frontiers | Frontiers in Ecology and Evolution

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

OPEN ACCESS

EDITED BY
Stelios Katsanevakis,
University of the Aegean, Greece

REVIEWED BY
Christophe Botella,
Institut National de Recherche en Informatique et en Automatique (INRIA), France
Bharat Babu Shrestha,
Tribhuvan University, Nepal

*CORRESPONDENCE
Amy J. S. Davis
amy.davis@uni-konstanz.de

†PRESENT ADDRESS
Rozemien De Troch,
Belgian Climate Centre, Brussels, Belgium

RECEIVED 20 January 2023
ACCEPTED 17 January 2024
PUBLISHED 09 February 2024

CITATION
Davis AJS, Groom Q, Adriaens T, Vanderhoeven S, De Troch R, Oldoni D, Desmet P, Reyserhove L, Lens L and Strubbe D (2024) Reproducible WiSDM: a workflow for reproducible invasive alien species risk maps under climate change scenarios using standardized open data. *Front. Ecol. Evol.* 12:1148895. doi: 10.3389/fevo.2024.1148895

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]

[1]Terrestrial Ecology Unit TEREC, Department of Biology, Ghent University, Gent, Belgium, [2]Ecology, Department of Biology, University of Konstanz, Konstanz, Germany, [3]Research Institute for Nature and Forest (INBO), Brussels, Belgium, [4]Belgian Biodiversity Platform, Département du Milieu Naturel et Agricole, Service Public de Wallonie, Gembloux, Belgium, [5]Royal Meteorological Institute of Belgium, Brussels, Belgium, [6]Meise Botanic Garden, Meise, Belgium

Introduction: Species distribution models (SDMs) are often used to produce risk maps to guide conservation management and decision-making with regard to invasive alien species (IAS). However, gathering and harmonizing the required species occurrence and other spatial data, as well as identifying and coding a robust modeling framework for reproducible SDMs, requires expertise in both ecological data science and statistics.

Methods: We developed WiSDM, a semi-automated workflow to democratize the creation of open, reproducible, transparent, invasive alien species risk maps. To facilitate the production of IAS risk maps using WiSDM, we harmonized and openly published climate and land cover data to a 1 km^2 resolution with coverage for Europe. Our workflow mitigates spatial sampling bias, identifies highly correlated predictors, creates ensemble models to predict risk, and quantifies spatial autocorrelation. In addition, we present a novel application for assessing the transferability of the model by quantifying and visualizing the confidence of its predictions. All modeling steps, parameters, evaluation statistics, and other outputs are also automatically generated and are saved in a R markdown notebook file.

Results: Our workflow requires minimal input from the user to generate reproducible maps at 1 km^2 resolution for standard Intergovernmental Panel on Climate Change (IPCC) greenhouse gas emission representative concentration pathway (RCP) scenarios. The confidence associated with the predicted risk for each 1km^2 pixel is also mapped, enabling the intuitive visualization and understanding of how the confidence of the model varies across space and RCP scenarios.

Discussion: Our workflow can readily be applied by end users with a basic knowledge of R, does not require expertise in species distribution modeling, and only requires an understanding of the ecological theory underlying species distributions. The risk maps generated by our repeatable workflow can be used to support IAS risk assessment and surveillance.

KEYWORDS

uncertainty in SDMs, conformal prediction, spatial sampling bias, ecological models, confidence assessment, invasive alien species

1 Introduction

Climate change and biological invasions represent two of the largest threats to biodiversity in the Anthropocene (Mazor et al., 2018; Urban, 2015). As a result of climate change, it is expected that a wide range of species will migrate to follow their shifting climatic niche and introduced species will find novel areas suitable for their establishment (Bellard et al., 2018). Some of these introduced species are likely to have negative impacts on native biodiversity and human well-being (Simberloff et al., 2013). Assessing the risk of invasion by alien species is a crucial step for proactive management, including identifying species for preventive actions such as legal bans on trade, transport, and possession, targeting early detection efforts both at entry points and in susceptible ecosystems, as well as risk management decisions to remove established populations or limit their further spread (Srivastava et al., 2019). Regardless of the specific protocol used, risk assessment is defined as the standardized evaluation of entry, exposure, and consequence of the introduction of an alien species (Vanderhoeven et al., 2017; González-Moreno et al., 2019). An evaluation of the risks of introduction, establishment, spread, and impact are the four main components of alien species risk assessments (Roy et al., 2017).

Species distribution models (SDMs) are the main tool for forecasting the risk of establishment of an alien species in a spatially explicit way (Guisan and Thuiller, 2005; Jeschke and Strayer, 2008). Correlative SDMs delineate the realized niche of the organism based on species-environmental relationships obtained from georeferenced species occurrence data (i.e., species presence located at specific geographic coordinates) and spatial environmental predictors. This way, SDMs predict the probability of species presence in unsampled areas. Additionally, SDMs also predict environmentally suitable areas where the species is currently absent, but can potentially be established in the future, depending on dispersal success. SDMs can be used to guide spatial decision-making, but recent critiques have highlighted how uncertainties in species distribution modeling practice have hindered their widespread uptake in decision-making workflows (Muscatello et al., 2021; Lee-Yaw et al., 2022; Nguyen and Leung, 2022; Liu et al., 2020). These issues include the impact of methodological choices on model outcomes including accuracy, ease of interpretation, and predicted distribution (Wenger et al., 2013; Sofaer et al., 2019; Brun et al., 2020). For example, algorithm choice is a major source of variability in model forecasts (Elith et al., 2006; Hallgren et al., 2019). Also, the choice of predictors, parameter settings, and spatial grain are all sources of variability that affect model predictions (Peterson et al., 2018; Fourcade, 2021; Chauvier et al., 2022). In addition to the uncertainty in model predictions stemming from the numerous choices to be made during model development, the failure to record and share these decisions prevents reproducibility (Feng et al., 2019a). Governmental and non-governmental nature conservation agencies often use SDMs to guide management and decision-making regarding invasive species but need transparent and reproducible workflows for acceptance by stakeholders and policy-makers (Schwartz et al., 2018; Ferraz et al., 2021; Baker et al., 2021).

There is an active debate on how to improve the reliability and transferability of invasive species distribution models, and new conceptual and methodological approaches are regularly published (e.g. Barbet-Massin et al., 2018; Bellard et al., 2018; Chapman et al., 2019; Hao et al., 2019; Sillero et al., 2023). However, as far as we are aware, most of these proposed innovations are not geared toward automated reproducibility (Kass et al., 2018; Feng et al., 2019a; Mostert et al., 2023).

To address this, we developed the WiSDM workflow to generate reproducible risk maps for potentially invasive alien species under scenarios of climate change at a high spatial resolution (1 km^2). Our workflow semi-automatically: 1) identifies highly correlated predictors; 2) mitigates the impact of sampling bias; 3) generates IAS risk maps for standard RCP scenarios using an ensemble of multiple machine learning algorithms; 4) quantifies spatial autocorrelation in the residuals to assess the impact of clustering of species occurrences; and 5) generates confidence maps for each IAS risk map. This species distribution modeling workflow is part of the Tracking Invasive Alien Species (TrIAS) project, a broader data-to-decision pipeline guiding alien species detection and management (Vanderhoeven et al., 2017). TrIAS encompasses the development and publication of alien species checklists (Reyserhove et al., 2020), identification of emerging species, and risk assessment. WiSDM is written in R

markdown and can be exported as an HTML or notebook file instantly recording all methodological decisions, parameter choices, and outputs, thereby facilitating reproducibility and transparency for risk assessments.

2 Methods

2.1 Overview of WiSDM

Our workflow (Figure 1) uses a hierarchical approach, whereby models are first created at a global scale and then integrated into the European-level models to characterize invasive species' realized niches as extensively as available occurrence data allow. This is achieved by using the model forecast derived from the global model as a probability surface to guide the selection of pseudoabsences for the European model(s). Our SDMs at both the global and European levels use an ensemble of machine learning (ML) algorithms: random forests (RF), gradient boosted machine (GBM), generalized linear model (GLM), and multivariate adaptive regression splines (MARS). These algorithms were chosen because they use distinct approaches: bagging, boosting, linear- and piecewise regression (Table 1). The resulting predictions from each model are stacked together using a GLM as a meta-model to combine the predictions in a weighted combination that optimizes model accuracy (Van der Laan et al., 2007). We used a GLM-based meta-model, instead of a simple averaging of the invasion risk predictions produced by the different modeling algorithms, so that the more accurate models are given a higher weight in the final model while minimizing the risk of overfitting (Hao et al., 2019). The meta-model refers to any statistical or ML model used to combine the information gained from each model's prediction in an ensemble, producing the final model for baseline conditions. Individual country-level maps are a subset of the European model.

