scientific data

OPEN

DATA DESCRIPTOR

Native range estimates for red-listed vascular plants

Jan Borgelt[1]✉, Jorge Sicacha-Parada[2], Olav Skarpaas[3] & Francesca Verones[1]

Besides being central for understanding both global biodiversity patterns and associated anthropogenic impacts, species range maps are currently only available for a small subset of global biodiversity. Here, we provide a set of assembled spatial data for terrestrial vascular plants listed at the global IUCN red list. The dataset consists of pre-defined native regions for 47,675 species, density of available native occurrence records for 30,906 species, and standardized, large-scale Maxent predictions for 27,208 species, highlighting environmentally suitable areas within species' native regions. The data was generated in an automated approach consisting of data scraping and filtering, variable selection, model calibration and model selection. Generated Maxent predictions were validated by comparing a subset to available expert-drawn range maps from IUCN (n = 4,257), as well as by qualitatively inspecting predictions for randomly selected species. We expect this data to serve as a substitute whenever expert-drawn species range maps are not available for conducting large-scale analyses on biodiversity patterns and associated anthropogenic impacts.

Background & Summary

Life on Earth is essential to human society as it forms the foundation of present welfare[1]. The growing human population, modern lifestyles and associated pressures on the planet have already resulted in a significant loss of natural habitat and are threatening biodiversity[2–6]. Different initiatives promote the protection of biodiversity and aim to halt its loss, such as the UN Sustainable Development Goals[7], the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services[8] and the International Union for the Conservation of Nature (IUCN). Different decision-support tools can contribute to this by assessing environmental performances of products, strategies and policies[2,9–11]. For the development of such tools, but also for the implementation of global conservation strategies and policies itself, spatial data, e.g. in the form of distribution maps of individual species[12], are crucial. However, besides many species remaining undiscovered or undescribed, we still lack spatial information for most of the ones we know[13]. Consequently, comprehensive and ready-to-use datasets for large-scale analyses are only available for a few vertebrate groups[14–16]. This is concerning, as global conservation strategies and biodiversity impact assessments are limited to these groups, while some hyperdiverse species groups, such as plants, are often not considered[17,18].

Here, we provide spatial distribution data for a large fraction of red-listed terrestrial vascular plant species at different levels of spatial detail (Fig. 1), i.e. native regions (n = 47,675), occurrence records (n = 30,906) and modelled range estimates (i.e. a predicted relative environmental suitability[19] within native regions; n = 27,208). The workflow included data scraping and filtering, as well as variable selection, model calibration and model selection, aiming for best practice[20–22] but within the constraints of data limitations and computational feasibility at this scale. Species-specific native regions were retrieved from a scheme specifically developed to challenge the lack of distributional knowledge for plant species[23]. Available native occurrence records were retrieved from the Global Biodiversity Information Facility (GBIF)[24] and subsequently filtered. Range estimates were generated using maximum entropy modelling[19,25–27], and show where environmentally suitable conditions exist within each species' native regions (Fig. 2a–d).

The underlying occurrence data is known to be highly spatiotemporally aggregated and variable across administrative borders for some species[28–31]. We aimed at counteracting a potential sampling bias by using three differently treated occurrence data types (i.e. different degree of spatial filtering: no filter, presence cells, thinned presence cells), and by dividing occurrence data in equally-sized bins during model calibration[32]. Up

[1]Industrial Ecology Programme, Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [2]Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [3]Natural History Museum, University of Oslo, Oslo, Norway. ✉e-mail: jan.borgelt@ntnu.no

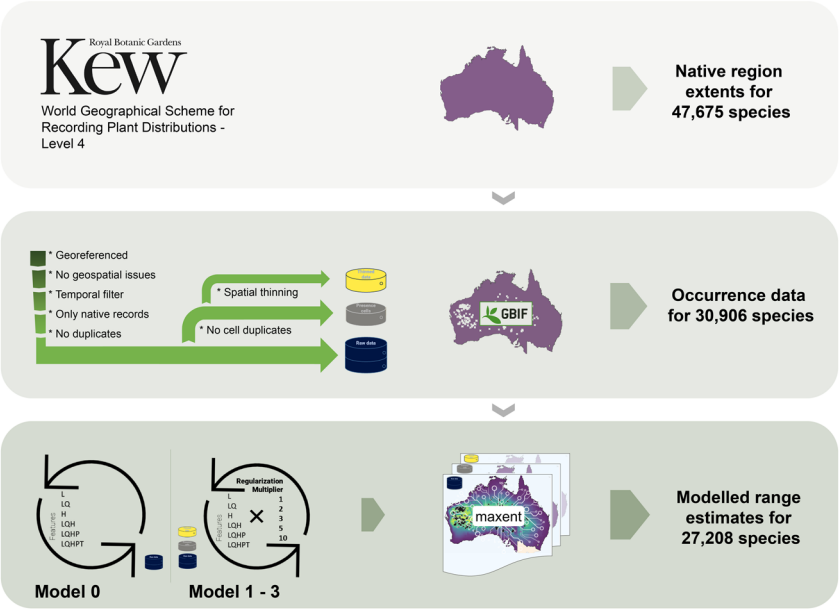

Fig. 1 Schematic summary of the dataset. Top: Native region extents were retrieved from Kew's Plants of the World online. Middle: Occurrence data was retrieved from the Global Biodiversity Information Facility (GBIF)[24] and filtered into three different occurrence data types: raw data (blue), presence cells (grey) and thinned data (yellow). Bottom: The different occurrence data types were used in Maxent models to predict relative environmental suitability indices within native regions (i.e. range estimates). Differences between Model 0 and Model 1 to 3. Model 0 was trained to support variable selection using raw data in k-fold cross validated Maxent models (one model for each combination of feature classes, i.e. linear (L), quadratic (Q), hinge (H), product (P) and threshold (T)). The selected variables and each of the three occurrence data types were used to train a set of separate k-fold cross validated Maxent models (one model for each possible combination of feature classes, regularization multipliers and occurrence data type). The overall best performing model was selected for each species based on performance metrics.

to 96 different models were fitted per species to find optimal variables, model settings and data type. The best prediction was selected for each species based on common performance metrics (i.e. AUC and AUC_{PR}).

However, some predictions will undoubtedly remain flawed by underlying biases. Based on comparisons to expert-drawn range maps available from IUCN (n = 4,257) and qualitative inspection of predictions for randomly selected species, we expect this to mainly influence widespread and common species, and hence, only affect the smallest proportion of global biodiversity[33]. In addition, the species most vital for assessing anthropogenic impacts or for defining conservation priorities, are more likely to be small-ranged and endemic. Although validating each prediction was not feasible, we found most individually inspected predictions to either offer an improvement compared to elsewhere available data or an acceptable substitute, although at a coarser spatial resolution and less detailed.

We want to stress that the presented dataset is generated for the purpose of global spatial screening studies and for building a basis for future, global biodiversity impact assessment models. In concert with powerful, species-specific trait and conservation-related databases, the provided data can benefit future work, such as assessing global extinction probabilities[34], effects of terrestrial acidification[35], drivers of invasion success[36], progress towards reaching global conservation goals[37] and act as pre-assessment prior to expert-based range map generation and red list assessments[38–41]. With a continuously increasing availability of species occurrence records, the presented dataset can be updated frequently to illustrate the state of knowledge at any time. With more data becoming available, precision is likely to increase in the future.

