A Julia toolkit for species distribution data

Timothée Poisot
Département de Sciences Biologiques, Université de Montréal, Montréal, Canada
timothee.poisot@umontreal.ca

Abstract LATER

Introduction

Species Distribution Models (SDMs) are one of the most effective predictive approach to study the global distribution of biodiversity (Elith & Leathwick, 2009). The training and evaluation of a SDM requires many steps, governing both its design and reporting (Zurell et al., 2020) and ultimate use and interpretation (Araújo et al., 2019). In the recent years, there has been an increase in the number of software packages and tools to assist ecologists with the development of species distribution models.

Because the practice of species distribution modeling and analysis usually involve many different data types, tools that can provide an integrated environment are important: many existing packages have been designed independently, and therefore may suffer when it comes to interoperability. In this manuscript, we present **SpeciesDistributionToolkit** (abbreviated as **SDT**), a meta-package for the **Julia** programming language, offering an integrated environment for the retrieval, formatting, and interpretation of data relevant to the modeling of species distributions.

As Kass et al. (2024) point out, this increase in the diversity of packages (most of them in the **R** language) is a good thing, as it can accommodate multiple workflows, and contributes to the adoption of good practices in the field. Yet, Kellner et al. (2025) highlight that about 20% of publications for abundance or distribution models are not reproducible because of issues in package dependencies. A leading design consideration for **SDT** was to prevent this issue from happening, both by relying on strict semantic versioning, but also through the use of interfaces rather dependencies between the components of **SDT**.

The **SDT** package is now used as part of the BON-in-a-Box project (Griffith et al., 2024), which seeks to facilitate the calculation and reporting of biodiversity indicators supporting the Kunming-Montréal Global Biodiversity Framework, to remove barriers to biodiversity data analysis (Gonzalez et al., 2023). In this manuscript, we describe (i) the high-level functionalities of the package, (ii) core design principles that facilitate long-term maintenance and development, and (iii) illustrative case studies with fully reproducible Jupyter notebooks.

Application description

SpeciesDistributionToolkit is released as a package for the **Julia** programming language (Bezanson et al., 2017), licensed under the open-source initiative approved MIT license. It has evolved from a previous collection of packages to handle GBIF data (Dansereau & Poisot, 2021), and now provides extended functionalities and improved performances. The package is registered in the **Julia** package repository and can be downloaded and installed anonymously. It is compatible with version 1.8 and above. The full source and complete edition history is available at `https://github.com/PoisotLab/SpeciesDistributionToolkit.jl`. This page additionally has a link to the documentation, containing a full reference for the package functions, a series of briefs how-to examples, and longer vignettes showcasing more integrative examples.

Component packages: An overview of the **SDT** package is given in Figure 1. The project is organized as a "monorepo", in which multiple packages live. This allows expanding the scope of the package by moving functionalities into new component packages, without complexifying the installation process. As **SDT** is registered in the **Julia** package repository, it can be installed by using `add SpeciesDistributionToolkit` when in package mode at the **Julia** prompt.

When loading the **SDT** package with `using SpeciesDistributionToolkit`, all component packages are automatically and transparently loaded. Therefore, users do not need to know where a specific method or function resides to use it. In the next section, we discuss how this modular design ensure that we can grow the functionality of the toolkit over time, while maintaining strict backward compatibility *and* allowing full reproducibility of an analysis.

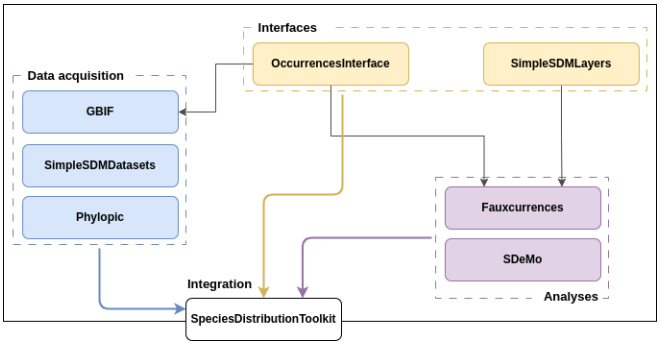

Figure 1: Overview of the packages included in **SpeciesDistributionToolkit**. The packages are color-coded by intended use, and their more specific content is presented in the main text. Note that because the package relies on *interfaces* to facilitate code interoperability, there are only three dependency relationships.

