Introduction to tools for mapping and modelling

Giuseppe Brundu (UNISS – Italy & EPPO Panel IAS)

OEPP/EPPO
21 bd Richard Lenoir, Paris,
75011, France
www.eppo.int

Outline of the presentation

- Mapping and Modelling IAS
- When modelling is appropriate (and when it may not be – limited datasets etc.)
- Which are the sources of uncertainty included in the model?
- Tools and methods (e.g., CLIMEX, R)
- Datasets and how to download (e.g., GBIF, EPPO GD)
- Interpreting models in PRAs
- Take home message

Mapping vs. Modelling in EPPO PRA

6. Distribution – *more guidance*

Continent	***Distribution*** *(list countries, or provide a general indication, e.g. present in West Africa)*	*Provide comments on the pest status in the different countries where it occurs e.g. widespread, native, introduced….)*	***Reference***
Africa			
America			
Asia			
Europe			
Oceania			

9. Likelihood of establishment outdoors in the PRA area – *more guidance*

Consider in particular the presence of host plants/habitats and climatic suitability and describe the area where establishment is most likely (area of potential establishment). Reference to maps such as Köppen-Geiger climate zones, day degrees and hardiness zones may help assess the likelihood of establishment (see e.g. https://www.eppo.int/media/uploaded_images/RESOURCES/special_projects/03_rating_guidance_for_climatic_suitability.pdf)

Rating of the likelihood of establishment outdoors	*Low* ☐	*Moderate* ☐	*High* ☐
Rating of uncertainty	*Low* ☐	*Moderate* ☐	*High* ☐

Major shortfalls in biodiversity and SDM knowledge

- **Linnaean shortfall** – How many species are there?-
- **Wallacean shortfall** – Where do species occur?-
- **Darwinian shortfall** – How are species related to one another?-
- **Mendelian shortfall** – How are individuals within the same species related to one another?
- **Grinnellian shortfall** – How are species limited by their environment?
- **Eltonian shortfall** – How do species interact with one another?

Miguel B. Araújo

maraujolab.eu
MIGUEL B.ARAÚJO

Major Shortfalls (SF) and Assumptions (AS) in SDMs

- SF 01: Availability of (high) **quality occurrences** (presence and absence data) in the native and invaded range;
- SF 02: **Lacking mechanisms** and independent **validation**;
- AS 01: **Environmental Equilibrium** (Quasi-Equilibrium) (The IAS has fully dispersed across its accessible range and occupies all suitable habitats while being absent from unsuitable ones);
- AS 02: **Niche Conservatism** / Niche Stability in Space and Time (The IAS maintains the same ecological and climatic requirements in a new geographic region as it does in its native range);
- AS 03: **Absence of Biotic Interference** and Unlimited/Neutral Dispersal (Distributions are shaped primarily by abiotic (climatic/environmental) gradients rather than transient biotic interactions or severe human-mediated transport barriers).

Major Shortfalls (SF) and Assumptions (AS) in SDMs - REFs

Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2012) **21**, 1126–1136

MACROECOLOGICAL METHODS

Invasive species distribution models – how violating the equilibrium assumption can create new insights

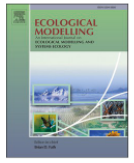
Laure Gallien[1]*, Rolland Douzet[2], Steve Pratte[1], Niklaus E. Zimmermann[3] and Wilfried Thuiller[1]

Ecological Modelling 493 (2024) 110754

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



How sensitive are species distribution models to different background point selection strategies? A test with species at various equilibrium levels

Bart Steen [a,b,*], Olivier Broennimann [a,c], Luigi Maiorano [b,1], Antoine Guisan [a,c,1]

[a] Institute of Earth Surface Dynamics, University of Lausanne, CH-1015 Lausanne Switzerland
[b] Department of Biology and Biotechnology "Charles Darwin", Sapienza University of Rome, viale dell'Università 32, 00185 Roma, Italy
[c] Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne Switzerland

ECOGRAPHY

Review

Powerful yet challenging: mechanistic niche models for predicting invasive species potential distribution under climate change

Erola Fenollosa[1], Sean E. H. Pang[2], Natalie J. Briscoe[3], Antoine Guisan[4,5], Roberto Salguero-Gómez[1]

CONCEPTS & SYNTHESIS
EMPHASIZING NEW IDEAS TO STIMULATE RESEARCH IN ECOLOGY

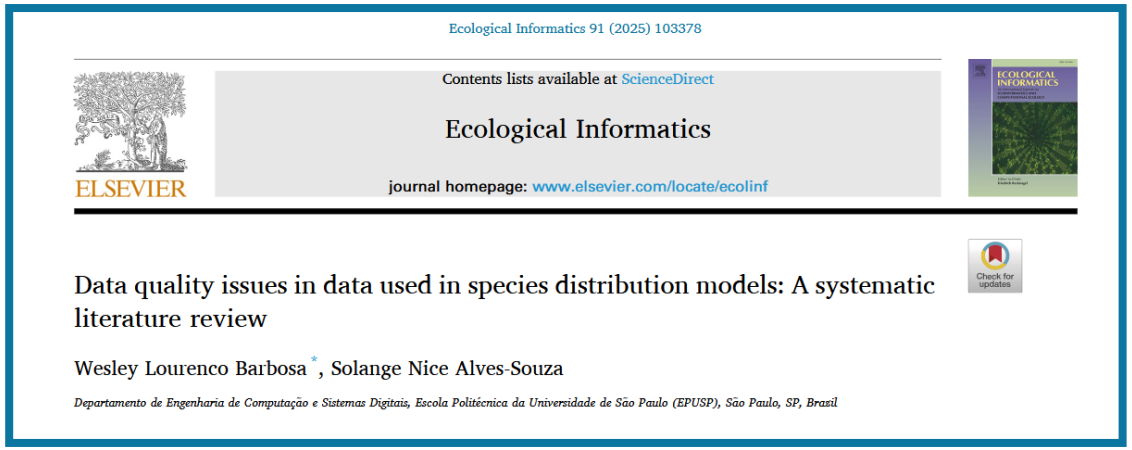
Ecology, 93(7), 2012, pp. 1527–1539
© 2012 by the Ecological Society of America

Uses and misuses of bioclimatic envelope modeling

Miguel B. Araújo[1,2,3,5] and A. Townsend Peterson[4]

Major Shortfalls (SF) and Assumptions (AS) in SDMs

- SF 01: Availability of (high) **quality occurrences** (presence and absence data) in the native and invaded range;



Ecological Informatics 91 (2025) 103378

Contents lists available at ScienceDirect

Ecological Informatics

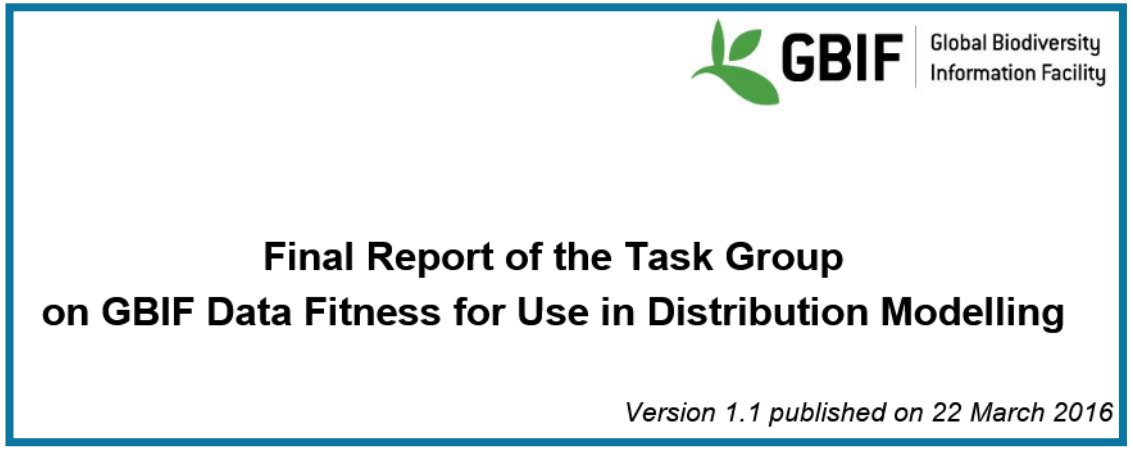
journal homepage: www.elsevier.com/locate/ecolinf

ELSEVIER

Data quality issues in data used in species distribution models: A systematic literature review

