N-SDM: a high-performance computing pipeline for Nested Species Distribution Modelling

Adde, Antoine ; Rey, Pierre-Louis ; Brun, Philipp ; Külling, Nathan ; Fopp, Fabian ; Altermatt, Florian ; Broennimann, Olivier ; Lehmann, Anthony ; Petitpierre, Blaise ; Zimmermann, Niklaus E ; Pellissier, Loïc ; Guisan, Antoine

DOI: https://doi.org/10.1111/ecog.06540

ECOGRAPHY

Software Note

N-SDM: a high-performance computing pipeline for Nested Species Distribution Modelling

Antoine Adde[1], Pierre-Louis Rey[1], Philipp Brun[2], Nathan Külling[3], Fabian Fopp[2,4], Florian Altermatt[5,6], Olivier Broennimann[1,7], Anthony Lehmann[3], Blaise Petitpierre[1,8], Niklaus E. Zimmermann[2], Loïc Pellissier[2,4] and Antoine Guisan[1,7]

[1]Inst. of Earth Surface Dynamics, Faculty of Geosciences and Environment, Univ. of Lausanne, Lausanne, Switzerland
[2]Land Change Science Research Unit, Swiss Federal Inst. for Forest, Snow and Landscape Research, WSL, Birmensdorf, Switzerland
[3]EnviroSPACE, Inst. for Environmental Sciences, Univ. of Geneva, Geneva, Switzerland
[4]Ecosystems Landscape Evolution, Inst. for Terrestrial Ecosystems, Dept of Environmental System Sciences, Zurich, Switzerland
[5]Dept of Evolutionary Biology and Environmental Studies, Faculty of Science, Univ. of Zurich, Zurich, Switzerland
[6]Dept of Aquatic Ecology, Eawag, Swiss Federal Inst. of Aquatic Science and Technology, Duebendorf, Switzerland
[7]Dept of Ecology and Evolution, Univ. of Lausanne, Lausanne, Switzerland
[8]InfoFlora, Chambésy-Geneva, Switzerland

Correspondence: Antoine Adde (antoine.adde@unil.ch)

Ecography
2023: e06540
doi: 10.1111/ecog.06540
Subject Editor: Brody Sandel
Editor-in-Chief:
Christine N. Meynard
Accepted 03 February 2023

Predicting contemporary and future species distributions is relevant for science and decision making, yet the development of high-resolution spatial predictions for numerous taxonomic groups and regions is limited by the scalability of available modelling tools. Uniting species distribution modelling (SDM) techniques into one high-performance computing (HPC) pipeline, we developed *N-SDM*, an SDM platform aimed at delivering reproducible outputs for standard biodiversity assessments. *N-SDM* was built around a spatially-nested framework, intended at facilitating the combined use of species occurrence data retrieved from multiple sources and at various spatial scales. *N-SDM* allows combining two models fitted with species and covariate data retrieved from global to regional scales, which is useful for addressing the issue of spatial niche truncation. The set of state-of-the-art SDM features embodied in *N-SDM* includes a newly devised covariate selection procedure, five modelling algorithms, an algorithm-specific hyperparameter grid search, and the ensemble of small-models approach. *N-SDM* is designed to be run on HPC environments, allowing the parallel processing of thousands of species at the same time. All the information required for installing and running *N-SDM* is openly available on the GitHub repository https://github.com/N-SDM/N-SDM.

Keywords: biodiversity, ensemble modelling, habitat suitability, high-performance computing, modelling pipeline, R package, species distribution models

NORDIC SOCIETY OIKOS

www.ecography.org

Background

Nature management and conservation decisions need to be guided by standardized biodiversity data (Araujo et al. 2019, Jetz et al. 2019, Kays et al. 2020). The largest fraction of raw biodiversity data exists in the form of billions of species occurrence records, an ever-increasing number boosted by the boom in citizen science initiatives and their compilation with scientific surveys or natural history collections (Amano et al. 2016, Pocock et al. 2017, Kays et al. 2020). To translate these spatially discrete data into continuous biodiversity maps, species distribution models (SDMs) relate species occurrences to environmental covariates (Franklin 2010, Peterson et al. 2011, Guisan et al. 2017). SDMs have developed extensively over the last two decades and are now key tools to predict the state and fate of biodiversity (Guisan et al. 2013, Ferrier et al. 2016, Araujo et al. 2019). Delivering SDM outputs that are relevant for biodiversity assessments requires modelling pipelines equipped with best-available techniques and that are scalable enough to process the large quantities of available data.

To help streamlining key SDM steps, several platforms have been developed, mainly in the form of R packages, with the best-known including Biomod2 (Thuiller et al. 2009), enmEval (Muscarella et al. 2014), dismo (Hijmans et al. 2017), Wallace (Kass et al. 2018), or SDMtune (Vignali et al. 2020). Although existing SDM platforms have widely contributed to SDM popularity (Guisan et al. 2013, Araujo et al. 2019, Hao et al. 2019), they have not been explicitly designed to cope with the challenge of high-dimensional input data, either in relation with the number of species or candidate covariates available for modelling them. To date and to our knowledge, an SDM platform specifically designed for high-performance computing (HPC) environments that are capable of efficiently handling large volumes of data was lacking.

The precision and coverage of species occurrence data can vary as regards to their sources, and modelling frameworks should accommodate these characteristics (Pagel et al. 2014, Fletcher et al. 2019, Pacifici et al. 2019). For instance, high-precision data can be obtained from regional databases, while larger scale data are useful for providing a global overview of the species distributions, but they are generally recorded with coarser precision. Currently no end-to-end SDM platform is including features for facilitating the combination of occurrence records retrieved from various sources and at different spatial scales (e.g. regional and global). Such features would be useful for addressing the issue of spatial niche truncation, which is likely to occur when models are fitted based on geographically restricted occurrence data that do not encompass the whole species range (Mateo et al. 2019a, Chevalier et al. 2021, Scherrer et al. 2021).

With the increasing need to support systematic conservation practices with informative spatial predictions for large sets of species, there was a pressing need for an automated modelling tool equipped with features that are lacking in existing platforms, including an advanced covariate selection procedure, solutions for rare or infrequent species, and a strategy for identifying best model parameter combinations. First, although rare species account for a large proportion of many taxonomic groups and are key in biodiversity assessments, their small sample size makes them challenging to model (Hernandez et al. 2006, Galante et al. 2017, Enquist et al. 2019). Approaches like the 'ensemble of small models' (ESM) (Lomba et al. 2010, Breiner et al. 2015, Breiner et al. 2018) have been developed to overcome this rare species modelling paradox (Lomba et al. 2010), by allowing for more predictors than in traditional SDM approaches. Similarly, selecting the best set of covariates out of many candidates is a key and highly influential step of the SDM process (Araujo and Guisan 2006, Fourcade et al. 2018, Cobos et al. 2019). Yet, there is currently no widely adopted reference approach by which to perform the covariate selection, and most of the existing SDM platforms do not include any feature for doing it. Last, identifying the optimal combination of model hyperparameters is important for delivering accurate predictions, but the exercise is not straightforward, especially when working with multiple algorithms. Parallel grid-search approaches have been proven useful for automatic hyperparameter tuning and should be present in any new SDM platform (Kuhn and Johnson 2013, Chicco 2017, Vignali et al. 2020).

