












# An integrated high-resolution mapping shows congruent biodiversity patterns of Fagales and Pinales

Lisha Lyu<sup>1,2</sup> , Flurin Leugger<sup>1,2</sup> , Oskar Hagen<sup>1,2</sup> , Fabian Fopp<sup>1,2</sup> , Lydian M. Boschman<sup>1,2</sup> ,  
Joeri Sergej Strijk<sup>3,4</sup> , Camille Albouy<sup>5</sup> , Dirk N. Karger<sup>2</sup> , Philipp Brun<sup>2</sup> , Zhiheng Wang<sup>6</sup>,  
Niklaus E. Zimmermann<sup>1,2</sup>  and Loïc Pellissier<sup>1,2</sup> 

<sup>1</sup>Department of Environmental System Science, ETH Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland; <sup>2</sup>Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland; <sup>3</sup>Institute for Biodiversity and Environmental Research, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong BE1410, Brunei Darussalam; <sup>4</sup>Alliance for Conservation Tree Genomics, Pha Tad Ke Botanical Garden, PO Box 959, 06000 Luang Prabang, Lao PDR; <sup>5</sup>IFREMER, Unité Écologie et Modèles pour l'Hallieutique, rue l'Île d'Yeu, BP21105, 44311 Nantes Cedex 3, France; <sup>6</sup>Institute of Ecology and Key Laboratory for Earth Surface Processes of the Ministry of Education, College of Urban and Environmental Sciences, Peking University, 100871 Beijing, China

## Summary

Author for correspondence:

Lisha Lyu

Email: [lisha.lyu@gmail.com](mailto:lisha.lyu@gmail.com)

Received: 20 October 2021

Accepted: 21 March 2022

New Phytologist (2022) 235: 759–772

doi: 10.1111/nph.18158

**Key words:** biodiversity, Fagales, mapping, Pinales, polygon (hull), range map, species distribution modelling (SDM), species richness.

- The documentation of biodiversity distribution through species range identification is crucial for macroecology, biogeography, conservation, and restoration. However, for plants, species range maps remain scarce and often inaccurate.
- We present a novel approach to map species ranges at a global scale, integrating polygon mapping and species distribution modelling (SDM). We develop a polygon mapping algorithm by considering distances and nestedness of occurrences. We further apply an SDM approach considering multiple modelling algorithms, complexity levels, and pseudo-absence selections to map the species at a high spatial resolution and intersect it with the generated polygons.
- We use this approach to construct range maps for all 1957 species of Fagales and Pinales with data compiled from multiple sources. We construct high-resolution global species richness maps of these important plant clades, and document diversity hotspots for both clades in southern and south-western China, Central America, and Borneo. We validate the approach with two representative genera, *Quercus* and *Pinus*, using previously published coarser range maps, and find good agreement.
- By efficiently producing high-resolution range maps, our mapping approach offers a new tool in the field of macroecology for studying global species distribution patterns and supporting ongoing conservation efforts.

## Introduction

Changes in climate (IPCC, 2019) and land use (Meyer *et al.*, 1994) rapidly alter environmental conditions and suitability for species (Walther *et al.*, 2002; Tittensor *et al.*, 2014). As a result, species extinction rates are up to hundreds of times higher than historic background rates, making effective measures to protect the remaining biodiversity urgent (De Vos *et al.*, 2015; Pimm & Joppa, 2015). Such protective measures rely on accurate knowledge of current species ranges, as well as predictions of changes therein under future climatic scenarios (Araújo & Williams, 2000; Heller & Zavaleta, 2009; Bellard *et al.*, 2012). Knowledge on current species ranges provides insight into the factors that shape these ranges (Wang *et al.*, 2010), which is crucial for the prediction of future change (Heller & Zavaleta, 2009; Bellard *et al.*, 2012). Furthermore, by combining accurate species range maps with knowledge on the geographical patterns of specific threats (e.g. climate