Wesley Lourenco Barbosa *, Solange Nice Alves-Souza

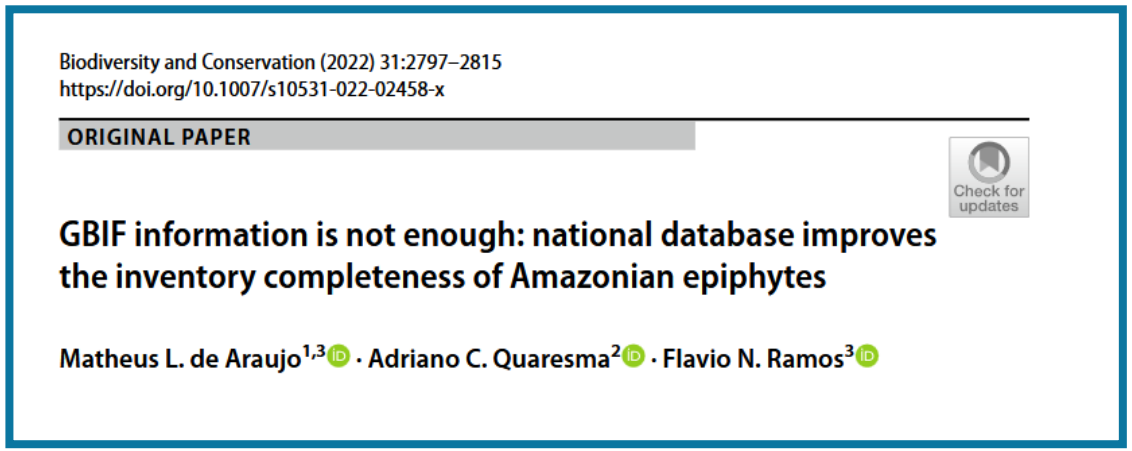
Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da Universidade de São Paulo (EPUSP), São Paulo, SP, Brasil



GBIF Global Biodiversity Information Facility

Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling

Version 1.1 published on 22 March 2016



Biodiversity and Conservation (2022) 31:2797–2815
https://doi.org/10.1007/s10531-022-02458-x

ORIGINAL PAPER

GBIF information is not enough: national database improves the inventory completeness of Amazonian epiphytes

Matheus L. de Araujo[1,3] · Adriano C. Quaresma[2] · Flavio N. Ramos[3]



Diversity and Distributions Open Access — A Journal of Conservation Biogeography

DATA PAPER | Open Access

A Global, Taxon-Stratified, High-Resolution Sampling-Effort Dataset From GBIF for Bias-Aware Ecological Modelling

Ahmed El-Gabbas

First published: 19 May 2026 | https://doi.org/10.1111/ddi.70205 | VIEW METRICS

Major Shortfalls (SF) and Assumptions (AS) in SDMs

- SF 01: Availability of (high) **quality occurrences** (presence and absence data) in the native and invaded range;

- **1. Taxonomic quality**
- **2. Record type and status**
- **3. Spatial quality**
- **4. Temporal quality**
- **5. Environmental quality**
- **6. Duplicates and redundancy**
- **7. Quantity**
- **8. Sampling bias**
- **9. Provenance and reporting**

Major Shortfalls (SF) and Assumptions (AS) in SDMs

- SF 01: Availability of (high) **quality occurrences** (presence and absence data) in the native and invaded range;

- **1. Taxonomic quality**
- 2. Record type and status
- 3. Spatial quality
- 4. Temporal quality
- 5. Environmental quality
- 6. Duplicates and redundancy
- 7. Quantity
- 8. Sampling bias
- 9. Provenance and reporting

- **Backbone match** (Taxonomic and nomenclatural references)
- **Synonyms and spitted or merged taxa**
- **Infraspecific ranks**: Decide whether subspecies, varieties or lower taxa (or CULTIGENS) are pooled or modelled separately
- **Identification reliability** of each single occurrence (e.g., if transferred to GBIF from iNaturalist)

Tools & Methods (CLIMEX)

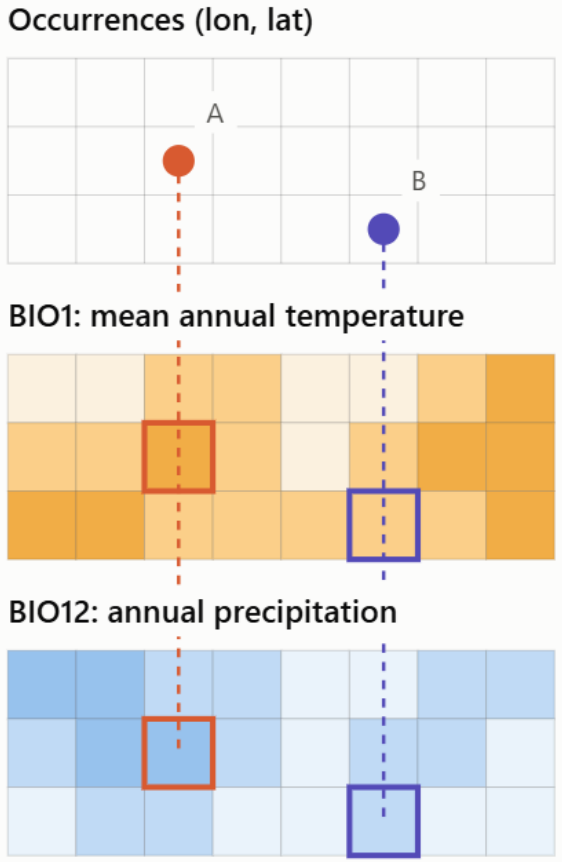




In 2015 EPPO organized a CLIMEX training in Paris.

CLIMEX is usually described as a **semi-mechanistic** (process-based, eco-physiological) model. It is not a probabilistic or correlative one like the GLM and RF commonly used in SDMs. The parameters are seldom all measured in the laboratory. In practice, species' climatic requirements are inferred from its known geographical distribution (either in its native range or in another region where it has been established for a long time), relative abundance and seasonal phenology. Some laboratory data, such as developmental threshold temperatures, can be used to fit or fine tune CLIMEX parameter values. Initial estimates of parameter values are fine-tuned by comparing the indices with the known presence or absence, seasonal phenology and, preferably, relative abundance of the species in each location.

Tools & Methods (SDMs with R)



Extracted values

Pt	lon	lat	BIO1	BIO12
A	8.9	46.0	11.8 °C	1750 mm
B	14.5	46.1	10.5 °C	1380 mm

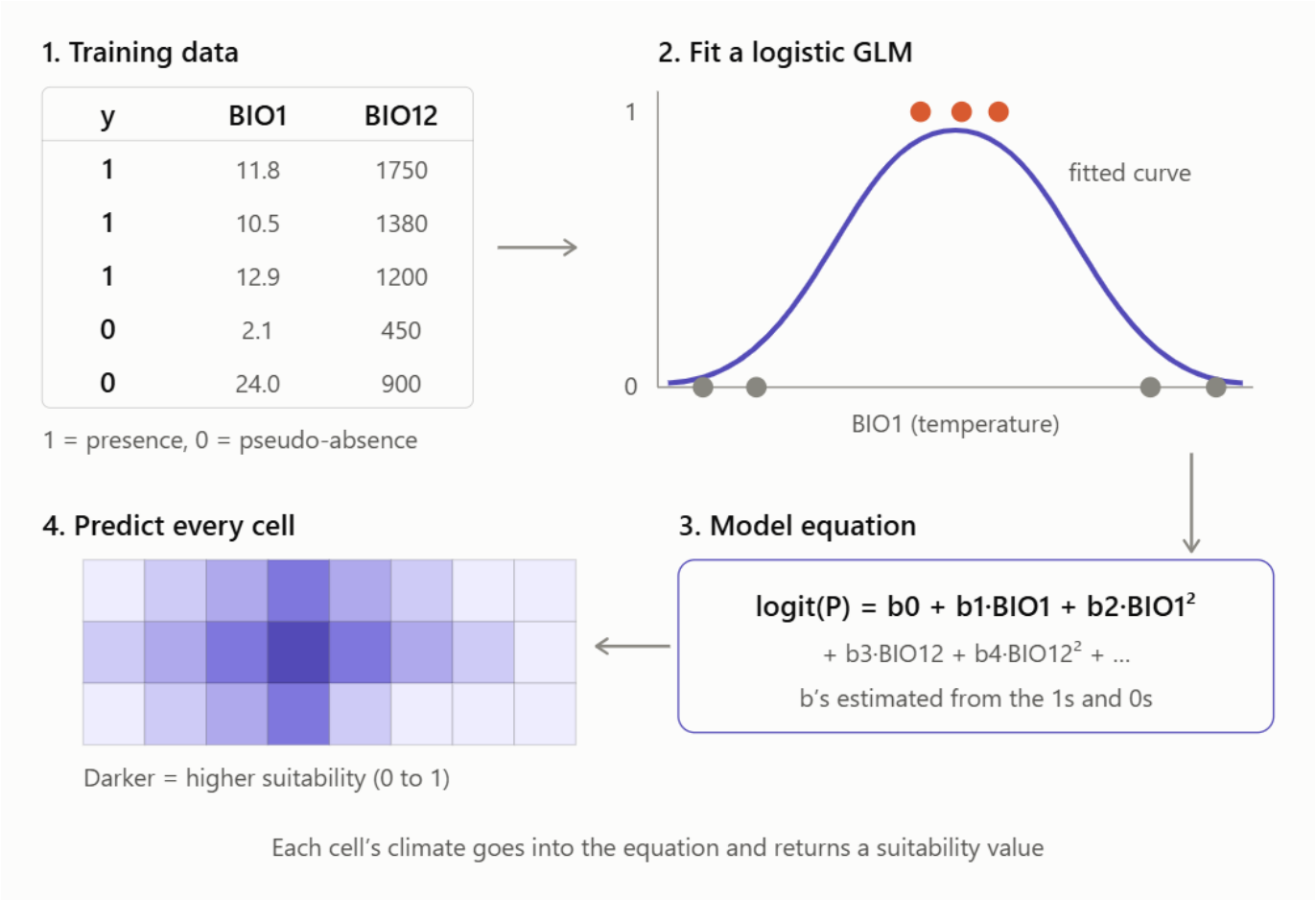
One row per occurrence,
one column per BIO layer