Here we introduce *N-SDM*, a flexible HPC-oriented SDM platform built around a scale-nesting framework intended at facilitating the combined use of species occurrence data retrieved from multiple sources and at various spatial scales. Integrating state-of-the-art SDM techniques into one HPC pipeline, *N-SDM* aims at delivering scalable and reproducible outputs for standard biodiversity assessments. After providing a description of the structure and methods included in *N-SDM*, we ran an example application for 1500 species at 25 m resolution across Switzerland to illustrate the main operations and performances.

Methods and features

N-SDM overview

N-SDM is an HPC-oriented SDM platform designed for Linux clusters equipped with the Slurm workload manager. Most of the *N-SDM* code is written in R (92.2%), where the modelling steps are handled. The core of this R code are custom functions wrapped together into the companion R package 'nsdm'. The remaining 7.8% of the code is written in Bash language, which is primarily used to efficiently distribute the modelling tasks on the computing cluster.

Figure 1 provides an overview of the main tasks performed during an *N-SDM* run. A key characteristic of *N-SDM* is the spatially-nested framework. Two models fitted to scale-specific species and covariate data are combined: 1) a 'global' model intended at quantifying the species response to the bioclimatic conditions that can be found across its full distributional range, and 2) a 'regional' model fitted with fine-scale occurrence data and habitat covariates. The 'global' vs. 'regional' terminology is the one used by Gallien et al. (2012), a study that mainstreamed the combination of models fitted

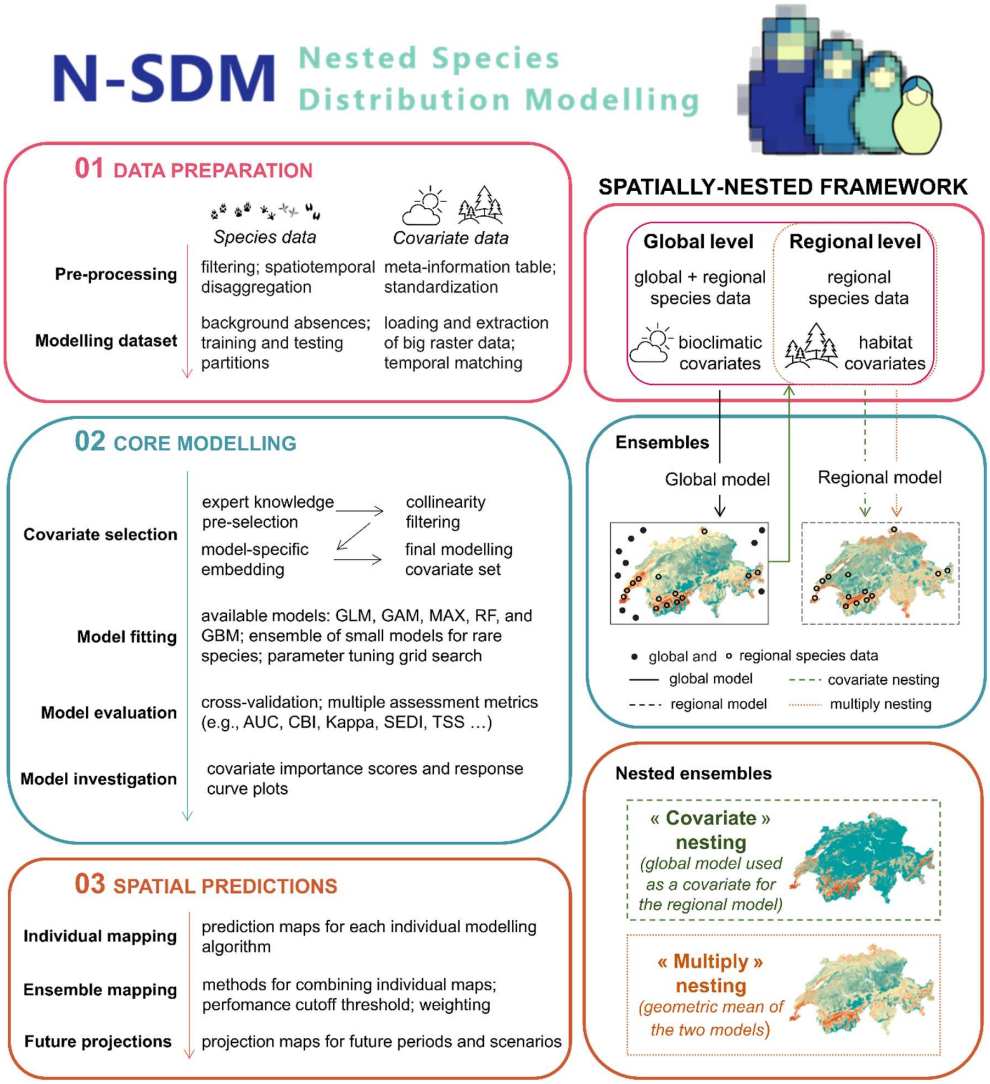

Figure 1. *N-SDM* overview. Panels on the left display the sequence of tasks performed during each of the three main stages (Data preparation, Core modelling, and Spatial predictions). Panels on the right provide a graphical representation of the spatially-nested framework used in *N-SDM*. All the complementary information required for the installation and use of *N-SDM*, along with data for an example *N-SDM* run, are available on the *N-SDM* GitHub repository https://github.com/N-SDM/N-SDM. GLM: Generalized Linear Model; GAM: Generalized Additive Model; MAX: Maxnet; RF: Random Forest; GBM: (light) Gradient Boosted Machine; AUC: Area Under the Curve; CBI: Continuous Boyce Index; Kappa: Cohen's Kappa coefficient; SEDI: Symmetric Extremal Dependence Index; TSS: True Skill Statistic.

with scale-specific species and covariate data (see Pearson et al. (2004) for one of the first studies). Further details on the spatially-nested modelling framework are provided in section 'Highlighted features'. *N-SDM* tasks are divided into three main stages: data preparation, core modelling, and spatial predictions. An extended version of the '*N-SDM* overview' section, with details on the custom functions used at the three stages, is provided in the Supporting information.

