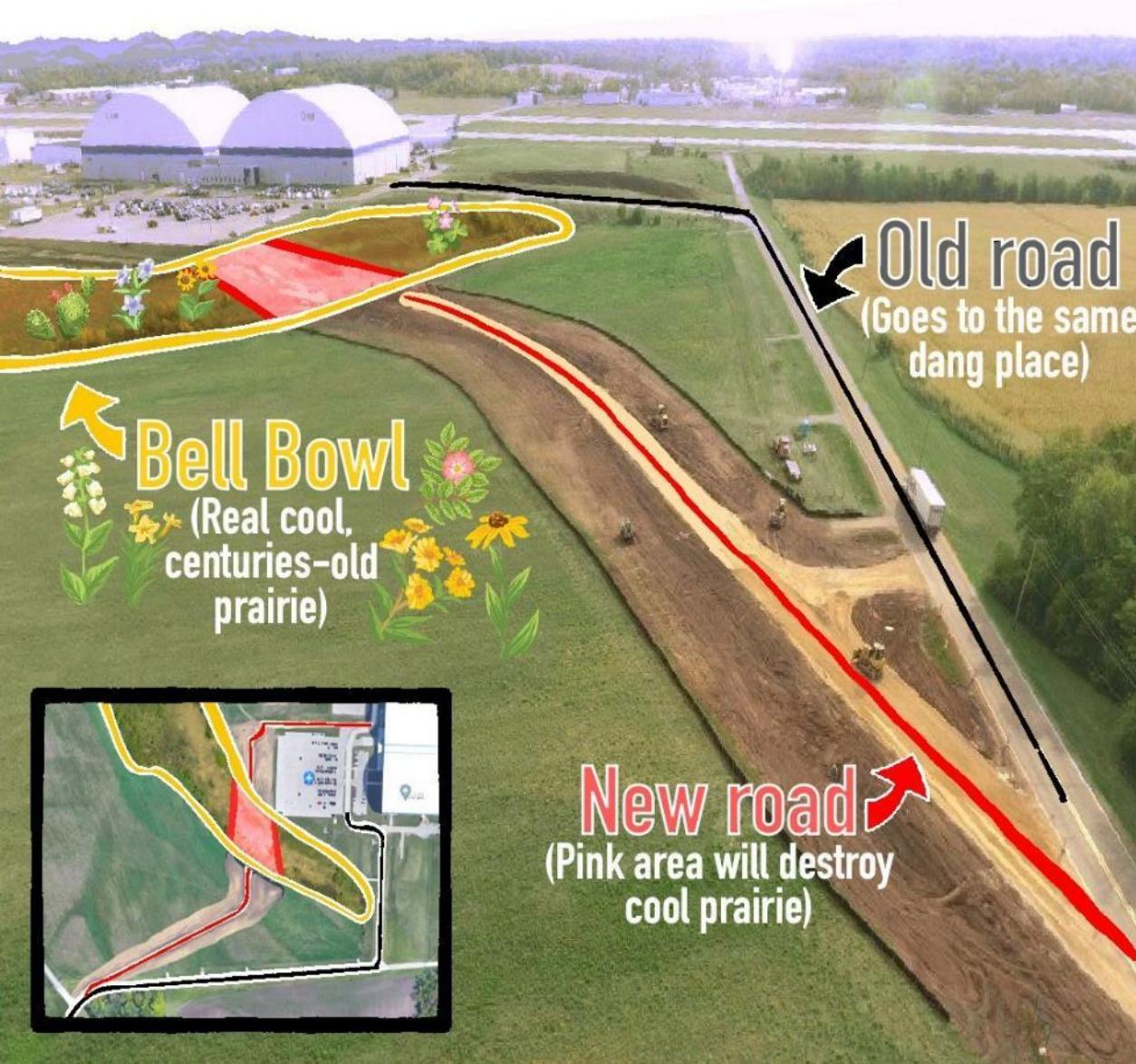


Hierarchical models



Complexity





“Bottom up” biodiversity conservation

Key concepts for SDMs

ECOGRAPHY

Review and synthesis

A standard protocol for reporting species distribution models

Damaris Zurell, Janet Franklin, Christian König, Phil J. Bouchet, Carsten F. Dormann, Jane Elith, Guillermo Fandos, Xiao Feng, Gurutzeta Guillera-Arroita, Antoine Guisan, José J. Lahoz-Monfort, Pedro J. Leitão, Daniel S. Park, A. Townsend Peterson, Giovanni Rapacciulo, Dirk R. Schmaltz, Boris Schröder, Josep M. Serra-Díaz, Wilfried Thuiller, Katherine L. Yates, Niklaus E. Zimmermann and Cory Merow

EDITOR'S
CHOICE

GENERAL SPECIFICATIONS

Overview / Conceptualisation

- Model objective
- Taxon, location, predictors, scale
- Conceptual underpinning
- Software, codes and data availability

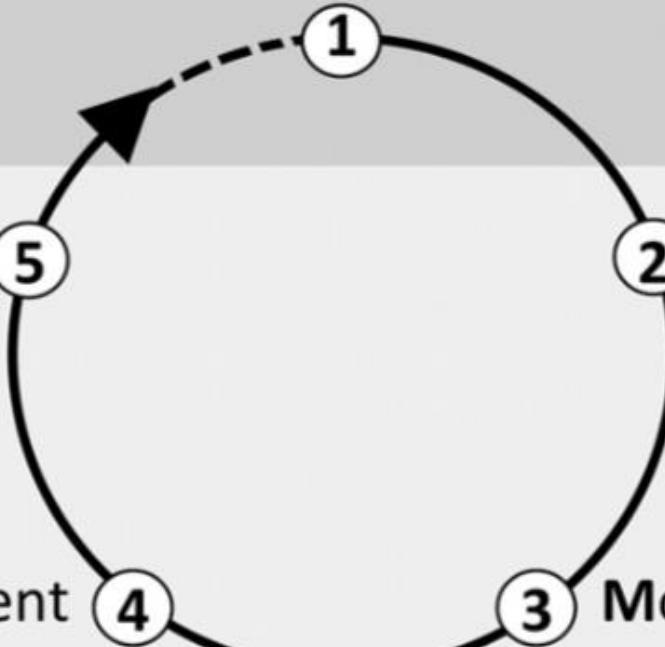
TECHNICAL DETAILS

Predictions

- Prediction output
- Uncertainty quantification

Assessment

- Performance statistics
- Plausibility: response shapes, expert judgement



Data

- Biodiversity data
- Data partitioning
- Environmental data
- Transfer data

Model fitting

- Variable selection
- Model settings and model complexity
- Model estimates, variable importance
- Model selection, averaging, ensembles
- Non-independence analyses
- Threshold selection

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?
Data	Model workflow	Brief description of modelling steps
	Software, codes and data	Specify software, availability of codes, availability of 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 Multicollinearity Model settings/model complexity Model estimates Model selection/model averaging/ensembles Non-independence correction/analyses Threshold selection Performance statistics	Details on pre-selection of variables Methods for identifying and dealing with multicollinearity Models settings for all selected algorithms and for extrapolation beyond sample range Model coefficients, variable importance Model selection strategy, method for model averaging, ensemble method Spatial autocorrelation in residuals, temporal autocorrelation in residuals, nested data Details on threshold selection Performance statistics estimated on training data, on validation data and on test (truly independent) data
Assessment		
Prediction	Plausibility check Prediction output Uncertainty quantification	Response plots; expert judgements (e.g. map display) Prediction unit; post-processing steps Uncertainty through algorithms, input data, parameters, scenarios; visualisation/treatment of novel environments

[Grey box] Obligatory; [Green box] Objective: mapping/interpolation; [Purple box] Objective: forecast/transfer; [White box] Optional/context dependent.

Standards for distribution models in biodiversity assessments

Miguel B. Araújo^{1,2,3*}, Robert P. Anderson^{4,5,6}, A. Márcia Barbosa³, Colin M. Beale⁷, Carsten F. Dormann⁸, Regan Early⁹, Raquel A. Garcia^{2,3,10,11}, Antoine Guisan^{12,13}, Luigi Maiorano^{14,15}, Babak Naimi², Robert B. O'Hara^{16,17}, Niklaus E. Zimmermann^{18,19}, Carsten Rahbek^{2,20}

Demand for models in biodiversity assessments is rising, but which models are adequate for the task? We propose a set of best-practice standards and detailed guidelines enabling scoring of studies based on species distribution models for use in biodiversity assessments. We reviewed and scored 400 modeling studies over the past 20 years using the proposed standards and guidelines. We detected low model adequacy overall, but with a marked tendency of improvement over time in model building and, to a lesser degree, in biological data and model evaluation. We argue that implementation of agreed-upon standards for models in biodiversity assessments would promote transparency and repeatability, eventually leading to higher quality of the models and the inferences used in assessments. We encourage broad community participation toward the expansion and ongoing development of the proposed standards and guidelines.