change, human activities), the conservation status of species can be quantified (e.g. International Union for Conservation of Nature (IUCN) Red List of Threatened Species; Bland *et al.*, 2015). Full documentation of species ranges is challenging, as the ecology of many species remains unknown or poorly documented (Pimm *et al.*, 2014), global distribution information is often missing or incomplete (Wiszniewski *et al.*, 2008; Duputié *et al.*, 2014), and regional information is scattered across diverse datasets and sources (Serra-Diaz *et al.*, 2017). Data collection is especially challenging for diverse taxa with many (but poorly monitored) species, such as plants (Butchart *et al.*, 2005). As a result, plant species ranges are much less documented compared with many animal clades, and existing documentation is often restricted to specific regions or clades (Miller *et al.*, 2012). This lack of documentation limits the study of global macroecological factors shaping plant diversity and slows down the process of designing global conservation priority settings (Miller *et al.*, 2012; Bland *et al.*, 2015).

The recent sharp increase in freely accessible online data opens the possibility for increased automation in the production of global distribution maps (Wüest *et al.*, 2020). For example, the Global Biodiversity Information Facility (GBIF; <http://www.gbif.org/>) is the largest database for species occurrence records (Beck *et al.*, 2014) and contains data from diverse sources, including museum records, inventory campaigns, and citizen science projects such as iNaturalist (<http://www.inaturalist.org/>) and Les Herbonautes (<http://lesherbonautes.mnhn.fr/>). Furthermore, an increasing proportion of natural history collections are being digitized and integrated into data networks (e.g. Botanical Information and Ecology Network (BIEN); Maitner *et al.*, 2018) and the construction of regional atlases and datasets (e.g. Global Inventory of Floras and Traits (GIFT); Weigelt *et al.*, 2020) continues, building up extensive records of (past) specimen occurrences, national forest inventories, and species checklists. This digitization process is ongoing and some of these datasets are far from being complete, with urban regions or areas along roads more likely to be surveyed than more remote or inaccessible areas (Kadmon *et al.*, 2004; Araujo & Guisan, 2006). Furthermore, the different datasets are not integrated and available data formats vary. Nevertheless, these growing public databases are providing useful high-quality data, which may be used to map species ranges, especially when datasets are combined (Duputié *et al.*, 2014). Given the large amount of available data and the ongoing improvement in the quantity and quality of these datasets over time, the generation of accurate and open-access range maps will benefit from the development of an automated mapping pipeline.