The **SDT** package primarily provides integration between the other packages, through the mechanism of method overloading, allowing to efficiently join packages together (Roesch et al., 2023). Additional functionalities that reside in the top-level package are the generation of pseudo-absences inspired by Barbet-Massin et al. (2012), access to the `gadm.org` database, handling of polygon data and zonal statistics, and various quality of life methods. Because of the modular nature of the code, any of these functions can be transparently moved to their own packages in the future.

The **SimpleSDMLayers** package offers a series of types to represent raster data in various projections, and a series of functions to operate on these layers. This package provides the main data representation for most functionalities that **SDT** supports, and handles saving and loading data.

The **OccurrencesInterface** is a light-weight package to provide a common interface for occurrence data. It implements abstract and concrete types to define a single occurrence and a collection thereof, and a series of methods allowing any occurrence data provider or data representation to become fully interoperable with the rest of **SDT**. All **SDT** methods that handle occurrence data do so through the **OccurrencesInterface** interface, allowing future data sources to be integrated without the need for new code.

The **GBIF** package offers access to the `gbif.org` streaming API (GBIF: The Global Biodiversity Information Facility, 2025), including the ability to retrieve, filter, and restart downloads. Although this package returns a rich data representation for occurrence data, all the objects it returns adhere to the **OccurrencesInterface** interface.

SimpleSDMDatasets implements an interface to retrieve and locally store raster data, which can be extended by users to support additional data sources. In addition, it offers access to a series of data sources, including the biodiversity mapping project (Jenkins et al., 2013), the EarthEnv collection for land cover (Tuanmu & Jetz, 2014) and habitat heterogeneity (Tuanmu & Jetz, 2015), Copernicus land cover 100m data (Buchhorn et al., 2020), the PaleoClim (Brown et al., 2018) data, the WorldClim 1 and 2 data (Fick & Hijmans, 2017) and their projections under various RCPs and SSP, and part of the CHELSA 1 and 2 data (Karger et al., 2017) and their projections under various RCPs and SSPs.

Phylopic offers a wrapper around the `phylopic.org` API to download silhouettes for taxonomic entities. It also provides utilities for citation of the downloaded images. Its functionalities are similar to the **rphylopic** package (Gearty & Jones, 2023).

The **Fauxcurrences** packages is inspired by the work of Osborne et al. (2022), and allows generating a series of simulated occurrence data that have the same statistical structure as observed ones. The package supports multi-species data, with user-specified relative weight of intra and inter-specific distances conservation.

The **SDeMo** package is aimed at providing tools to use as part of training and education material on species distribution modeling. By providing a series of data transformation (PCA, Whitening, z-score) and classifiers (BIOCLIM, Naive Bayes, and decision trees), it offers the basic elements to demonstrate training and evaluation of SDMs, as well as techniques related to ensembles and bagging. In addition, to promote the use of interpretable techniques, the package supports regular (Elith et al., 2005) and inflated (Zurell et al., 2012) partial responses, as well as the calculation and mapping of Shapley values (Mesgaran et al., 2014; Wadoux et al., 2023), and the generation of counterfactuals (Van Looveren & Klaise, 2019, Karimi et al. (2019)).

Software information: **SDT** uses the built-in **Julia** package manager to ensure that the version of all dependencies are kept up to date. Furthermore, we use strict semantic versioning: major versions correspond to no breaking changes in user-developped code, minor versions increase with additional functionalities, and patch releases cover minor bug fixes or documentation changes. All packages have a *CHANGELOG* file, which documents what changes are included in each release. Following a constructive cost model analysis (Kemerer, 1987) of the version described in this publication, the package represents approx. 11k lines of active code (no blank lines, no comments), for an estimated development cost of approx. 325k USD.

This strict reliance on semantic versioning solves the issues of maintaining compatibility when new functionalities are added: all releases in the *v1.x.x* branch of **SDT** depend on component packages in their respective *v1.x.x* branch, and users can benefit from now functionalities without risking to break existing code. This behavior is extensively tested, both using unit tests, and through integration testing generated as part of the online documentation.