All layers share the same grid, so a coordinate falls in the same cell in every layer

This is what terra::extract(predictors, occ_xy) does in Section 6 of your script provided as an example.

Each GBIF record has a longitude and latitude.
All the WorldClim BIO layers are rasters on the same 10' grid, so a given coordinate falls in the same cell in every layer.
For each point, extract() finds that cell and reads its value in each layer. The result is one row per occurrence and one column per BIO variable.
Points that fall in the sea, where cells are NA, get missing values. That is why the script drops them as “with climate data (on land)”.

Tools & Methods (SDMs with R)

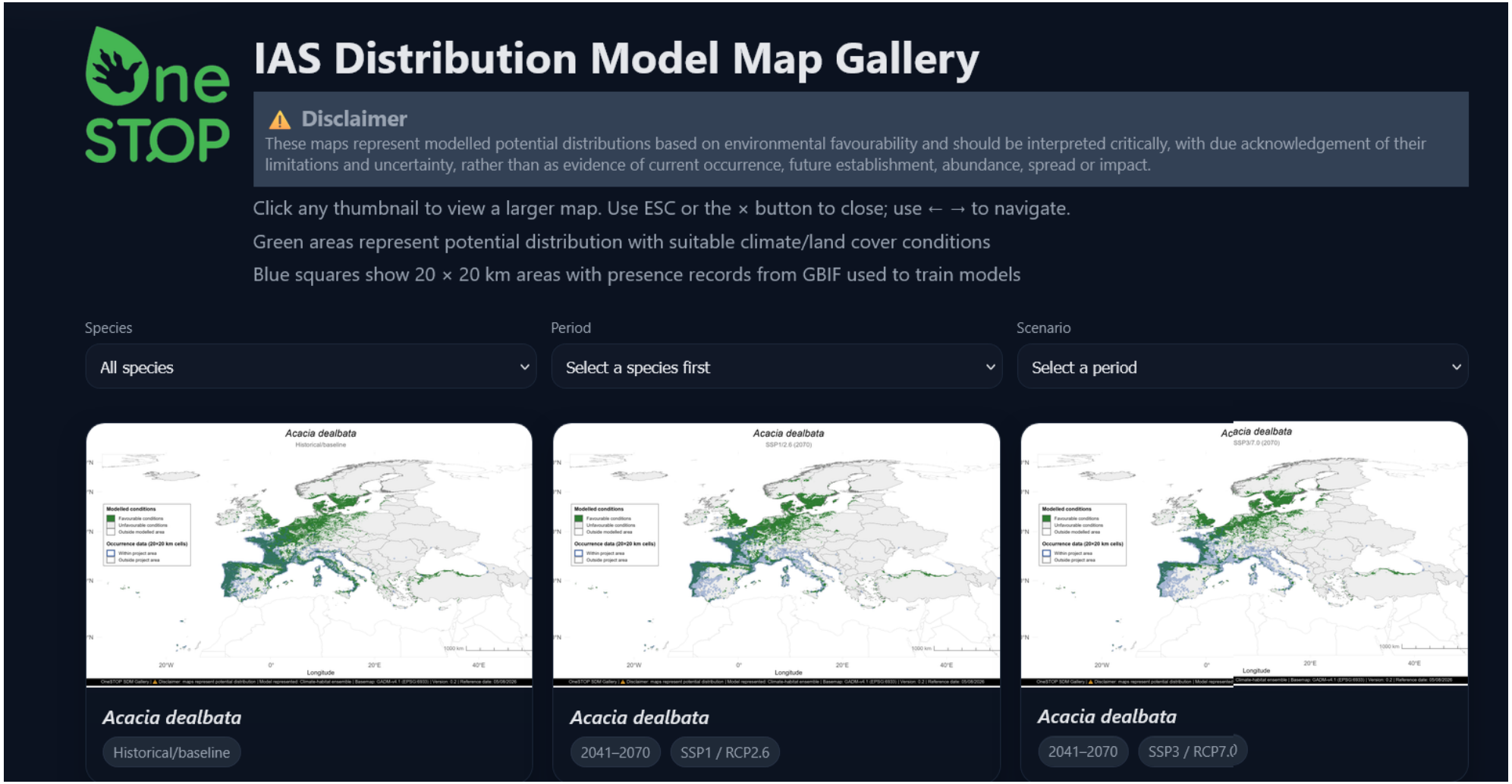


The result is relative climatic suitability, not a true probability of occurrence, because the 0s are pseudo-absences, not confirmed absences.

The RF model in your script uses the same table but learns the relationship with decision trees instead of an equation.

Tools & Methods (SDMs)

https://onestop-cibio.github.io/OneSTOP_SDM_Map_Gallery/



Tools & Methods (SDMs)

scientific data

View al

Explore content | About the journal | Publish with us

Data Descriptor | Open access | Published: 06 November 2025

SDMapCH: a Comprehensive database of >7,500 modelled species habitat suitability maps for Switzerland

Antoine Adde, Pierre-Louis Rey, Nathan Külling, Yohann Chauvier-Mendes, Fabian Fopp, Manuel R. Popp, Olivier Broennimann, Blaise Petitpierre, Nicolas Strebel, Andrin Gross, Silvia Stofer, Anthony Lehmann, Niklaus E. Zimmermann, Loïc Pellissier, Antoine Guisan & Florian Altermatt

Scientific Data **12**, Article number: 1752 (2025) | Cite this article

Tools & Methods (SDMs)

Reproducible WiSDM: a workflow for reproducible invasive alien species risk maps under climate change scenarios using standardized open data

Amy J. S. Davis[1,2*], Quentin Groom[3], Tim Adriaens[3], Sonia Vanderhoeven[4], Rozemien De Troch[5†], Damiano Oldoni[3], Peter Desmet[3], Lien Reyserhove[6], Luc Lens[3] and Diederik Strubbe[1]