Methods

Taxonomic scope. A species list containing all terrestrial vascular plants (n = 52,372) of the global IUCN red list was retrieved from IUCN in April 2021, IUCN version 2021-1[16]. We retrieved each species' accepted name from Plants of the World Online (POWO)[42] to facilitate communication to various data portals using the package *taxize*[43] in R[44]. Plant family, order and class were retrieved from the Integrated Taxonomic Information System[45] using the package *taxize*[43] in R. Only species outside the IUCN threat categories "Extinct" and "Extinct in the Wild" were kept, and all species considered as subspecies or varieties according to POWO removed. We attempted to assemble spatial data for each of the remaining 48,144 species.

Native regions. Species-specific native regions (Fig. 1) were retrieved from POWO using a customized web-scraper function (see section *Code Availability*) and the packages *taxize*[43] and *rvest*[46] in R. The data follows the World Geographical Scheme for Recording Plant Distributions (WGSRPD)[23] and includes a continental,

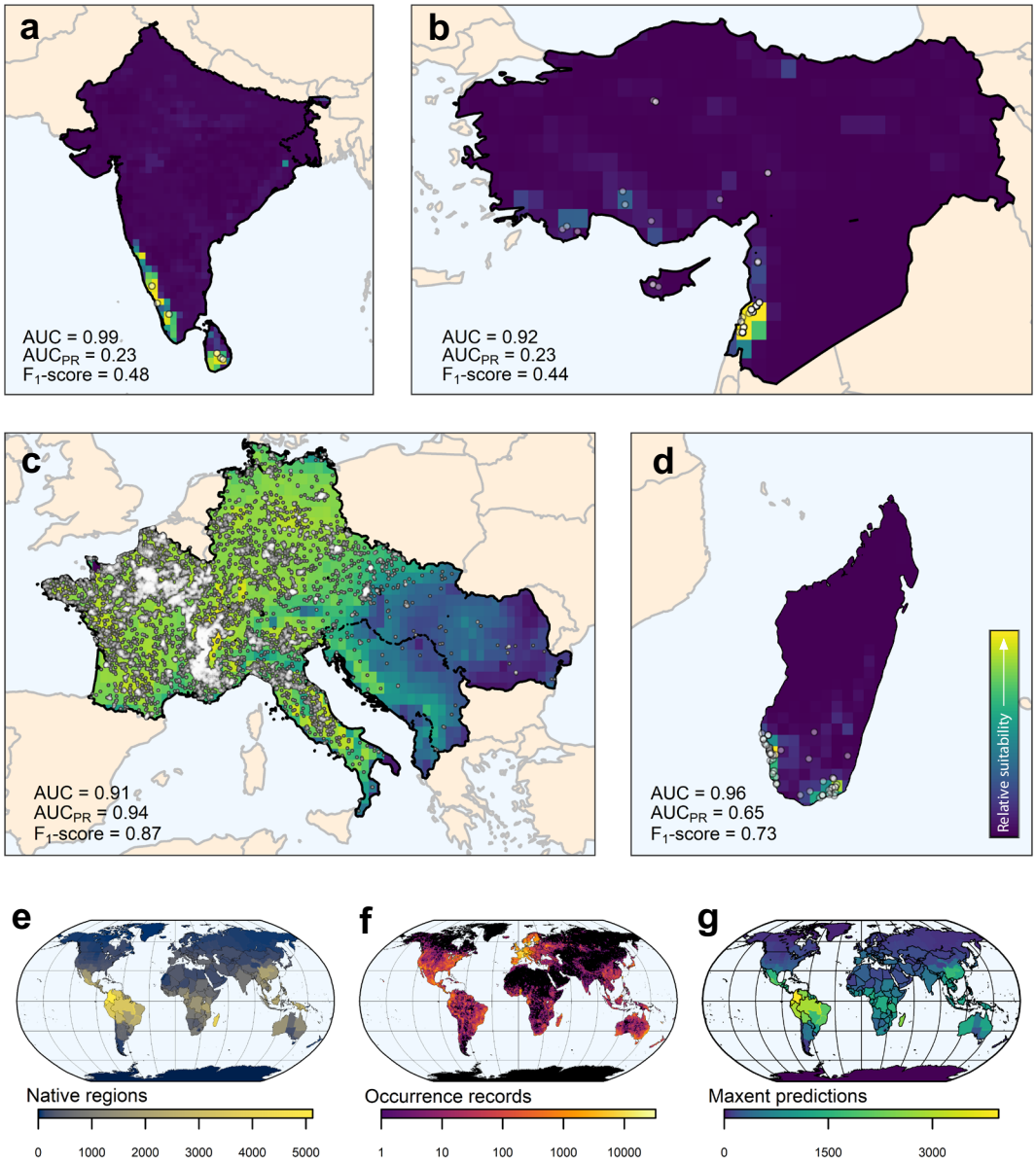

Fig. 2 Data examples for randomly selected species and spatial coverage of the dataset. Best performing Maxent prediction, highlighting environmentally suitable conditions within the species native regions (i.e. modelling extent) along retrieved occurrence records (white points) for (**a**) *Amomum pterocarpum*, (**b**) *Cedrus libani*, (**c**) *Laburnum anagyroides*, (**d**) *Megistostegium nodulosum*. Performance of the shown predictions indicated by maximum F_1-score and the area under the receiver operating characteristics curve for true vs. false positive rate (AUC) and recall vs. precision (AUC_{PR}). Bottom: number of (**e**) retrieved native regions, (**f**) retrieved occurrence records, and (**g**) generated Maxent predictions across the globe.

country and regional level. Retrieved WGSRPD-*regions* were matched to its corresponding shapefile at level 4, available from the Biodiversity Information Standards GitHub repository[47] and rasterized at 30 arc minutes spatial resolution (approximately 56 km at the equator).

Occurrence records. For species with given native extents in POWO, the maximum number of most recent occurrence points (i.e. 100,000) per native WGSRPD-*country* was retrieved from the GBIF application programming interface (API) using the package *rgbif*[48] in R (the equivalent full dataset[49] is available at https://doi.org/10.15468/dl.uvd56q). The considered environmental variables have changed tremendously in the past decades[50,51] and only cover a limited period of time, i.e. the years 1979–2013 and 2015 respectively (see section *Environmental data*). Therefore, only records between the years 2000 and 2020 were considered to temporally align occurrence data to both sets of environmental variables as best as possible. If less than 25 records were available for a given species after the year 2000, no temporal filter was set to maximize data retrieval. GBIF records without specified coordinates and with flagged geospatial issues[48] were not considered. As such, we expect

inaccurate coordinate notations as well as records of specimens preserved in museums or other biodiversity facilities to be typically detected. Only points inside reported native WGSRPD-*regions* were kept and duplicated records were removed (hereafter: raw data). The number of raw data records was counted per cell (30 arc min.) using the package *raster*[52] in R.