Data preparation

Two main types of input data are feeding *N-SDM*: georeferenced species occurrence data and environmental raster layers. Occurrence data from different sources can be used for fitting the global-level (full species range) and regional-level (study area extent) models. For the global level, where the aim is to cover as much as possible the bioclimatic conditions encountered throughout the species' range, the increasing quantity and extent of data available from the Global Biodiversity Information Facility (GBIF; www.gbif.org) is a valuable resource. For the regional level, more precise data, such as those extracted from national biodiversity databases or monitoring programs, can be used. Both global and regional occurrence data are used for global model fitting, but only regional data are used for the regional model. Additional details about species data formatting are provided in the Supporting information. Covariate data should be provided as raster layers, standardized to a common spatial grid

with consistent resolution, extent, and projection system. All the details for covariate data formatting are provided in the Supporting information. The two-level nesting framework allows using in the regional-level model covariates that are not available at the extent of the global level. However, the same covariate cannot be used in both global and regional models. If available, a pre-filtering correspondence table indicating which covariates are candidates for modelling a target group of species can be provided (see Supporting information for an example pre-filtering table). Such a table is useful for incorporating expert knowledge on species ecology. If temporally-dynamic covariate and species data are provided to *N-SDM*, it is possible to perform temporal matching such that *N-SDM* will extract covariate values at the time steps that best match with species occurrences. The data preparation stage also includes the spatial-temporal thinning of occurrence data and the generation of background absences. In the first *N-SDM* ver. (ver. 1.0.0), the only option available for generating background absences is 'random'. For each species and level, 10 000 background absences (user-customizable number) are randomly generated across the target areas.

Core modelling

The modelling stage starts with a newly-devised covariate selection procedure aimed at optimizing the predictive abilities and parsimony of SDMs fitted in a context of high-dimensional candidate covariate space. Selected covariates are used for fitting the modelling algorithms chosen among the five options available at the time of *N-SDM* release (ver. 1.0.0): Generalized Linear Model (GLM) (McCullagh and Nelder 1989), Generalized Additive Model (GAM) (Hastie 2017), Maxnet (MAX) (Phillips et al. 2017), Random Forest (RF) (Breiman 2001), and light Gradient Boosted Machine (GBM) (Ke et al. 2017). In addition, the ESM approach (Breiner et al. 2015, Breiner et al. 2018) can be specified as a sixth option. Model assessment and selection is based on cross-validated evaluation metrics including the Area Under the Curve (AUC) (Fielding and Bell 1997), the Continuous Boyce Index (CBI) (Hirzel et al. 2006), the maximized Cohen's Kappa coefficient (maxKappa) (Kappa 1960), or the maximized True Skill Statistic (maxTSS) (Allouche et al. 2006). More details on available model assessment metrics can be found in the Supporting information. See Guisan et al. (2017) for further explanations on the maximization approach. In addition, we developed a consensus 'Score' metric averaging the AUC′ (or Somers' D, such as $AUC' = AUC * 2 - 1$) (Somers 1962), the maxTSS, and the CBI.

Spatial predictions

Four main types of prediction surfaces are mapped by *N-SDM*: 1) the predicted response from each modelling algorithm at global and regional levels, 2) the ensemble average of these predictions for the two levels (optionally weighted averages using the value of the user-specified model evaluation metric for each target algorithm), 3) their coefficients of variation at the two levels, which provides a measure of uncertainty, and 4) the final nested ensemble combining global- and regional-level predictions according to the scale-nesting strategy specified by the user. No ensemble of the two nesting strategies is computed. However, if the two options are specified, *N-SDM* will produce individual predictions for each of the two strategies, which is useful for comparison. At this time, two nesting strategies are provided with *N-SDM*: 'multiply' and 'covariate' (Fig. 1). Details on the scale-nesting strategies are provided in the 'Highlighted features' section. If raster layers for future scenarios (e.g. climate and/or land-use changes) are provided to *N-SDM*, the prediction and mapping processes are repeated. To account for potential observational biases in occurrence data (Fithian et al. 2015, Robinson et al. 2018, Isaac et al. 2020), *N-SDM* is equipped with an option that allows setting a constant value (e.g. zero or median) to the set of specified covariates expected to be related to these biases, for example distance from the transportation network or human population density. By doing so, the observer bias is assumed to be removed from the prediction area (Warton et al. 2013, Bonnet-Lebrun et al. 2020, Chauvier et al. 2021).

Highlighted features

In addition to being the first end-to-end SDM platform explicitly designed for HPC environments, *N-SDM* also pioneers in integrating within a single platform: 1) a spatially-nested modelling framework, 2) an 'embedded' covariate selection procedure, 3) the ESM approach, and 4) an hyperparameter grid-search strategy.

Spatially-nested modelling framework

An increasing number of ecological records has recently become available to the scientific community, particularly in connection with the rise of citizen science projects (Dickinson et al. 2010, Amano et al. 2016, Pocock et al. 2017). These sources have proven useful in resolving spatial gaps in scientific surveys and tackling several modelling challenges, such as niche truncation (Titeux et al. 2017, Mateo et al. 2019a, Chevalier et al. 2021). The niche truncation issue is particularly relevant for regional and regional study areas, for which it is likely that target species can also be found in a wider range of bioclimatic conditions. Consequently, models fitted on the basis of regional occurrences only are likely to result in truncated estimates of the species' response to environmental covariates, failing at providing accurate predictions for areas or periods with conditions outside the range of those used for model calibration (Barbet-Massin et al. 2010, Sanchez-Fernandez et al. 2011, Scherrer et al. 2021). A data-oriented solution is to use additional species occurrences from outside the regional study area, with the objective of covering as much as possible of the full species niche (Hannemann et al. 2016, Fernandes et al. 2019, Chevalier et al. 2021).

To integrate data arising from multiple sources into a single model, recent studies have developed hierarchical frameworks,

most of which having been built around GLM and estimated in a Bayesian context (Fletcher et al. 2019, Isaac et al. 2020, Adde et al. 2021). However, being restricted to GLM, these methods are not compatible with the ensemble modelling strategy used in *N-SDM*. Moreover, the computational costs of the Bayesian approach make it inappropriate for modelling thousands of taxa. Based on more straightforward methods, the spatially-nested modelling framework used in *N-SDM* consists in combining global and regional level model outputs using the specified nesting strategy. At the time of its release, two nesting strategies are provided with *N-SDM*: 'covariate' and 'multiply'. The 'covariate' strategy consists in using the global model output as an additional covariate for fitting the regional model (Mateo et al. 2019a, Bellamy et al. 2020). The global model output covariate cannot be discarded during the covariate selection step. This strategy allows to directly provide the regional model with larger-scale information on the conditions that are favorable, or not, for the species. The 'multiply' strategy is calculating the geometric mean of the two model outputs (Fournier et al. 2017, Mateo et al. 2019b) to produce a consensual indicator of the habitat suitability values obtained at the two levels. The main advantage of the geometric mean compared to the arithmetic mean is that a very low habitat suitability value at either one of the two levels will be directly reflected in the combined indicator.