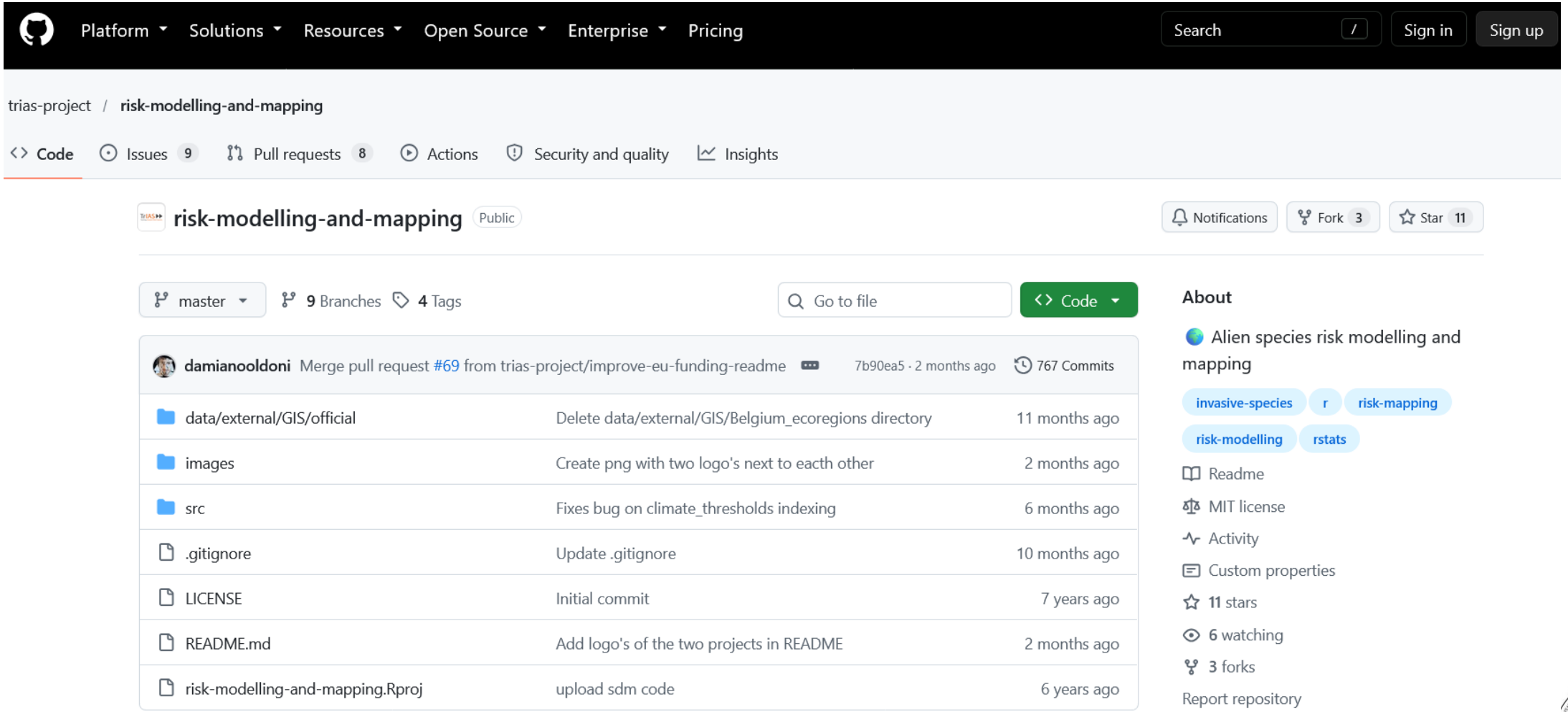
frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 09 February 2024
DOI 10.3389/fevo.2024.1148895

TABLE 1 Classification algorithms used in the WiSDM workflow.

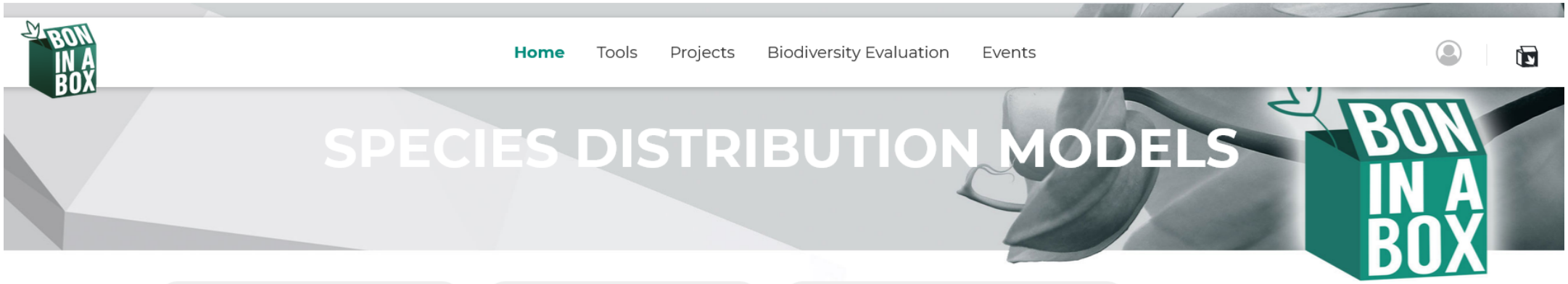
Algorithm	Type	Technique	Reference
Random Forests (RF)	Supervised	Bagging	Breiman, 2001
Gradient Boosted Machine (GBM)	Supervised	Boosting	Friedman, 2001
Logistic regression (GLM)	Supervised	Regression	Cox, 1958
Multivariate adaptive regression splines (MARS)	Supervised	Piece wise regression	Friedman, 1991

Tools & Methods (SDMs)



Tools & Methods (SDMs)

https://boninabox.geobon.org/indicator?i=SDM



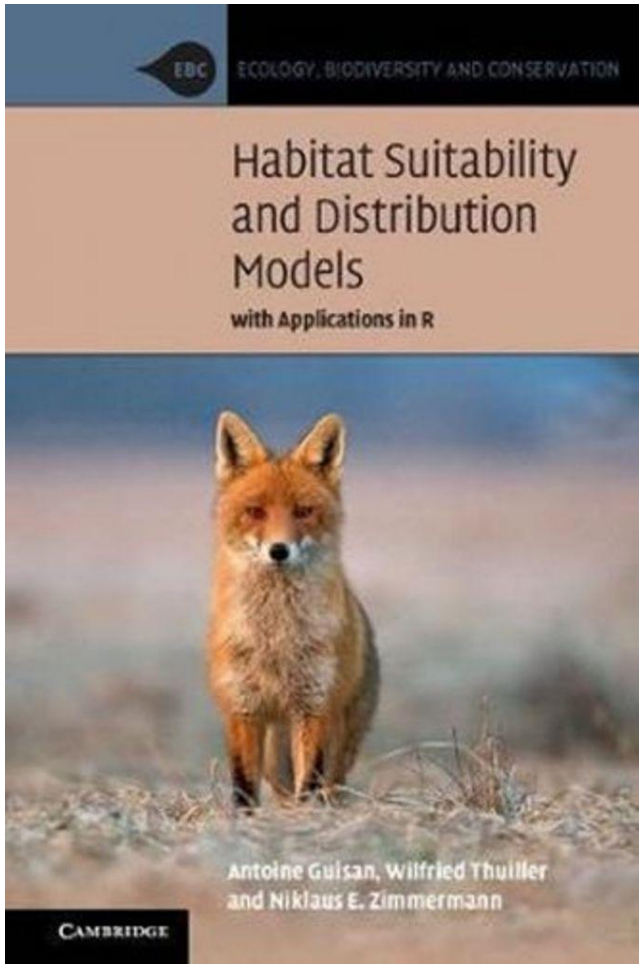
Example of Maxent SDM outputs | Example of BRT SDM outputs | Example of ewlgcpSDM SDM outputs

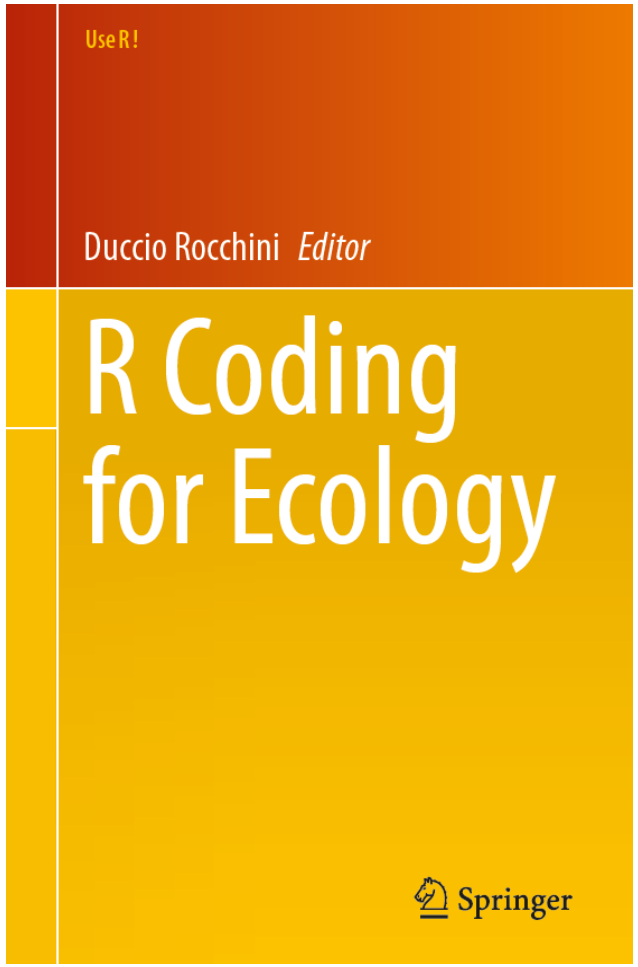
Species distributions are an important EBV in the 'species populations' class. Knowing where species are is essential for understanding biodiversity patterns and informing conservation efforts. However, less than 10% of the world is well sampled, and even the longest running and well-sampled biodiversity observation networks have substantial data gaps. Information on species occurrences is often sparse and heavily spatially and taxonomically biased, necessitating the need for species distribution models (SDMs) to fill these data gaps and provide a better, less biased idea of where species are. SDM outputs be used as key base layers for a wide variety of purposes including: creating maps for sampling prioritization, quantifying the impact of environmental stressors on species, mapping habitat suitability for at-risk species, mapping biodiversity hotspots across the landscape, identifying the locations of conservation priorities and protected area expansion, identifying sampling gaps and the needed locations of future sampling, and calculating a range of biodiversity indicators including the Species Habitat Index (SHI), the Species Protection Index (SPI)