Maxent predictions. We generated spatial predictions within species' native WGSRPD-*regions* at 30 arc min. resolution (approximately 56 km at the equator) using maximum entropy modelling (Maxent)[19,26,27], for all species with at least 5 raw data records[53,54] that were distributed across at least 3 cells, and a native region extent of at least 9 cells. Although an arbitrary threshold, we attempted to allocate computational resources to more meaningful predictions, modelled across larger extents. Maxent is a probability density estimation approach widely used for predicting species distributions based on presence-only data[55]. Background information, required to fit response curves[56], was collected from each cell within each species' native regions[57]. For generating models we utilized a high-performance computing infrastructure[58] allowing for parallel computations using the Maxent software[25] via R packages *dismo*[59] and *ENMeval*[60].

Environmental data. We downloaded all CHELSA bioclimatic variables[61,62] (n = 19, see Table 1 for full list) in 30 arc seconds resolution and aggregated, for computational efficiency, to the chosen modelling resolution (30 arc min.) by averaging. CHELSA bioclimatic variables are a set of modelled, biologically relevant, climatic variables based on data collected during the years 1979–2013[61]. In addition, fractions for different natural land cover types, including different types and mosaics of forest, shrubland, grassland and sparse vegetation, (n = 17, see Table 1 for full list) were calculated based on the European Space Agency's land cover product for the year 2015 in 300 m resolution[63]. Each land cover class was transformed into a binary raster depicting presence (=1) and absence (=0) of the land cover type. The binary raster was then aggregated to modelling resolution by averaging, resulting in one raster for each land cover class, representing the proportion of land covered by that class per pixel.

Occurrence data types. For some species, several raw data records can be in the same cell at the given spatial resolution (30 arc min.). Although pseudo-replication can inflate model performance (here: during model calibration) and, hence, increases the risk of overfitting, we argue that these occurrence points still contain valid information if they are discrete observations and therefore kept this data. However, we henceforth applied two filters to counteract potential spatial biases, as well as pseudo-replication (Fig. 1). We removed all cell-duplicates from the raw data (hereafter: presence cells), and we applied spatial thinning with a minimum distance of two cells on the presence cells (hereafter: thinned data). Occurrence data was spatially filtered using the R package *spThin*[64].

Model training. A set of Maxent models was fitted for each species using the differently treated occurrence data types. All models were calibrated using k-fold cross validation. The employed occurrence data was partitioned into training and testing bins. For species with only few data points ($n < 25$), we used k - 1 Jackknife partitioning ($k = n$)[54]. For species with more data points ($n \geq 25$) we used block partitioning ($k = 4$) to account for spatial autocorrelation of occurrence points in larger datasets[32]. This partitioning splits the occurrence data at a longitudinal and latitudinal line, resulting in approximately equally sized bins[60].

An initial model (Fig. 1; Model 0) was trained to support the selection of uncorrelated environmental variables using the raw data and all environmental variables (n = 36) for each species. Separate models, one for each possible combination out of all included feature classes (i.e. environmental variables and transformations thereof), were trained. We included linear (l), quadratic (q), product (p), hinge (h) and threshold (t) transformations, resulting in 6 possible combinations (i.e. l, lq, h, lqh, lqhp, and lqhpt). The best performing model was selected based on the corrected Akaike information criterion (AICc)[65–67]. However, if no model performed best in terms of AICc, or if this metric was unavailable for 50% of fitted models, the average testing area under the receiver operating characteristics curve (AUC; see section *Technical Validation*) during model calibration was used instead. Permutation importance was retrieved for all variables in Model 0. Correlated variables were identified using Spearman's rank correlation coefficient (ρ) and defined as $\rho \geq |\pm 0.7|$. In any set of correlated variables, only the variable with the greatest permutation importance was kept.

The selected environmental variables were used to train separate models for each of the three differently treated occurrence data types: raw data (Model 1), presence cells (Model 2), and thinned data (Model 3). Model 1 was trained if at least 5 raw data records were available, distributed across at least 3 cells (see above). Model 2 and Model 3 were trained if at least 3 records of the corresponding data type were available to avoid computational failure. Although a smaller sample size, we argue that if those models performed better than Model 1, the threshold of 5 records becomes arbitrary and the assessed performance indicators (see section *Technical Validation*) more valuable. The same model architecture as in Model 0 was utilized, including model calibration and selection of the best performing model. However, this time, we added five different regularization multipliers (RM; i.e. 1, 2, 3, 5 and 10; based on previous studies[68–70]) to counteract overfitting[20,56] and for building simpler, ecologically more relevant, models[60]. Hence, separate models for each possible combination out of feature classes and RMs were trained (Fig. 1; Model 1–3), resulting in 30 trained models for each data type and up to 90 models per species.

Metadata. Metadata was assembled for all data and includes general information about species (taxonomy and red list status), provided data type (native regions, occurrence records or Maxent prediction), bounding box of native regions, and if relevant, information about the occurrence data (number of raw data records, Moran's

Variable	Code
Annual Mean Temperature	CHELSA_BIO1
Mean Diurnal Range	CHELSA_BIO2
Isothermality	CHELSA_BIO3
Temperature Seasonality	CHELSA_BIO4
Max Temperature of Warmest Month	CHELSA_BIO5
Min Temperature of Coldest Month	CHELSA_BIO6
Temperature Annual Range	CHELSA_BIO7
Mean Temperature of Wettest Quarter	CHELSA_BIO8
Mean Temperature of Driest Quarter	CHELSA_BIO9
Mean Temperature of Warmest Quarter	CHELSA_BIO10
Mean Temperature of Coldest Quarter	CHELSA_BIO11
Annual Precipitation	CHELSA_BIO12
Precipitation of Wettest Month	CHELSA_BIO13
Precipitation of Driest Month	CHELSA_BIO14
Precipitation Seasonality	CHELSA_BIO15
Precipitation of Wettest Quarter	CHELSA_BIO16
Precipitation of Driest Quarter	CHELSA_BIO17
Precipitation of Warmest Quarter	CHELSA_BIO18
Precipitation of Coldest Quarter	CHELSA_BIO19
Fraction of mosaic cropland/natural vegetation	X30_ESA_CCI
Fraction of mosaic natural vegetation/cropland	X40_ESA_CCI
Fraction of broadleaved evergreen, closed to open, tree cover	X50_ESA_CCI
Fraction of broadleaved deciduous, closed to open, tree cover	X60_ESA_CCI
Fraction of needleleaved evergreen, closed to open, tree cover	X70_ESA_CCI
Fraction of needleleaved deciduous, closed to open, tree cover	X80_ESA_CCI
Fraction of mixed leaf type tree cover	X90_ESA_CCI
Fraction of mosaic tree and shrub/herbaceous cover	X100_ESA_CCI
Fraction of mosaic herbaceous cover/tree and shrub	X110_ESA_CCI
Fraction of shrubland	X120_ESA_CCI
Fraction of grassland	X130_ESA_CCI
Fraction of lichens and mosses	X140_ESA_CCI
Fraction of sparse vegetation	X150_ESA_CCI
Fraction of tree cover, flooded, fresh or brakish water	X160_ESA_CCI
Fraction of tree cover, flooded, saline water	X170_ESA_CCI
Fraction of shrub or herbaceous cover, flooded, fresh/saline/brakish water	X180_ESA_CCI
Fraction of bare areas	X200_ESA_CCI

Table 1. Environmental data used in this study. The layers (n = 36) are based on Karger *et al.*[62] and the European space agency's land cover product[63].