The code necessary to run the workflow is available on GitHub (https://github.com/trias-project/risk-modelling-and-mapping).

The global models are climate-only (Pearson and Dawson, 2003), and use high-resolution climate data layers (30 arc second, ~ 1 km) which are available from CHELSA (Karger et al., 2017). The European model uses climate data layers developed specifically for Europe as part of the TrIAS project (De Troch et al., 2020). The TrIAS climate data have been bias-corrected to be compatible with the CHELSA data layers. In order to use our workflow, we have made available the environmental and climate data layers developed for TrIAS via Zenodo (De Troch et al., 2020). The climate layers summarize 30-year climate data (1976-2005), and for three emission scenarios of future climate (the representative concentration pathways (RCP 2.6, RCP 4.5, RCP 8.5) as defined by the IPCC with coverage for Europe. They are based on an ensemble of regional climate models from the EURO-CORDEX archive (Kotlarski et al., 2014), that have been statistically downscaled from a 12.5 × 12.5 km to a 1 km^2 spatial resolution. WiSDM includes predictors characterizing land use/land cover for Europe derived from the CORINE (Coordination of Information on the Environment) landcover product, anthropogenic pressure from the global terrestrial human footprint dataset (Venter et al., 2016), the distance to the nearest freshwater body, and climate variables based on historical (1976-2005) and future (2040-2070) scenarios. These data have been aligned with the 1 km^2 EEA Reference Grid (European Environment Agency, 2011). The outputs of WiSDM include 1) a risk map for Europe produced by global ensemble model based on historical climate conditions; 2) risk map(s) produced by European ensemble model based on historical climate conditions; 3) assessment of the predictive performance of all models; 4) country-level risk maps based on European ensemble model for historical climate conditions and under RCP scenarios; 5) country level maps that visualize differences in current *vs* projected risk under each of the three RCP scenarios; 6) country-level confidence maps; 7) table of variable importance; 8) response curves for all predictors used in the European model; and 9) HTML file from R markdown document saving all code including, decisions, parameters, thresholds, and model outputs (GBIF Secretariat, 2022b). Currently, WiSDM is suitable for modelling plants, mammals, reptiles, amphibians, and birds in Europe, but it can be adapted for other regions. A list of all the predictors used in the workflow and links to download via Zenodo is available (see Data Availability statement). We provide a list of known ecologically relevant predictors for each taxonomic group as Supplementary Information (Supplementary Table 1).

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

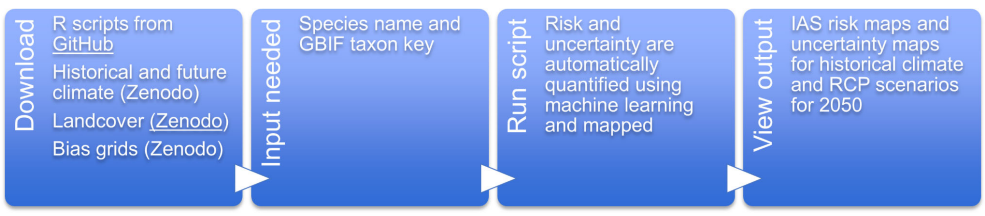

FIGURE 1
Main steps of the WiSDM workflow.

2.2 Occurrence data preparation

The WiSDM workflow utilizes GBIF as it is the largest collector of occurrence data in the world (Waller et al., 2021) with over 2 billion species occurrence records (GBIF, 2023) and follows FAIR data principles Wilkinson et al. 2016. GBIF has taxonomic and geographic data gaps, notably for insects and Asia, respectively which have been the focus of data mobilization efforts (GBIF Secretariat, 2022a). We recommend that users check for the availability of additional occurrence datasets from regional and national environmental agencies if they are not already present on GBIF. In Belgium, data from the relevant agencies (e.g. the Institute for Nature and Forest Research (INBO), Waarnemingen.be, Florabank) are already contributed to GBIF and regularly updated.

Species names are matched with GBIF taxon keys to download only those occurrences with accepted or synonymous names, minimizing taxonomic uncertainty (GBIF Secretariat, 2022b). All species occurrences that have geographic coordinates and are within the time frame specified were downloaded. The default end dates of 1971 and 2010 were chosen to maximize the number of available observations while staying with +/- 5 years of the end dates used for the climate data to minimize a temporal mismatch between the two datasets (Davis et al., 2017). Data with spatial uncertainty greater than 1 km, and duplicate occurrences in the same grid cell are removed. Occurrences that correspond to geographic centroids, biodiversity institutions, and invalid coordinates are flagged and removed using the Coordinate Cleaner package (Zizka et al., 2019). If most of these occurrences are outside of Europe, and there are fewer than ~ 80-100 occurrences in Europe, we recommend running only the global model until more occurrence data become available. Although it is possible to obtain accurate SDMs with low numbers of occurrences as few as five (Pearson et al., 2007; van Proosdij et al., 2016), we recommend a minimum of 30 and restricting the number of predictors used to the number of occurrences divided by 10 to reduce the risk of overfitting.

2.2.1 Mitigating spatial bias in occurrence data

To achieve large geographic coverage, species occurrence databases that are composed of aggregated species data collections such as those provided by GBIF are often used. A drawback to using these databases is their potential for geographic sampling bias (Beck et al., 2013). Uneven sampling or search effort can mislead conclusions about the extent and drivers of species distributions (Gotelli and Colwell, 2001; Lobo, 2008). Sampling bias in our workflow is mitigated by using taxonomic occurrence grids to exclude areas of low sampling effort from the background when randomly placing pseudoabsences (Phillips et al., 2009). The occurrence grids have a 1-degree spatial resolution in the WGS84 coordinate system (EPSG:4326). Each 1-degree grid cell contains the number of records present in GBIF corresponding to a specific taxonomic group: plants, mammals, reptiles, amphibians, and birds. These are also available for download via Zenodo (Davis et al., 2023).

2.3 Global model

The global climate SDM is constructed using all available species occurrence data, employing CHELSA high-resolution climate data layers to delineate the complete range of suitable climate conditions for each species. The number of pseudoabsences equal to the number of species occurrences (Barbet-Massin et al., 2012) are randomly located in the same ecoregions (Olson et al., 2001) inhabited by occurrences, but not in areas of low sampling effort as indicated by the taxonomic occurrence grid (see below). Ecoregions are hypothesized to delineate the area considered theoretically accessible to the organism (Barve et al., 2011; Guisan et al., 2014). To avoid inflating model performance metrics, pseudoabsences are sampled within relevant ecoregions rather than over a large, unrealistic area (Lobo et al., 2008). These pseudoabsences are then combined with the occurrences to form a presence-pseudoabsence spatial point dataset. We use an equal number of pseudoabsences and presences in both the global and European models because large numbers of pseudoabsences relative to presences bias the model towards predicting absences (maximizing specificity). Reported gains in model performance and accuracy as measured by ROC and AUC are due to gains in specificity (Lobo et al., 2008). For models with 800-1000 presences, one draw of an equal number of pseudoabsences is sufficient, otherwise, at least 10 draws are needed (Barbet-Massin et al., 2012). Highly correlated predictors are identified using the 'findCorrelation' command of the 'caret' package, which identifies the predictor(s) with the highest mean correlation with all other predictors (Kuhn, 2022). A global risk map is produced at 1 km resolution based on historical climate conditions using ensemble modeling as described above. The spatial extent of the risk map is limited to Europe to reduce computational processing times. The risk map generated from the global model is used as input into the European model so that the placement of pseudoabsences is restricted to areas with a probability of presence less than 0.5.

2.4 European model

As emerging invasive alien species are unlikely to have many occurrences in a particular European country, WiSDM constructs SDMs using occurrences from all of Europe, and then country-level risk maps are a subset of the European model. The European occurrences are a subset of the cleaned global occurrence data used to build the global SDMs. The European level model incorporates the climatic suitability map generated by the global SDM to locate pseudoabsences in areas of predicted low habitat suitability. The pseudoabsences are randomly located in the same ecoregions inhabited by the alien species (as described above) that overlap with the areas of low habitat suitability predicted by the global model. While introduced species may not have had the chance to fully colonize the ecoregions into which they are introduced, restricting the invasive range of pseudoabsence selection to these regions minimizes the chance of selecting pseudoabsences

corresponding to inaccessible environmental conditions (Chapman et al., 2019). As with the global model, taxonomic occurrence grids are used to avoid locating pseudoabsences in areas with low sampling effort. The pseudoabsences are joined to the occurrences to create a European presence-pseudoabsence dataset. The baseline European level risk model uses the historical climate data for Europe described above, LULC cover data, anthropogenic pressure, and distance to water, described above. From this model, risk maps for specific European countries can be obtained. This model is then projected onto future climate data according to the three RCP scenarios only for the country of interest for faster computational processing times. Country-level risk maps are generated automatically for baseline conditions and the RCPs. Difference maps in the current baseline risk as compared to the future risk under the RCP scenarios are also generated.