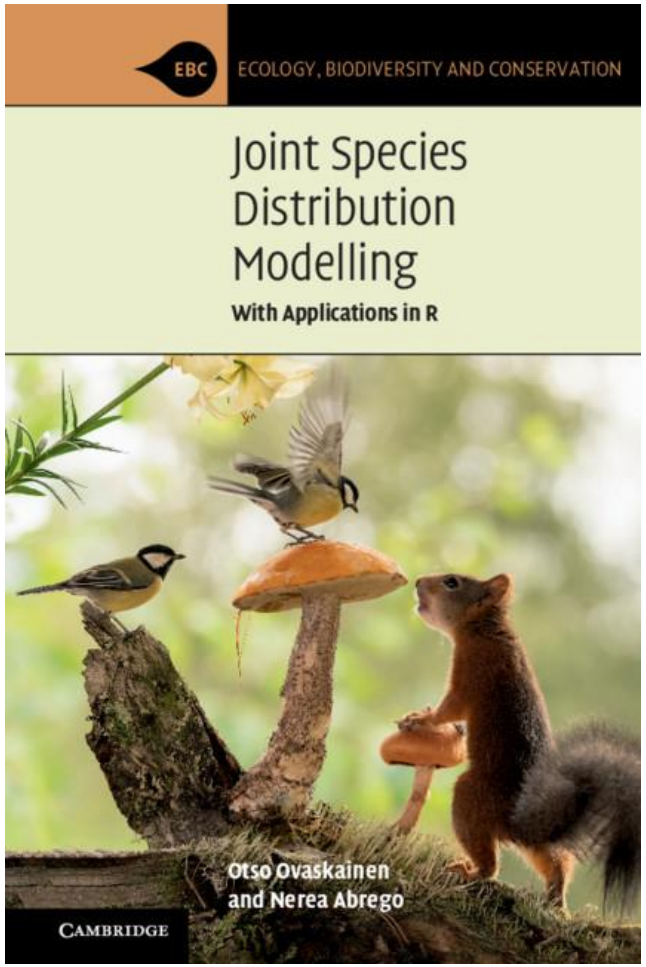
Take home message

- Species distribution modelling is advancing rapidly, and SDMs can provide valuable support to experts preparing a Pest Risk Analysis.
- They have become a fundamental tool in PRA.
- However, experts must always be fully aware of the limitations of modelling and of SDMs in particular, so that models are used to support their decisions, not to replace them.

Suggested readings







Suggested readings



Suggested readings



Software note

Achieving higher standards in species distribution modeling by leveraging the diversity of available software

Jamie M. Kass[1,2], Adam B. Smith[3], Dan L. Warren[4,5], Sergio Vignali[6], Sylvain Schmitt[7,8], Matthew E. Aiello-Lammens[9], Eduardo Arlé[10,11], Ana Márcia Barbosa[12], Olivier Broennimann[13], Marlon E. Cobos[14], Maya Guéguen[15], Antoine Guisan[13], Cory Merow[16], Babak Naimi[17], Michael P. Nobis[18], Ian Ondo[19,20], Luis Osorio-Olvera[21,22], Hannah L. Owens[23,24], Gonzalo E. Pinilla-Buitrago[25], Andrea Sánchez-Tapia[26], Wilfried Thuiller[15], Roozbeh Valavi[27], Santiago José Elías Velazco[28,29], Alexander Zizka[30] and Damaris Zurell[31]

Suggested readings

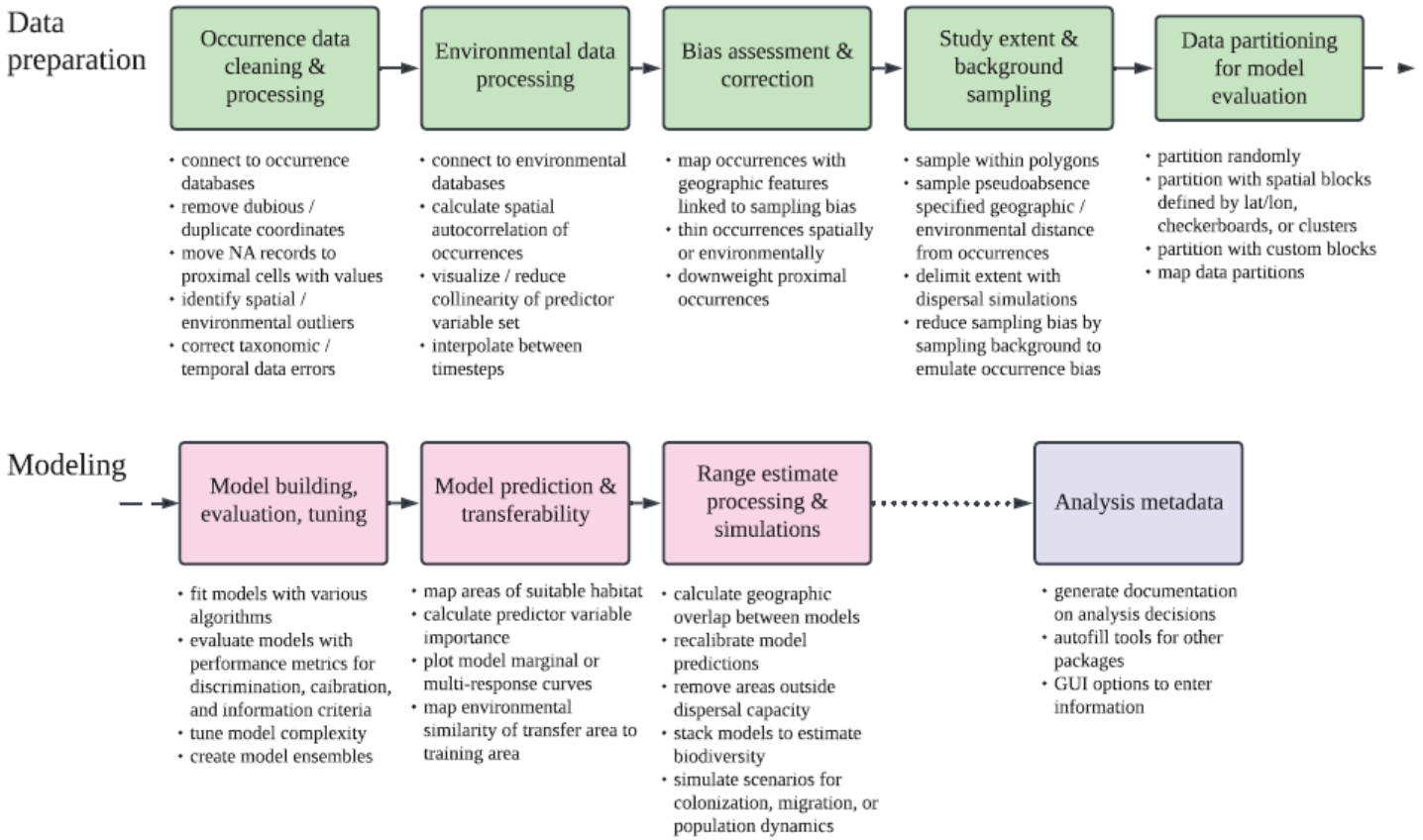


Figure 1. Species distribution model (SDM) workflow as described in this paper, separated into steps within two main phases: "Data preparation" (green) and "Modeling" (pink). Analysis metadata (purple) is collated after the modeling phase is complete. A summary of analyses, visualizations, and processes enabled by R packages is found below each step.