Index[71], calculated as a measure of spatial autocorrelation and based on the number of raw occurrence points obtained per cell), and Maxent metadata: training data (filter treatment, number of training data points), thresholds for converting the prediction into binary range maps[59], model settings (features, parameters, transformations, regularization multiplier, variables) and out of the box[60] model performance, including degree of overfit (DOO) quantified as the difference between calibration and testing AUC during k-fold cross validation[70], as well as self-assessed model performance metrics as described in the section *Technical Validation*.

Data Records

Dataset. The presented dataset is stored in a stable Dryad Digital Repository[72] and can be explored at https://plant-ranges.indecol.no. The dataset includes spatial information for 47,675 species at different levels of detail. In total, range estimates (i.e. relative environmental suitability within native regions) have been predicted for 27,208 species using Maxent, for 30,906 species native occurrence records are provided, and for 47,675 species the spatial extent of its native WGSRPD-*regions* is provided.

All gathered and generated data are stored in netCDF files and can be called by specifying a *varname*. Spatial predictions are provided in Maxent's raw as well as default output (i.e. complementary log-log (cloglog) transformed, but see section *Usage Notes*)[27,59,60]. The suggested data is stored in folder *basic*. These netCDF files (default output and raw output) assemble the best performing Maxent prediction (*varname*: Maxent prediction) for each species selected based on the highest harmonic mean between AUC and AUC_{PR} (see *Technical Validation*), along with number of occurrence records per cell (*varname*: Presence cells) and rasterized native WGSRPD-*regions* (*varname*: Native region).

	Reference		Red list category						
			DD	LC	NT	VU	EN	CR	Total
AUC	Presence - background	Mean	0.939	0.937	0.95	0.96	0.971	0.957	0.945
		Median	0.961	0.951	0.977	0.985	0.994	0.989	0.964
	Reference range	Mean	0.817	0.89	0.927	0.931	0.929	0.915	0.902
		Median	0.852	0.925	0.972	0.974	0.98	0.987	0.943
AUC_{PR}	Presence - background	Mean	0.576	0.529	0.656	0.69	0.749	0.7	0.589
		Median	0.603	0.535	0.717	0.755	0.833	0.797	0.617
	Reference range	Mean	0.516	0.664	0.686	0.653	0.655	0.592	0.658
		Median	0.527	0.702	0.737	0.712	0.699	0.626	0.702

Table 2. Performance of Maxent predictions in the suggested dataset. Mean and median values of area under the receiver operating characteristics curve for true vs. false positive rate (AUC) and recall vs. precision (AUC_{PR}) for all species and across different IUCN threat categories (i.e. data-deficient (DD), least concern (LC), near-threatened (NT), vulnerable (VU), endangered (EN) and critically endangered (CR)). Calculations are based on presence-background data (n = 27,208) and on comparison to expert-based range maps retrieved from IUCN (i.e. reference range, n = 4,257).

The netCDF files in folder *advanced* contain one Maxent prediction for each occurrence data type (*varname*: Model 1, Model 2 or Model 3), instead of best performing Maxent prediction (i.e. *varname* Maxent prediction is not applicable). Number of occurrence records per cell (*varname*: Presence cells) and rasterized native WGSRPD-*regions* (*varname*: Native region) are identical in all netCDF files.

Each band in the netCDF files assembles the mentioned variables for one species. The corresponding bands can be looked up in the metadata (i.e. *speciesID*). Furthermore, the metadata can be used to select appropriate cut-off thresholds for generating binary range maps, filter models based on species, performance, or desired datatypes, and to lookup the relevant study extent for masking individual predictions (see *Usage Notes*).

Technical Validation

Maxent predictions. We calculated performance metrics for model 1 to 3 for each species using its corresponding presence cells to validate the Maxent predictions. Receiver operating characteristic curves and the corresponding area under the curve for *recall* (i.e. *true positive rate, sensitivity*) versus *false positive rate* (AUC) as well as *precision* versus *recall* (AUC_{PR}) were generated using the packages *ROCR*[73] and *PRROC*[74] in R. *Recall* was calculated as the fraction of correctly predicted presence cells compared to all presence cells of the reference (Eq. 1), the *false positive rate* as the fraction of falsely assigned presence cells compared to all true absence cells (Eq. 2), and *precision* as the fraction of correctly assigned presence cells compared to all predicted presence cells (Eq. 3). In addition, F_1-scores (Eq. 4) were calculated as harmonic mean between *recall* and *precision* at all possible cut-off thresholds to transform the Maxent prediction into a binary range map. The maximum obtained F_1-score indicates how well a potential binary range map performs at equal importance of *recall* and *precision*.

$$Recall = \frac{True\ Presence}{True\ Presence + False\ Absence} \tag{1}$$

$$False\ positive\ rate = \frac{False\ Presence}{False\ Presence + True\ Absence} \tag{2}$$

$$Precision = \frac{True\ Presence}{True\ Presence + False\ Presence} \tag{3}$$

$$F_1 = 2\left(\frac{precision \cdot recall}{precision + recall}\right) \tag{4}$$

AUC and AUC_{PR} are threshold-independent performance measures for binary classifiers. An AUC value of 1 indicates a perfect model, an acceptable AUC value (>0.7)[75] indicates the ability to predict many true presences at a low false positive rate, and an AUC value of 0.5 indicates the model performing as good as a random guess. The average AUC obtained across the suggested dataset was 0.95 when comparing predictions to its corresponding presence cells (Table 2), indicating well-performing models for the majority of species. For 26,977 species (99%), at least one Maxent prediction had an AUC value above 0.7[75].

AUC_{PR} is not affected by true negatives (i.e. true absence) which often dominated our dataset. A higher AUC_{PR} value indicates a relatively higher ability to correctly predict a high proportion of presumably true range while maintaining a high precision compared to a lower AUC_{PR}. However, the AUC and AUC_{PR} values, as well as max. F_1-score, described here were calculated based on presence-background data and are highly influenced by class balances. Strictly speaking, both false presences and true absences cannot be determined