2.4.1 Addressing multicollinearity

Multicollinearity in SDMs can increase uncertainty and obscure the most salient predictor in driving species distributions, as well as inhibit model transferability (Yates et al. 2018; Feng et al., 2019b; Liu et al., 2020). WiSDM records and removes highly correlated predictors in the European model using the same method described for the global model. After adding habitat and anthropogenic predictors (heretofore referred to as "habitat" predictors) to the filtered climate dataset, the climate and habitat predictors are examined together for multicollinearity. If an ecologically relevant predictor is flagged, we suggest users consult the correlation matrix to identify alternative predictors for removal as there could be a less crucial predictor contributing to the collinearity. While dimension reduction techniques such as principal component analysis can be used to reduce multicollinearity and improve model transferability for invasive species (Petitpierre et al., 2017), the effects of individual predictors on species distributions are obscured. An understanding of the relationships between invasive species and their environment can inform decision-making, hence our choice not to use data reduction methods.

2.4.2 Assessing spatial autocorrelation

WiSDM assesses the residuals from the European level ensemble model for spatial autocorrelation, using Moran's I. Values of Moran's I greater than 0.1 indicate that the occurrence data may be highly clustered and thinning before model fitting should be employed (Boria et al., 2014; Diniz-Filho and Bini, 2005). The option to thin occurrence data is provided in the workflow via the rarefy command from the Humboldt R package (Brown, 2023).

2.5 Model evaluation and validation

WiSDM reports both threshold-independent (AUC) and threshold-dependent (accuracy, sensitivity, specificity, and kappa) measures of model performance for each algorithm using cross-validation (Kuhn, 2008) for both the global and European-level models. The AUC (Area Under the Curve) statistic quantifies the overall ability of a binary classification model to distinguish between positive and negative classes, with an AUC of 0.5 indicating random performance and an AUC of 1.0 representing perfect discrimination. Accuracy is measured as the number of True Positives + True Negatives/Total Observations. The kappa statistic is measured on a scale of -1 to +1, with 0 indicating the predictive ability of the model is no better than as expected by chance. Sensitivity (true positive rate) and specificity (true negative rate) are reported on a scale of 0-1, with a value of "1" indicating a perfect score. A variety of methods exist to choose a threshold to convert the predicted probabilities to classify a location as either "species present" or "species absent" (Liu et al., 2005). The threshold value can be determined based on ecological knowledge or by optimizing a specific evaluation metric, such as the true positive rate (sensitivity) or the true negative rate (specificity). WiSDM identifies and applies a threshold where sensitivity is equal to specificity with the assumption that the cost of predicting false presences and false absences is the same (Lobo et al., 2008).

2.6 Quantifying and visualizing confidence of model predictions

Uncertainty associated with model predictions and their transferability to new biogeographic areas or novel climate conditions presents another barrier to effective decision-making with SDMs (Brodie et al., 2022). Typically, the accuracy of SDMs is assessed based on how well the model has performed using cross-validation or independent data sets. The dominant methods for quantifying the uncertainty of SDMs are model averaging of the predictive outputs from different algorithms, reporting the standard deviation of the predictions, or using the consensus of the outputs (Thuiller et al., 2019). However, this does not show how good our individual-level predictions are, or how confident we are of them, especially outside the conditions that the model has been calibrated on ('extrapolation'). Our workflow includes code developed by the authors to implement the conformal prediction algorithm to quantify confidence in model predictions. Conformal prediction is a method that leverages past experience to estimate the level of confidence associated with individual predictions, providing a measure of how likely a prediction is to be correct based on historical data. Conformal prediction is distribution-free and has a guaranteed error rate (Shafer and Vovk, 2008). Using a prediction from any method, conformal prediction produces a conformity measure so that strange or unlikely observations are assigned lower conformity scores as compared to more likely observations. The conformity score also known as a p-value (not to be confused with the P values used for statistical hypothesis testing), is the probability estimate that the observation belongs to a class label, with a $1 - \epsilon$ error rate prediction region, a set that contains y with a probability of at least $1 - \epsilon$ percent. The smaller the error rate, the larger the prediction region becomes. WiSDM defaults to an error rate of 20% to balance confidence levels with prediction region size. In binary classification problems, both classes are often included in the prediction region, regardless of the size of the error rate. To address this, probability estimates for each prediction belonging to a class are obtained separately, yielding p-value A that the prediction belongs to class A, and p-value B that it belongs to

class B. Thus the most likely class label is based on the class with the highest p-value. The confidence that the prediction belongs to that class is 1 - the second highest p-value. (Vovk et al., 2005). Conformal prediction has been successfully applied in other fields including Computational biology (Norinder et al., 2014), Medicine (Pereira et al., 2017), and Drug discovery (Alvarsson et al., 2021) but is surprisingly absent from ecological applications. The confidence of each prediction can be visualized in maps, providing an intuitive understanding of how model confidence varies across space and climate scenarios. This can help to identify areas or scenarios where model predictions are less reliable, or where additional data are needed to improve the model. To facilitate the interpretation of the confidence maps, WiSDM can optionally show only those predictions that meet or exceed a user-defined minimum threshold of confidence.

2.7 Use case

We applied WiSDM to a case study species: *Vaccinium corymbosum* L. (North American blueberry, Ericaceae family). This species was identified as a species of potential conservation concern by the TrIAS automated early warning pipeline for prioritizing emerging alien species (Adriaens et al., 2022). North American blueberry is a deciduous shrub that typically grows in moist forests, bogs, and swamps, was introduced to Belgium in the early 1950s and was recently observed to escape from nurseries (Adriaens et al., 2019). This species and its hybrid *Vaccinium corymbosum* × *angustifolium* is considered invasive in Germany and the Netherlands and is known to be problematic in protected areas (wet heathlands, peatlands) there (Schepker and Kowarik, 1998; Penninkhof et al., 2018).

1678 georeferenced occurrences of *V. corymbosum* from 1971-2010 were downloaded from GBIF. After removing centroids, duplicates occurring in the same grid cell, and occurrences with a spatial uncertainty greater than 1 km, 1064 occurrences remained. The majority of these occurrences are located in North America (Figure 2). Of these, only 66 occurrences were located in Europe, with 3 occurrences in Belgium. To account for variability resulting from the location of pseudoabsences, we ran 10 models for Europe, each with a different draw of pseudoabsences equal to the number of presences (Barbet-Massin et al., 2012). The 10 models were evaluated using 4-fold cross-validation. The model with the highest sensitivity, specificity, Kappa, and AUC was selected and projected onto the RCP scenarios. WiSDM automatically generated confidence maps using a minimum confidence threshold of 0.7 for the best predictive model and RCP scenarios. The R markdown document published from this workflow shows in detail all settings, data, algorithms, parameters used, and model validation results and is included as Supplementary Information (Supplementary S1).

3 Results

3.1 Global model

After filtering for multicollinearity, five climate predictors remained and were used in the models. Annual precipitation, the maximum temperature of the warmest month, amount of precipitation (mm) during the driest month, the annual variation of precipitation, and the range of annual temperature °C. The mean predictive accuracy assessed by 10-fold cross-validation for the algorithms ranged from 0.66 to 0.78 and kappa from 0.32 to 0.55. After ensembling, the final model had a mean accuracy of 0.77, and a Kappa of 0.54. Details regarding the performance of each algorithm, model correlation, and variable importance, as well as maps of the area used for sampling pseudoabsences, are available as Supplementary Information (Supplementary S1). The risk map is shown in Figure 3.