Suggested readings

CORRECTED 11 APRIL 2019; SEE FULL TEXT

SCIENCE ADVANCES | REVIEW

ECOLOGY

Standards for distribution models in biodiversity assessments

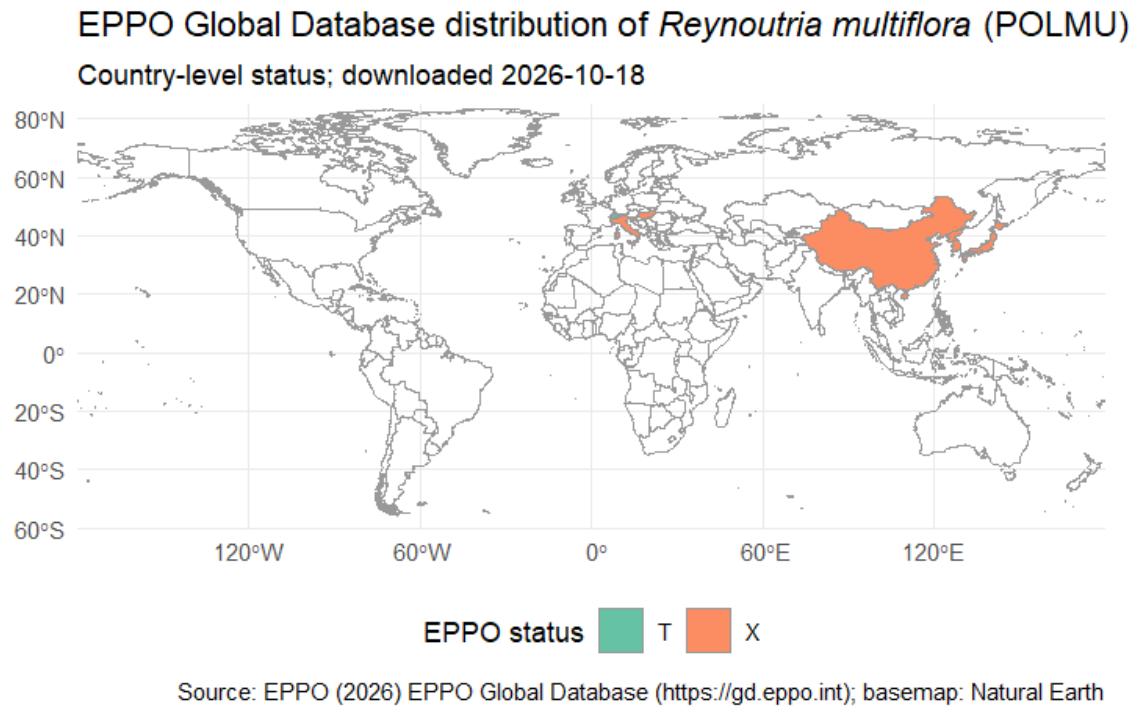
Miguel B. Araújo[1,2,3]*, Robert P. Anderson[4,5,6], A. Márcia Barbosa[3], Colin M. Beale[7], Carsten F. Dormann[8], Regan Early[9], Raquel A. Garcia[2,3,10,11], Antoine Guisan[12,13], Luigi Maiorano[14,15], Babak Naimi[2], Robert B. O'Hara[16,17], Niklaus E. Zimmermann[18,19], Carsten Rahbek[2,20]

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.

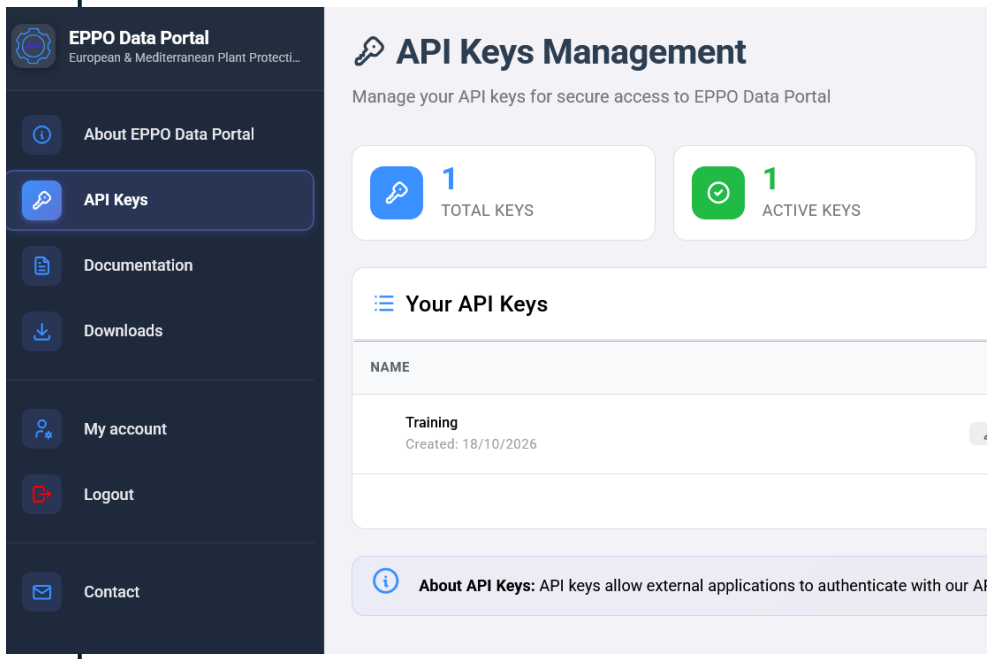
Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Downloaded fr

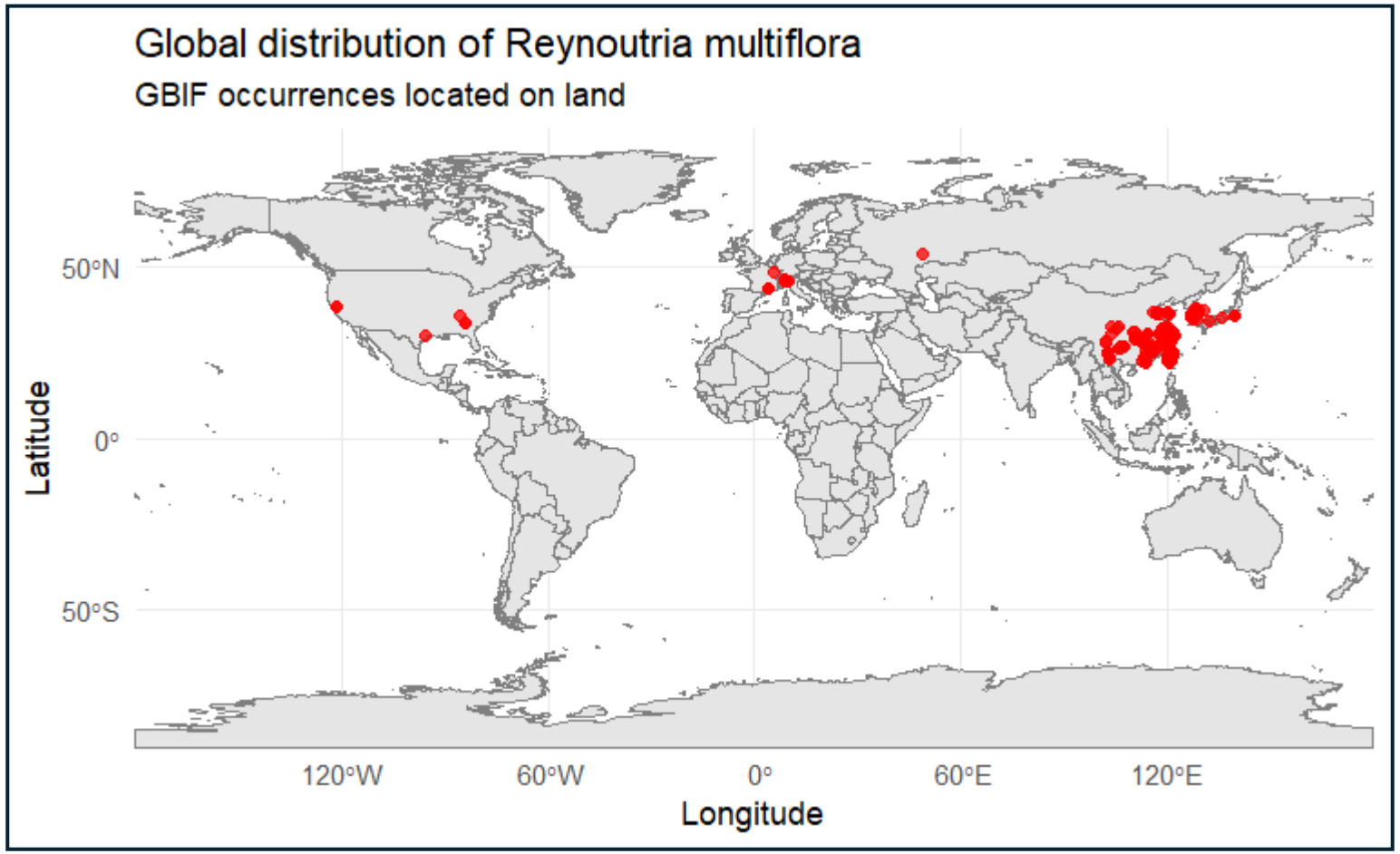
Optional group exercise (EPPO GDB API)