Embedded covariate selection procedure

With Earth observation data made available at an unprecedented rate, species distribution modelers are increasingly challenged by high-dimensional spaces of candidate covariates to define realistic niches (Kuenzer et al. 2014, Soille et al. 2018, Sudmanns et al. 2020). For optimizing the predictive abilities and parsimony of the models fitted by *N-SDM*, we devised an innovative 'embedded' covariate selection method developed around three main algorithms: GLM (McCullagh and Nelder 1989), GAM (Hastie 2017), and RF (Breiman 2001). These algorithms are among the most used in SDM studies (Hao et al. 2019) and are covering a gradient of flexibility and fitting methods that makes the ensemble of their results generalizable to many modelling frameworks. Thus, the covariate subset selected after applying the embedded procedure can also be used for fitting other popular SDM algorithms, such as Maxent (Phillips et al. 2006) or Gradient Boosting (Elith et al. 2008), even if they are not directly related to any of the three target algorithms.

The embedded covariate selection consists of two main steps: Step A 'Collinearity filtering', and Step B 'Model-specific embedding'. In 1) we reduce the dimensionality of the candidate set by eliminating the less informative covariates among collinear pairs, based on correlation matrices and univariate GLM p-values. In 2), selected covariates are used to fit models with embedded selection procedures. Specifically, we use GLM with elastic-net regularization (Zou and Hastie 2005), GAM with null-space penalization (Marra and Wood 2011), and guided regularized RF (Deng and Runger 2013). For each algorithm, the n covariates retained after regularization are ranked from 1 ('best') to n ('worst'). The algorithm-specific ranking is done based on the absolute value of the regularized regression coefficients for GLM, the chi-square statistic for GAM, and the Mean Decrease Gini index for RF, all to be maximized. The final overall ranking is obtained by ordering the sum of the three ranks for each covariate, starting with the covariates that were commonly selected by all the three models, and then adding the remaining. The top k covariates were selected as the final modelling set, with k being a user-specifiable setting and ceiling (log2(number of occurrences)) the default value. The covariate selection is applied individually to each species at both the global and regional levels. Additional details on the covariate selection method and evaluation results are provided in the Supporting information.

Ensemble of small models (ESM)

Former studies have shown that an increase in the number of species occurrences tended to be associated with an increase in model accuracy (Guisan et al. 2007, van Proosdij et al. 2016, Fernandes et al. 2018). Among the main reasons proposed are that species niches are less well covered with smaller sample sizes, while the risk of model overfitting increases, especially when many covariates are considered. The ESM method (Breiner et al. 2015, Breiner et al. 2018) has been developed to address the challenge of modelling species with small sample sizes, as it is often the case for rare and threatened organisms. This method consists in fitting every possible combinations of two covariates (i.e. fitting $n(n-1)/2$ models, with n being the number of covariates) and averaging individual predictions from selected bivariate models into an ensemble. Results from Breiner et al. (2015) showed that predictive skills and transferability of ESMs were higher than those obtained under a more standard SDM approach, especially for the rarest species. More specifically, Breiner et al. showed in a second paper that ensembling different algorithms was not improving the accuracy of the ESM predictions (likely because the ESM is already itself an ensembling approach, even when using a single algorithm) and that the simplest algorithms were advantageous in terms of both model performance and computation time (Breiner et al. 2018). This was notably the case for GLM, which is currently the sole algorithm available for ESM fitting in *N-SDM*, but others could be added on request if justified. Three main settings need to be defined for running the ESM approach with *N-SDM*: the number of occurrences below which ESMs are run, the total number of covariates evaluated, and the threshold value for the specified evaluation metric below which a bivariate model is discarded from the final ensemble.

Hyperparameter grid search

Each modelling algorithm comes with a specific set of hyperparameters, whose optimal values depend on the characteristics of the data to be modelled (Chicco 2017, Vignali et al. 2020, Valavi et al. 2022). The fine tuning of these hyperparameters is decisive for maximizing the predictive value of the model

and avoiding overfitting. However, when working with multiple algorithms and species at the same time, manually setting the optimal combination of hyperparameter values is unrealistic. In *N-SDM*, an exhaustive automatic grid-search strategy aimed at finding the best combination of hyperparameter values is used. The grid-search procedure starts by generating all the possible combinations of candidate hyperparameter values. Each combination is evaluated and ranked using the specified cross-validated assessment metric and the best setting is kept for fitting the final model. More details on available model assessment metrics can be found in the Supporting information. The hyperparameter tuning grid is provided in a form of an editable 'param-grid.csv' where default values can easily be modified (Supporting information). The list of algorithm-specific hyperparameters that are tunable in *N-SDM* along with their default values is provided in the Supporting information.

Installing and running *N-SDM*

N-SDM (ver. 1.0.0) can be downloaded from the GitHub repository https://github.com/N-SDM/N-SDM, where complementary instructions for installation and example data are provided. Prerequisites for running *N-SDM* include 1) Linux cluster computer equipped with the Slurm workload manager, 2) availability of the modules (with versions used for *N-SDM* development) gcc (9.3.0), r (4.0.5), proj (5.2.0), perl (5.32.1), curl (7.76.1), geos (3.8.1) and gdal (2.4.4), 3) a clone of the N-SDM/N-SDM GitHub repository in the working directory, and 4) an installation of the 'nsdm' R package. Once these prerequisites are satisfied, and the user-specific settings of the 106 editable options in the 'settings.csv' table (Supporting information) are defined, *N-SDM* can be run by executing the main *N-SDM* script 'nsdm.sh'. From this script, 14 encapsulated R scripts are automatically run. The Supporting information provides an overview of these 14 R scripts, along with the relative distribution of computational resources (memory and time) needed for each of them. The list of outputs generated during an *N-SDM* run is provided in the Supporting information.

Applied example

We ran an example aimed at illustrating the main operations and performances of *N-SDM* by modelling the habitat suitability of 1500 plant and vertebrate species (Supporting information) at 25 m resolution across Switzerland and projecting it to future conditions. An extended version of the 'Applied example' section, with complementary details and analyses, is provided in the Supporting information.