3.2 European model

All occurrences found in Europe (n=66) were used in the European models. The following predictors were used: annual variation of precipitation, maximum temperature of the warmest month, range of annual temperature °C, and percent wetland per 1 km^2. 10 models were constructed with 10 unique draws of

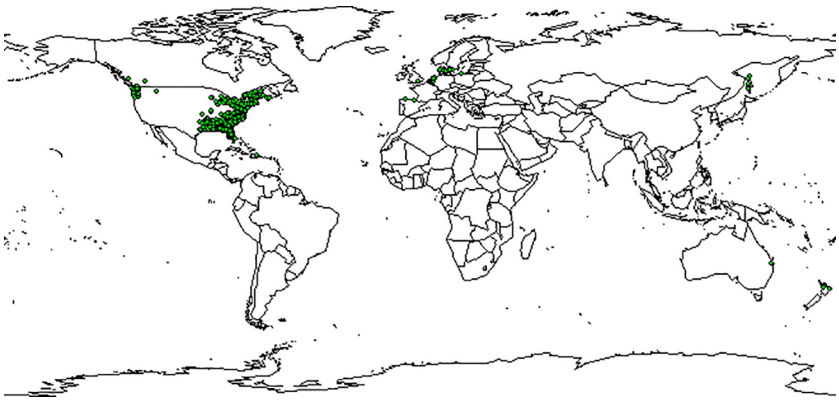

FIGURE 2
Global distribution *of V. corymbosum* occurrences (shown in green) used in the global model.

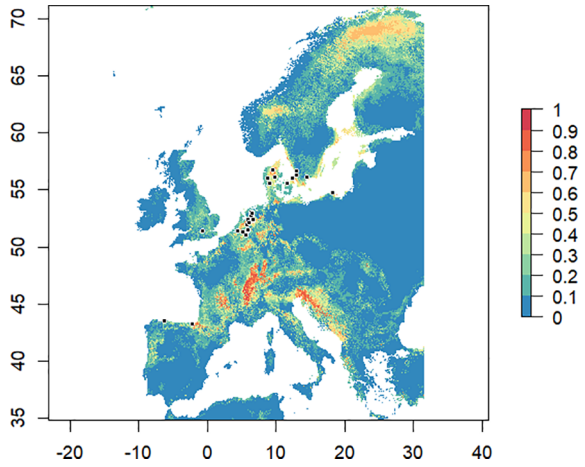

FIGURE 3
Current risk map based on historical climate conditions for Europe generated by the global model for *Vaccinium corymbosum*. The black circles indicate occurrences.

pseudoabsences. The results of the 10-fold cross-validation of these models and the mean of the predicted probabilities demonstrated consistently good performance, with model 6 having the best performance (Table 2). The Moran's I of the residuals from model 6 was very low (- 0.007) indicating that the occurrences did not need thinning. To further test and evaluate the model, *Vaccinium corymbosum* occurrences located in Belgium from 2011-2021 were downloaded (n=111) from GBIF. We regard these data as independent, as they were not used in model building and date after model calibration (> 2010). This model correctly classified 90% of the occurrences as present (Supplementary S1). Model 6 was used for forecasting risk under the RCP scenarios and for the remaining steps in the workflow. The risk map for Europe generated from model 6 is shown in Figure 4.

The current risk map based on historical climate indicates that much of northern Belgium and the Ardennes region(located along the southeast border) is highly suitable for North American blueberry (Figure 5). Risk forecasts for RCP scenarios 2.6, 4.5, and 8.5 suggest that environmental suitability for North American blueberry will greatly decrease in the future for Northern Belgium, but will remain for the Ardennes in RCPs 2.6 and 4.6 (Figures 5, 6).

Confidence in the predicted risk values is highest by area for the current risk map (Figures 7A, 8A). The majority of the predicted risk values under the RCP scenarios are of low confidence (< 0.4) (Figures 7B–D), with very few predicted risk values having high confidence (> 0.7) (Figures 8B–D).

The maximum temperature (°C) of the warmest month followed by the annual range in temperature (°C) had the highest overall variable importance (Table 3). The response curves indicate that the probability of occurrence decreases with increasing maximum temperature of the warmest month and that the species prefers habitats with both warm and cold seasons with yearly temperature differences of approximately 22°C (Figure 9).

4 Discussion

WiSDM constitutes an open, reproducible, and flexible workflow for generating invasion risk forecasts for use in invasive species risk assessment and management. Our framework is ideally suited for agencies, consultants, or environmental planners where fast and easily updatable information on species invasion risk is needed, e.g., for answering to legal reporting requirements such as those mandated by the EU (1143/2014) regulation on invasive alien species or to identify areas where early-detection and rapid response measures preventing invader establishment should be prioritized. Uncertainties in the use of SDM outcomes can lead stakeholders to question the usefulness of invasion risk forecasts for conservation planning (Kujala et al., 2013).

Given the uncertainty associated with extrapolating risk to novel climates, WiSDM produces maps of confidence associated with each risk forecast, allowing identification of where and when model predictions are the most confident. Notably, the majority of predictions for the RCP scenarios have low confidence (Figure 7). For both the historical climate-based and RCP scenarios, the areas predicted as highly suitable for North American blueberry in

TABLE 2 Results of 4-fold cross-validation for European models.

model	threshold	sensitivity	specificity	Kappa	AUC
1	0.48	0.85	0.85	0.70	0.89
2	0.51	0.85	0.85	0.70	0.88
3	0.50	0.85	0.85	0.70	0.89
4	0.47	0.82	0.82	0.64	0.87
5	0.52	0.77	0.76	0.53	0.85
6	0.42	0.88	0.88	0.76	0.89
7	0.46	0.83	0.83	0.67	0.91
8	0.54	0.85	0.83	0.68	0.87
9	0.55	0.82	0.82	0.64	0.88
10	0.55	0.79	0.77	0.56	0.85

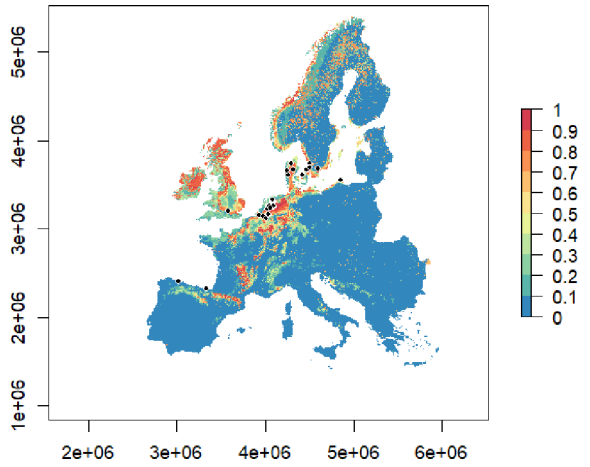

FIGURE 4
Current risk map based on historical climate conditions for Europe generated by the best European model for *Vaccinium corymbosum*. Occurrences are represented by black circles.

Belgium have high confidence and areas predicted to be absent or of low suitability have low confidence (Figures 5, 7). This suggests that in addition to monitoring high-risk areas, surveillance efforts should potentially also include "predicted to be absent but low confidence areas", particularly if they overlap with protected areas or suspected dispersal pathways. Overall, the high uncertainty of the forecasts under the RCP scenarios observed in this study warrants future investigation to determine what steps, if any, can be taken to decrease it. For example, conformal prediction can be used to examine the impacts of variable selection or algorithm choice on model confidence in SDMs. Multivariate environmental similarity surface (MESS) maps (Elith et al., 2010) provide a spatially explicit visualization of the correlation between different climate regimes or scenarios but leave the user to infer how robust their model is. Conformal prediction goes beyond mapping correlation by quantifying the confidence of predictions using a statistical framework with a guaranteed error rate (Vovk et al., 2005). Thus, the user can immediately assess the robustness of their model based on confidence rather than guessing based on climate (dis)similarity.

The European model predicted new areas (Ireland, northern UK, and the coast at risk of invasion as compared to the global model (Figures 2, 3) suggesting the existence of regional niches that would not be observed using only the global model. WiSDM uses the global model to decrease the likelihood of having false absences in the European model by not locating pseudoabsences in areas predicted as suitable by the global model. Furthermore, the European level model is constrained to regional climate and land use data which can help to uncover a regional niche (Gallien et al., 2012).

Response curves provided by WiSDM visualize the relationship between climate and habitat and invasion risk. They can be used to evaluate the ecological realism of the model forecasts as well as to help formulate optimal surveillance efforts in response to changing environmental conditions. For example, the response curve for the annual range of temperature and probability of North American blueberry occurrence shows that invasion is more probable as the difference between the coldest and warmest temperatures increases (Figure 9). This suggests that when annual climate extremes occur (i.e. an unusually cold winter and warm summer), additional monitoring is warranted (Johnstone, 1986).