Kellner, Doser, & Belant (2025) reported that about 20% of failures to reproduce species distribution or abundance modeling code was related to package issues. The strict reliance on semantic versioning, alongside technical choices in the **Julia** package manager and repository, means that it is possible to specify the full version of all dependencies used in

a project, which addresses this important obstacle to reproducibility.

Integration with other packages: The **SDT** package benefits from close integration with other packages in the **Julia** universe. Notably, this includes **Makie** (and all related backends) for plotting and interactive data visualisation, where usual plot types are overloaded for both layer and occurrence data. Most data handled by **SDT** can be exported using the **Tables** interface, which allows data to be consumed by other packages like **DataFrames** and **MLJ**, or directly saved as csv files.

Interfaces internal to **Julia** are also implemented whenever they make sense. In particular, **SimpleSDMLayers** objects behave like arrays, are iterable, and broadcastable; objects from **OccurrencesInterface** behave as arrays and are similarly iterable. The **SDeMo** package relies on part of the **StatsAPI** interface, allowing to easily define new data transformation and classifier types to support additional features.

Achieving integration with other packages through method overloading and the adherence to well-established interfaces is important, as it increases the chances that additional functionalities external to **SDT** can be used directly or fully supported with minimal addition of code.

Illustrative case studies

In this section, we provide a series of case studies, meant to illustrate the use of the package. The on-line documentation offers longer tutorials, as well as a series of how-to vignettes to illustrate the full scope of what the package allows. The code for each of these case studies is available as fully independent Jupyter notebooks, forming the supplementary material of this article. The example we use throughout is the distribution of *Akodon montensis* (Rodentia, family Cricetidae), a host or orthohantaviruses (Burgos et al., 2021; Owen et al., 2010), in Paraguay. As the notebooks accompanying this article cover the full code required to run these case studies, we do not present code snippets in the main text, and instead focus on explaining which component packages are used in each example.

Using data from GBIF: To illustrate the interactions between the component packages, we provide a simple illustration (Supp. Mat. 1) where we (i) request occurrence data using the **GBIF** package, (ii) download the silhouette of the species through **Phylopic**, and (iii) extract temperature and precipitation data at the points of occurrence. The results are presented in Figure 2. The full notebook includes information about basic operations on raster data, as well as extraction of data based on occurrence records.

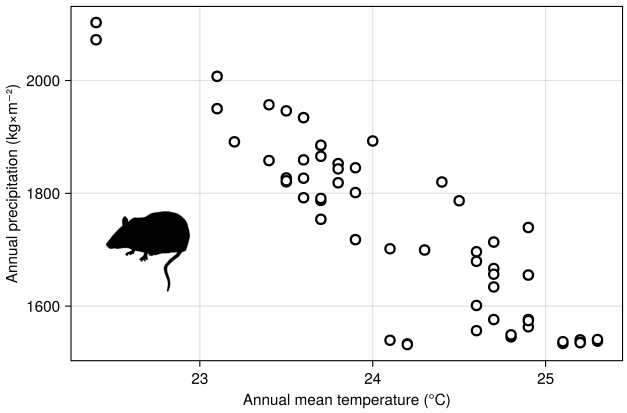

Figure 2: Relationship between temperature and precipitation (BIO1 and BIO12) at each georeferenced occurrence known to GBIF for *Akodon montensis*. The code to produce this figure is available as Supp. Mat. 1.

In practice, although the data are retrieved using the **GBIF** package, they are used internally by **SDT** through the **OccurrencesInterface** package. This package defines a small convention to handle georeferenced occurrence data, and allows to transparently integrate additional occurrence sources. By defining five methods for a custom data type, users can plug-in any occurrence data source and enjoy full compatibility with the entire **SDT** functionalities.

Landcover consensus map: In this case study (Supp. Mat. 2), we retrieve the land cover data from Tuanmu & Jetz (2014), clip them to a GeoJSON polygon describing the country of Paraguay (**SDT** can download data directly from `gadm.org`), and apply the `mosaic` operation to figure out which class is the most locally abundant. This case study uses the **SimpleSDMDatasets** package to download (and locally cache) the raster data, as well as the **SimpleSDMLayers** package to provide basic utility functions on raster data. The results are presented in Figure 3.