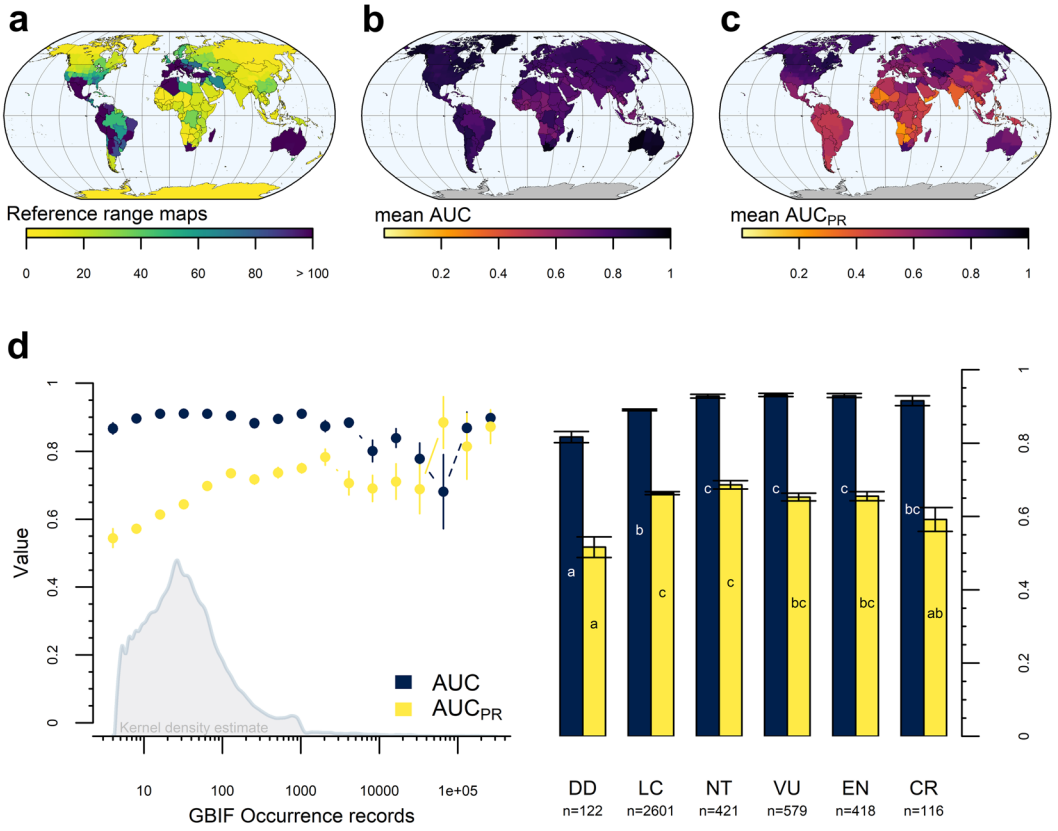

Fig. 3 Performance metrics for the suggested Maxent predictions. (**a**) Number of reference range maps available used for calculating performance metrics. Average values for species native to the corresponding regions of area under the receiver operating characteristics curve for (**b**) true vs. false positive rate (AUC) and (**c**) recall vs. precision (AUC_{PR}). (**d**) Mean and standard deviation of AUC (blue) and AUC_{PR} (yellow) per rounded log-transformed number of raw occurrence data points (left) and for species in different IUCN red list categories (right), i.e. data-deficient (DD), least concern (LC), near-threatened (NT), vulnerable (VU), endangered (EN) and critically endangered (CR). Significant differences across IUCN categories in d are indicated by different letters in bars for AUC (white text) and AUC_{PR} (black text).

with presence-only data. Hence, the performance metrics described here can only be used to compare different models for a given species, but not across different species[76,77].

Therefore, we evaluated the Maxent predictions by comparison to available expert-based range maps, as an additional evaluation dataset[32]. Expert-based range maps were retrieved from IUCN, if available (hereafter: reference ranges). Only reference ranges that were labelled as "native" and "extant (resident)" or "probably extant (resident)" were considered. For 4,257 species of our Maxent predictions, range maps were available at IUCN. These species were unevenly distributed in space (Fig. 3a), across IUCN red list categories (Fig. 3d) as well as the plant classes dicots (Magnoliopsida, n = 3,480), monocots (Liliopsida, n = 731), ferns (Polypodiopsida, n = 27), conifers (Pinopsida, n = 17), and lycopods (Lycopodiopsida, n = 2). Reference ranges were used to calculate the above described performance measures (i.e. max. F_1-score, AUC and AUC_{PR}). However, this time we dealt, presumably, with actual presences and absences of the given species, making the performance metrics comparable across species[76]. Maxent predictions for species classified as "data-deficient" (DD) obtained the lowest, and predictions for species classified as "near-threatened" (NT), "vulnerable" (VU) and "endangered" (EN) the highest AUC values (Fig. 3d). However, these differences were marginal and all average values consistently high across different IUCN categories (mean AUC: 0.9; Table 2) and across the globe (Fig. 3b). Although AUC is a strong indication of model performance[75], the predictions seem to rarely accommodate both a high *recall* and a high *precision* (represented in either max. F1-score or AUC_{PR} value) when compared to reference ranges. However, we found a large variation and no clear trend in AUC_{PR} values for species across different threat-level categories (Fig. 3d), and although the average AUC_{PR} was lowest for species native to parts of central Africa, India and south-eastern Asia (Fig. 3c), we expect these values to be of little explanatory power due to the limited sample sizes in these regions (Fig. 3a). Moreover, AUC_{PR} seems to increase with increasing data availability (Fig. 3d). We assume that low data coverage in sparsely populated areas influenced modelling performance for some, primarily widespread, species, highlighting that sometimes more spatially distributed occurrence data is required for making expert-alike range maps[78].

Furthermore, based on a qualitative assessment of predictions for twelve randomly selected species, we expect uncertainties due to differences in data availability across administrative borders as well as for highly naturalized species. For instance, the clustered occurrence records for *Cedrus libani* in Lebanon (Fig. 2b)

resulted in less precise data than elsewhere available for this species[79], while the prediction for *Laburnum anagyroides* (Fig. 2c) was affected by naturalized occurrence records outside its native origin[80] but still within its native WGSRPD-*regions*. However, this will be most problematic for abundant, widespread, and naturalized species, and hence only relevant for the smallest fraction of global biodiversity[33]. In addition, the predictions for more vulnerable species, presumably small-ranged or endemic, seem to perform better than species in the lowest red list category (i.e. least concern (LC)) in terms of AUC when compared to reference ranges (Fig. 3d).

In fact, the remaining randomly selected predictions were either consistent with point data (e.g. *Terminalia macrostachya*[81]), reflected the current knowledge of elsewhere available data, although at a coarser spatial resolution and less detailed (e.g. *Mammillaria grahamii*[82]), or offered an improvement compared to previously unavailable spatial data (e.g. *Eucalyptus elliptica*[83], *Megistostegium nodulosum*[84] (Fig. 2d), *Memecylon elegantulum*[85], *Psidium salutare*[86,87], *Siparuna conica*[88,89], *Trisetaria dufourei*[90]). However, the prediction of *Pyracantha angustifolia* was difficult to evaluate due to poorly understood range dynamics[91], highlighting the need for more data for vascular plant species.

We want to stress that our predictions indicate environmentally suitable conditions even if isolated from known species occurrence locations. For instance, *Amomum pterocarpum* seems to be restricted to southern India and Sri Lanka[92,93] while our prediction indicates environmentally suitable conditions in north-eastern India (Fig. 2a), which in fact, supports a possible observation nearby[94]. We further detected several expert-based range maps with a substantial mismatch to our data, confirming that some of the expert-based data may be too conservative[95] (e.g. *Magnolia pugana*)[96]. However, we also found expert-based ranges being smaller (e.g. *Vallesia glabra* or *Tetraclinis articulata*)[97,98] than predicted environmental suitability indicates, or being incorrectly georeferenced (e.g. *Corylus cornuta*)[99]. Hence, besides highlighting mismatches to expert-based range maps, we expect this dataset to be of sufficient quality to serve as time- and cost-efficient range map substitutes and pre-assessed range estimates for currently unmapped species.