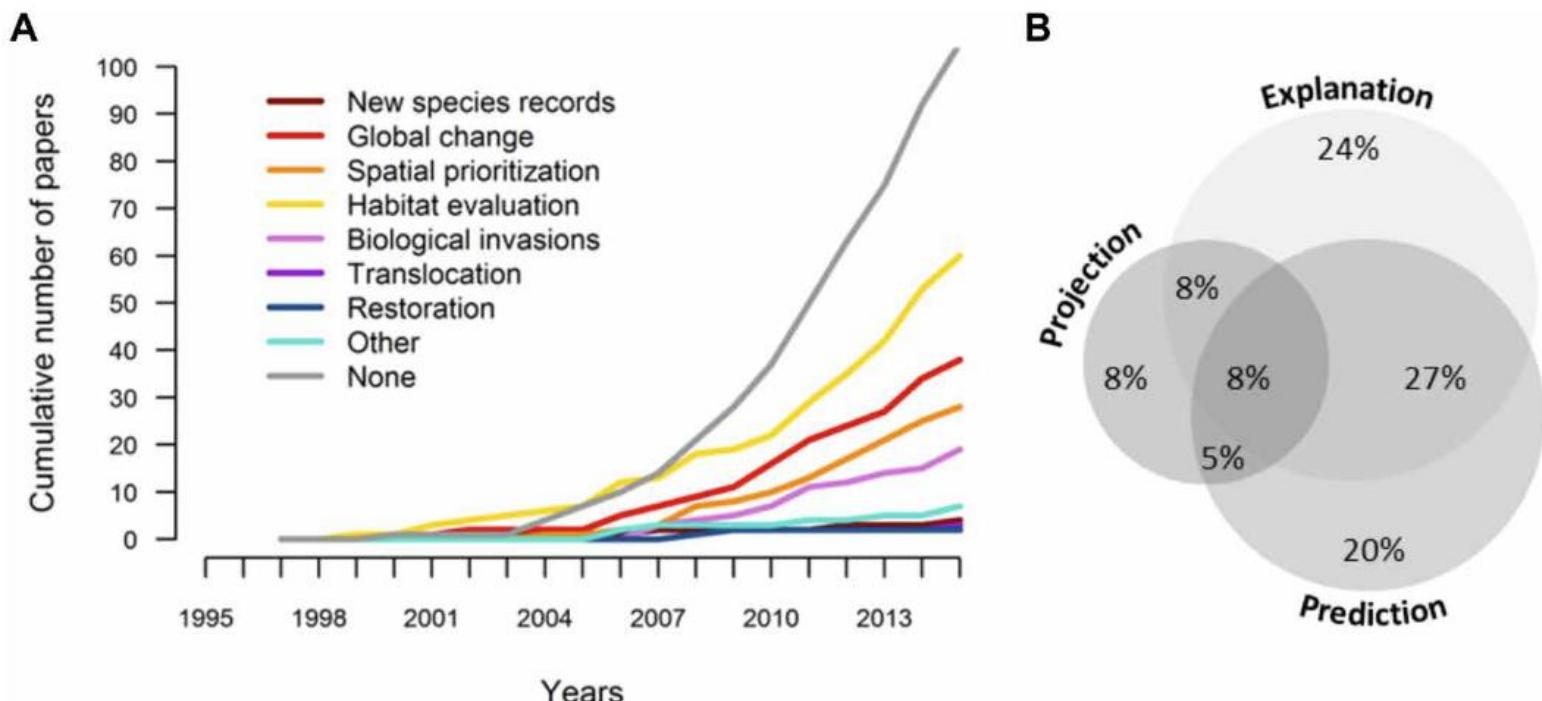
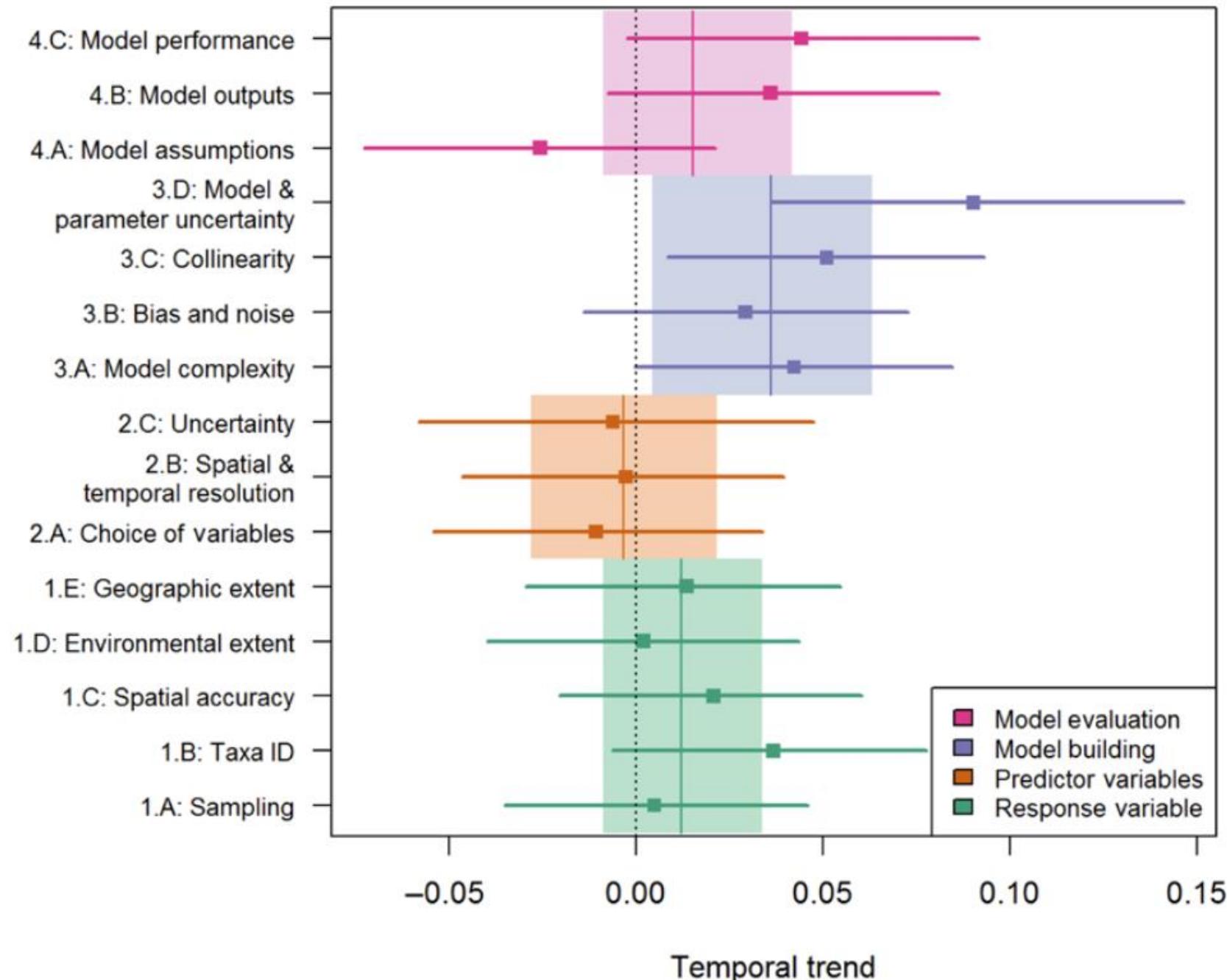
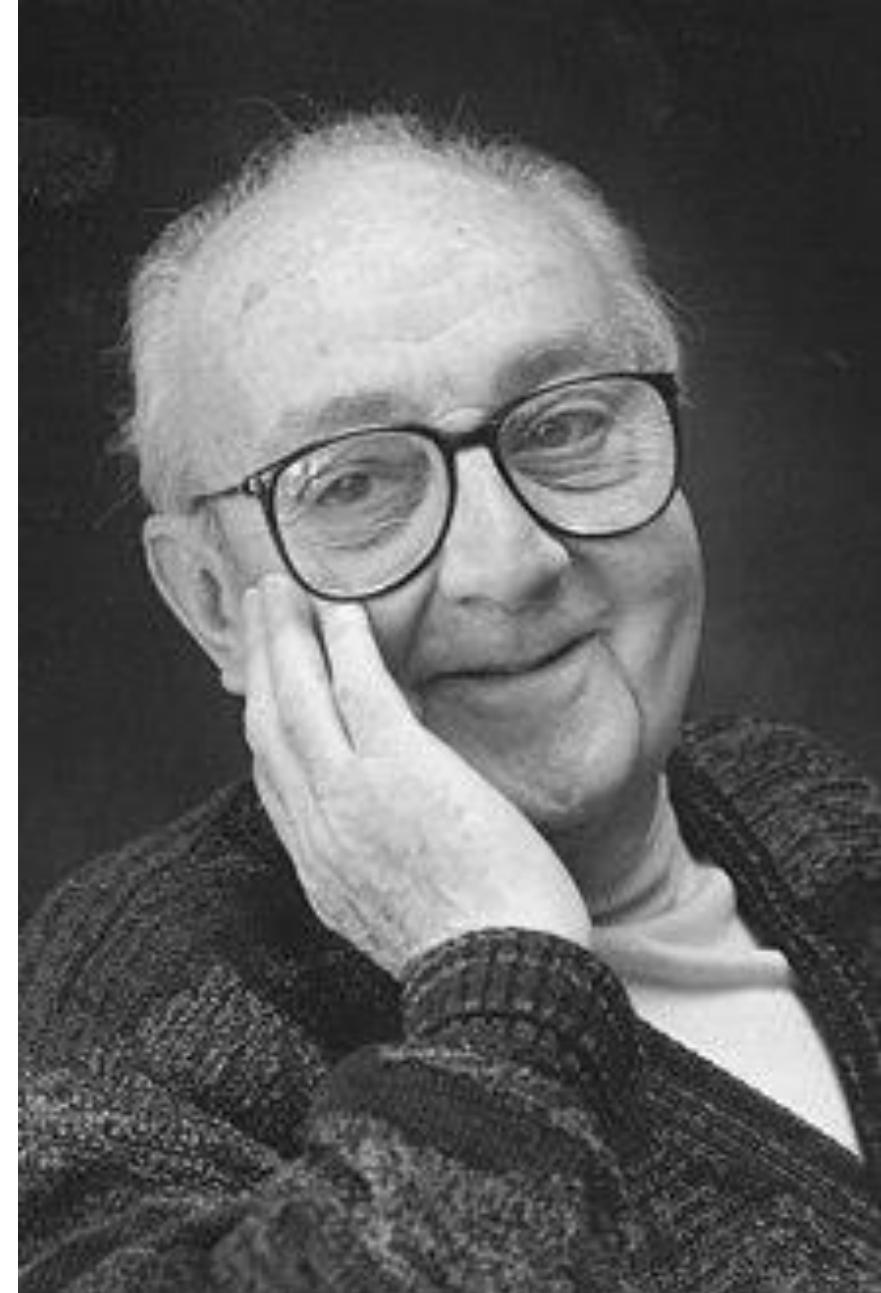


Fig. 1. Uses of SDMs. Classification of published species distribution modeling studies by (A) type of biodiversity assessment accomplished with the trend in the numbers of studies shown over time and (B) purpose of the model (see glossary in text S4). In (A), the trend for translocation is very similar to that of restoration, and hence is hardly visible. The classification is based on a random sample of 400 papers (of 6483 identified articles mentioning statistical models of species distributions); 238 of the randomly selected papers used SDMs and were included in this analysis. Details on the literature search and analyses appear in text S1, figs. S1.1, S1.2, S1.3, and S1.4, and tables S1.1, S1.2, and S1.3.



**All models are wrong,
but some are useful**

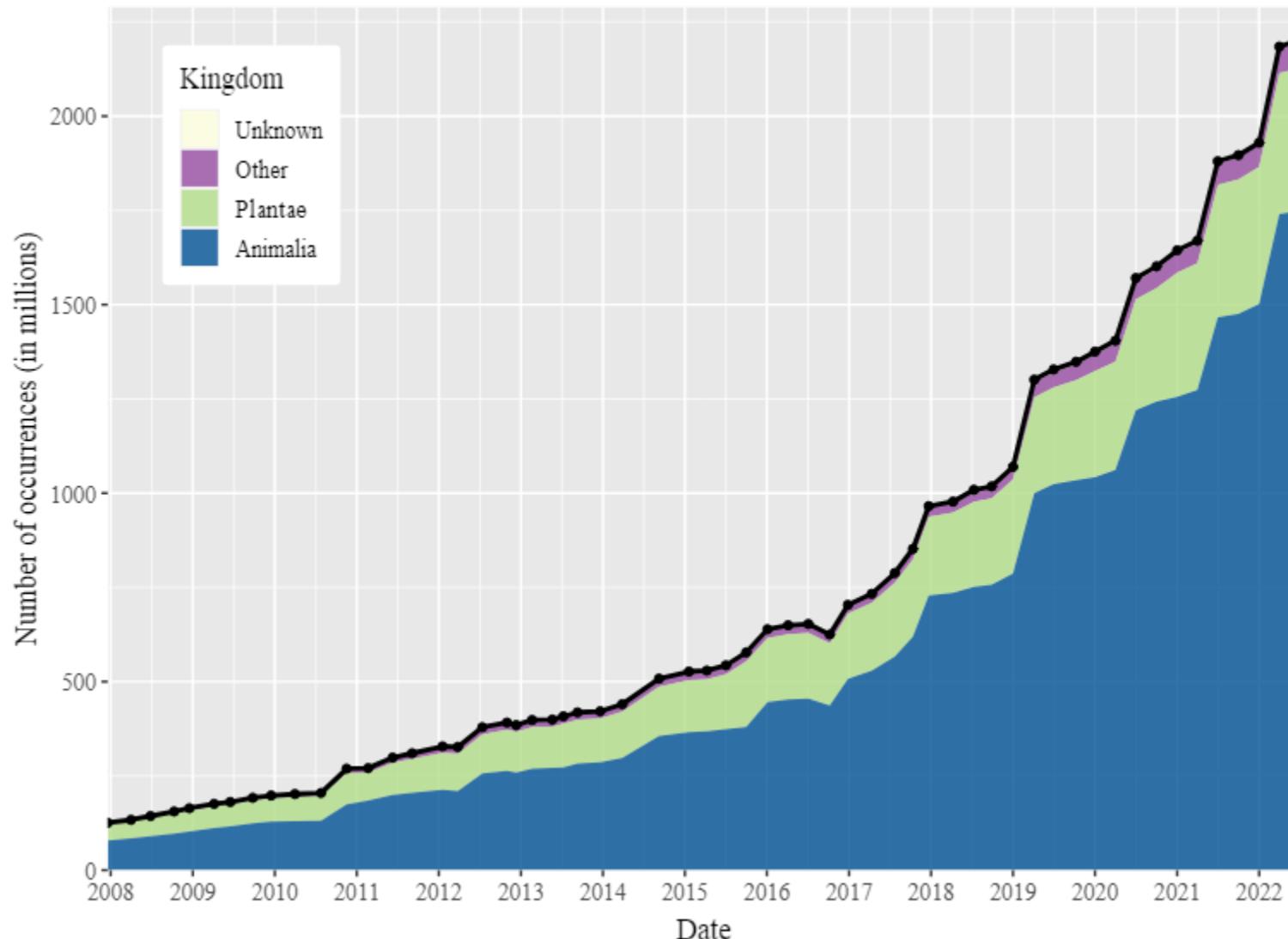
George E.P. Box



Preparing biodiversity data

Lots of data

Species occurrence records accessible through GBIF over time



Not enough data

Editor's choice and Research |  Open Access |  

Sampling biases shape our view of the natural world

Alice C. Hughes , Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, Huijie Qiao 