RStudio®

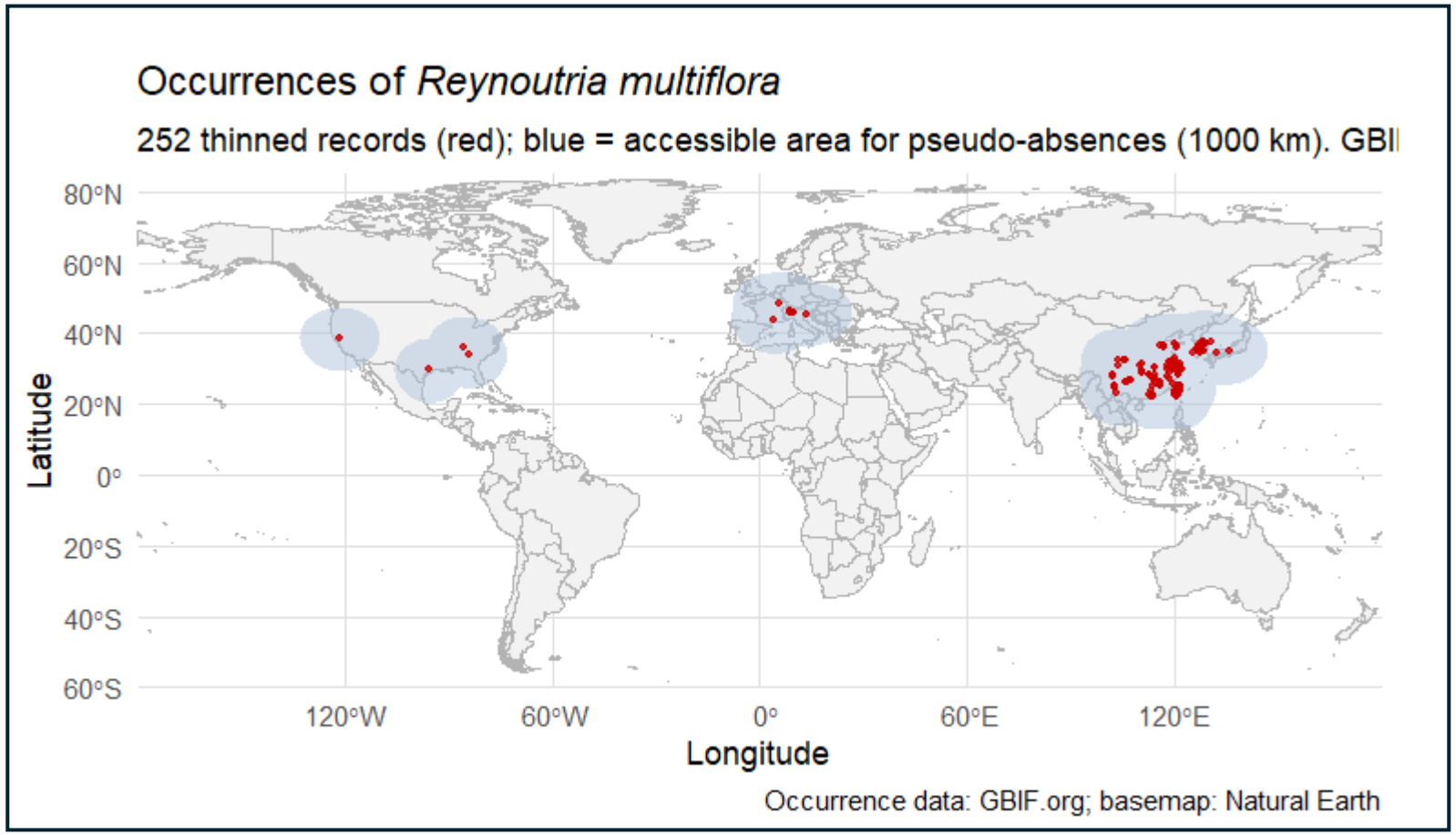


Optional group exercise



ChatGPT (OpenAI) was used to assist in drafting and refining the R code for downloading and processing GBIF occurrence data, producing (1) the species distribution maps here on the left, and (2) implementing the species distribution modelling workflow using the biomod2 package (see the next slide). The code and resulting outputs should be independently verified by the users/authors.

Optional Group exercise



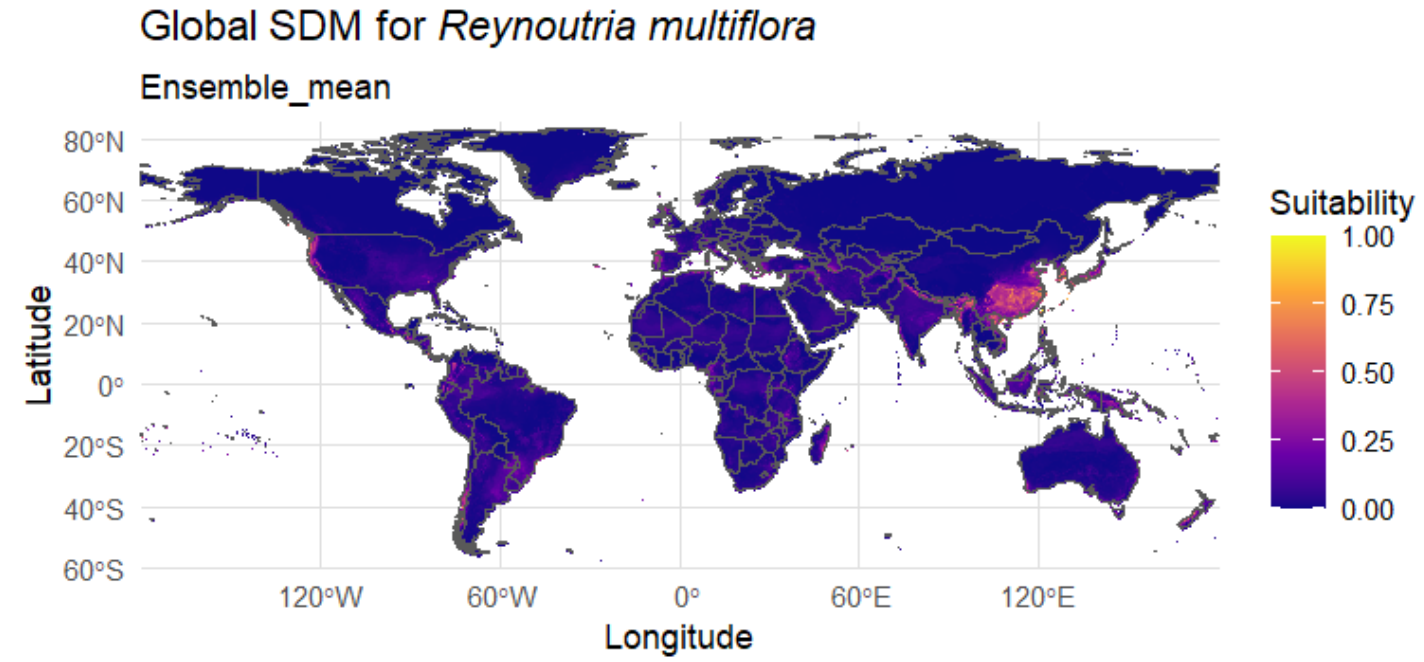
The first draft of this workflow was generated with ChatGPT (OpenAI).# It was subsequently reviewed and revised with Claude Sonnet 5.5# (Anthropic, https://claude.ai), October 2026.

The revisions covered GBIF retrieval (taxon-key queries, record cleaning, spatial thinning), a VIF check of the predictors, and the pseudo-absence design (accessible-area background).

All code and outputs should be checked by the trainers, who take full responsibility for the analysis.

For publication, use rgbif::occ_download() (needs a GBIF account) so that the data set has a citable DOI.

Optional Group exercise



BIO4, BIO6, BIO12, BIO15. Calibrated within a 1000 km accessible area; elsewhere = extrapolation.