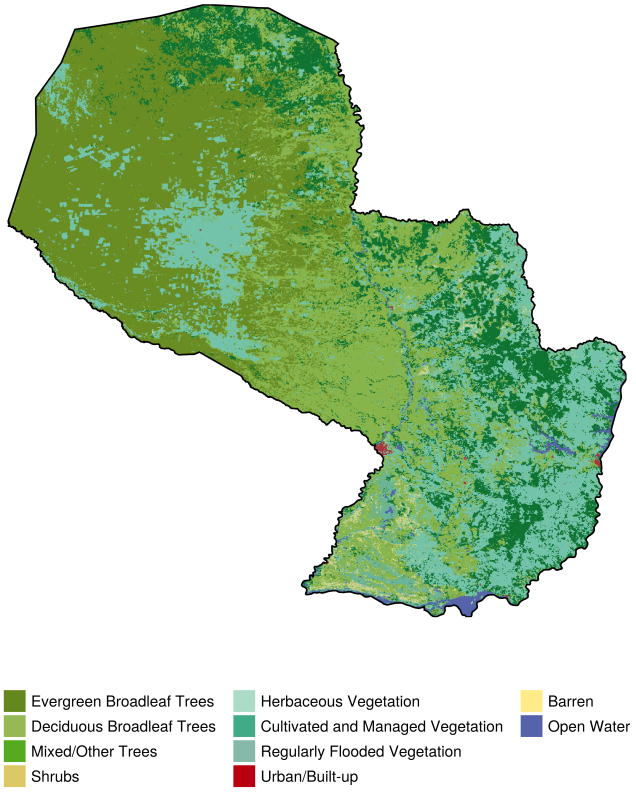

Figure 3: Land cover consensus (defined as the class with the strongest local representation) in the country of Paraguay. Only the classes that were most abundant in at least one pixel are represented. The code to produce this figure is available as Supp. Mat. 2.

When first downloading data through **SimpleSDMDatasets**, they will be stored locally for future use. When the data are requested a second time, they are read directly from the disk, speeding up the process massively. Note that the location of the data is (i) standardized by the package itself, making the file findable to humans, and (ii) changeable by the user to, *e.g.*, store the data within the project folder rather than in a central location. As much as possible, **SDT** will only read the part of the raster data that is required given the region of interest to the user. This is done by providing additional context in the form of a bounding box (in WGS84, regardless of the underlying raster data projection). **SDT** has methods to calculate the bounding box for all the objects it supports.

Training a species distribution model: In this case study, we illustrate the integration of **SDeMo** and **SimpleSDMLayers** to train a species distribution model. We specifically train a rotation forest (Bagnall et al., 2018), an homogeneous ensemble of PCA followed by decision trees. The results are presented in Figure 4. The model is built by selecting an optimal suite of BioClim variables, then predicted in space, and the resulting predicted species range is finally clipped by the elevational range observed in the occurrence data.

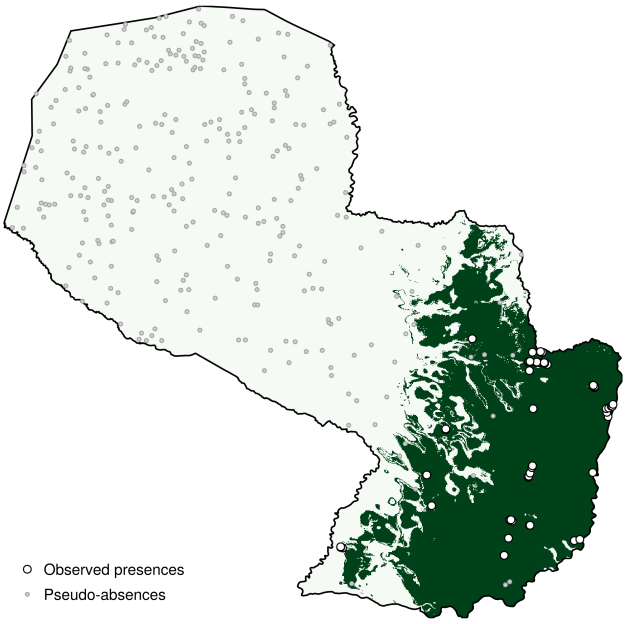

Figure 4: Predicted range of *Akodon montensis* in Paraguay based on a rotation forest trained on GBIF occurrences and the BioClim variables. The code to produce this figure is available as Supp. Mat. 3.