First published: 21 June 2021 | <https://doi.org/10.1111/ecog.05926> | Citations: 13

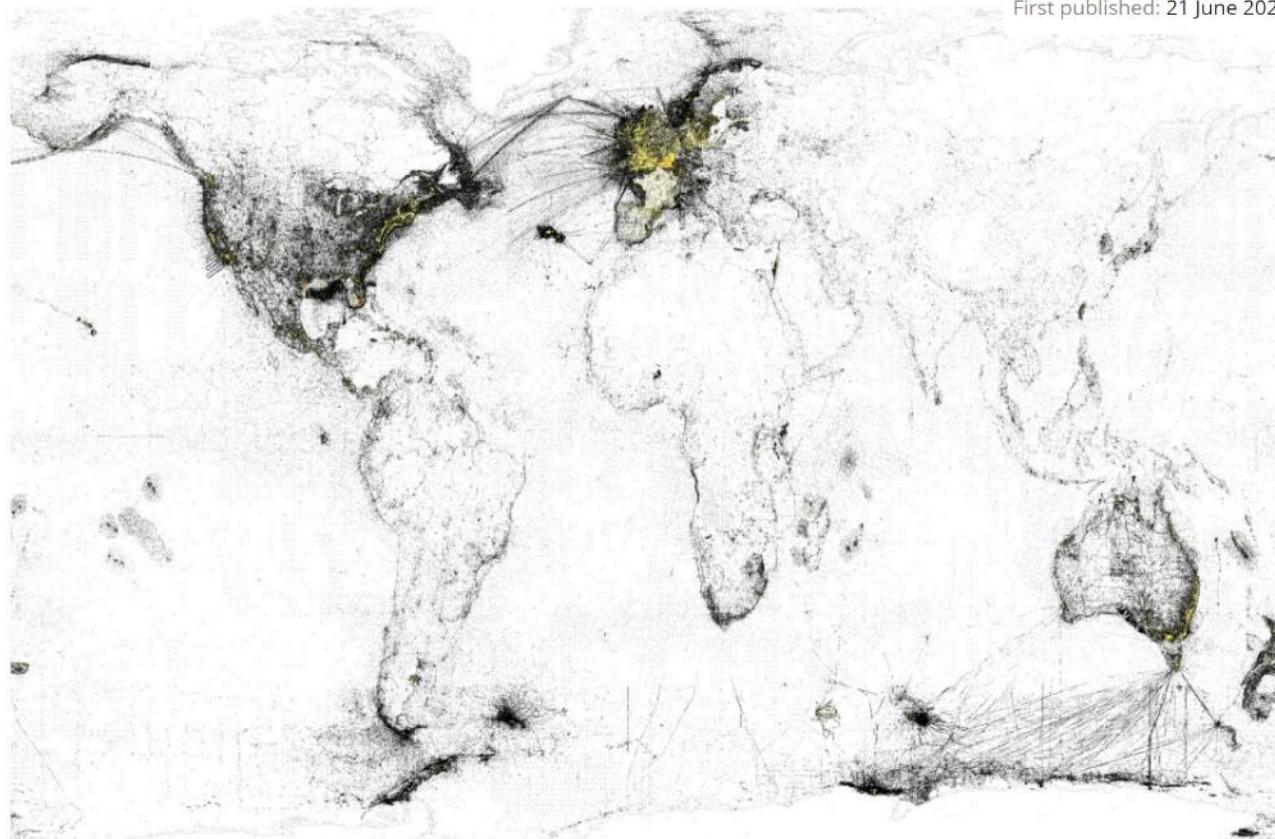


Figure 1. Areas with high numbers of records in GBIF and OBIS databases. Black 1–50 records, Yellow-red > 50 records at a 5 km resolution.

- We have a ton of data, but unevenly spaced
- SDMs can be used to address this “Wallacean shortfall”
- Statistical approaches can account for limitations in data, but typically data quality >> method

Different types of data

- Occurrence locations
- Presence and absence data
- Detection and non-detection data

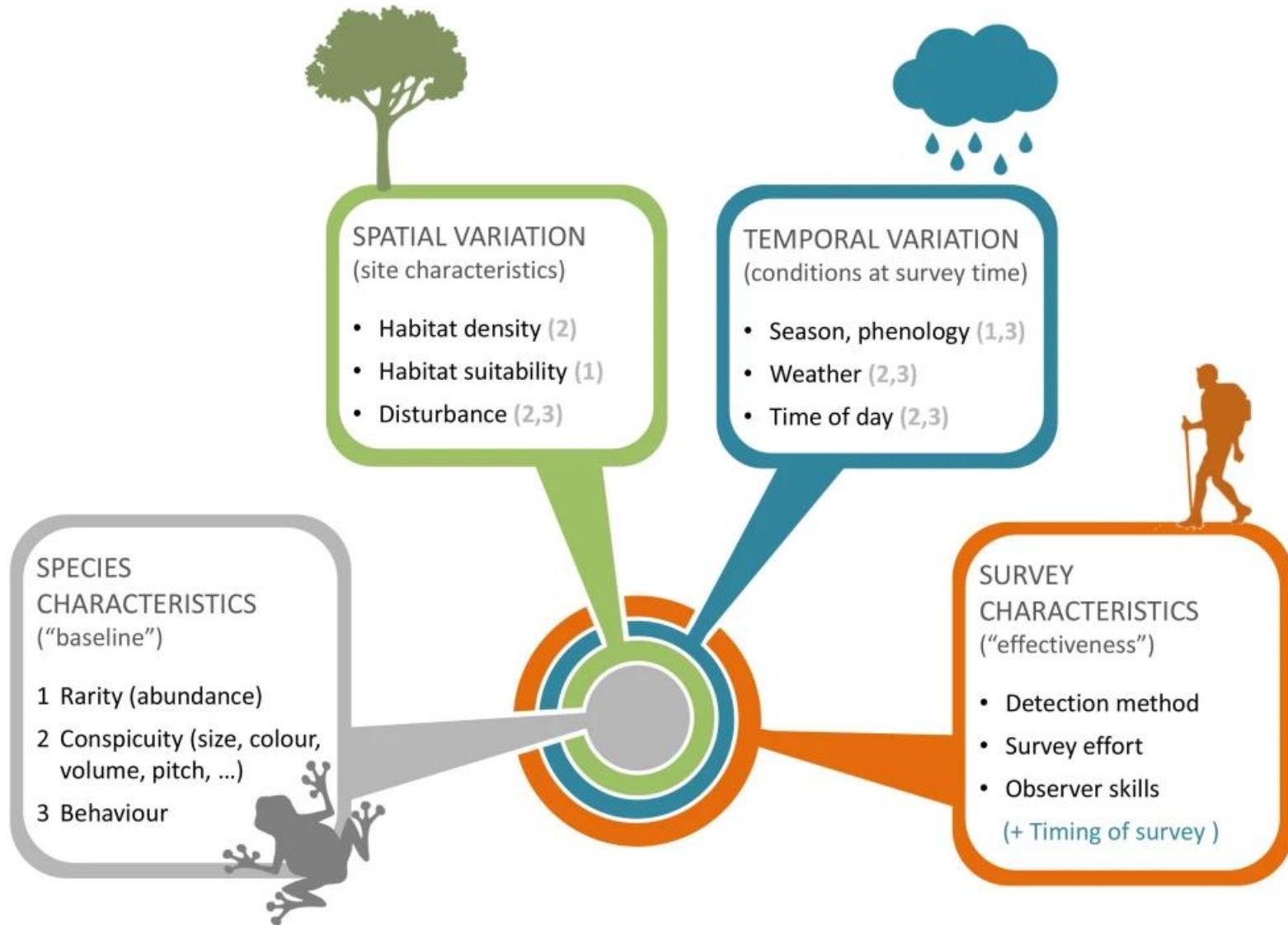
Review & synthesis |  Free Access

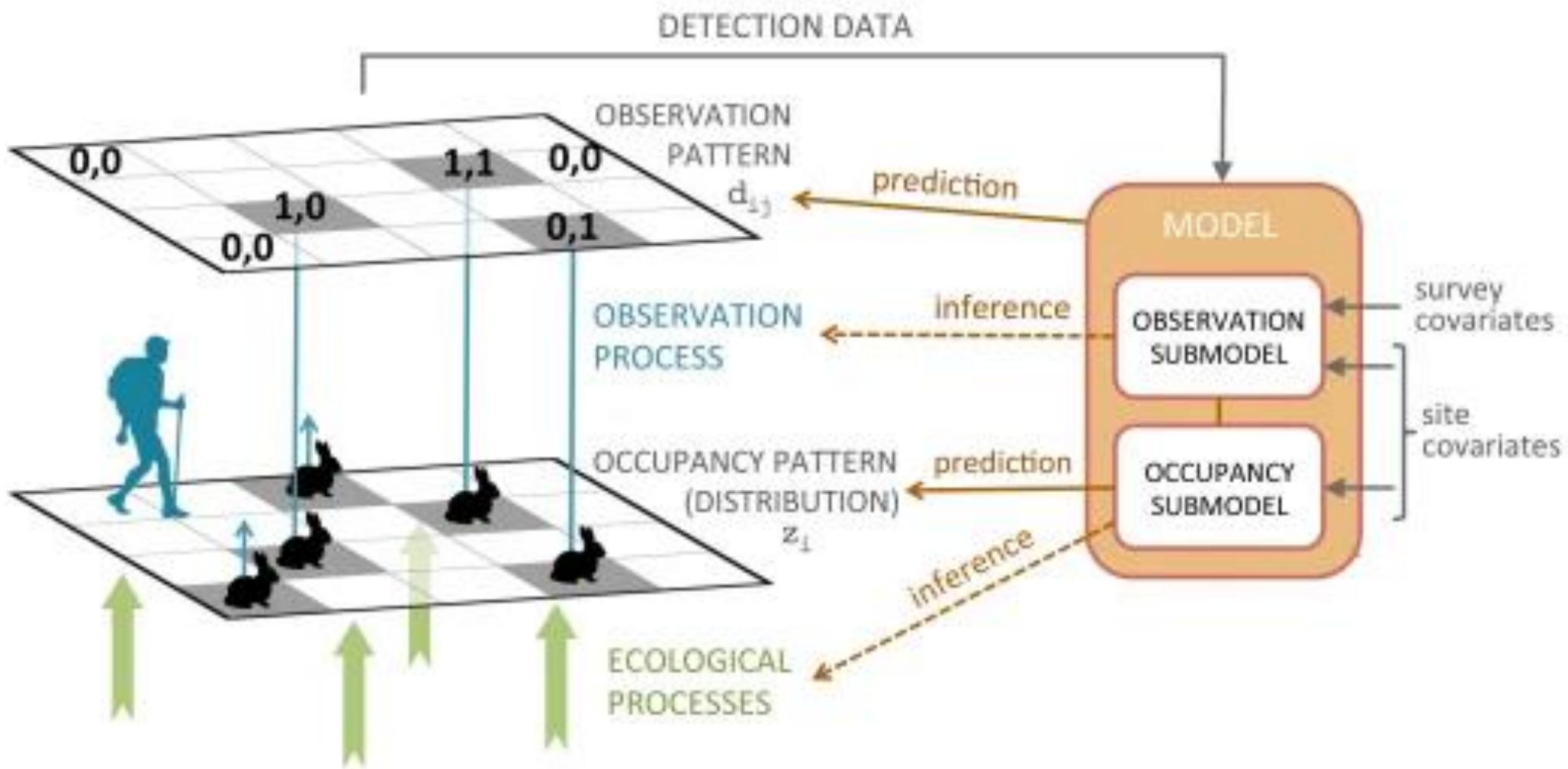
Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities

Gurutzeta Guillera-Arroita 

First published: 20 June 2016 | <https://doi.org/10.1111/ecog.02445> | Citations: 190

Now you see me, now you don't





Different types of data

- Occurrence locations
- Presence and absence data
- Detection and non-detection data