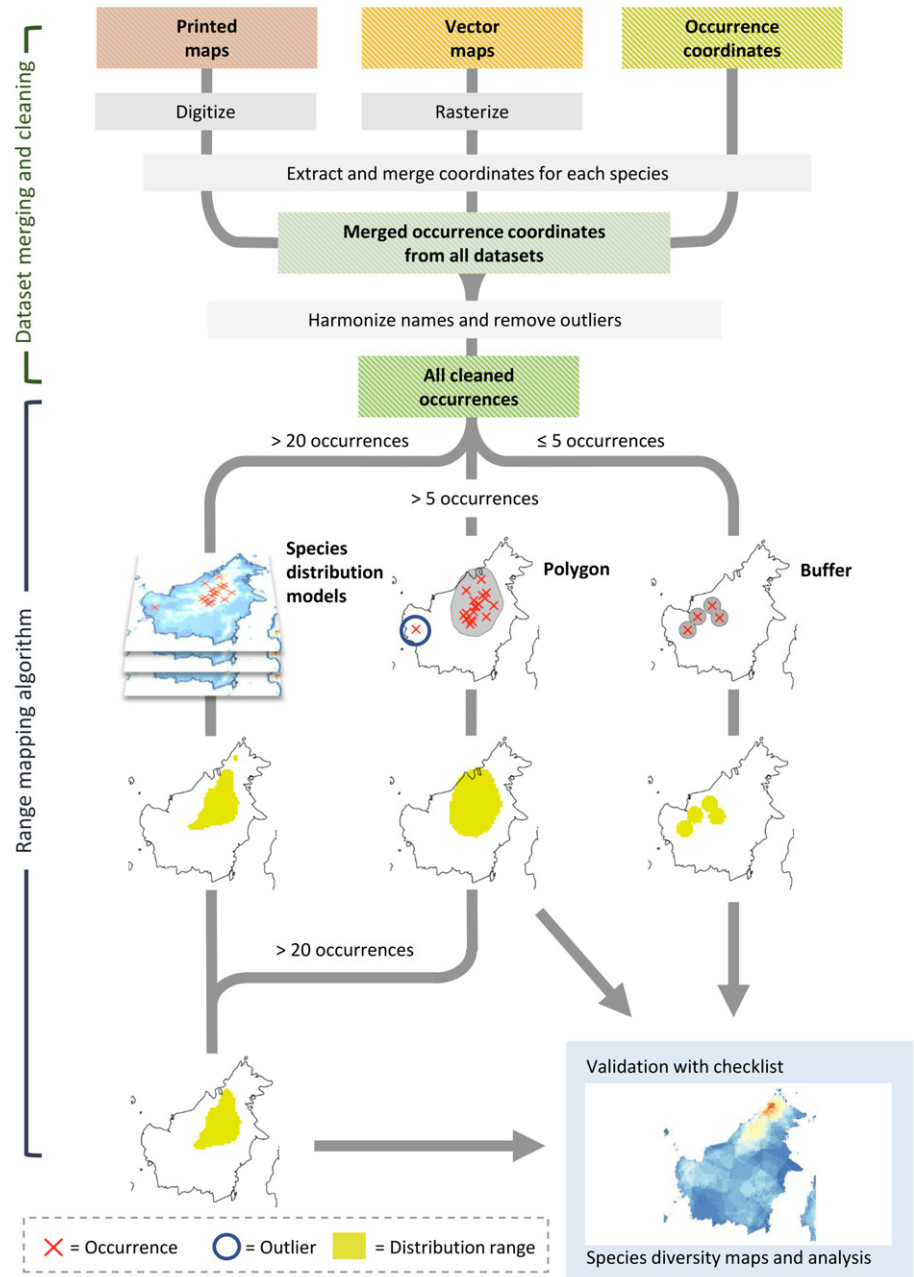
Current estimates list over 380 000 species of vascular plants (Cheek *et al.*, 2020), which is far more than any other existing group for which mapping efforts have been attempted. Because mapping so many species at the global scale is challenging, previous mapping attempts have mainly focused on direct mapping of species richness using statistical models (Kier *et al.*, 2005; Kreft & Jetz, 2007). Otherwise, the enormous challenge of mapping individual species ranges has, so far, been approached using methods in three categories (Graham & Hijmans, 2006; Rocchini *et al.*, 2011): (1) expertise-based mapping (e.g. Rahbek & Graves, 2001); (2) mapping based on predictions derived from species distribution modelling (SDM; e.g. Vasconcelos *et al.*, 2012); and (3) mapping based on polygons or hulls (convex hulls or concave hulls) derived from occurrence records (e.g. Morueta-Holme *et al.*, 2013). Each of these methods has its benefits and drawbacks. Expert-drawn range maps are usually coarsely resolved, are limited to well-known taxa or regions, often overestimate or underestimate distribution ranges (Graham & Hijmans, 2006; Hurlbert & Jetz, 2007), and are usually time-consuming to create. SDM typically account for abiotic conditions but not for historical dispersal and connectivity (Guisan & Thuiller, 2005; Pollock *et al.*, 2014). As a result, their outcome represents the potential niche of a species rather than the actual distribution range (Guisan & Thuiller, 2005; Merow *et al.*, 2017), which may include nonnative ranges. Polygons around known observation points may underestimate the range of a species if observations do not cover its range well (Burgman & Fox, 2003) or overestimate it if unsuitable areas

among observation points are not masked out (Meyer *et al.*, 2017). For coarse-resolution (> 100 km) range maps, polygon-mapping (e.g. Rodríguez-Casal & López-Pateiro, 2010; Hagen *et al.*, 2019) is useful. For instance, Sundaram *et al.* (2019) mapped conifer assemblages in 100 km × 100 km grid cells across the globe using the  $\alpha$ -hull approach. Applying the same approach for 43 635 tree species, Xu *et al.* (2020) quantified global patterns in tree diversity and found correlations between (spatially varying) temperature changes since the Last Glacial Maximum (LGM) and global diversity patterns such as species turnover and nestedness. Mapping approaches can be improved significantly by taking into account both general distribution limits via polygon mapping and the suitability of local abiotic conditions using SDM, thereby minimizing the limitations of the individual approaches (Graham & Hijmans, 2006; Merow *et al.*, 2017; Di Febbraro *et al.*, 2018).