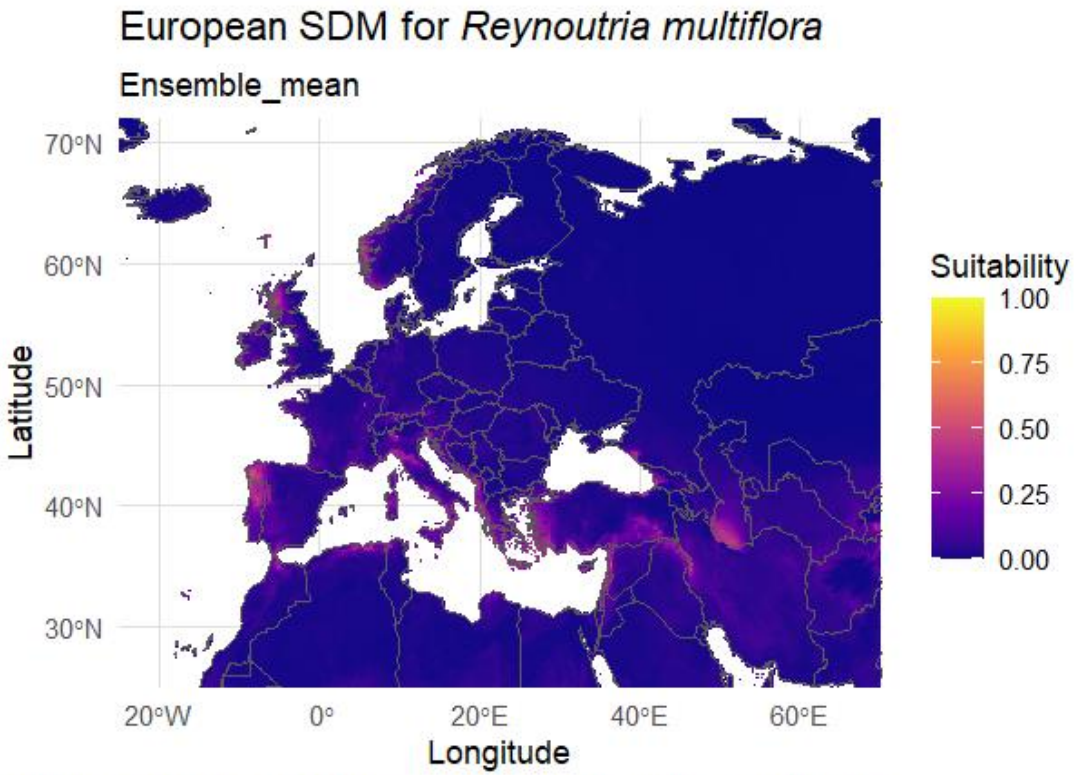
The first draft of this workflow was generated with ChatGPT (OpenAI).# It was subsequently reviewed and revised with Claude Sonnet 5.5# (Anthropic, https://claude.ai), October 2026.

The revisions covered GBIF retrieval (taxon-key queries, record cleaning, spatial thinning), a VIF check of the predictors, and the pseudo-absence design (accessible-area background).

All code and outputs should be checked by the trainers, who take full responsibility for the analysis.

For publication, use rgbif::occ_download() (needs a GBIF account) so that the data set has a citable DOI.

Optional Group exercise



BIO12, BIO15. Calibrated within a 1000 km accessible area; elsewhere = extrapolation.

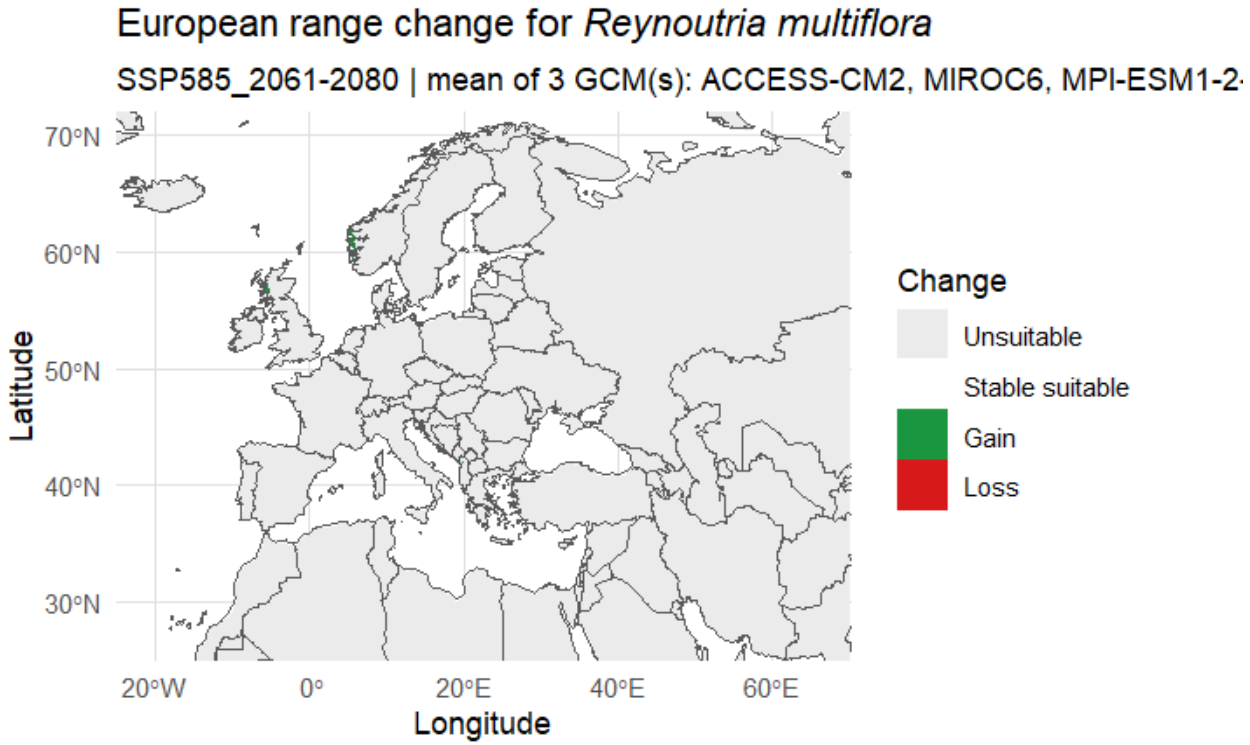
The first draft of this workflow was generated with ChatGPT (OpenAI).# It was subsequently reviewed and revised with Claude Sonnet 5.5# (Anthropic, https://claude.ai), October 2026.

The revisions covered GBIF retrieval (taxon-key queries, record cleaning, spatial thinning), a VIF check of the predictors, and the pseudo-absence design (accessible-area background).

All code and outputs should be checked by the trainers, who take full responsibility for the analysis.

For publication, use rgbif::occ_download() (needs a GBIF account) so that the data set has a citable DOI.

Optional Group exercise



in climatic suitability relative to the current climate; no dispersal or land-use effects.



Optional Group exercise

Section 16 uses CMIP6 future climate projections from WorldClim 2.1, downloaded with geodata::cmip6_world() at 10 arc-minutes. These are the same 19 bioclimatic variables as the current climate, of which the model uses BIO1, BIO4, BIO6, BIO12 and BIO15.
The settings in Section 1 define 4 scenarios, each projected with 3 climate models:
Emission pathways (FUTURE_SSPS)
SSP2-4.5 ("245"): a "middle of the road" pathway, with intermediate emissions and roughly 2.5–3 °C of global warming by 2100.
SSP5-8.5 ("585"): a fossil-fuel-intensive, high-emission pathway, with roughly 4–5 °C by 2100. It is usually treated as a worst case.
Time periods (FUTURE_PERIODS)
2041–2060, around the 2050s
2061–2080, around the 2070s
Global climate models (FUTURE_GCMS)
ACCESS-CM2 (Australia)
MIROC6 (Japan)
MPI-ESM1-2-HR (Germany)
That gives 2 SSPs × 2 periods × 3 GCMs = **12 future projections**. For each SSP × period combination, the script averages suitability across the 3 GCMs and records how many of them agree a cell is suitable. It also maps gain, loss and stable areas against the current climate, which is the 1970–2000 WorldClim baseline.
To change them, edit the three vectors in Section 1:
SSPs can be "126", "245", "370" or "585". SSP1-2.6 is a common low-emission contrast.
Periods can be "2021-2040", "2041-2060" or "2061-2080".
More GCMs make the average more robust. Any combination WorldClim doesn't provide is skipped with a message.
All of these combinations, including the GCMs that actually succeeded, are recorded in Reynoutria_multiflora_sessionInfo.txt and ..._future_projection_folders.csv.

Thank you!

linkedin.com/company/eppo-oepp/

bsky.app/profile/epponews.bsky.social

www.facebook.com/EPPOsecretariat/