It should be noted that an inherent limitation to correlative SDMs, including ours, is that the area at risk of invasion may be greater than what is predicted by the model due to the ability of the species to potentially occupy climates and regions that it does not currently inhabit. Failures to accurately predict the full invasive

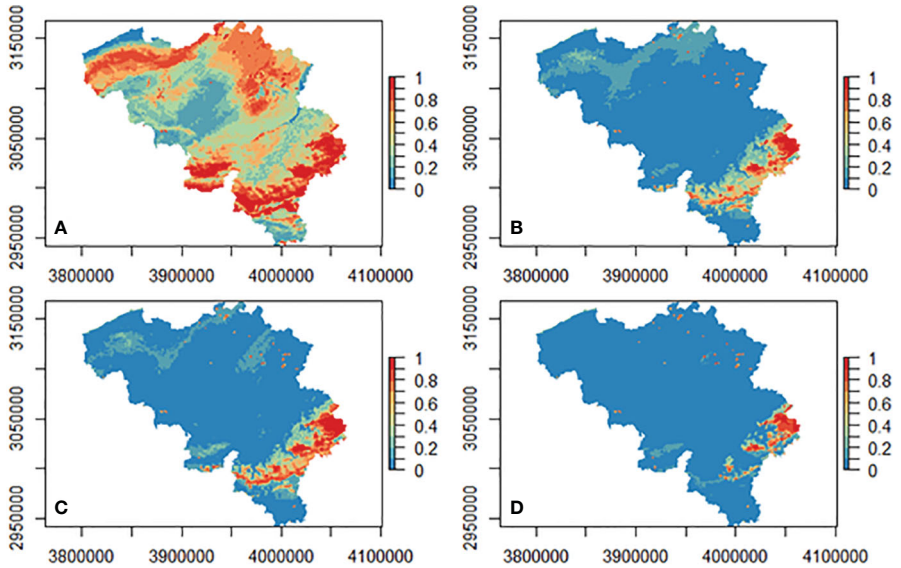

FIGURE 5
Current risk map **(A)** based on historical climate conditions for Belgium generated by the best European model for *Vaccinium corymbosum* and under RCP scenarios 2.6 **(B)**, 4.5 **(C)**, and 8.5 **(D)**.

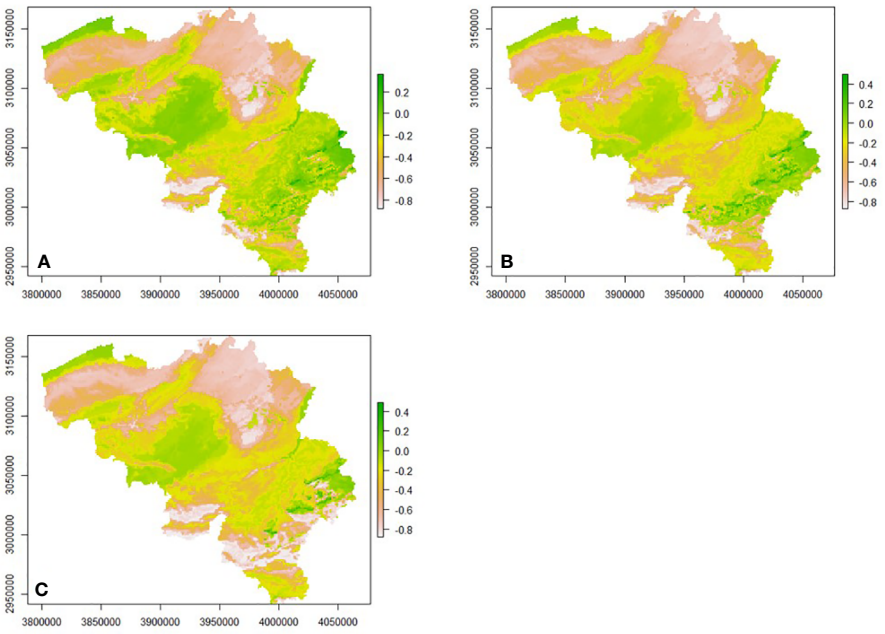

FIGURE 6
Difference maps illustrate the spatial difference between historical climate and RCP scenarios 2.6 **(A)**, 4.5 **(B)**, and 8.5 **(C)**. Green areas indicate where the highest positive differences are observed, and beige and white areas indicate the highest negative differences.

distribution of introduced species are frequently attributed to the violation of a core assumption of SDMs: that the species being modeled is in equilibrium with the environment. The violation of this assumption can occur when the species realized niche is substantially different from its fundamental niche, or in the case of niche expansion when introduced species colonize 'novel' environments in their introduced range, which may not be apparent during the early stages of invasion (Václavík and Meentemeyer, 2012). Apparent niche expansion can occur when eco-evolutionary changes (e.g., genetic adaptations) result in changes in species' fundamental niches, or because, for example, biotic interactions and dispersal limitations prevent species from occupying all suitable areas available to them across their native range. Characterizing species' fundamental niches is generally

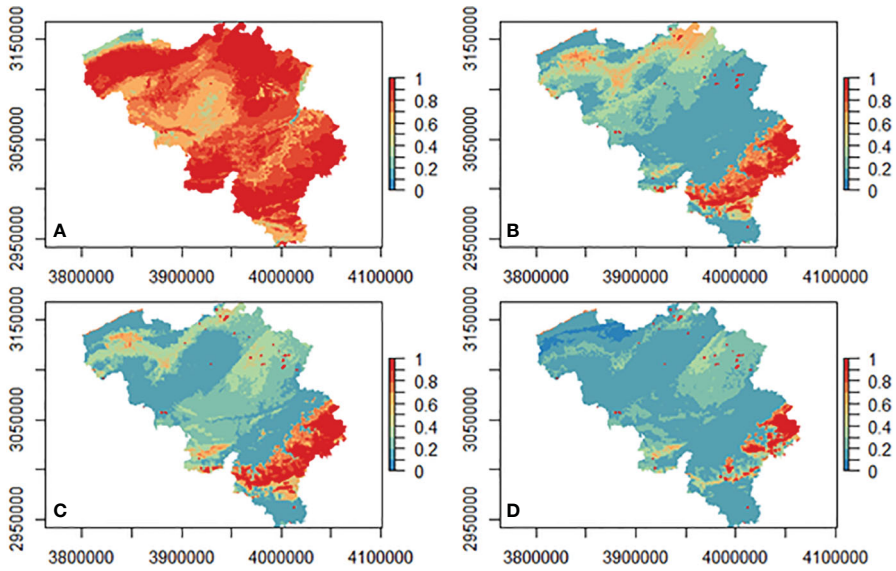

FIGURE 7
Confidence maps for the predicted distribution of *Vaccinium corymbosum* based on historical climate **(A)**, and RCP scenarios 2.6 **(B)**, 4.5 **(C)**, and 8.5 **(D)**, with confidence values ranging between 0 (no confidence) and 1 (maximum confidence). A value of 0 indicates that the prediction is completely nonconforming and not supported by previous data while 1 indicates the prediction is identical to a previous observation in the data.

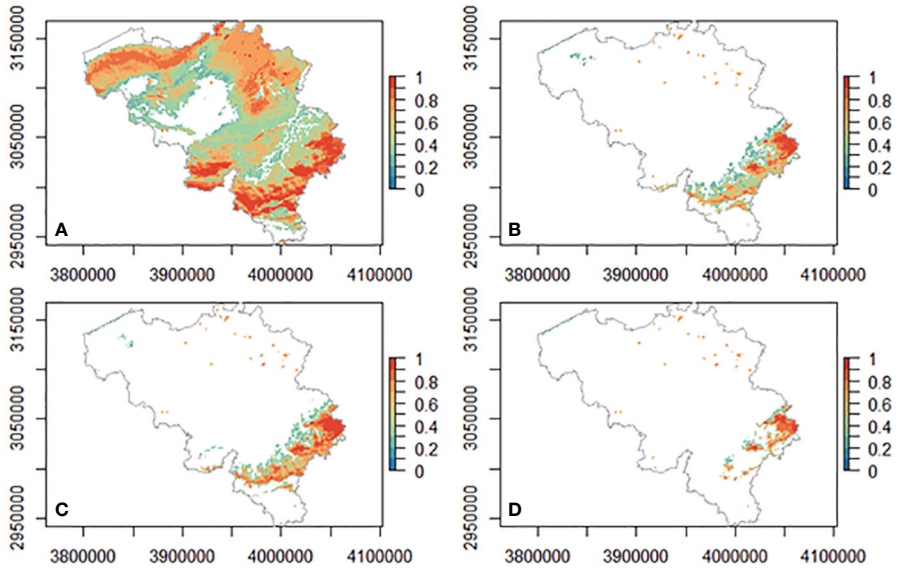

FIGURE 8
Maps of predicted risk based on historical climate **(A)**, and RCP scenarios 2.6 **(B)**, 4.5 **(C)**, and 8.5 **(D)**, with confidence levels equal to or greater than 0. 7. Pixels containing risk values with confidence levels less than 0.7 are not shown.

considered impossible without information on ecophysiological tolerances. Still, there is an active debate about whether certain model settings or algorithms are better able to approximate fundamental niches – and thus species' full potential distribution - using occurrence data only (Jiménez et al., 2019). In addition, missing ecological and/or anthropogenic predictors and gaps in occurrence data that span ecoregions or larger, can also lead to the under-prediction of the full distribution of a species. WiSDM is not set up as a bespoke ecological niche modeling framework to test hypotheses about the factors governing species distributions across native and invasive ranges and how to best model them. Instead, WiSDM produces data-driven SDMs, taking a pragmatic approach by combining GBIF occurrence data from both native and invasive ranges, into a single modeling framework.