In this study, we present an integrated mapping approach to construct standardized global species range maps by combining polygon mapping with SDM. We develop a new polygon mapping algorithm by introducing new parameters considering distances and nestedness of occurrences. We explore SDM features related to modelling algorithm and complexity settings, and pseudo-absence selection. We integrate maps from both algorithms and take the intersection as the final species range map. To validate the performance of our method, we integrate occurrence data from a large variety of sources and map the distribution of species from two major plant lineages: the orders of Fagales and Pinales. Both are globally distributed (Govaerts & Frodin, 1998; Yang *et al.*, 2017), are locally dominant in a wide range of ecosystems and environments (Manos & Stanford, 2001; Brodribb *et al.*, 2012), and include both widely distributed and rare or endemic species (Fragnière *et al.*, 2015; Yang *et al.*, 2017). Moreover, these orders are well suited for the purpose of our study, as occurrence data are relatively abundant. These two clades are of high ecological and economic value and are often the protagonists in ecological and evolutionary studies (e.g. Wang & Ran, 2014; Xing *et al.*, 2014; Xu *et al.*, 2019), but high-resolution distribution and richness maps are not yet available. Improved global mapping of species in these clades would significantly contribute to the macroecology and biogeography studies, while improving the chances of successful *in situ* conservation (Ferrier, 2002), and also supporting efforts to conserve genetic diversity in viable *ex situ* populations (Huamán *et al.*, 2000).

## Materials and Methods

The workflow includes five main parts (Fig. 1): data collection, data cleaning, parameter optimization, mapping by integration of SDM and polygons, and map validating. The working environment is in R (R Core Team, 2013), and the scripts for data cleaning, parameter optimization, and mapping are accessible online (<https://gitlab.ethz.ch/gdplants/gdplants/>). This code can be flexibly applied to any plant clade or region of interest. Illustrated here for Fagales and Pinales, the species range and richness maps can be efficiently constructed for other clades following the data science workflow.



**Fig. 1** Diagram of the workflow of data collection, data cleaning, parameter optimization, map construction, and map validation.

### Data collection and merging

We retrieved occurrence information for Fagales (including Betulaceae, Casuarinaceae, Fagaceae, Juglandaceae, Myricaceae, Nothofagaceae, and Ticodendraceae families) and Pinales (including Araucariaceae, Cephalotaxaceae, Cupressaceae, Phyllocladaceae, Pinaceae, Podocarpaceae, Sciadopityaceae, and Taxaceae families) from 48 databases (see Supporting Information Table S1 for details). To reduce the risk of underestimation of species ranges in regions for which observational data are scarce, we included not only text-based datasets but also existing distribution maps, which were either already available in the form of raster or shape files or were digitized by our team. The 48 databases used in this study consist of online data sources, journal

articles, and books containing regional checklists, expert-drawn maps, and occurrence points. For two large online data sources, specific packages in the R environment are available: we used the RGBIF package (Chamberlain *et al.*, 2017) to access the GBIF and the BIEN package (Maitner *et al.*, 2018) for the BIEN (data downloaded in October 2018). We retrieved data from all other sources manually. We converted all occurrence data into decimal longitude/latitude format in the World Geodetic System 1984 (EPSG 4326).