Study areas

Following the spatially-nested framework of *N-SDM*, we distinguished between regional- and global-level study areas. The regional-level area included all of Switzerland. For the global-level area, we used a bounding box covering the European continent. The extent of the global level was chosen to accommodate most of the 1500 species modelled in this example. However, for species with narrow distributions (e.g. restricted to the Alps), this extent could have been reduced for optimizing model performances. For this example, global and regional levels were combined by using the 'covariate' nesting strategy.

Data and methods

Species occurrence data

The species occurrence records used for the regional-level model were provided by the Swiss Species Information Center InfoSpecies (www.infospecies.ch) on 23 August 2021. For the global level model, occurrence records were obtained from GBIF (www.gbif.org) on 27 October 2021 (https://doi.org/10.15468/dl.zwp3dx). The minimum (maximum) number of occurrence available per species at global and regional levels were 65 (750 038) and 50 (109 211), respectively. For each species and level, 10 000 background absences were randomly generated across the target areas.

Covariate data

We applied the embedded covariate selection procedure described in the 'Highlighted features' section on a suite of 1508 candidate covariates derived from 109 individual variables and belonging to eight main categories (bioclimatic, edaphic, hydrologic, land use and cover, population, topographic, transportation, and vegetation) (Supporting information). Only bioclimatic covariates (n = 19) were used for fitting global-level models and only habitat covariates (n = 1489 covariates) were used at the regional level. Since the 'covariate' nesting strategy was used, the global-model output was forced as an additional covariate in all regional models.

Model fitting and assessment

GLM, GAM, MAX, RF, and GBM models were fitted using their default values for hyperparameter tuning (Supporting information). Given the number of occurrence records per species, no ESMs were run for this example application. Hyperparameter selection and model accuracy were evaluated using the average 'Score' of the AUC′, maxTSS, and CBI values obtained through a split-sample procedure repeated 100 times with 30% of the data kept for validation. Because of large differences in the sample size or spatial coverage of species occurrence data, no between-species analyses of model accuracy were done.

Spatial predictions

Ensemble predictions were mapped over a 25 m resolution grid covering Switzerland. Climate projections derived from the representative concentration pathways (RCPs) RCP4.5 ('Low Carbon') and RCP8.5 ('High Carbon') (van Vuuren et al. 2011) were used for projecting habitat suitability by the end of the century (2070–2100). For illustrative purposes, we reported maps obtained for eight example species sampled from the taxonomical group represented in the set of 1500 species.

Running time and memory usage

We reported the average *N-SDM* run time and maximum memory usage requirements for 100 species, applying the parallelization approach described in the Supporting information and using a 10-core central processing unit strategy with AMD® EPYC 7402 on the University of Lausanne HPC cluster.

Results

N-SDM was successfully run for all 1500 species. The average run time for 100 species was ≈ 7 h and the maximum memory usage ≈ 250 GB.

Covariate selection

Out of the 1489 habitat covariates available at the regional level, 833 were selected in at least one of the 1500 models. Among the eight main categories of environmental covariates, 'hydrologic' was the most often selected relative to the overall number of covariates available in this category, whereas 'land use and cover' was the least (Fig. 2A).

Model fitting

All the models obtained for the 1500 species and five algorithms had a median Score value above 0.85 (Fig. 2B), indicating high predictive performances. GLM was the algorithm with the lowest median and the one with the highest interquartile range. On the other hand, GBM was the algorithm with the highest median.

Spatial predictions

Figure 2C shows the spatial predictions obtained for eight example species. Projections for future periods obtained

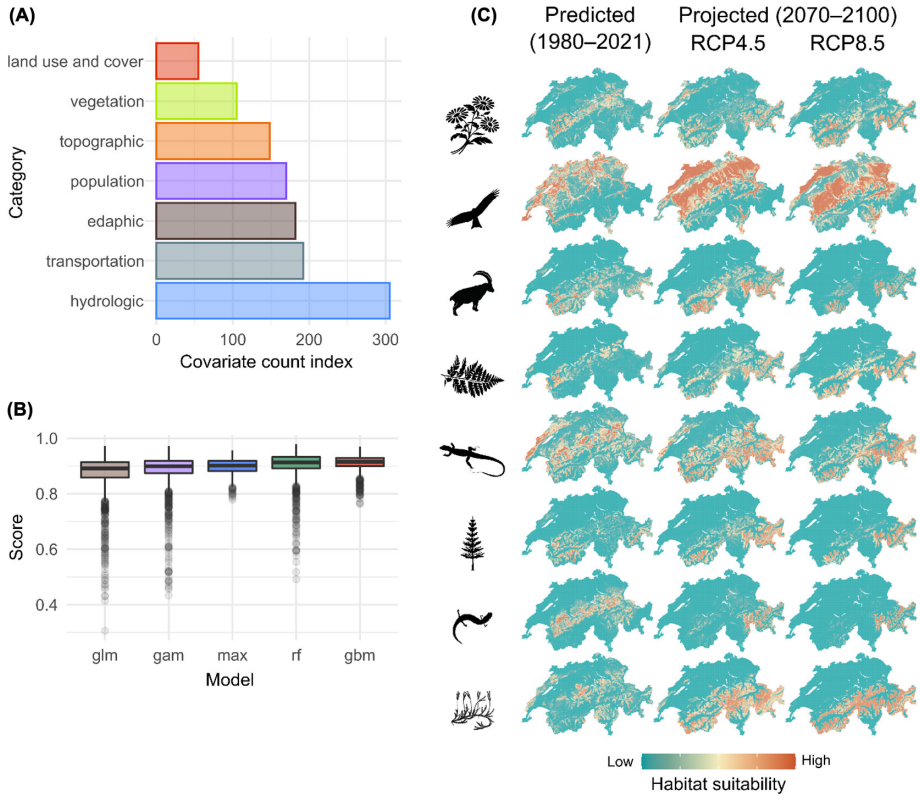

Figure 2. Applied *N-SDM* example results. (A) Covariate categories representation in the 1,500 regional-level models. The covariate count index for a category was equal to the count of covariates from this category selected in the final models divided by the overall number of covariates available in this category. For readability, the global model output covariate being the only representative of his category and forced into all models (covariate count index of 1,500) was excluded from this analysis. (B) Cross-validated model "Score" values (average of Somers' D, maximized True Skill Statistic, and Continuous Boyce Index) for the 1,500 species and five algorithms (Generalized Linear Model, Generalized Additive Model, Maxnet, Random Forest, light Gradient Boosted Machine). For each boxplot, the central box represents the 1st quartile, the median, and the 3rd quartile. The two whiskers extend to the furthest non-outlier points. (C) Habitat suitability for eight species sampled from the groups represented in the set of 1,500 species, with from top to bottom *Primula auricula* (Angiospermae), *Milvus migrans* (Aves), *Capra ibex* (Mammalia), *Dryopteris villarii* (Pteridophyta), *Zootoca vivipara* (Reptilia), *Pinus cembra* (Pinophyta), *Salamandra atra* (Amphibia), and *Lycopodium annotinum* (Lycopodiophyta). Left column: predicted habitat suitability for the current period (1980–2021). Middle column: projected habitat suitability for 2070–2100 under the "low carbon" (RCP4.5) scenario. Right column: projected habitat suitability for 2070–2100 under the "high carbon" (RCP8.5) scenario.

under the two RCP scenarios indicated extensive changes in habitat suitability patterns, with some species likely to have their areas of high habitat suitability values increased (e.g. *Milvus migrans*), while for others these areas could be considerably reduced (e.g. *Salamandra atra*). The magnitude of changes, either in terms of elevational, latitudinal, or longitudinal shifts, was exacerbated under RCP8.5.