Module: Query Database

sPOCC : Interface to Species Occurrence Data

Sources

Choose Database

GBIF VertNet BISON

Enter species scientific name

Papilio canadensis

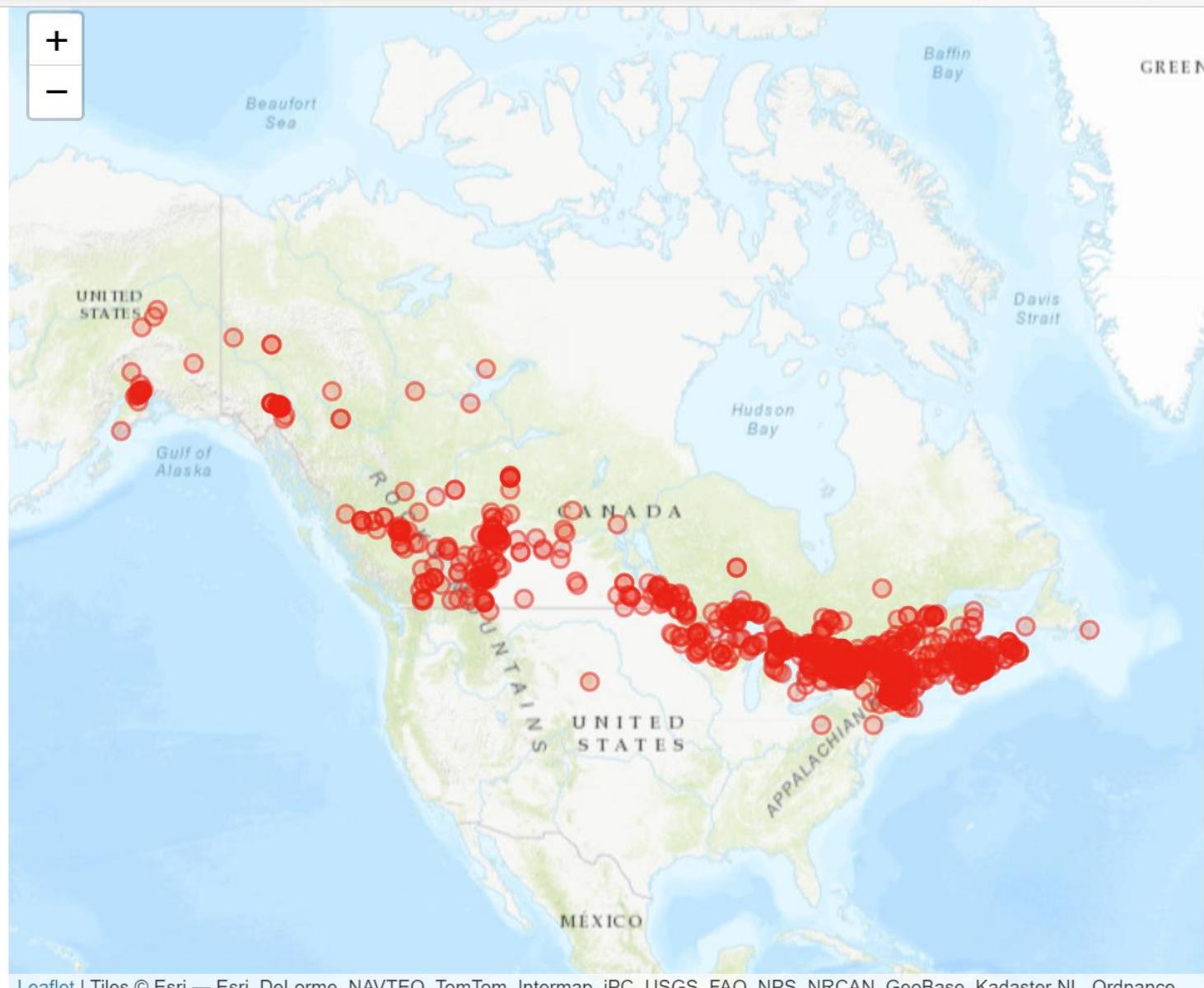
Set maximum number of occurrences

1000

Query Database

Download database occurrence localities (.csv)

Download



Spatial thinning

Remove observations to reduce bias towards highly-sampled locations

If you want to know more about thinning:



Software note | Free Access

spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models

Matthew E. Aiello-Lammens , Robert A. Boria, Aleksandar Radosavljevic, Bruno Vilela, Robert P. Anderson

First published: 06 February 2015 | <https://doi.org/10.1111/ecog.01132> | Citations: 597

Wallace x +

Not syncing ...

Step 1: Choose Background Extent

Background Extents:

- Bounding box
- Minimum convex polygon
- Point buffers

Study region buffer distance (degree)

20

Select

Step 2: Sample Background Points

Mask predictor rasters by background extent and sample background points

No. of background points

10000

...

Arctic Ocean

North America

Europe

Africa

Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

Partitioning occurrences

Divide the occurrence data in subsets and fit the model multiple times to assess how spatial distribution of points affect the model performance

If you want to know more about partitioning:

Received: 4 January 2021 | Accepted: 30 March 2021

DOI: 10.1111/2041-210X.13628

APPLICATION



ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions

Jamie M. Kass^{1,2,3} | Robert Muscarella⁴ | Peter J. Galante⁵ | Corentin L. Bohl⁶ |
Gonzalo E. Pinilla-Buitrago^{2,3} | Robert A. Boria⁷ | Mariano Soley-Guardia⁸ |
Robert P. Anderson^{2,3,9}

Module: Non-spatial Partition

ENMeval : Automated Runs and Evaluations of Ecological Niche Models

Options Available:

Random k-fold

Number of Folds

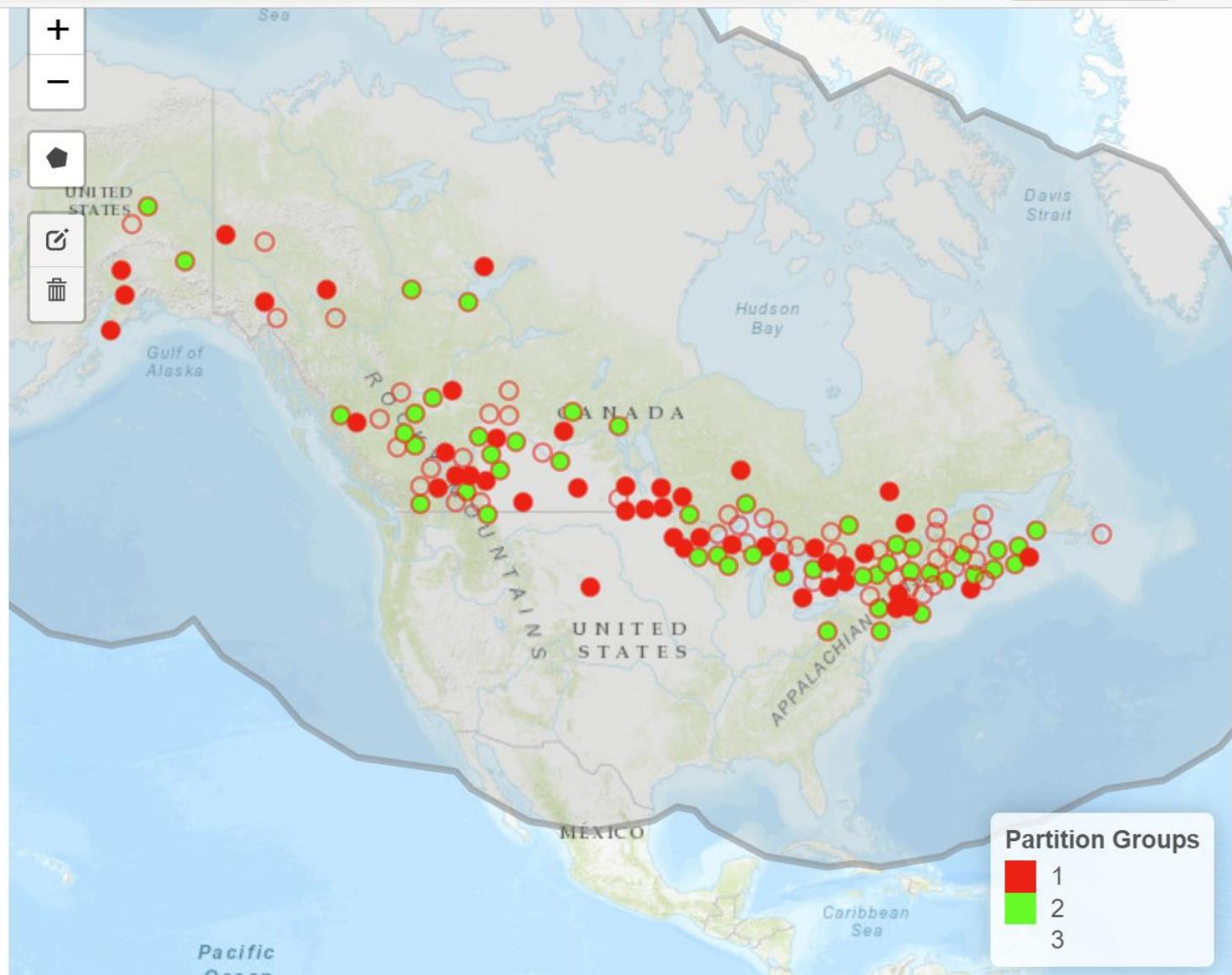
3

Partition

Download occurrence and background localities with partition values (.csv)

Download

Module Developers: Jamie M. Kass, Bruno Vilela, Robert P. Anderson



Modeling

Model

The algorithm used to infer species-environment relationships, and thus to generate spatial predictions of species distributions

Wallace allows two choosing between two algorithms:

- Bioclim
- Maxent

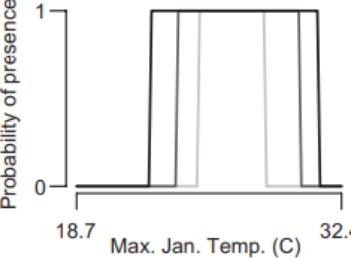
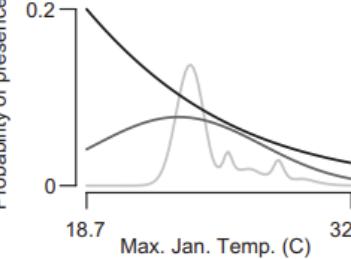
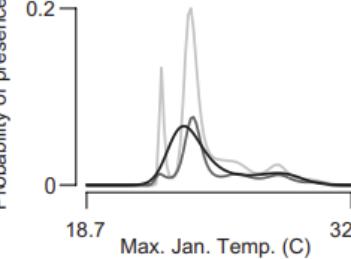
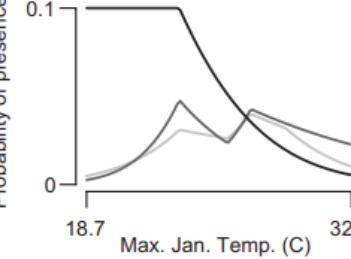


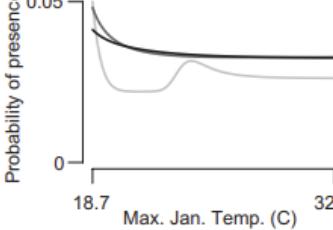
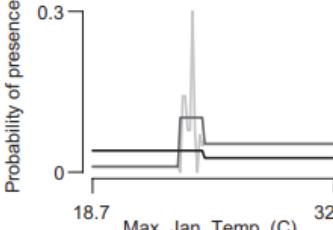
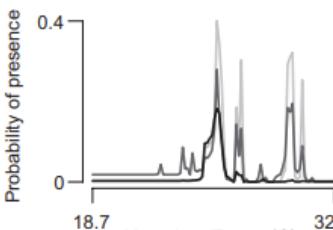
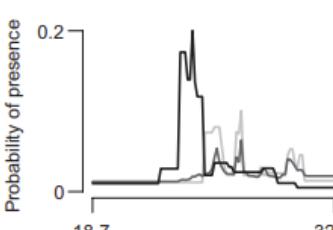
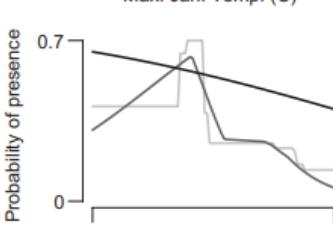
Ecography 37: 1267–1281, 2014
doi: 10.1111/ecog.00845

© 2014 The Authors. Ecography published by John Wiley & Sons Ltd on the behalf of Nordic Society Oikos.
Subject Editor: Heike Lischke. Editor-in-Chief: Jens-Christian Svenning. Accepted 24 July 2014

What do we gain from simplicity versus complexity in species distribution models?