External data. The retrieved native WGSRPD-*regions* are provided by POWO under a CC BY 3.0 license (https://creativecommons.org/licenses/by/3.0/) and have been checked for consistency to assure proper workflow of data retrieval from POWO and feature matching to the WGSRPD level 4 shapefile. However, the data provider, POWO, cannot warrant the quality or accuracy of the WGSRPD data[42]. In addition, other data (e.g. ecoregions[100]) may ecologically be more relevant than administrative boundaries. However, WGSRPD offers the most detailed data on species' native origins available on a large-scale, to the best of our knowledge. An attempt in matching native WGSRPD-*regions* to ecoregions was discontinued after loss of information due to incompatible geographical boundaries. Hence, we consider the utilized WGSRPD-*regions*, currently, as the best compromise between level of detail and availability of data on species' native origins. Furthermore, spatial inaccuracies and biases in the occurrence data retrieved from GBIF were counteracted by the implemented filtering steps, the coarse spatial resolution, by avoiding non-native occurrence records and the model calibration techniques. However, any unforeseen misclassified or misreported records may flaw predictions for individual species. In addition, data retrieval via GBIF's API was limited to 100,000 occurrence records per request. We extended this limit by sending one request per native country for each species, and hence, expect this issue to be irrelevant for our study. We further want to stress that most of the generated predictions have not been validated individually, and that some predictions may be erroneous either due to data limitations or simply because digitally stored data can contain minor but crucial blunders. For instance, in terms of nomenclature, the red-listed species *Cotoneaster cambricus* is endemic to Wales[101], but also seems to be a synonym for a widespread species according to POWO[42]. Consequently, either our spatial prediction or the expert-based range for this species is incorrect.

Usage Notes

All data handling, modelling and visualization was done using R version 4.0.3[44] in RStudio version 1.4.1103[102]. Handling of all spatial data was done using the R packages *raster*, *rgdal*, *maptools*, *rgeos* and *sp*[52,103–106]. A showcase for opening the different data types for individual species, is available at https://github.com/jannebor/plant_range_estimates. Although functionality of the code may be given at newer, or older, versions, we expect the best user-experience using the versions specified in this descriptor.

Maxent predictions are given as raw and cloglog transformed output. These outputs are related monotonically, meaning that the performance metrics described in this study, as well as a potential binary range map (excluding prevalence dependent thresholds), will be identical for both raw and cloglog output[56]. For users mostly interested in qualitative analyses, both predictions can simply be interpreted as indices of environmental suitability[20]. However, due to rescaling, the exact interpretation and appearance of each output differs. In general, Maxent's output interpretation depends on the underlying data, and differs, in our case between Model 1 (raw data including pseudo-replicates = abundance) compared to Model 2 and 3 (presence), but gives an estimate of the abundance, or presence, of the species in relation to the true modelled quantity (either abundance or presence). Maxent's raw output reflects the exponential Maxent model itself, and can be interpreted as a relative occurrence (or presence) rate summing up to 1[20]. The raw output does not rely on any assumptions[20], however, it may not perform well in visualizing actual differences in suitability[107]. Being rescaled on a more common range from 0 to 1, the cloglog transformation compresses extreme values, and hence facilities visualization and comparison amongst predictions[27]. It can, arguably, be interpreted as a relative probability of presence under certain assumptions[27]. However, as these assumptions are rarely met, we strongly discourage users from this interpretation and suggest interpreting the cloglog output values as an estimate of relative environmental suitability[20] instead.

We further suggest using Maxent predictions with an AUC below 0.7 only in exceptions, and in large-scale studies. In general, our predictions may overestimate true range extents of endemic species and underestimate

ranges of widespread species. However, in worst case, the entire native WGSPRD-*regions* are outlined as being environmentally suitable, which may be acceptable in some cases, but not in others.

In addition, Model 1 has been fitted with the suggested minimum number of records for generating meaningful distributions models[53,54], but Model 2 and 3 were in some cases trained with less records. Whether this low sample size as well as its implied uncertainty is acceptable or not will differ between users and applications and needs to be considered.

The full data, including Maxent predictions (cloglog transformed), underlying occurrence records, native regions and corresponding metadata, can be explored at https://plant-ranges.indecol.no. Here, the predictions based on individual models (Model 1 to 3) as well as a suggested (i.e. best performing) prediction highlight environmentally suitable conditions, if available for the selected species. Predictions can potentially be transformed into a map indicating where the species is most certainly found, as required for local management and conservation actions[95], or into a conservative range map, best suited for analysing global patterns[108] and highlighting where a species is certainly absent[109]. However, the choice of an appropriate cut-off threshold is highly application specific. We outlined "potential range maps" in the data explorer for illustrational purposes only and based on the best performing prediction. We applied different cut-off thresholds to represent different levels of confidence using the R package *dismo*[59]. The threshold at which there was no omission (possibly suitable), the threshold at which the F_1-score is highest (probably suitable) and presence cells (presence).

Code availability

All data and code is available without restrictions under the terms of a Creative Commons Zero (CC0) waiver (https://creativecommons.org/share-your-work/public-domain/cc0/). R code for retrieving and filtering data from POWO and GBIF, and for generating and evaluating Maxent models is available on GitHub (https://github.com/jannebor/plant_range_estimates). Any further requests can be directed to the corresponding author.

Received: 2 June 2021; Accepted: 24 February 2022;
Published online: 29 March 2022