### Data cleaning

To account for synonymous, unresolved, misspelled, or wrong species names and wrong or missing family names, we

standardized, corrected, or added names, following the Catalogue of Life (<https://www.catalogueoflife.org/>; accessed in April 2021). We kept only records with standardized species names, and we removed all duplicate records. We attributed all subspecies to species, and we removed hybrid species. To account for the records of cultivated species or records that were assigned to incorrect coordinates, we removed records falling within a 10 km radius around country capitals, within a 5 km radius around country centres, within a 1 km radius around biodiversity institutions, within a 1° radius around the GBIF headquarters (Copenhagen, Denmark), and within a 0.5° radius around longitudinal/latitudinal coordinates 0,0 using the R-package COORDINATE-CLEANER (Zizka *et al.*, 2019). We evaluated whether observations were made in the species' native range using the regional-level distribution database, Royal Botanic Gardens, Kew, UK (POWO, 2019; accessed in February 2019), which includes most families of Fagales (except Juglandaceae and Myricaceae) and all Pinaceae families. For each species, we generated a 2° buffer around the Kew distribution range and removed records outside this buffer. We manually checked species for which the cleaning process resulted in more than 50% of records being deleted, and we manually retrieved erroneous treatments. As uneven distribution of occurrence records may increase the uncertainty on both the shapes and connectedness of hulls, and may cause underestimation of species ranges in regions with sparse occurrences due to the deviation of weight in SDM mapping, for species with > 50 occurrences, we removed occurrences closer to each other than 0.1° using the 'desaggregation' function in the R-package ECOSPAT (Di Cola *et al.*, 2017).

### Species distribution modelling

For SDM, we used nine environmental variables related to temperature, precipitation, and soil conditions as predictive variables. Climate variables included average annual temperature, aridity (annual precipitation divided by annual potential evapotranspiration), frost change frequency, precipitation in the driest quarter, mean diurnal temperature range, and precipitation seasonality. These factors represent basic resource requirements, metabolic modifiers, or disturbance constraints to plant growth and survival. We extracted these climate variables from Climatologies at High resolution for the Earth's Land Surface Areas (CHELSA v.2.1; Karger *et al.*, 2017). We downloaded the soil variables organic carbon content, pH, and clay content from SoilGrids (Hengl *et al.*, 2014, 2017; <http://soilgrids.org>). We extracted all variables at a 30 arc-s resolution and converted them to the World Geodetic System 1984 (EPSG 4326) projection. The total set of nine variables has a rather low multicollinearity (Pearson's  $r < |0.78|$ , highest correlation is between precipitation in the driest quarter and precipitation seasonality).

We considered four algorithms for modelling: generalized linear models (GLMs; Nelder & Wedderburn, 1972), generalized additive models (GAMs; Hastie & Tibshirani, 1990), generalized boosting machines (GBMs; Friedman, 2001), and random forest (RF) models (Breiman, 2001). For each algorithm, we implemented three complexity levels with regard to model

formulation, resulting in 12 different models (Brun *et al.*, 2020). We ran the SDM analyses in R using the packages GAM (Hastie, 2018), RANDOMFOREST (Liaw & Wiener, 2002), and GBM (Greenwell *et al.*, 2018). The number of predictors we considered per species was constrained by the number of available occurrence data, such that the number of observations available was at least 10 times the number of predictors used (Harrell Jr *et al.*, 1996). If the final number of presences was between 20 and 30, we fitted bivariate models based only on the mean annual temperature and aridity; if the number of presences was between 30 and 40, we also added the third most important predictor, organic carbon content; if the number of presences was between 40 and 90, with every increase of 10 occurrences we added frost change frequency, precipitation in the driest quarter, soil pH, mean diurnal temperature range, and precipitation seasonality one after another to the predictor set. If 90 or more filtered presence observations were available, we considered the full predictor set. For species with fewer than 20 occurrence records, we did not execute the SDM mapping. For each species, we projected the environmental suitability across the study area based on the six models achieving the highest scores in the true skill statistic (TSS; Allouche *et al.*, 2006), as evaluated by a three-fold random cross-validation. We converted model-based projections to binary presence/absence using the threshold that maximized TSS. We then summed the binary projections and assumed the species to be present in areas where all six models predicted presence. We generated the SDM maps at 1 km resolution.