The full notebook (Supp. Mat. 3) has additional information on routines for variable selection, stratified cross-validation, as well as the construction of the ensemble from a single PCA and decision tree. In addition, Supp. Mat. 3 presents the partial responses and Shapley values for the most important predictor.

Distribution of a virtual species: (Leroy et al., 2016)

The results are presented in Figure 5.

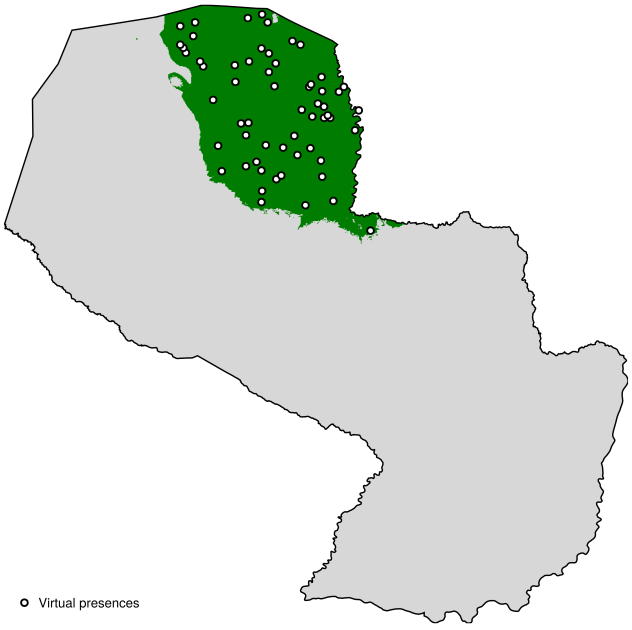

Figure 5: yeah

Bibliography

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Bagnall, A., Flynn, M., Large, J., Line, J., Bostrom, A., & Cawley, G. (2018). Is rotation forest the best classifier for problems with continuous features?. *Arxiv [Cs.lg]*.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many?: How to use pseudo-absences in niche modelling?. *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210x.2011.00172.x

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review. Society for Industrial and Applied Mathematics*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Brown, J. L., Hill, D. J., Dolan, A. M., Carnaval, A. C., & Haywood, A. M. (2018). PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. *Scientific Data*, *5*(1), 180254–180255. https://doi.org/10.1038/sdata.2018.254

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., & Fritz, S. (2020). *Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3939050

Burgos, E. F., Vadell, M. V., Bellomo, C. M., Martinez, V. P., Salomon, O. D., & Gómez Villafañe, I. E. (2021). First evidence of Akodon-borne orthohantavirus in northeastern Argentina. *Ecohealth*, *18*(4), 429–439. https://doi.org/10.1007/s10393-021-01564-6

Dansereau, G., & Poisot, T. (2021). SimpleSDMLayers.Jl and GBIF.Jl: A framework for species distribution modeling in Julia. *Journal of Open Source Software*, *6*(57), 2872–2873. https://doi.org/10.21105/joss.02872

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, And Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, *186*(3), 280–289. https://doi.org/10.1016/j.ecolmodel.2004.12.007

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

GBIF: The Global Biodiversity Information Facility. (2025). *\textit{What is GBIF?}*.

Gearty, W., & Jones, L. A. (2023). rphylopic: An R package for fetching, transforming, and visualising PhyloPic silhouettes. *Methods in Ecology and Evolution*, *14*(11), 2700–2708. https://doi.org/10.1111/2041-210x.14221

Gonzalez, A., Vihervaara, P., Balvanera, P., Bates, A. E., Bayraktarov, E., Bellingham, P. J., Bruder, A., Campbell, J., Catchen, M. D., Cavender-Bares, J., Chase, J., Coops, N., Costello, M. J., Dornelas, M., Dubois, G., Duffy, E. J., Eggermont, H., Fernandez, N., Ferrier, S., … Wright, E. (2023). A global biodiversity observing system to unite monitoring and guide action. *Nature Ecology & Evolution*, 1–5. https://doi.org/10.1038/s41559-023-02171-0