References

1. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis*. (World Resources Institute, 2005).
2. Moran, D. & Kanemoto, K. Identifying species threat hotspots from global supply chains. *Nat. Ecol. Evol.* **1**, 0023 (2017).
3. Newbold, T. Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proc. R. Soc. B Biol. Sci.* **285**, 20180792 (2018).
4. Newbold, T. *et al*. Global effects of land use on local terrestrial biodiversity. *Nature* **520**, 45–50 (2015).
5. Newbold, T. *et al*. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science (80-.).* **353**, 288–291 (2016).
6. Verones, F., Moran, D., Stadler, K., Kanemoto, K. & Wood, R. Resource footprints and their ecosystem consequences. *Sci. Rep.* **7**, 40743 (2017).
7. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development*. A/RES/70/1 (United Nations, 2015).
8. Díaz, S. *et al*. Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science (80-.).* **366**, eaax3100 (2019).
9. Lenzen, M. *et al*. International trade drives biodiversity threats in developing nations. *Nature* **486**, 109–112 (2012).
10. Hellweg, S. & Milà i Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science (80-.).* **344**, 1109–1113 (2014).
11. Chaudhary, A. & Brooks, T. M. National Consumption and Global Trade Impacts on Biodiversity. *World Dev.* **121**, 178–187 (2019).
12. Pereira, H. M., Ziv, G. & Miranda, M. Countryside Species-Area Relationship as a Valid Alternative to the Matrix-Calibrated Species-Area Model. *Conserv. Biol.* **28**, 874–876 (2014).
13. Lomolino, M. V & Heaney, L. R. *Frontiers of Biogeography: New Directions in the Geography of Nature*. (Sinauer Associates Inc. Publishers, 2004).
14. World Wildlife Fund. *WildFinder: Online database of species distributions*. http://www.worldwildlife.org/WildFinder (2006).
15. BirdLife International. *IUCN Red List for birds*. http://www.birdlife.org (2019).
16. IUCN. *The IUCN Red List of Threatened Species*. *Version 2021-1* https://www.iucnredlist.org (2021).
17. Curran, M. *et al*. Toward Meaningful End Points of Biodiversity in Life Cycle Assessment. *Environ. Sci. Technol.* **45**, 70–79 (2011).
18. Woods, J. S. *et al*. Ecosystem quality in LCIA: status quo, harmonization, and suggestions for the way forward. *Int. J. Life Cycle Assess.* **23**, 1995–2006 (2018).
19. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190**, 231–259 (2006).
20. Merow, C., Smith, M. J. & Silander, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography (Cop.).* **36**, 1058–1069 (2013).
21. Araújo, M. B. *et al*. Standards for distribution models in biodiversity assessments. *Sci. Adv.* **5**, eaat4858 (2019).
22. Zurell, D. *et al*. A standard protocol for reporting species distribution models. *Ecography (Cop.).* **43**, 1261–1277 (2020).
23. Brummitt, R. K., Pando, F., Hollis, S. & Brummitt, N. A. World Geographical Scheme for Recording Plant Distributions. *International Working Group on Taxonomic Databases (TDWG)* https://www.tdwg.org/standards/wgsrpd/ (2001).
24. GBIF. The Global Biodiversity Information Facility: What is GBIF? https://www.gbif.org/what-is-gbif (2021).
25. Phillips, S. J., Dudík, M. & Schapire, R. E. Maxent software for modeling species niches and distributions (Version 3.4.0). http://biodiversityinformatics.amnh.org/open_source/maxent/ (2016).
26. Phillips, S. J., Dudík, M. & Schapire, R. E. A maximum entropy approach to species distribution modeling. *Proc. Twenty-first Int. Conf. Mach. Learn.* 655–662 (2004).
27. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: an open-source release of Maxent. *Ecography (Cop.).* **40**, 887–893 (2017).
28. Reddy, S. & Dávalos, L. M. Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30**, 1719–1727 (2003).
29. Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M. & Baselga, A. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847–858 (2008).
30. Isaac, N. J. B. & Pocock, M. J. O. Bias and information in biological records. *Biol. J. Linn. Soc.* **115**, 522–531 (2015).

31. Feeley, K. J. & Silman, M. R. Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Divers. Distrib.* **17**, 1132–1140 (2011).
32. Radosavljevic, A. & Anderson, R. P. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J. Biogeogr.* **41**, 629–643 (2014).
33. ter Steege, H. *et al*. Hyperdominance in the Amazonian Tree Flora. *Science (80-.).* **342**, 1243092 (2013).
34. Kuipers, K. J. J., Hellweg, S. & Verones, F. Potential Consequences of Regional Species Loss for Global Species Richness: A Quantitative Approach for Estimating Global Extinction Probabilities. *Environ. Sci. Technol.* **53**, 4728–4738 (2019).
35. Gade, A. L., Hauschild, M. Z. & Laurent, A. Globally differentiated effect factors for characterising terrestrial acidification in life cycle impact assessment. *Sci. Total Environ.* **761**, 143280 (2021).
36. Géron, C. *et al*. Urban alien plants in temperate oceanic regions of Europe originate from warmer native ranges. *Biol. Invasions* **23**, 1765–1779 (2021).
37. Mair, L. *et al*. A metric for spatially explicit contributions to science-based species targets. *Nat. Ecol. Evol.* **5**, 836–844 (2021).
38. Bachman, S., Moat, J., Hill, A., de la Torre, J. & Scott, B. Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. *Zookeys* **150**, 117–126 (2011).
39. Cardoso, P. red - an R package to facilitate species red list assessments according to the IUCN criteria. *Biodivers. Data J.* **5**, e20530 (2017).
40. Lee, C. K. F., Keith, D. A., Nicholson, E. & Murray, N. J. Redlistr: tools for the IUCN Red Lists of ecosystems and threatened species in R. *Ecography (Cop.).* **42**, 1050–1055 (2019).
41. Bachman, S., Walker, B., Barrios, S., Copeland, A. & Moat, J. Rapid Least Concern: towards automating Red List assessments. *Biodivers. Data J.* **8** (2020).
42. POWO. Plants of the World Online. *Facilitated by the Royal Botanic Gardens, Kew*. http://www.plantsoftheworldonline.org/ (2021).
43. Chamberlain, S. *et al*. *taxize: Taxonomic information from around the web*. *R package version 0.9.98*. https://github.com/ropensci/taxize (2020).
44. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* https://www.r-project.org/ (2021).
45. ITIS. Integrated Taxonomic Information System. https://www.itis.gov/ (2021).
46. Wickham, H. *rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5*. https://cran.r-project.org/package=rvest (2019).
47. Desmet, P. & Page, R. WGSRPD. *GitHub repository* https://github.com/tdwg/wgsrpd (2018).
48. Chamberlain, S. *et al. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.6.0*. https://cran.r-project.org/package=rgbif (2021).
49. GBIF. GBIF Occurrence Download. https://doi.org/10.15468/dl.uvd56q (2021).
50. Winkler, K., Fuchs, R., Rounsevell, M. & Herold, M. Global land use changes are four times greater than previously estimated. *Nat. Commun.* **12**, 2501 (2021).
51. Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E. & Knutti, R. Climate change now detectable from any single day of weather at global scale. *Nat. Clim. Chang.* **10**, 35–41 (2020).
52. Hijmans, R. J. *raster: Geographic Data Analysis and Modeling. R package version 3.0-7*. https://cran.r-project.org/package=raster (2019).
53. Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography (Cop.).* **29**, 773–785 (2006).
54. Pearson, R. G., Raxworthy, C. J., Nakamura, M. & Townsend Peterson, A. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *J. Biogeogr.* **34**, 102–117 (2006).
55. Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography (Cop.).* **31**, 161–175 (2008).
56. Elith, J. *et al*. A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
57. Anderson, R. P. & Raza, A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *J. Biogeogr.* **37**, 1378–1393 (2010).
58. Själander, M., Jahre, M., Tufte, G. & Reissmann, N. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure. *arXiv* 1–4 (2019).
59. Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. *dismo: Species Distribution Modeling. R package version 1.1-4*. https://cran.r-project.org/package=dismo (2017).
60. Muscarella, R. *et al*. 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 (2014).
61. Karger, D. N. *et al*. Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).
62. Karger, D. N. *et al*. Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad, Dataset* https://doi.org/10.5061/dryad.kd1d4 (2018).
63. ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. http://maps.elie.ucl.ac.be/CCI/viewer/download.php (2017).
64. Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography (Cop.).* **38**, 541–545 (2015).
65. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in *2nd International Symposium on Information Theory* (eds. Petrov, B. N. & Csaki, F.) 267–281 (Akademia Kiado, 1973).
66. Hurvich, C. M. & Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika* **76**, 297–307 (1989).
67. Sugiura, N. Further analysts of the data by akaike' s information criterion and the finite corrections. *Commun. Stat. - Theory Methods* **7**, 13–26 (1978).
68. Morales, N. S., Fernández, I. C. & Baca-González, V. MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. *PeerJ* **5**, e3093 (2017).
69. Shcheglovitova, M. & Anderson, R. P. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecol. Modell.* **269**, 9–17 (2013).
70. Warren, D. L. & Seifert, S. N. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol. Appl.* **21**, 335–342 (2011).
71. Moran, P. A. P. Notes on Continuous Stochastic Phenomena. *Biometrika* **37**, 17 (1950).
72. Borgelt, J., Sicacha-Parada, J., Skarpaas, O. & Verones, F. Native range estimates for red-listed vascular plants. *Dryad, Dataset* https://doi.org/10.5061/dryad.qbzkh18h9 (2022).
73. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
74. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
75. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression*. *The Statistician* **45** (Wiley, 2013).
76. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2008).