The models underlying WiSDM do not account for dispersal, thus the maps produced by WiSDM indicate where a species can potentially colonize once introduced to a region. Furthermore, risk assessments for IAS are often conducted for species that are not yet (widely) present in a country or region thus quantifying the geographic area suitable for the species is an essential step in the risk assessment process. Until a consensus emerges on how potential distributions are best obtained using correlative SDMs (e.g. algorithms, choice of background area, parameter settings) or until alternative (data demanding) process-based models (e.g. based on ecophysiological mechanisms and/or demography and dispersal) can be upscaled to apply to modeling large numbers of species, the WiSDM approach represents a robust and informed tool for use in invasive species risk assessment and management. Furthermore, the modeling workflow can easily be rerun when new occurrence data become available, e.g., through increased biodiversity monitoring, to potentially improve the prediction of the area at risk of invasion. Models can also be run using different baseline climate and habitat predictor layers. WiSDM currently defaults to using a 1976-2005 climate average for model training, which may lead to some uncertainty in estimating species occurrence - environment relationships especially for the most recent occurrence data (Milanesi et al., 2020). The amount of uncertainty introduced by our choice to use a 'static' baseline depends on the rate of change of the predictor variables over time and on how important individual predictor variables are for each species' distribution (Bracken et al., 2022). While no consensus currently exists about how to best ensure optimal correspondence between available occurrence data and predictor variables (Steen et al., 2019), users may decide to use more detailed annual predictor

TABLE 3 Percent variable importance for each algorithm for the best European model (model 3).

	overall	GLM	GBM	RF	MARS
Percent wetland	1.3	8.6	0.0	0.0	0.3
Variation in annual precipitation (coefficient of variation)	28.2	**53.2**	**36.0**	32.1	0.0
Temperature annual range °C	32.3	0.0	32.1	27.0	**63.1**
Maximum temperature warmest month °C	**38.2**	38.2	31.9	**40.9**	36.6

The number corresponding to the most important variable is shown in bold for each algorithm.

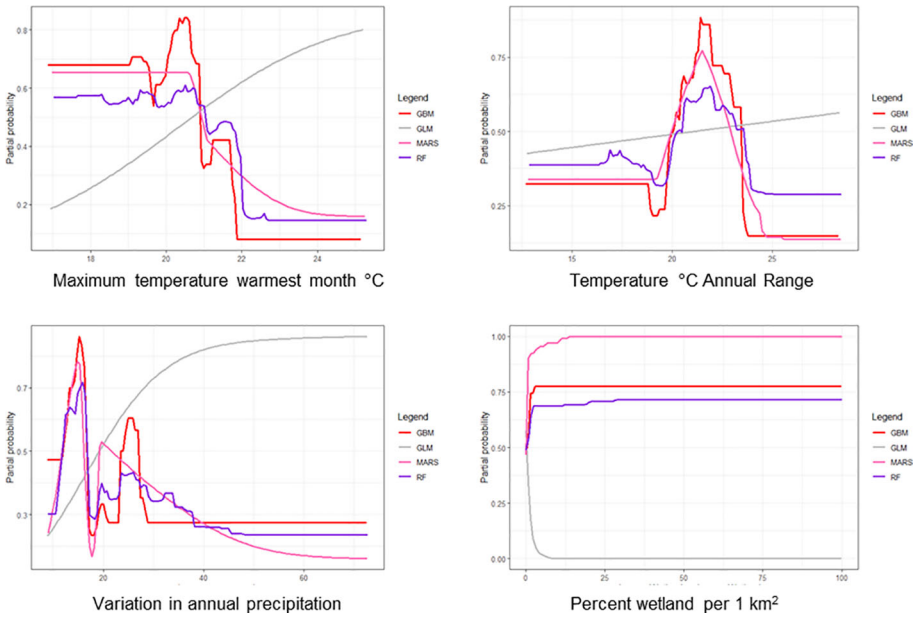

FIGURE 9
Response curves for each algorithm in the best model.

variables (e.g. such as the ERA5 and ERA5-Land time series), effectively turning WiSDM into a dynamic species distribution model (Abrahms et al., 2019).

The climate data currently used by WiSDM were generated using the RCP scenarios from the CMIP5 (Coupled Model Intercomparison Project Phase 5). The RCP scenarios have since been updated with the new SSP (Shared Socioeconomic Pathway) based scenarios from CMIP6 (Coupled Model Intercomparison Project Phase 6). The updated scenarios in CMIP6 that correspond to RCP2.6, RCP4.5, and RCP8.5 from CMIP5 are called SSP1-2.6, SSP2-4.5, and SSP5-8.5, respectively. The SSP scenarios result in similar 2100 radiative forcing levels used by their RCP counterparts, but use different assumptions and improved models with more recent emissions data (Tebaldi et al, 2021). In contrast to RCP scenarios, the SSP scenarios provide economic and social reasons for the assumed emission pathways and land use changes. The SSP scenarios start with emissions data from 2014 (the RCPs start with data from 2007), thus the scenarios start with a higher emissions level and also show a slower decline. When interpreting the results from SDMs using either RCP or SSP scenarios, it is important to consider the assumptions used such as the expected levels of greenhouse gases, population growth, and mitigation as these in addition to the climate models used can influence the results and introduce uncertainty into the projections (Thuiller et al., 2019).

Open, transparent, data-driven risk assessments, with clear indications of uncertainties, foster credibility, which is vital for acceptance by stakeholders and uptake by policy-makers (McGeoch et al., 2012; Groom et al., 2019; Sofaer et al., 2019). The WiSDM approach fits well with recent trends towards transparency and repeatability in ecological forecasting, such as encapsulated in the 'best practice standards' for SDM model development (e.g. Araújo et al., 2019; Zurell et al., 2020). WiSDM further promotes the uptake of SDM modeling into policy and conservation actions by its adoption of the FAIR principles of 'Findability, Accessibility, Interoperability, and Reuse' by making the workflow freely available on GitHub and publishing all data layers needed to run the workflow on Zenodo. The flexible nature of WiSDM also makes it possible for users to customize our code to match the specific demands of the assessment under consideration (e.g. use of alternative climate scenarios and habitat predictors, or model algorithms and settings). The customized settings used are automatically recorded in an R markdown document that can be shared to ensure transparency and reproducibility. Thus, the reproducible workflow presented here maximizes the usefulness of available open data and provides a structured framework for obtaining and interpreting forecasts of the invasion risk of introduced species.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://zenodo.org/communities/trias/?page=1&size=20.

Author contributions

DS and AD conceptualized the research. All authors contributed to the development of the modelling workflow, AD, DS and QG wrote the manuscript with input from all authors. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Belgian Science Policy Office under the TrIAS project (BR/165/A1/TrIAS).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2024.1148895/full#supplementary-material

References

Abrahms, B., Welch, H., Brodie, S., Jacox, M. G., Becker, E. A., Bograd, S. J., et al. (2019). Dynamic ensemble models to predict distributions and anthropogenic risk exposure for highly mobile species. *Divers. Distrib.* 25, 1182–1193. doi: 10.1111/ddi.12940

Adriaens, T., Van Daele, T., Groom, Q., Vanderhoeven, S., Davis, A. J., Strubbe, D., et al. (2022). "Automated early warning: a pipeline for feeding headline indicators on the state of invasions and to prioritize emerging alien species. In Biological Invasions in a Changing World. Book of Abstracts," in *Neobiota 2022–12th International Conference on Biological Invasions*, Tartu, Estonia, 12–16 September 2022. 34.

Adriaens, T., Van Valkenburg, J., Verloove, F., and Groom, Q. (2019). Trosbosbes, probleemsoort in wording? *Natuur. Focus.* 2019 (2), 75–76.