Cory Merow, Mathew J. Smith, Thomas C. Edwards Jr, Antoine Guisan, Sean M. McMahon, Signe Normand, Wilfried Thuiller, Rafael O. Wüest, Niklaus E. Zimmermann and Jane Elith

Algorithm	Response curves	Responses are built from	Complexity controlled by
Bioclimatic envelope models (BIOCLIM)		quantiles, between which occurrence probability is 1	<ul style="list-style-type: none"> • Features: step functions • Quantiles
Generalized linear models (GLM)		parametric terms specified by user	<ul style="list-style-type: none"> • Features: polynomials, piecewise functions, splines • Feature complexity specified by user
Generalized additive models (GAM)		combination of parametric terms and flexible smooth functions suggested by the data or the user	<ul style="list-style-type: none"> • Features: parametric terms as in GLMs and various smoothers (e.g. splines, loess) • Number of nodes • Penalties
Multivariate adaptive regression splines (MARS)		the sum of multiple piecewise basis functions of predictors suggested by the data	<ul style="list-style-type: none"> • Features: splines • Number of knots • Cost per degree of freedom • Pruning

Algorithm	Response curves	Responses are built from	Complexity controlled by
Artificial neural networks (ANN)		networks of interactions between simple functions of predictors suggested by the data	<ul style="list-style-type: none"> Number of hidden layers
Classification and regression trees (CART)		repeated partitioning of predictors into different categories, suggested by the data, associated with different occurrence probabilities	<ul style="list-style-type: none"> Features: threshold, with implicit interactions Minimum observations for split/terminal node Maximum node depth Complexity threshold to attempt a split
Random forests (RF)		an average of multiple CARTs, each constructed on bootstrapped samples of the data and using different random subsets of the full predictor set	<ul style="list-style-type: none"> Features: threshold, with implicit interactions See CARTs Number of trees
Boosted regression trees (BRT)		regression trees at multiple steps; at each, models the residuals from the sum of all previous models weighted by the learning rate 2	<ul style="list-style-type: none"> Features: threshold, with implicit interactions See CARTs Number of trees Learning rate
Maximum entropy (MAXENT)		a GLM with a large number of features , which are suggested by the data or the user	<ul style="list-style-type: none"> Features: linear, quadratic, interaction, hinge, threshold Feature classes used Regularization penalty

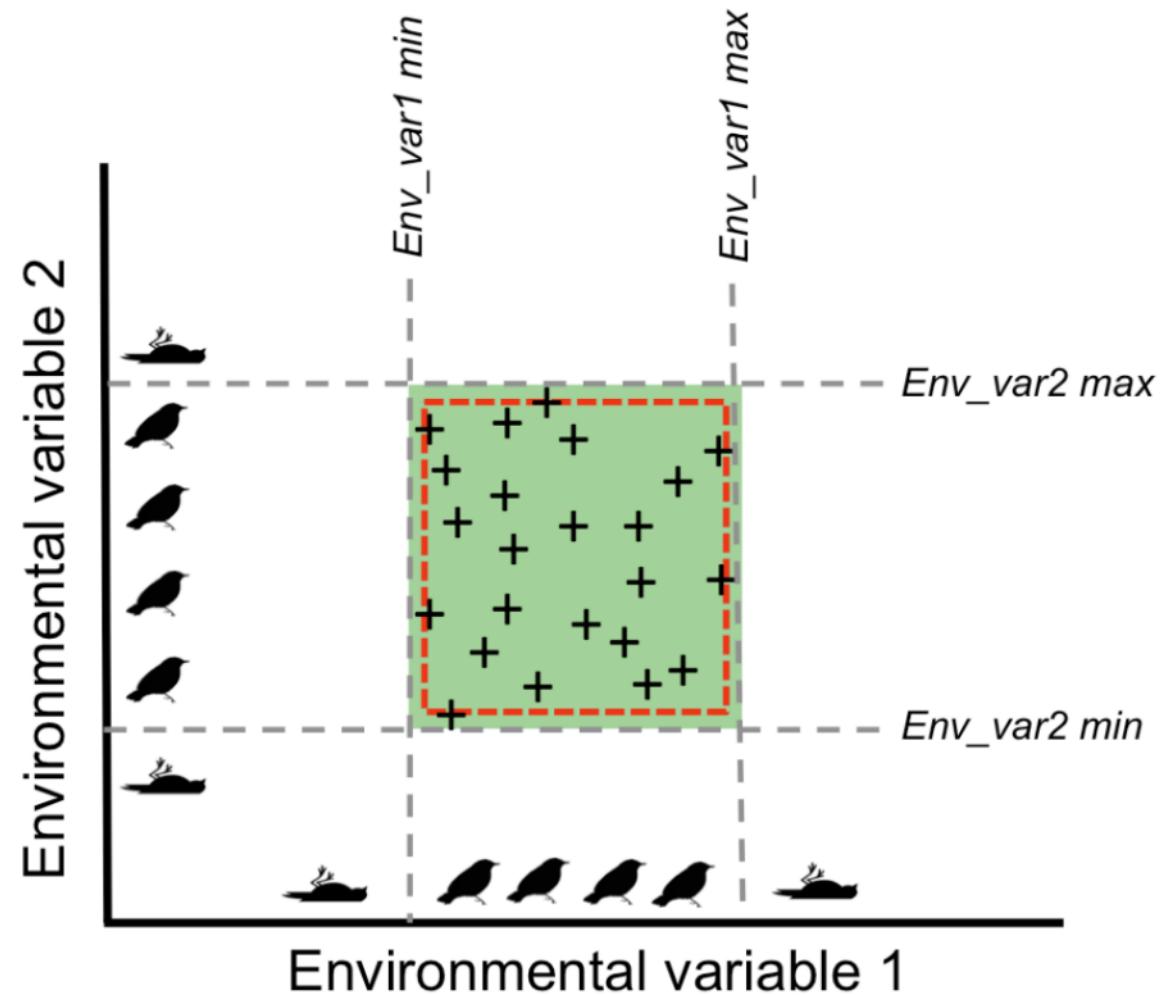
Bioclim

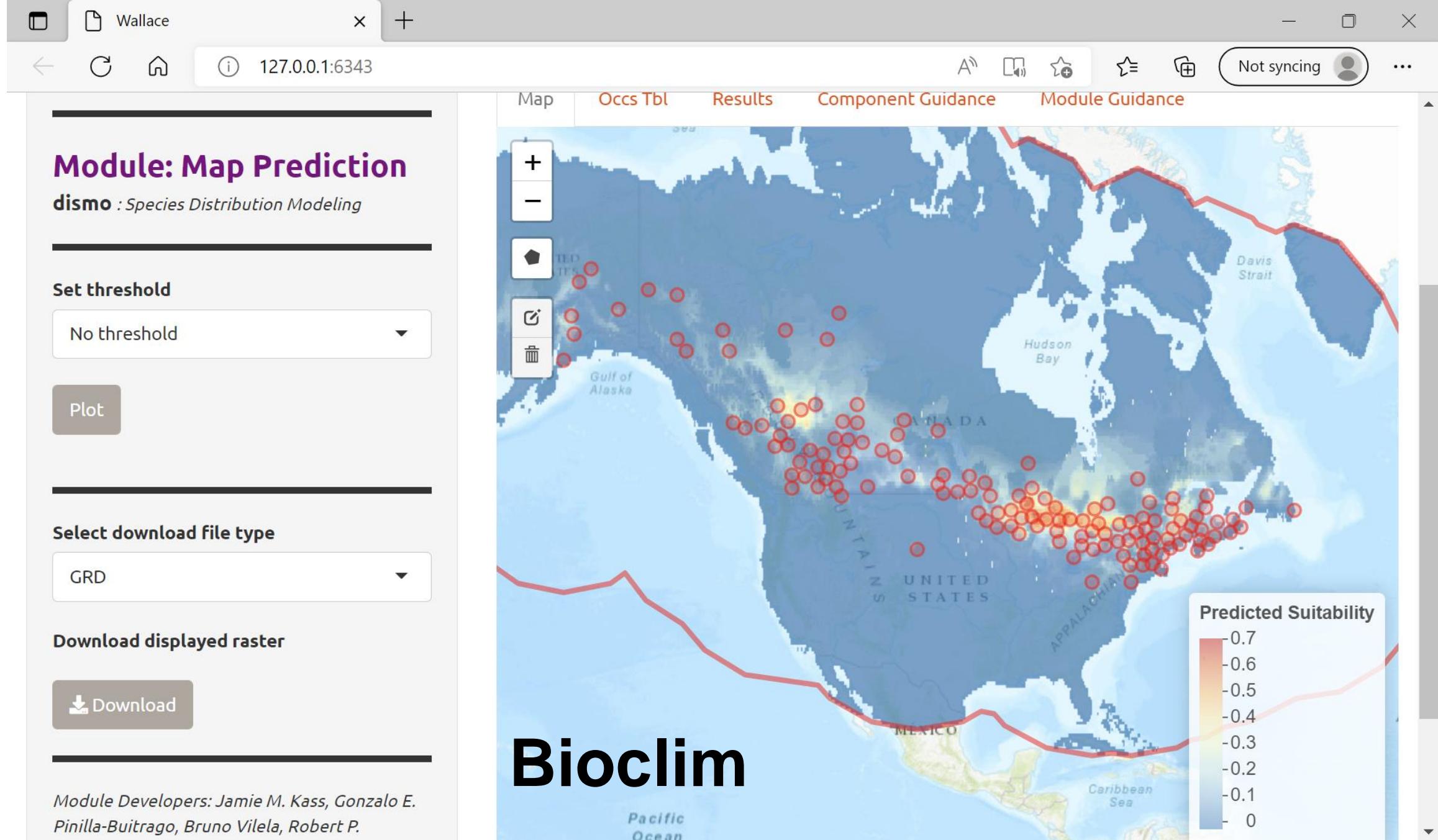
<https://csiropedia.csiro.au/bioclim/>

Modern SDM began in January 1984 with the release of the BIOCLIM package on the CSIRONET computer network. Researchers input information on individual sites where a species had been recorded, its environmental requirements were determined by the BIOCLIM package and climatically-suitable locations mapped. The ease-of-use of BIOCLIM meant it could readily be used by anyone who had latitude, longitude and elevation data describing a species' distribution. The first BIOCLIM release included a coarse 0.5-degree digital elevation model (DEM) for prediction, soon updated with a 0.1-degree DEM. DEMs with much greater precision subsequently became available.

Bioclim

Bioclim is a so-called envelope-style method, which uses only occurrence data to **define a multi-dimensional environmental space in which a species can occur**. This environmental space is constructed as a bounding box around the minimum and maximum values of the environmental variables for all occurrences, resulting in a multi-dimensional rectilinear envelope.