Conclusions

By uniting leading-edge SDM methods into one HPC modelling pipeline, *N-SDM* facilitates the delivery of scalable and reproducible outputs for standard biodiversity assessments. Among the set of features that make *N-SDM* a powerful SDM platform, three attributes give it a unique and innovative character: 1) the spatially-nested framework, 2) the HPC design, and 3) its customizability and collaborative nature.

1. *N-SDM* is an end-to-end SDM platform facilitating the use of species occurrence data retrieved from multiple sources and at different scales, which is useful for addressing the issue of niche truncation. This was achieved by building *N-SDM* around a spatially-nested framework allowing for the combination of two models fitted with their respective global- and regional-level data. The two methods currently available in *N-SDM* for combining the two levels are inspired from the existing literature and are straightforward enough to be compatible with the exigence of the ensemble modelling context, which involves working with several algorithms. Moreover, these methods are fast enough for delivering results for possibly thousands of species in a reasonable amount of time. This potential massive production of SDMs is backed by the integration of the latest available methodological standards and reporting tools, which should help to enhance the quality of *N-SDM* outputs, the capacity to critically review model setups, and to assess the potential of the results for policy planning. Nevertheless, it comes with the risk of yielding misleading predictions if not conducted with appropriate data and in close partnership with experts on the modelled species to help build and validate the prediction maps.
2. *N-SDM* fills the gap in the availability of SDM platforms specifically designed for HPC environments, which are crucial for handling big data contexts in relation with the increasing amount of species occurrence data and environmental layers for modelling them. Moreover, to aim for completeness, biodiversity assessments need to consider and model as many species as possible, which is difficult to achieve with the computational frameworks of existing SDM platforms. By specifically targeting HPC environments and incorporating cutting-edge parallelization strategies, *N-SDM* directly meets these requirements. Through an applied example, we demonstrated the abilities of *N-SDM* to fit models with high predictive values and to deliver high resolution maps in a relatively short period of time and with comparably low memory usage, given the large number of species, occurrences, and candidate covariates considered.
3. Provided with a set of 106 tunable parameters, *N-SDM* has been designed for high customization flexibility, so it can be adapted to anyone's objectives and computing environment. This first *N-SDM* version (ver. 1.0.0) was developed in and for Linux cluster computers equipped with the Slurm workload manager, because they are among the most common resources used for HPC. In addition, *N-SDM* has been designed to incorporate inputs from both the computational modelling community and species experts whose guidance can be particularly useful at the crucial stage of pre-selecting covariates that are relevant for the distribution of target organisms. On the modelling community side, *N-SDM* is hosted on a public GitHub repository where anyone interested in contributing to its improvement is invited to suggest optimizations, or to create new features that could be added to the pipeline. *N-SDM* is intended to be improved and updated at regular intervals. Anticipated improvements could include the addition of background-absences generation techniques, cross-validation procedures (e.g. spatial blocking), and modelling algorithms or algorithm versions.

To cite *N-SDM* or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 1.0':

Adde, A. et al. 2023. *N-SDM*: a high-performance computing pipeline for Nested Species Distribution Modelling. – Ecography 2023: e06272 (ver. 1.0).

Acknowledgements – The Swiss Species Information Center InfoSpecies (www.infospecies.ch) supplied Swiss-level species occurrence data and expertise on species' ecology, and we acknowledge their support regarding the database. This research was enabled in part by the support provided by the Scientific Computing and Research Unit of Lausanne University (www.unil.ch/ci/dcsr).
Funding – We gratefully acknowledge financial support through the Action Plan of the Swiss Biodiversity Strategy by the Federal Office for the Environment (FOEN) for financing the ValPar.CH and SwissCatchment projects.

Author contributions

Antoine Adde: Conceptualization (equal); Data curation(lead); Methodology (lead); Software (lead); Validation (lead); Writing – original draft (lead); Writing – review and editing (lead). **Pierre-Louis Rey**: Conceptualization-Supporting, Data curation (equal); Methodology (equal); Software-Supporting, Validation-Supporting, Writing – original draft-Supporting, Writing – review and editing (equal). **Philipp Brun**: Methodology-Supporting, Software (equal); Validation (equal); Writing – review and editing (equal). **Nathan Külling**: Data curation (equal); Software-Supporting, Validation (equal); Writing – review and editing (equal). **Fabian Fopp**: Conceptualization-Supporting, Data curation (equal); Methodology-Supporting, Software-Supporting, Validation-Supporting, Writing

– review and editing (equal). **Florian Altermatt**: Conceptualization-Supporting, Methodology-Supporting, Validation-Supporting, Writing – review and editing (equal). **Olivier Broennimann**: Conceptualization-Supporting, Data curation-Supporting, Methodology (equal); Validation-Supporting, Writing – review and editing (equal). **Anthony Lehmann**: Conceptualization-Supporting, Data curation-Supporting, Methodology-Supporting, Validation-Supporting, Writing – review and editing (equal). **Blaise Petitpierre**: Conceptualization-Supporting, Methodology (equal); Validation-Supporting, Writing – review and editing (equal). **Niklaus E. Zimmermann**: Conceptualization-Supporting, Methodology-Supporting, Validation-Supporting, Writing – review and editing (equal). **Loic Pellissier**: Conceptualization (equal); Data curation-Supporting, Methodology-Supporting, Supervision (equal); Validation-Supporting, Writing – review and editing (equal). **Antoine Guisan**: Conceptualization (lead); Data curation-Supporting, Funding acquisition (lead); Methodology (equal); Supervision (lead); Validation (equal); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://publons.com/publon/10.1111/ecog.0XXXX.

Data availability statement

N-SDM is available from the GitHub Repository, https://github.com/N-SDM/N-SDM, where complementary instructions for installation and example data are provided. Current and previous *N-SDM* releases can be downloaded from https://doi.org/10.5281/zenodo.7659995.