Alvarsson, J., McShane, S. A., Norinder, U., and Spjuth, O. (2021). Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* 110 (1), 42–49. doi: 10.1016/j.xphs.2020.09.055

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., et al. (2019). Standards for distribution models in biodiversity assessments. *Sci. Adv.* 5, eaat4858. doi: 10.1126/sciadv.aat4858

Baker, D. J., Maclean, I. M. D., Goodall, M., and Gaston, K. J. (2021). Species distribution modelling is needed to support ecological impact assessments. *J. Appl. Ecol.* 58, 21–26. doi: 10.1111/1365-2664.13782

Barbet-Massin, M., Rome, Q., Villemant, C., and Courchamp, F. (2018). Can species distribution models really predict the expansion of invasive species? *PloS One* 13 (3), e0193085. doi: 10.1371/journal.pone.0193085

Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* 3, 327–338. doi: 10.1111/j.2041-210X.2011.00172.x

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol. Model.* 222 (11), 1810–1819. doi: 10.1016/j.ecolmodel.2011.02.011

Beck, J., Holloway, J. D., and Schwanghart, W. (2013). Undersampling and the measurement of beta diversity. *Methods Ecol. Evol.* 4, 370–382. doi: 10.1111/2041-210x.12023

Bellard, C., Jeschke, J. M., Leroy, B., and Mace, G. M. (2018). Insights from modeling studies on how climate change affects invasive alien species geography. *Ecol. Evol.* 8, 5688–5700. doi: 10.1002/ece3.4098

Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol. Modell.* 275, 73–77. doi: 10.1016/j.ecolmodel.2013.12.012

Bracken, J. T., Davis, A. Y., O'Donnell, K. M., Barichivich, W. J., Walls, S. C., and Jezkova, T. (2022). Maximizing species distribution model performance when using historical occurrences and variables of varying persistency. *Ecosphere* 13 (3), e3951. doi: 10.1002/ecs2.3951

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brodie, S., Smith, J. A., Muhling, B. A., Barnett, L. A. K., Carroll, G., Fiedler, P., et al. (2022). Recommendations for quantifying and reducing uncertainty in climate projections of species distributions. *Global Change Biol.* 28. 6586–6601. doi: 10.1111/gcb.16371

Brown, J. L., and Carnaval, A. C. (2019). A tale of two niches: methods, concepts and evolution. *Front. Biogeogr.* 11, e44158. doi: 10.21425/F5FBG44158

Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z., et al. (2020). Model complexity affects species distribution projections under climate change. *J. Biogeogr.* 47, 130–142. doi: 10.1111/jbi.13734

Chapman, D., Pescott, O. L., Roy, H. E., and Tanner, R. (2019). Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. *J. Biogeogr.* 46 (5), 1029–1040. doi: 10.1111/jbi.13555

Chauvier, Y., Descombes, P., Guéguen, M., Boulangeat, L., Thuiller, W., and Zimmermann, N. E. (2022). Resolution in species distribution models shapes spatial patterns of plant multifaceted diversity. *Ecography* 2022, e05973. doi: 10.1111/ecog.05973

Cox, D. R. (1958). The regression analysis of binary sequences. *J. R. Stat. Socie.: Ser. B (Methodological)* 20 (2), 215–232. doi: 10.1111/j.2517-6161.1958.tb00292.x

Davis, A., Strubbe, D., and Groom, Q. (2023). Global taxonomic occurrence grids using GBIF data for species distribution models. (1.0.0) [Data set]. *Zenodo*. doi: 10.5281/zenodo.7556851

Davis, A. J. S., Thill, J.-C., and Meentemeyer, R. K. (2017). Multi-temporal trajectories of landscape change explain forest biodiversity in urbanizing ecosystems. *Landsc. Ecol.* 32, 1789–1803. doi: 10.1007/s10980-017-0541-8

De Troch, R., Termonia, P., and Van Schaeybroeck, B. (2020). High-resolution future climate data for species distribution models in Europe [Data set]. *Zenodo*. doi: 10.5281/zenodo.3694065

Diniz-Filho, J. A. F., and Bini, L. M. (2005). Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecol. Biogeogr.* 14 (2), 177–185. doi: 10.1111/j.1466-822X.2005.00147.x

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods Ecol. Evol.* 1, 330–342. doi: 10.1111/j.2041-210X.2010.00036.x

Elith, J. H., Graham, C. P., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29, 129–151. doi: 10.1111/j.2006.0906-7590.04596.x

European Environment Agency. (2011). *EEA reference grid for Europe (1km)*. Available at: https://sdi.eea.europa.eu/catalogue/srv/api/records/d9d4684e-0a8d-496c-8be8-110f4b9465f6.

Feng, X., Park, D. S., Walker, C., Peterson, T., Merow, C., and Papeş, M. (2019a). A checklist for maximizing reproducibility of ecological niche models. *Nat. Ecol. Evol.* 3, 1382–1395. doi: 10.1038/s41559-019-0972-5

Feng, X., Park, D. S., Liang, Y., Pandey, R., and Papeş, M. (2019b). Collinearity in ecological niche modeling: Confusions and challenges. *Nat. Ecol. Evol.* 3, 1382–1395. doi: 10.1002/ece3.5555

Ferraz, K. M. P. M., Morato, R. G., Bovo, A. A. A., da Costa, C. O. R., Ribeiro, Y. G. G., de Paula, R. C., et al. (2021). Bridging the gap between researchers, conservation planners, and decision makers to improve species conservation decision-making. *Conserv. Sci. Pract.* 3, e330. doi: 10.1111/csp2.330

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian Obama nungara. *Ecol. Model.* 457, 109686. doi: 10.1016/j.ecolmodel.2021.109686

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Stat* 19 (1), 1–67. doi: 10.1214/aos/1176347963

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat* 29 (5), 1189–1232. doi: 10.1214/aos/1013203451

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. *Glob. Ecol. Biogeogr.* 21, 1126–1136. doi: 10.1111/j.1466-8238.2012.00768.x

GBIF. (2023). *About species counts in GBIF* Copenhagen, Global Biodiversity Information Facility. Available at: https://www.gbif.org/about-species-counts.

GBIF Secretariat. (2022a). *GBIF Backbone Taxonomy*. Copenhagen, Global Biodiversity Information Facility. doi: 10.15468/39omei

GBIF Secretariat. (2022b). *GBIF Work Programme 2022: Annual Update to Implementation Plan 2017–2022* Copenhagen, Global Biodiversity Information Facility. doi: 10.35035/doc-jjrz-b144

González-Moreno, P., Lazzaro, L., Vilà, M., Preda, C., Adriaens, T., Bacher, S., et al. (2019). Consistency of impact assessment protocols for non-native species. *NeoBiota* 44, 1–25. doi: 10.3897/neobiota.44.31650

Gotelli, N. J., and Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4 (4), 379–391. doi: 10.1046/j.1461-0248.2001.00230.x

Groom, Q., Strubbe, D., Adriaens, T., Davis, A. J. S., Desmet, P., Oldoni, D., et al. (2019). Empowering citizens to inform decision-making as a way forward to support invasive alien species policy. *Citizen Science: Theory Pract.* 4 (1), 33. doi: 10.5334/cstp.238

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., and Kueffer, C. (2014). Unifying niche shift studies: insights from biological invasions. *Trends Ecol. Evol.* 29, 260–269. doi: 10.1016/j.tree.2014.02.009

Guisan, A., and Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009. doi: 10.1111/j.1461-0248.2005.00792.x

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., and Mackey, B. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecol. Model.* 408, 108719. doi: 10.1016/j.ecolmodel.2019.108719

Hao, T., Elith, J., Guillera-Arroita, G., and Lahoz-Monfort, J. J. A. (2019). review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Divers. Distrib.* 25, 839–852. doi: 10.1111/ddi.12892

Jeschke, J. M., and Strayer, D. L. (2008). Usefulness of bioclimatic models for studying climate change and invasive species. *Ann. N.Y. Acad. Sci.* 1134, 1–24. doi: 10.1196/annals.1439.002

Jiménez, L., Soberón, J., Christen, J. A., and Soto, D. (2019). On the problem of modeling a fundamental niche from occurrence data. *Ecol. Model.* 397, 74–83. doi: 10.1016/j.ecolmodel.2019.01.020

Johnstone, I. M. (1986). Plant invasion windows: a time-based classification of invasion potential. *Biol. Rev.* 61 (4), 369–394. doi: 10.1111/j.1469-185X.1986.tb00659.x

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the Earth land surface areas. *Sci. Data.* 4, 170122. doi: 10.1038/sdata.2017.122

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. doi: 10.1111/2041-210X.12945

Kotlarski, S., Keuler, K., Christensen, O. B., Colette, A., Déqué, M., Gobiet, A., et al. (2014). Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geoscie. Model. Dev.* 7 (4), 1297–1333. doi: 10.5194/gmd-7-1297-2014