MaxEnt

Follows the maximum entropy principle, i.e., predicts species occurrences by finding the distribution that is most spread out, or closest to uniform, while taking into account the constraints provided by the environmental variables of occurrence locations

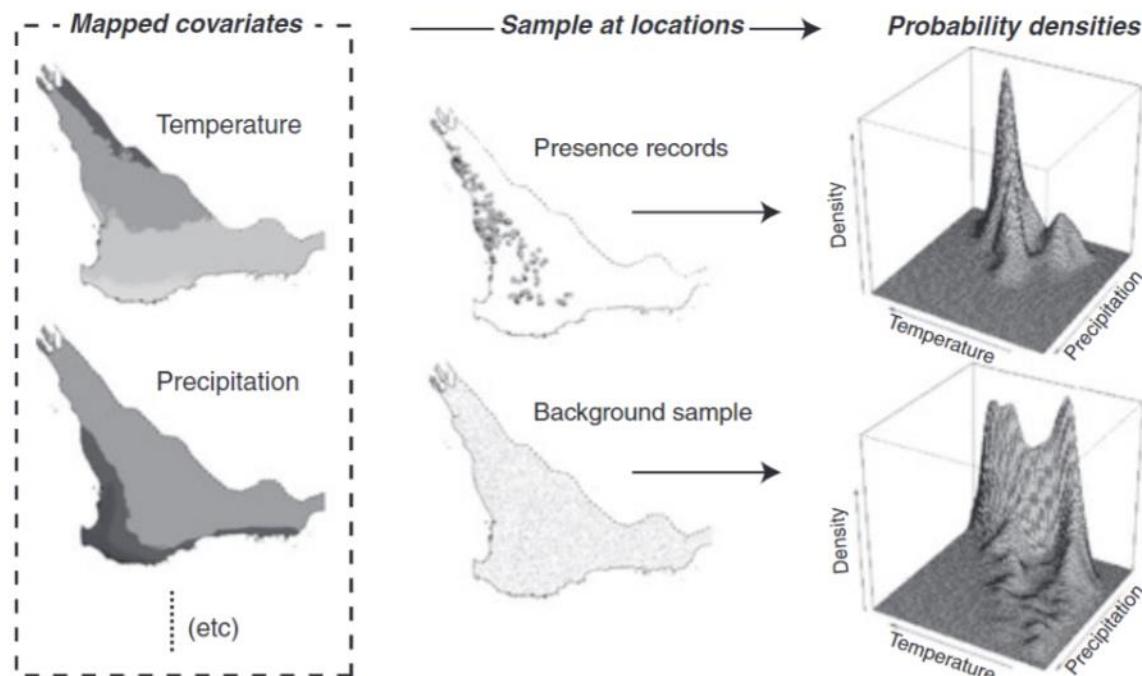


Figure 1 A diagrammatic representation of the probability densities relevant to our statistical explanation, using data presented in case study 1. The maps on the left are two example mapped covariates (temperature and precipitation). In the centre are the locations of the presence and background samples. The density estimates on the right are not in geographic (map) space, but show the distributions of values in covariate space for the presence (top right) and background (bottom right) samples. These could represent the densities $f_1(\mathbf{z})$ and $f(\mathbf{z})$ for a simple model with linear features.

Opening the black box: an open-source release of Maxent

Steven J. Phillips, Robert P. Anderson, Miroslav Dudík, Robert E. Schapire and Mary E. Blair

Maxent estimates the distribution (geographic range) of a species by finding the distribution which has maximum entropy (i.e. is closest to geographically uniform) subject to constraints derived from environmental conditions at recorded occurrence locations.

The constraints are defined in terms of ‘features’ (environmental variables such as temperature, and simple functions of those variables such as quadratic terms), and require that the mean of each feature should match the sample mean.

This formulation is equivalent to maximizing the likelihood of a parametric exponential distribution. More recently, it was noted that the exact same maximum likelihood exponential model can be obtained from an inhomogeneous Poisson process (IPP)

|  Open Access

A statistical explanation of MaxEnt for ecologists

Jane Elith , Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, Colin J. Yates

First published: 25 November 2010 | <https://doi.org/10.1111/j.1472-4642.2010.00725.x> |

Citations: 3,270

ECOGRAPHY

Open Access

A JOURNAL OF SPACE
AND TIME IN ECOLOGY

Forum |  Free Access

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

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

First published: 18 June 2013 | <https://doi.org/10.1111/j.1600-0587.2013.07872.x> |

Citations: 1,603

MaxEnt

- Features: the bioclimatic predictors in the model
- Feature types: the shape of the relationship that MaxEnt is allowed to explore estimating the species-environment relationships
- Regularization: automatic procedure that retains only features relevant for predicting the species
- Regularization multipliers: setting that determines penalization in the model for additional parameters

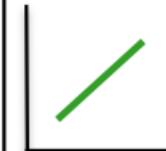
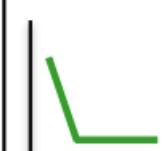
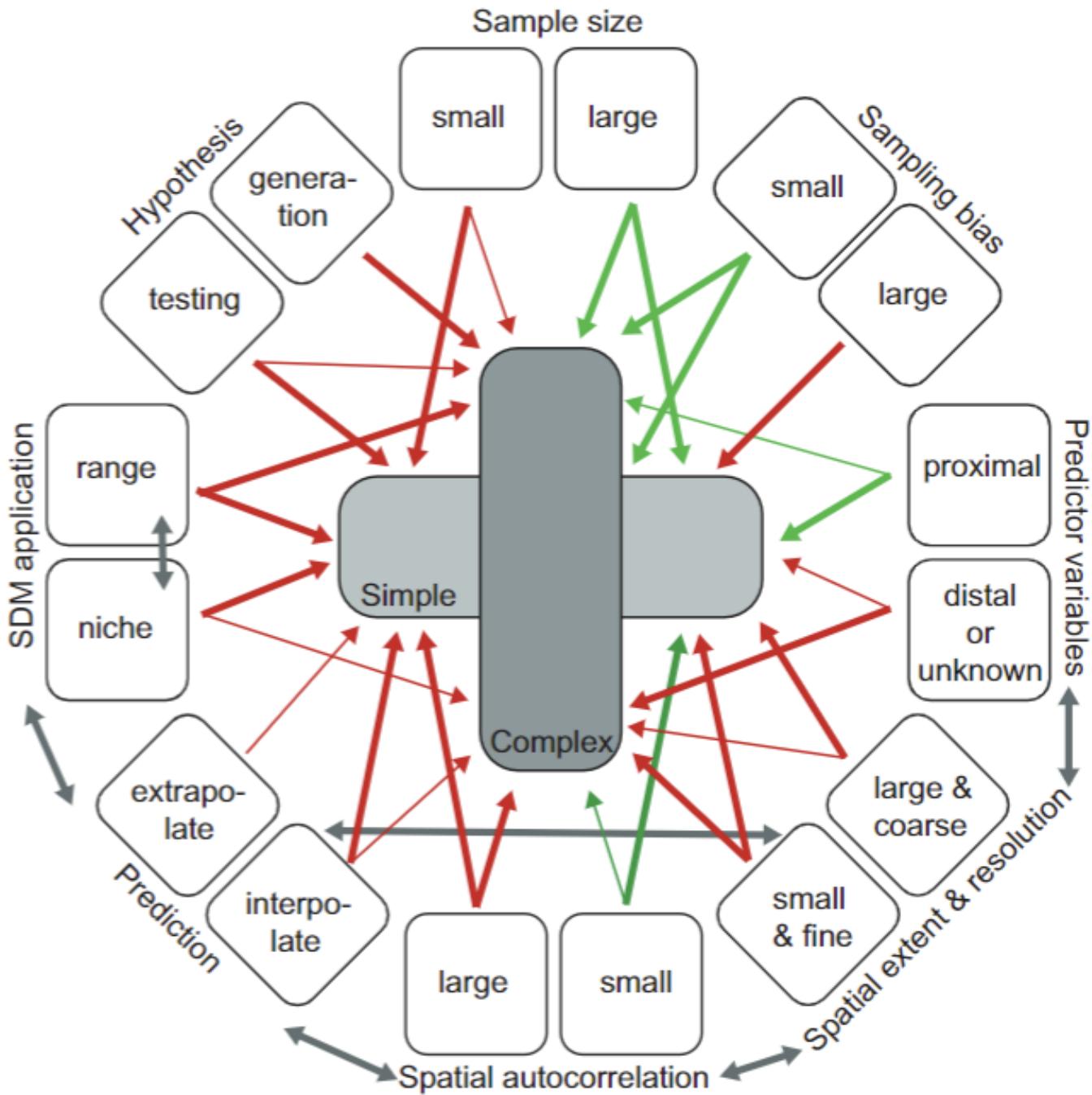
Feature type	Interpretation	Constraint	Shape
Linear	Continuous variable	The <i>mean</i> of each environmental variable at an unknown location should be close to the mean of that variable in known occurrence locations.	
Quadratic	Square of the variable	The <i>variance</i> of each environmental variable 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 <i>co-variance</i> 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.	

Figure 1. Influence of attributes of study objectives and data attributes on the choice of model complexity. Green arrows illustrate attributes where the choice of complexity is of no particular concern. Red arrows illustrate the situations where caution and/or experimentation with model complexity is needed. Gray arrows indicate decisions that involve interactions with other study goals or data attributes. The thickness of the arrows illustrates the strength of the arguments in favor of choosing a specific level of complexity, with thicker arrows indicating stronger arguments.