Supporting information

The Supporting information associated with this article is available with the online version.

References

Adde, A., Casabona, C., Amat, I., Mazerolle, M. J., Darveau, M., Cumming, S. G. and O'Hara, R. B. 2021. Integrated modeling of waterfowl distribution in western Canada using aerial survey and citizen science (eBird) data. – Ecosphere 12: 20.

Allouche, O., Tsoar, A. and Kadmon, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Amano, T., Lamming, J. D. L. and Sutherland, W. J. 2016. Spatial gaps in global biodiversity information and the role of citizen science. – BioScience 66: 393–400.

Araujo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Araujo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E. and Rahbek, C. 2019. Standards fordistribution models in biodiversity assessments. – Sci. Adv. 5: eaat4858.

Barbet-Massin, M., Thuiller, W. and Jiguet, F. 2010. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? – Ecography 33: 878–886.

Bellamy, C., Boughey, K., Hawkins, C., Reveley, S., Spake, R., Williams, C. and Altringham, J. 2020. A sequential multi-level framework to improve habitat suitability modelling. – Landsc. Ecol. 35: 1001–1020.

Bonnet-Lebrun, A. S., Karamanlidis, A. A., Hernando, M. D., Renner, I. and Gimenez, O. 2020. Identifying priority conservation areas for a recovering brown bear population in Greece using citizen science data. – Anim. Conserv. 23: 83–93.

Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.

Breiner, F. T., Guisan, A., Bergamini, A. and Nobis, M. P. 2015. Overcoming limitations of modelling rare species by using ensembles of small models. – Methods Ecol. Evol. 6: 1210–1218.

Breiner, F. T., Nobis, M. P., Bergamini, A. and Guisan, A. 2018. Optimizing ensembles of small models for predicting the distribution of species with few occurrences. – Methods Ecol. Evol. 9: 802–808.

Chauvier, Y., Zimmermann, N. E., Poggiato, G., Bystrova, D., Brun, P. and Thuiller, W. 2021. Novel methods to correct for observer and sampling bias in presence-only species distribution models. – Global Ecol. Biogeogr. 30: 2312–2325.

Chevalier, M., Mod, H., Broennimann, O., Di Cola, V., Schmid, S., Niculita-Hirzel, H., Pradervand, J.-N., Schmidt, B. R., Ursenbacher, S., Pellissier, L. and Guisan, A. 2021. Low spatial autocorrelation in mountain biodiversity data and model residuals. – Ecosphere 12: e03403.

Chicco, D. 2017. Ten quick tips for machine learning in computational biology. – Biodata Mining 10.

Cobos, M. E., Peterson, A. T., Osorio-Olvera, L. and Jimenez-Garcia, D. 2019. An exhaustive analysis of heuristic methods for variable selection in ecological niche modeling and species distribution modeling. – Ecol. Informat. 53: 100983.

Deng, H. and Runger, G. 2013. Gene selection with guided regularized random forest. – Pattern Recogn. 46: 3483–3489.

Dickinson,J. L., Zuckerberg,B. and Bonter,D. N. 2010. Citizen-science as an ecological research tool: challenges and benefits. – Annu. Rev. Ecol. Evol. Syst. 41: 149–172.

Elith, J., Leathwick, J. R. and Hastie, T. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.

Enquist, B. J., Feng, X., Boyle, B., Maitner, B., Newman, E. A., Jorgensen, P. M., Roehrdanz, P. R., Thiers, B. M., Burger, J. R., Corlett, R. T., Couvreur, T. L. P., Dauby, G., Donoghue, J. C., Foden, W., Lovett, J. C., Marquet, P. A., Merow, C., Midgley, G., Morueta-Holme, N., Neves, D. M., Oliveira-Filho, A. T., Kraft, N. J. B., Park, D. S., Peet, R. K., Pillet, M., Serra-Diaz, J. M., Sandel, B., Schildhauer, M., Simova, I., Violle, C., Wieringa, J. J., Wiser, S. K., Hannah, L., Svenning, J. C. and McGill, B. J. 2019. The commonness of rarity: global and future distribution of rarity across landplants. – Sci. Adv. 5: eaaz0414.

Fernandes, R. F., Scherrer, D. and Guisan, A. 2018. How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach. – Ecol. Inform. 48: 125–134.

Fernandes, R. F., Honrado, J. P., Guisan, A., Roxo, A., Alves, P., Martins, J. and Vicente, J. R. 2019. Species distribution models support the need of international cooperation towards suc-