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Kujala, H., Moilanen, A., Araújo, M. B., and Cabeza, M. (2013). Conservation planning with uncertain climate change projections. *PloS One* 8, e53315. doi: 10.1371/journal.pone.0053315

Lee-Yaw, A. J., McCune, L. J., Pironon, S., and Sheth, N. S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography* 2022, e05877. doi: 10.1111/ecog.05877

Liu, C., Berry, P. M., Dawson, T. P., and Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28 (3), 385–393. doi: 10.1111/j.0906-7590.2005.03957.x

Liu, C., Wolter, C., Xian, W., and Jeschke., J. M. (2020). Species distribution models have limited spatial transferability for invasive species. *Ecol. Lett.* 23, 1682–1692. doi: 10.1111/ele.13577

Lobo, J. M. (2008). Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodivers. Conserv.* 17 (4), 873–881. doi: 10.1007/s10531-008-9333-4

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr.* 17 (2), 145–151. doi: 10.1111/j.1466-8238.2007.00358.x

Mazor, T., Doropoulos, C., Schwarzmueller, F., Gladish, D. W., Kumaran, N., Merkel, K., et al. (2018). Global mismatch of policy and research on drivers of biodiversity loss. *Nat. Ecol. Evol.* 2, 1071–1074. doi: 10.1038/s41559-018-0563-x

McGeoch, M. A., Spear, D., Kleynhans, E. J., and Marais, E. (2012). Uncertainty in invasive alien species listing. *Ecol. Appl.* 22, 959–971. doi: 10.1890/11-1252.1

Milanesi, P., Mori, E., and Menchetti, M. (2020). Observer-oriented approach improves species distribution models from citizen science data. *Ecol. Evol.* 10 (21), 12104–12114. doi: 10.1002/ece3.6832

Mostert, P., Bjørkås, R., Bruls, A. J. H. M., Koch, W., Martin, E. C., and Perrin, S. W. (2023). intSDM : an R package for building a reproducible workflow for the field of integrated species distribution models. *bioRxiv*, 2022.09.15.507996. doi: 10.1101/2022.09.15.507996

Muscatello, A., Elith, J., and Kujala, H. (2021). How decisions about fitting species distribution models affect conservation outcomes. *Conserv. Biol.* 35, 1309–1320. doi: 10.1111/cobi.13669

Nguyen, D., and Leung, B. (2022). How well do species distribution models predict occurrences in exotic ranges? *Global Ecol. Biogeogr.* 31, 1051–1065. doi: 10.1111/geb.13482

Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. (2014). Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. model.* 54 (6), 1596–1603. doi: 10.1021/ci5001168

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., et al. (2001). Terrestrial ecoregions of the world: a new map of life on Earth. *Bioscience* 51 (11), 933–938. doi: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Pearson, R. G., Raxworthy, C. J., Nakamura, M., and Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *J. Biogeogr.* 34, 102–117. doi: 10.1111/j.1365-2699.2006.01594.x

Pearson, R. G., and Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.* 12, 361–371. doi: 10.1046/j.1466-822X.2003.00042.x

Penninkhof, J., Boosten, M., and de Groot, C. (2018). Effect bestrijding trosbosbes in de Pelen. Resultaten van de monitoring in de periode 2015-2017. *Probos Wageningen.* 41.

Pereira, T., Cardoso, S., Silva, D., Mendonça, A. D., Guerreiro, M., and Madeira, S. C. (2017). Towards trustworthy predictions of conversion from mild cognitive impairment to dementia: a conformal prediction approach, in *International Conference on Practical Applications of Computational Biology & Bioinformatics*, eds. F. Fdez-Riverola, M. S. Mohamad, M. Rocha, J. F. De Paz and T. Pinto (Cham: Springer International Publishing, 155–163. doi: 10.1007/978-3-319-60816-7_19

Peterson, A. T., Cobos, M. E., and Jiménez-García, D. (2018). Major challenges for correlational ecological niche model projections to future climate conditions. *Ann. N.Y. Acad. Sci.* 1429, 66–77. doi: 10.1111/nyas.13873

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., and Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global Ecol. Biogeogr.* 26, 275–287. doi: 10.1111/geb.12530

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197. doi: 10.1890/07-2153.1

Reyserhove, L., Desmet, P., Oldoni, D., Adriaens, T., Strubbe, D., Davis, A. J. S., et al. (2020). A checklist recipe: making species data open and FAIR. *Database* 2020, baaa084. doi: 10.1093/database/baaa084

Roy, H., Rabitsch, W., Scalera, R., Stewart, A., Gallardo, B., Genovesi, P., et al. (2017). Developing a framework of minimum standards for the risk assessment of alien species. *J. Appl. Ecol.* 55, 526–538. doi: 10.1111/1365-2664.13025

Schepker, J., and Kowarik, I. (1998). "Invasive North American blueberry hybrids Vaccinium corymbosum x angustifolium in Northern Germany," in *Plant invasions. Ecology and human response*. Eds. U. Starfinger, K. Edwards, I. Kowarik and M. Williamson (Leiden: Backhuys Publisher).

Schwartz, M. W., Cook, C. N., Pressey, R. L., Pullin, A. S., Runge, M. C., Salafsky, N., et al. (2018). Decision support frameworks and tools for conservation. *Conserv. Lett.* 11, e12385. doi: 10.1111/conl.12385

Shafer, G., and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* 9 (3), 371–421.

Sillero, N., Campos, J. C., Arenas-Castro, S., and Barbosa, A. M. (2023). A curated list of R packages for ecological niche modelling. *Ecol. Model.* 476, 110242. doi: 10.1016/j.ecolmodel.2022.110242

Simberloff, D., Martin, J. L., Genovesi, P., Maris, V., Wardle, D. A., Aronson, J., et al. (2013). Impacts of biological invasions: what's what and the way forward. *Trends Ecol. Evol.* 28, 58–66. doi: 10.1016/j.tree.2012.07.013

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., et al. (2019). Development and delivery of species distribution models to inform decision-making. *BioScience* 69 (7), 544–557. doi: 10.1093/biosci/biz045

Srivastava, V., Lafond, V., and Griess, V. C. (2019). Species distribution models (SDM): applications, benefits and challenges in invasive species management. *CABI Rev.* 2019), 1–13. doi: 10.1079/PAVSNNR201914020

Steen, V. A., Elphick, C. S., and Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity Distrib.* 25 (12), 1857–1869. doi: 10.1111/ddi.12985

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., et al. (2021). Climate model projections from the scenario model intercomparison project (ScenarioMIP) of CMIP6. *Earth Syst. Dyn.* 12, 253–293. doi: 10.5194/esd-12-253-2021

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., and Zimmermann, N. E. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nat. Commun.* 10 (1), 1446. doi: 10.1038/s41467-019-09519-w

Urban, M. C. (2015). Accelerating extinction risk from climate change. *Science* 348, 571–573. doi: 10.1126/science.aaa4984

Václavík, T., and Meentemeyer, R. K. (2012). Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity Distrib.* 18, 73–83. doi: 10.1111/j.1472-4642.2011.00854.x

Vanderhoeven, S., Adriaens, T., Desmet, P., Strubbe, D., Backeljau, T., Barbier, Y., et al. (2017). Tracking Invasive Alien Species (TrIAS): Building a data-driven framework to inform policy. *Res. Ideas Outcomes* 3, e13414. doi: 10.3897/rio.3.e13414

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* 6 (1). doi: 10.2202/1544-6115.1309

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. doi: 10.1111/ecog.01509

Venter, O., Sanderson, E., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., et al. (2016). Global terrestrial Human Footprint maps for 1993 and 2009. *Sci. Data* 3, 160067. doi: 10.1038/sdata.2016.67

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world* Vol. 29 (New York: Springer).

Waller, J., Volik, N., Mendez, F., and Hahn, A. (2021). GBIF data processing and validation. *Biodivers. Inf. Sci. Standards* 5, e75686. doi: 10.3897/biss.5.75686

Wenger, S. J., Som, N. A., Dauwalter, D. C., Isaak, D. J., Neville, H. M., Luce, C. H., et al. (2013). Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Change Biol.* 19 (11), 3343–3354. doi: 10.1111/gcb.12294

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3 (1), 1–9. doi: 10.1038/sdata.2016.18

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* 33 (10), 790–802. doi: 10.1016/j.tree.2018.08.001

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., et al. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* 10, 744–751. doi: 10.1111/2041-210X.13152

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., et al. (2020). A standard protocol for reporting species distribution models. *Ecography* 43, 1261–1277. doi: 10.1111/ecog.04960