Griffith, J., Lord, J.-M., Catchen, M. D., Arce-Plata, M. I., Bohorquez, M. F. G., Chandramohan, M., Diaz-Corzo, M. C., Gravel, D., Gonzalez, L. F. U., Gutiérrez, C., Helfenstein, I., Hoban, S., Kass, J. M., Laroque, G., Laikre, L., Leigh, D., Leung, B., Mastretta-Yanes, A., Millette, K., … Gonzalez, A. (2024). *BON in a Box: An Open and Collaborative Platform for Biodiversity Monitoring, Indicator Calculation, and Reporting*. https://doi.org/10.32942/X2M320

Jenkins, C. N., Pimm, S. L., & Joppa, L. N. (2013). Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(28), E2602–10. https://doi.org/10.1073/pnas.1302251110

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1), 170122–170123. https://doi.org/10.1038/sdata.2017.122

Karimi, A.-H., Barthe, G., Balle, B., & Valera, I. (2019). Model-agnostic counterfactual explanations for consequential decisions. *Arxiv [Cs.lg]*.

Kass, J. M., Smith, A. B., Warren, D. L., Vignali, S., Schmitt, S., Aiello-Lammens, M. E., Arlé, E., Márcia Barbosa, A., Broennimann, O., Cobos, M. E., Guéguen, M., Guisan, A., Merow, C., Naimi, B., Nobis, M. P., Ondo, I., Osorio-Olvera, L., Owens, H. L., Pinilla-Buitrago, G. E., … Zurell, D. (2024). Achieving higher standards in species distribution modeling by leveraging the diversity of available software. *Ecography*. https://doi.org/10.1111/ecog.07346

Kellner, K. F., Doser, J. W., & Belant, J. L. (2025). Functional R code is rare in species distribution and abundance papers. *Ecology*, *106*(1), e4475. https://doi.org/10.1002/ecy.4475

Kemerer, C. F. (1987). An empirical validation of software cost estimation models. *Communications of the ACM*, *30*(5), 416–429. https://doi.org/10.1145/22899.22906

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Mesgaran, M. B., Cousens, R. D., & Webber, B. L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity & Distributions*, *20*(10), 1147–1159. https://doi.org/10.1111/ddi.12209

Osborne, O. G., Fell, H. G., Atkins, H., Tol, J. van, Phillips, D., Herrera-Alsina, L., Mynard, P., Bocedi, G., Gubry-Rangin, C., Lancaster, L. T., Creer, S., Nangoy, M., Fahri, F., Lupiyaningdyah, P., Sudiana, I. M., Juliandi, B., Travis, J. M. J., Papadopulos, A. S. T., & Algar, A. C. (2022). Faux-currence: simulating multi-species occurrences for null models in species distribution modelling and biogeography. *Ecography*, *2022*(7), e5880. https://doi.org/10.1111/ecog.05880

Owen, R. D., Goodin, D. G., Koch, D. E., Chu, Y.-K., & Jonsson, C. B. (2010). Spatiotemporal variation in Akodon montensis (Cricetidae: Sigmodontinae) and hantaviral seroprevalence in a subtropical forest ecosystem. *Journal of Mammalogy*, *91*(2), 467–481. https://doi.org/10.1644/09-MAMM-A-152.1

Roesch, E., Greener, J. G., MacLean, A. L., Nassar, H., Rackauckas, C., Holy, T. E., & Stumpf, M. P. H. (2023). Julia for biologists. *Nature Methods*, *20*(5), 655–664. https://doi.org/10.1038/s41592-023-01832-z

Tuanmu, M.-N., & Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling: Consensus land cover. *Global Ecology and Biogeography: A Journal of Macroecology*, *23*(9), 1031–1045. https://doi.org/10.1111/geb.12182

Tuanmu, M.-N., & Jetz, W. (2015). A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling: Global habitat heterogeneity. *Global Ecology and Biogeography: A Journal of Macroecology*, *24*(11), 1329–1339. https://doi.org/10.1111/geb.12365

Van Looveren, A., & Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. *Arxiv [Cs.lg]*.

Wadoux, A. M. J.-C., Saby, N. P. A., & Martin, M. P. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. *SOIL*, *9*(1), 21–38. https://doi.org/10.5194/soil-9-21-2023

Zurell, D., Elith, J., & Schröder, B. (2012). Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. *Diversity & Distributions*, *18*(6), 628–634. https://doi.org/10.1111/j.1472-4642.2012.00887.x

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Townsend Peterson, A., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960