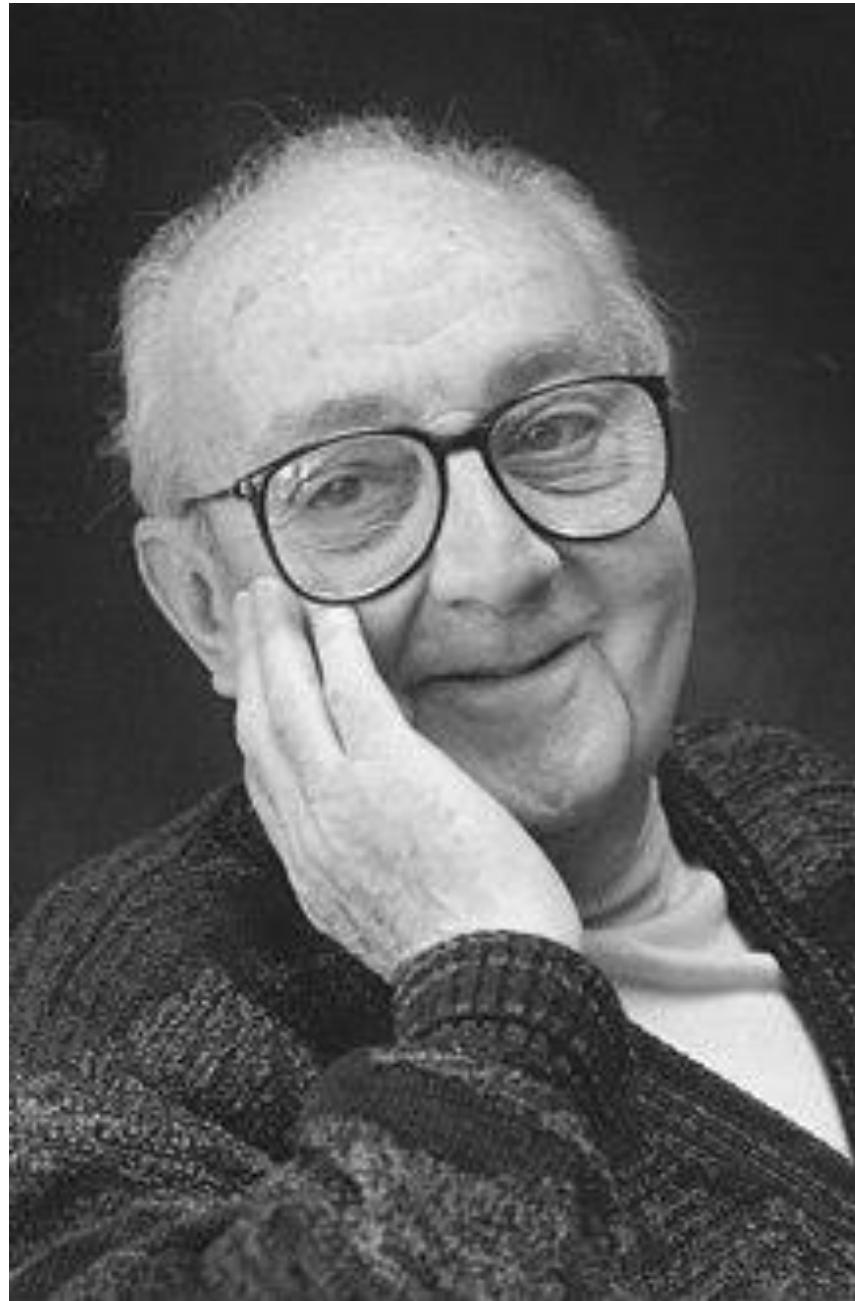


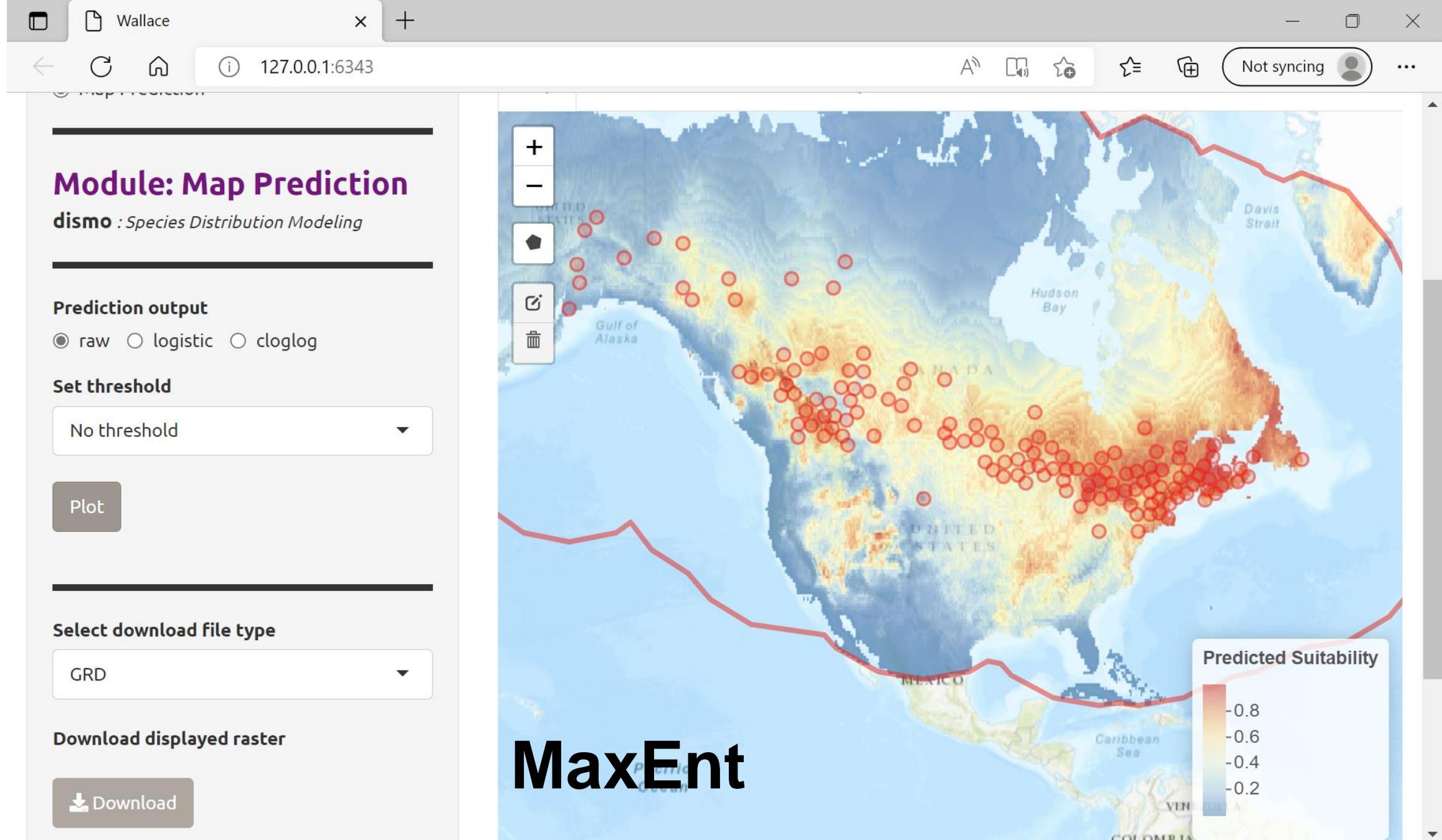
Rarity and prevalence in MaxEnt

Box 2 Consider the jaguar: reconciling logistic output and sampling effort

The jaguar (*Panthera onca*) and the collared peccary (*Pecari tajacu*) have very similar ranges in South and Central America, and MaxEnt models for the two species would therefore be similar using the default τ . However, the jaguar is much rarer than the peccary, so how can the outputs be compared? The answer is that probability of presence is only defined relative to a given definition of presence/absence (i.e., the temporal and spatial scale of a sample; see Preamble). For instance, for a rare species like the jaguar a presence record is likely to derive from sampling over a longer time and/or larger area (e.g., using camera traps over months) than it would for the peccary, which is fairly common and easier to observe. Since with presence-only data there is usually no information on sampling effort, this elasticity in definition is largely conceptual – it explains how to think about the meaning of the probabilities across species. When τ is 0.5 typical presence sites will have a logistic output near 0.5. This is reasonable as long as we can interpret logistic output as corresponding to a temporal and spatial scale of sampling that results in a 50% chance of the species being present in suitable areas. See Appendix S3 for more information.

Alternatively, if the value of τ is available for a given level of sampling effort, it could be used instead of the default and then the predictions for the two species would be directly comparable. Tau measures a form of rarity (Rabinowitz *et al.*, 1986). The jaguar has very low local abundance even in suitable areas within its range, so a very small value τ is appropriate for all but the most intensive sampling schemes. The estimate of τ could come from expert knowledge or targeted surveys. While τ is determined by prevalence, and vice versa, τ is arguably more ecologically intuitive, as it is a characteristic property of the species while prevalence strongly depends on the choice of study area.





Thresholding

Transforming a continuous model output into a binary map based on a classification criterion (e.g., the value that equalizes sensitivity and specificity)

If you want to know more about thresholding:



| Free Access

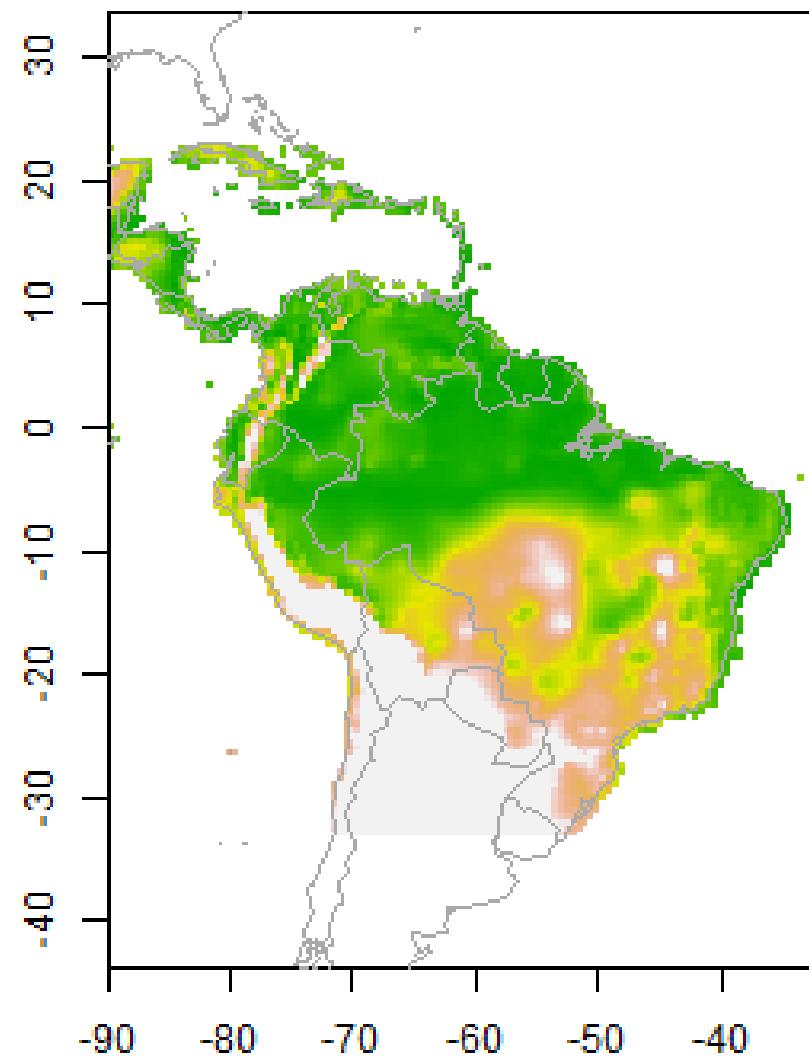
Selecting thresholds of occurrence in the prediction of species distributions

Canran Liu, Pam M. Berry, Terence P. Dawson, Richard G. Pearson

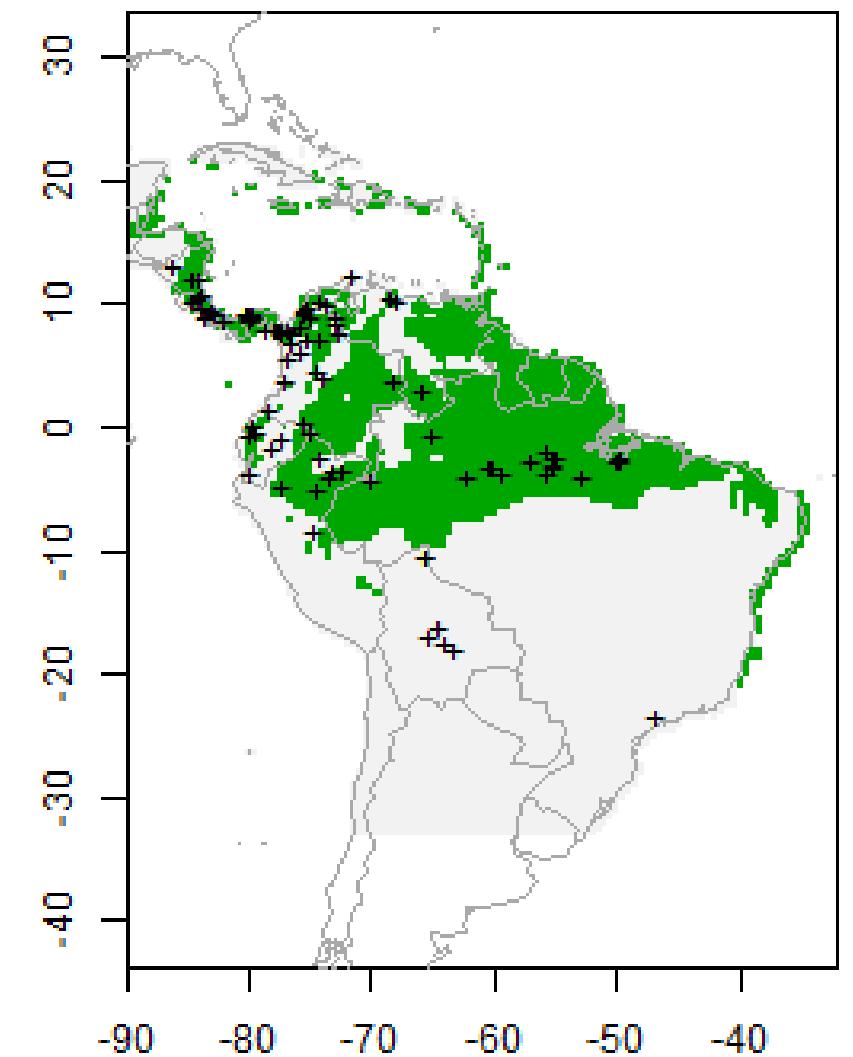
Table 1. Indices for assessing the predictive performance of species distribution models, a is true positives (or presences), b is false positives (or presences), c is false negatives (or absences), d is true negatives (or absences), n ($=a+b+c+d$) is the total number of sites and α is a parameter between 0 and 1 (inclusive).