cessful management of plant invasions. – J. Nat. Conserv. 49: 85–94.
Ferrier, S., Ninan, K. N., Leadley, P., Alkemade, R., Acosta, L. A., Akçakaya, H. R., Brotons, L., Cheung, W. W. L., Christensen, V., Harhash, K. A., Kabubo-Mariara, J., Lundquist, C., Obersteiner, M., Pereira, H. M., Peterson, G., Pichs-Madruga, R., Ravindranath, N., Rondinini, C. and Wintle, B. A. 2016. IPBES: the methodological assessment report on scenarios and models of biodiversity and ecosystem services. – Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES).
Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.
Fithian, W., Elith, J., Hastie, T. and Keith, D. A. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.
Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A. and Dorazio, R. M. 2019. A practical guide for combining data to model species distributions. – Ecology 100: e02710.
Fourcade, Y., Besnard, A. G. and Secondi, J. 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. – Global Ecol. Biogeogr. 27: 245–256.
Fournier, A., Barbet-Massin, M., Rome, Q. and Courchamp, F. 2017. Predicting species distribution combining multi-scale drivers. – Global Ecol. Conserv. 12: 215–226.
Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. – Cambridge Univ. Press.
Galante, P. J., Alade, B., Muscarella, R., Jansa, S. A., Goodman, S. M. and Anderson, R. P. 2017. The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. – Ecography 41: 726–736.
Gallien, L., Douzet, R., Pratte, S., Zimmermann, N. E. and Thuiller, W. 2012. Invasive species distribution models - How violating the equilibrium assumption can create new insights. – Global Ecol. Biogeogr. 21: 1126–1136.
Guisan, A., Graham, C. H., Elith, J. and Huettmann, F. 2007. Sensitivity of predictive species distribution models to change in grain size. – Divers. Distrib. 13: 332–340.
Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P. and Buckley, Y. M. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.
Guisan, A., Thuiller, W. and Zimmermann, N. E. 2017. Habitat suitability and distribution models, with applications in R. – Cambridge Univ. Press.
Hannemann, H., Willis, K. J. and Macias-Fauria, M. 2016. The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. – Global Ecol. Biogeogr. 25: 26–35.
Hao, T. X., Elith, J., Guillera-Arroita, G. and Lahoz-Monfort, J. J. 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. – Divers. Distrib. 25: 839–852.
Hastie, T. J. 2017. Generalized additive models. – In: Hastie, T. J. (ed.), Statistical models in S. Routledge, pp. 249–307.
Hernandez, P. A., Graham, C. H., Master, L. L. and Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.
Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J. and Hijmans, M. R. J. 2017. Package 'dismo'. – Circles 9: 1–68.
Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C. and Guisan, A. 2006. Evaluating the ability of habitat suitability models to predict species presences. – Ecol. Modell. 199: 142–152.
Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G. and O'Hara, R. B. 2020. Data integration for large-scale models of species distributions. – Trends Ecol. Evol. 35: 56–67.
Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S. and Turak, E. 2019. Essential biodiversity variables for mapping and monitoring species populations. – Nat. Ecol. Evol. 3: 539–551.
Kappa, C. 1960. Coefficient of concordance. – Educ. Psychol. Meas. 20: 37–46.
Kass, J. M., Vilela, B., Aiello-Lammens, M. E., Muscarella, R., Merow, C. and Anderson, R. P. 2018. WALLACE: a flexible platform for reproducible modeling of species niches and distributions built for community expansion. – Methods Ecol. Evol. 9: 1151–1156.
Kays, R., McShea, W. J. and Wikelski, M. 2020. Born-digital biodiversity data: millions and billions. – Divers. Distrib. 26: 644–648.
Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., Ye, Q. W. and Liu, T. Y. 2017. LightGBM: a highlyefficient gradient boosting decision tree. – Adv. Neural Inform. Process. Syst. 30: 3146–3154.
Kuenzer, C., Ottinger, M., Wegmann, M., Guo, H., Wang, C., Zhang, J., Dech, S. and Wikelski, M. 2014. Earth observation satellite sensors for biodiversity monitoring: potentials and bottlenecks. – Inter. J. Remote Sens. 35: 6599–6647.
Kuhn, M. and Johnson, K. 2013. Applied predictive modeling. – Springer.
Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J. and Guisan, A. 2010. Overcoming the rare species modelling paradox: a novel hierarchical framework applied to an Iberian endemic plant. – Biol. Conserv. 143: 2647–2657.
Marra, G. and Wood, S. N. 2011. Practical variable selection for generalized additive models. – Comput. Stat. Data Anal. 55: 2372–2387.
Mateo, R. G., Aroca-Fernández, M. J., Gastón, A., Gómez-Rubio, V., Saura, S. and García-Viñas, J. I. 2019a. Looking for an optimal hierarchical approach for ecologically meaningful nichemodelling. – Ecol. Modell. 409: 108735.
Mateo, R. G., Gaston, A., Aroca-Fernandez, M. J., Broennimann, O., Guisan, A., Saura, S. and Garcia-Vinas, J. I. 2019b. Hierarchical species distribution models in support of vegetation conservation at the landscape scale. – J. Veg. Sci. 30: 386–396.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. 2nd edition. – Chapman and Hall.
Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M. and Anderson, R. P. 2014. ENMeval: an R package for conducting spatially independent evaluations and

estimating optimal model complexity for MAXENT ecological niche models. – Methods Ecol. Evol. 5: 1198–1205.
Pacifici, K., Reich, B. J., Miller, D. A. W. and Pease, B. S. 2019. Resolving misaligned spatial data with integrated species distribution models. – Ecology 100: e02709.
Pagel, J., Anderson, B. J., O'Hara, R. B., Cramer, W., Fox, R., Jeltsch, F., Roy, D. B., Thomas, C. D. and Schurr, F. M. 2014. Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. – Methods Ecol. Evol. 5: 751–760.
Pearson, D., Dawson, T. P. and Liu, C. 2004. Modelling species distribution in Britain: a hierarchical integration of climate and land-cover data. – Ecography 27: 285–298.
Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R., Martínez-Meyer, E., Nakamura, M. and Araújo, M. P. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.
Phillips, S. J., Anderson, R. P. and Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Modell. 190: 231–259.
Phillips, S. J., Anderson, R. P., Dudik, M., Schapire, R. E. and Blair, M. E. 2017. Opening the black box: an open-source release of Maxent. – Ecography 40: 887–893.
Pocock, M. J., Tweddle, J. C., Savage, J., Robinson, L. D. and Roy, H. E. 2017. The diversity and evolution of ecological and environmental citizen science. – PLoS One 12: e0172579.
Robinson, O. J., Ruiz-Gutierrez, V. and Fink, D. 2018. Correcting for bias in distribution modelling for rare species using citizen science data. – Divers. Distrib. 24: 460–472.
Sanchez-Fernandez, D., Lobo, J. M. and Hernandez-Manrique, O. L. 2011. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. – Divers. Distrib. 17: 163–171.
Scherrer, D., Esperon-Rodriguez, M., Beaumont, L. J., Barradas, V. L. and Guisan, A. 2021. National assessments of species vulnerability to climate change strongly depend on selected data sources. – Divers. Distrib. 27: 1367–1382.
Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V. and Vasilev, V. 2018. A versatile data-intensive computing platform for information retrieval from big geospatial data. – Fut. Gen. Comp. Syst. 81: 30–40.
Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. – Am. Sociol. Rev. 799–811.
Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A. and Blaschke, T. 2020. Big Earth data: disruptive changes in Earth observation data management and analysis? – Int. J. Dig. Earth. 13: 832–850.
Thuiller, W., Lafourcade, B., Engler, R. and Araujo, M. B. 2009. BIOMOD – A platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.
Titeux, N., Maes, D., Van Daele, T., Onkelinx, T., Heikkinen, R. K., Romo, H., Garcia-Barros, E., Munguira, M. L., Thuiller, W., van Swaay, C. A. M., Schweiger, O., Settele, J., Harpke, A., Wiemers, M., Brotons, L. and Luoto, M. 2017. The need for large-scale distribution data to estimate regional changes in species richness under future climate change. – Divers. Distrib. 23: 1393–1407.
Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J. and Elith, J. 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. – Ecol. Monogr. 92: e01486.
van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J. and Raes, N. 2016. Minimum required number of specimen records to develop accurate species distribution models. – Ecography 39: 542–552.
van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V. and Lamarque, J.-F. 2011. The representative concentration pathways: an overview. – Clim. Change. 109: 5–31.
Vignali, S., Barras, A. G., Arlettaz, R. and Braunisch, V. 2020. SDMtune: an R package to tune and evaluate species distribution models. – Ecol. Evol. 10: 11488–11506.
Warton, D. I., Renner, I. W. and Ramp, D. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. – PLoS One 8: e79168.
Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. – J. R. Stat. Soc. B. 67: 301–320.