77. Sofaer, H. R., Hoeting, J. A. & Jarnevich, C. S. The area under the precision-recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* **10**, 565–577 (2019).
78. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
79. Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. Chorological maps for the main European woody species. *Data Br.* **12**, 662–666 (2017).
80. Rivers, M. C. Laburnum anagyroides. *The IUCN Red List of Threatened Species 2017: e.T79919483A79919650* https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T79919483A79919650.en (2017).
81. Botanic Gardens Conservation International Group & IUCN SSC Global Tree Specialist. Terminalia macrostachya. *The IUCN Red List of Threatened Species 2019: e.T150118895A150118897* https://doi.org/10.2305/IUCN.UK.2019-3.RLTS.T150118895A150118897.en (2019).
82. Heil, K., Terry, M. & Corral-Díaz, R. Mammillaria grahamii (amended version of 2013 assessment). *The IUCN Red List of Threatened Species 2017: e.T152723A121546147* https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T152723A121546147.en (2017).
83. Brooker, M. & Kleinig, D. *Field Guide to Eucalypts*. (Bloomings Books, 2006).
84. Koopman, M. M. A synopsis of the Malagasy endemic genus Megistostegium Hochr. (Hibisceae, Malvaceae). *Adansonia* **33**, 101–113 (2011).
85. World Conservation Monitoring Centre. Memecylon elegantulum. *The IUCN Red List of Threatened Species 1998: e.T32597A9713234* https://doi.org/10.2305/IUCN.UK.1998.RLTS.T32597A9713234.en (1998).
86. Landrum, L. R. A revision of the Psidium salutare complex (Myrtaceae). *SIDA, Contrib. to Bot.* **20**, 1449–1469 (2003).
87. Tropical Plants Database. Ken Fern. *tropical.theferns.info* https://tropical.theferns.info/viewtropical.php?id=Psidium+salutare (2021).
88. Bernal, R., Gradstein, S. R. & Celis, M. Siparuna conica S.S.Renner & Hausner. *Catálogo de plantas y líquenes de Colombia* http://catalogoplantasdecolombia.unal.edu.co (2015).
89. Renner, S. S. & Hausner, G. New Species of Siparuna (Monimiaceae) II. Seven New Species from Ecuador and Colombia. *Missouri Bot. Gard. Press* **6**, 103–116 (1996).
90. Melendo, M., Giménez, E., Cano, E., Mercado, F. G. & Valle, F. The endemic flora in the south of the Iberian Peninsula: taxonomic composition, biological spectrum, pollination, reproductive mode and dispersal. *Flora - Morphol. Distrib. Funct. Ecol. Plants* **198**, 260–276 (2003).
91. Chari, L. D., Martin, G. D., Steenhuisen, S.-L., Adams, L. D. & Clark, V. R. Biology of Invasive Plants 1. Pyracantha angustifolia (Franch.) C.K. Schneid. *Invasive Plant Sci. Manag.* **13**, 120–142 (2020).
92. Sasidharan, N. Amomum pterocarpum Thwaites. *India Biodiversity Portal* https://indiabiodiversity.org/species/show/258864#habitat-and-distribution (2013).
93. Contu, S. Amomum pterocarpum. *The IUCN Red List of Threatened Species 2013: e.T44393013A44450020* https://doi.org/10.2305/IUCN.UK.2013-1.RLTS.T44393013A44450020.en (2013).
94. Babyrose Devi, N., Das, A. & Singh, P. Amomum Pterocarpum (Zingiberaceae): a new record in the flora of Manipur. *Int. J. Adv. Res.* **6**, 546–549 (2018).
95. Jetz, W., Sekercioglu, C. H. & Watson, J. E. M. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110–9 (2008).
96. Gibbs, D. & Khela, S. Magnolia pugana. *The IUCN Red List of Threatened Species 2014: e.T194806A2363344* https://doi.org/10.2305/IUCN.UK.2014-1.RLTS.T194806A2363344.en (2014).
97. Sayer, C. Vallesia glabra. *The IUCN Red List of Threatened Species 2015: e.T62543A72668627* https://doi.org/10.2305/IUCN.UK.2015-2.RLTS.T62543A72668627.en (2015).
98. Sánchez Gómez, P., Stevens, D., Fennane, M., Gardner, M. & Thomas, P. Tetraclinis articulata. *The IUCN Red List of Threatened Species 2011: e.T30318A9534227* https://doi.org/10.2305/IUCN.UK.2011-2.RLTS.T30318A9534227.en (2011).
99. Stritch, L., Roy, S., Shaw, K. & Wilson, B. Corylus cornuta (errata version published in 2017). *The IUCN Red List of Threatened Species 2016: e.T194448A115337731* https://doi.org/10.2305/IUCN.UK.2016-1.RLTS.T194448A2336319.en (2016).
100. Olson, D. M. *et al*. Terrestrial ecoregions of the world: A new map of life on Earth. *Bioscience* **51**, 933–938 (2001).
101. Rivers, M. C. Cotoneaster cambricus. *The IUCN Red List of Threatened Species 2017: e.T102827479A102827485* https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T102827479A102827485.en (2017).
102. RStudio Team. RStudio: Integrated Development Environment for R. *RStudio, PBC, Boston, MA* http://www.rstudio.com/ (2021).
103. Bivand, R., Keitt, T. & Rowlingson, B. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. https://cran.r-project.org/package=rgdal (2019).
104. Bivand, R. & Lewin-Koh, N. *maptools: Tools for Handling Spatial Objects. R package version 0.9-5*. https://cran.r-project.org/package=maptools/ (2019).
105. Bivand, R. & Rundel, C. *rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.5-1*. https://cran.r-project.org/package=rgeos (2019).
106. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer New York, 2013).
107. Phillips, S. J. & Elith, J. POC plots: calibrating species distribution models with presence-only data. *Ecology* **91**, 2476–2484 (2010).
108. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci.* **104**, 13384–13389 (2007).
109. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).

Acknowledgements

We want to thank Radek Lonka and the IndEcol Digital Lab for facilitating the use of the high-performance computing infrastructure and hosting the online application. This study is part of the Transforming Citizen Science for Biodiversity project hosted by the Digital Transformation initiative of the Norwegian University of Science and Technology.

Author contributions

J.B. was responsible for study design, methodologies, code writing, code execution, and writing the manuscript. J.S.P. contributed to methods for technical validation of the data and writing the manuscript. O.S. contributed to methodologies, interpretation of the data, and writing the manuscript. F.V. contributed to study design, interpreting the results, and writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.