Index	Formula
Sensitivity (or Recall, R)	$a/(a+c)$
Specificity	$d/(b+d)$
Precision (P)	$a/(a+b)$
Overall prediction success (OPS)	$(a+d)/n$
Kappa	$\frac{(a + d) - [(a + c)(a + b) + (b + d)(c + d)]/n}{n - [(a + c)(a + b) + (b + d)(c + d)]/n}$
Odds ratio	$(ad)/(cb)$
Normalized mutual information statistic (NMI)	$\frac{-alna - blnb - clnc - dlnd + (a + b)\ln(a + b) + (c + d)\ln(c + d)}{nlnn - [(a + c)\ln(a + c) + (b + d)\ln(b + d)]}$
F	$\frac{1}{\alpha/P + (1 - \alpha)/R} \quad (0 \leq \alpha \leq 1)$

Domain, raw values

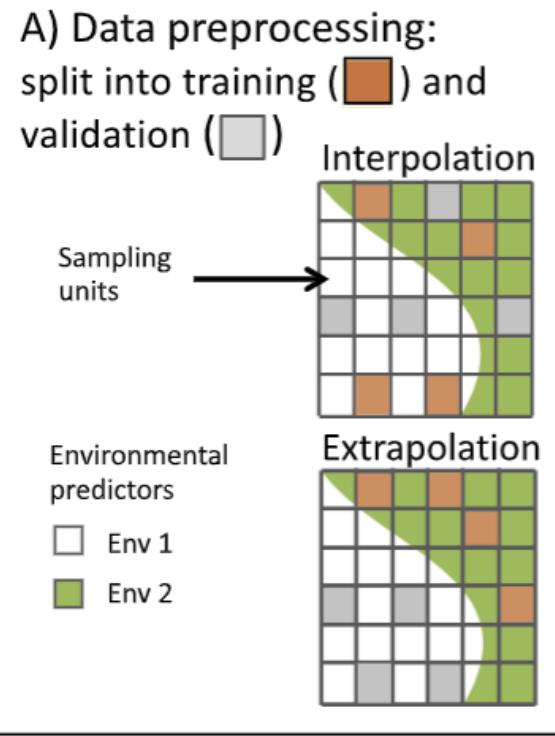


presence/absence



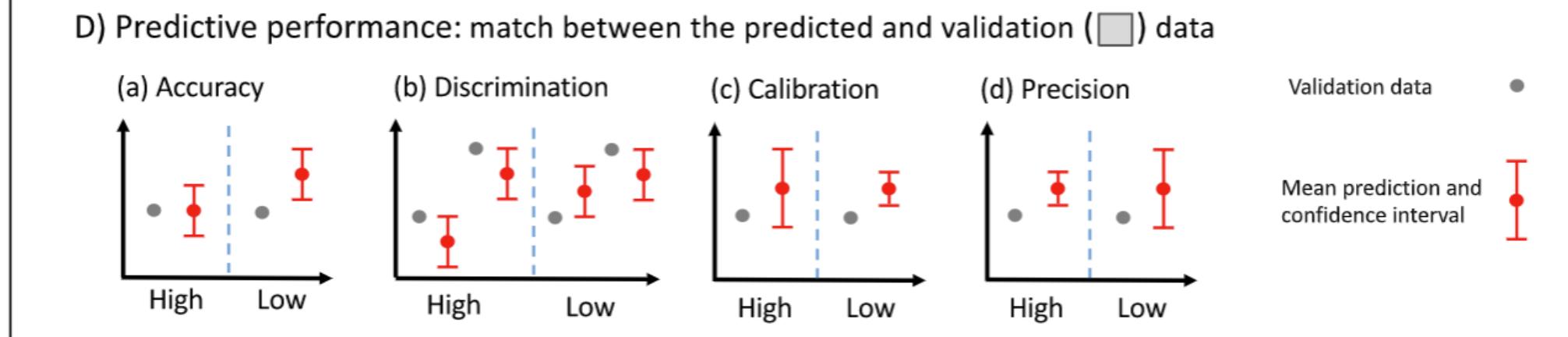
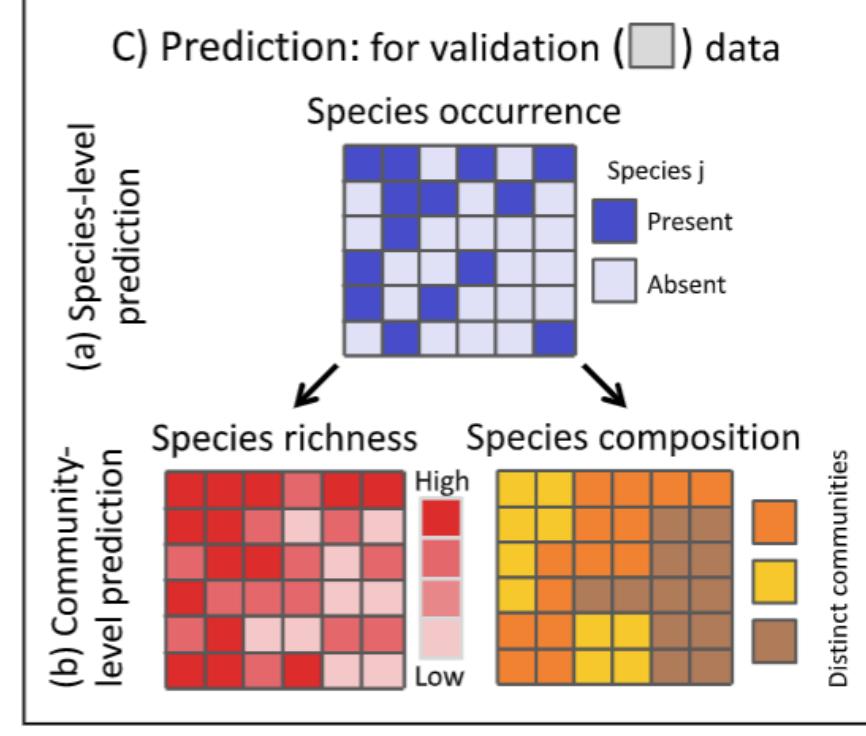
A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels

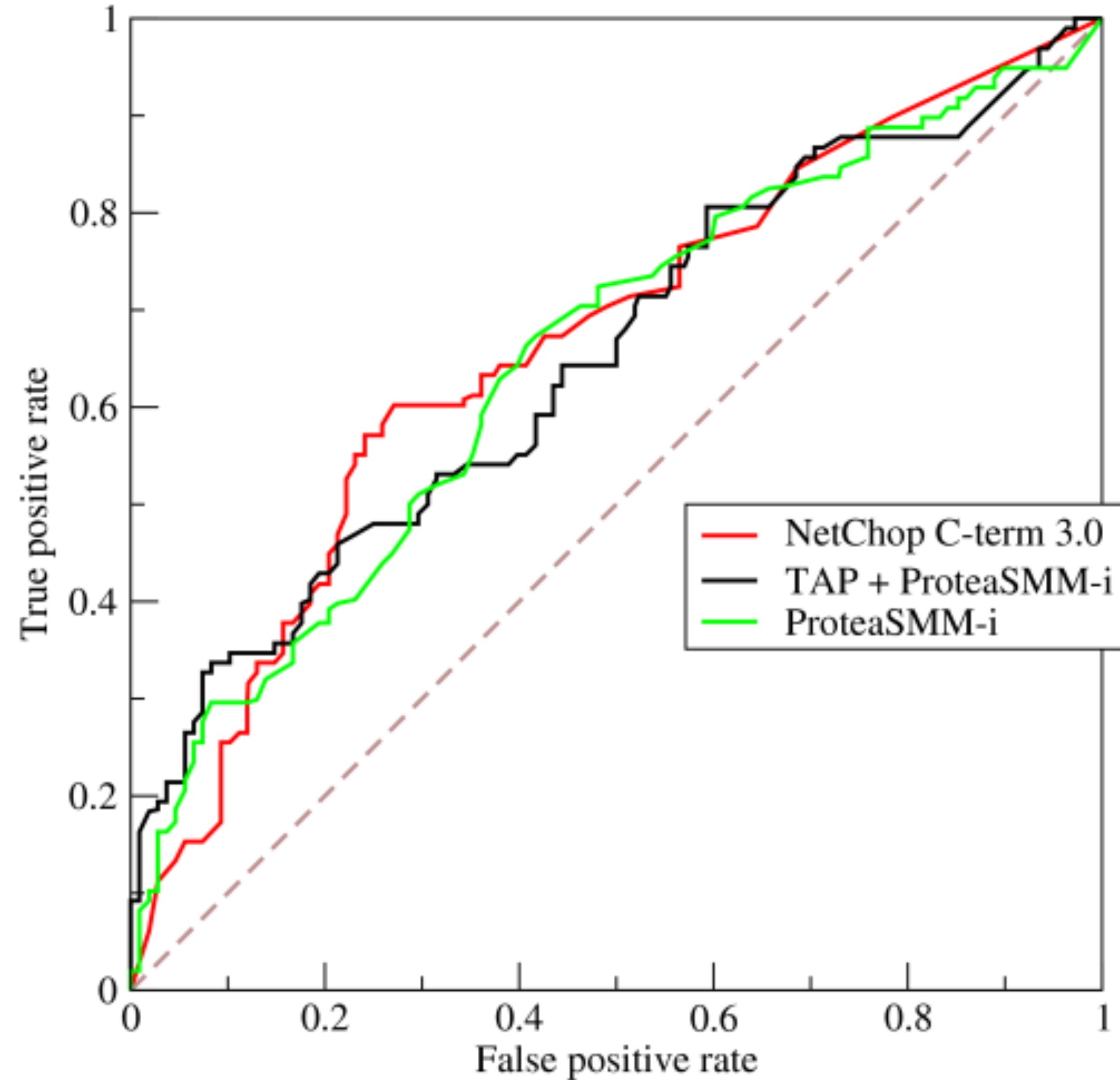
ANNA NORBERG ,^{1,34} NEREA ABREGO,^{2,3} F. GUILLAUME BLANCHET,⁴ FREDERICK R. ADLER,^{5,6} BARBARA J. ANDERSON,⁷ JANI ANTILA,¹ MIGUEL B. ARAÚJO,^{8,9,10} TAD DALLAS,¹ DAVID DUNSON,¹¹ JANE ELITH,¹² SCOTT D. FOSTER,¹³ RICHARD FOX,¹⁴ JANET FRANKLIN,¹⁵ WILLIAM GODSOE,¹⁶ ANTOINE GUISAN,^{17,18} BOB O'HARA,¹⁹ NICOLE A. HILL,²⁰ ROBERT D. HOLT,²¹ FRANCIS K. C. HUI,²² MAGNE HUSBY,^{23,24} JOHN ATLE KÅLÅS,²⁵ ALEKSI LEHIKOINEN,²⁶ MISKA LUOTO,²⁷ HEIDI K. MOD,¹⁸ GRAEME NEWELL,²⁸ IAN RENNER,²⁹ TOMAS ROSLIN ,^{3,30} JANNE SOININEN ,²⁷ WILFRIED THUILLER,³¹ JARNO VANHATALO,¹ DAVID WARTON,³² MATT WHITE,²⁸ NIKLAUS E. ZIMMERMANN,³³ DOMINIQUE GRAVEL,⁴ AND OTSO OVASKAINEN ^{1,2}



B) Model fitting: 33 variants of 15 SDMs fitted to training (■) data

Model Features	Description of the features with respect to which the models were classified
A	Parametric vs. semi-parametric models
B	Interactions among environmental predictors
C	Shared information on environmental responses
D	Species associations
E	Spatial vs. non-spatial models
F	Shrinkage
G	Parameter uncertainty





Applications in conservation



Biodiversity Data Journal 5: e20530
doi: [10.3897/BDJ.5.e20530](https://doi.org/10.3897/BDJ.5.e20530)



R Package

red - an R package to facilitate species red list assessments according to the IUCN criteria

Pedro Cardoso ‡, §

‡ Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

§ IUCN SSC Spider & Scorpion Specialist Group, Helsinki, Finland

Opinion

Measuring Terrestrial Area of Habitat (AOH) and Its Utility for the IUCN Red List

Thomas M. Brooks,^{1,2,3,*} Stuart L. Pimm,⁴ H. Resit Akçakaya,⁵ Graeme M. Buchanan,⁶ Stuart H.M. Butchart,^{7,8} Wendy Foden,^{9,10,11} Craig Hilton-Taylor,¹² Michael Hoffmann,¹³ Clinton N. Jenkins,¹⁴ Lucas Joppa,¹⁵ Binbin V. Li,^{4,16} Vivek Menon,¹⁷ Natalia Ocampo-Peñuela,¹⁸ and Carlo Rondinini¹⁹