To determine the most appropriate pseudo-absence sampling strategies and complexity levels, we explored 192 ensembles of the combination of four algorithms (GLM, GAM, GBM, and RF), six complexity levels, and seven sampling strategies to fit the SDM. We applied the parameterization following the method of Brun *et al.* (2020). Initially, we set up 24 modelling strategies by combining six levels of complexity in each of the four models: (1) in GLM, we set the polynomial degree to 1, 2, 3, 4, 5, or 6; (2) in GAM, we set the degrees of freedom to 1, 2, 3, 5, 10, or 15; (3) in GBM, we set the maximum number of trees to 100, 200, 300, 500, 1000, or 10 000; and (4) in RF, we set the minimum node size to 1000, 500, 20, 10, 3, or 1.

The seven different pseudo-absence strategies were: random, target-group, geographic, density, geographically stratified, environmentally stratified, and environmentally semi-stratified (see Notes S1 for details). At the exploration stage, we used each of these sampling strategies to draw 8000 pseudo-absences and complemented those with 2000 points sampled with the environmentally-stratified approach. Adding environmentally-stratified pseudo-absences guaranteed that the entire environmental space was considered for model training, and that uninformed model extrapolations were avoided. We combined presences and pseudo-absences of each species into a presence-absence dataset. For all presences and for all pseudo-absence sampling methods, we ensured that the final points selected were at least 5 arc-min apart from each other to avoid spatial autocorrelation and bias from overly dense sampling.

We used Kew's regional-level distribution maps as a reference, randomly drawing 2000 presence points (inside the distribution



ranges) and 10 000 absence points (as described earlier), which we used as independent validation data to evaluate the various parameterization settings. We assessed the results of different ensembles of pseudo-absence strategies and modelling strategies based on TSS values. We selected the pseudo-absence strategy with the highest mean TSS value among the 24 modelling methods (the geographically stratified approach; see Notes S2; Table S6; Figs S1, S2 for details). Then, for each of the four modelling methods, we selected the three complexity levels with the highest TSS values under the selected pseudo-absence strategy for all species: a polynomial degree of 1, 2, or 3 for GLM, degrees of freedom of 2, 3, or 5 for GAM, a maximum number of trees of 500, 1000, or 10 000 for GBM, and a minimum node size of 10, 3, or 1 for RF. For each species, we applied these 12 modelling strategies for the SDM mapping.

### Generating species geographic boundaries

We developed a polygon (hull) range mapping algorithm for the generation of species ranges from species occurrence data. We defined six main bioregions: Nearctic, Palearctic, Afrotropic, Indomalaya, Australasia, and Neotropic (Antarctica and Oceania were excluded from this analysis; Fig. S3; The Nature Conservancy, 2009). For rare species with fewer than four occurrences, we created a polygon by simply drawing a 0.5° buffer around the occurrences. For all other species, the algorithm identified cluster points (an occurrence or occurrences within a certain distance) and removed outliers (occurrence(s) isolated from cluster points). Within each bioregion, the cluster points were grouped into cluster(s) based on the *k*-means algorithm, and polygons (or a point buffer) were drawn surrounding these clusters. We then assembled the multiple polygons in each bioregion into a single shapefile per species, which we then converted to a raster. To define a cluster point and an outlier, we defined two parameters: (1) the minimum number of points needed to be considered a cluster (minimum cluster size); and (2) the minimum distance for a point or points (depending on cluster size) to be considered as an outlier (outlier distance). To test the parameters, we randomly sampled 200 species from all species as a subset. Using this subset, we explored the parameters by setting: (1) minimum cluster size to 1, 2, 3, 5, 7, or 9; and (2) outlier distance to 1°, 2°, 3°, 5°, or 7°. We used these 30 parameter sets to generate different polygon maps, and overlaid these polygons with the maps from SDM mapping for each species. We then calculated two indices to evaluate the results of this exploration: (1) the number of species with polygon maps generated; and (2) the fraction of occurrences falling within the combined map (overlap between range polygons and SDM).

Nearly half of the species had fewer than 20 occurrences whose ranges could not be further optimized by SDM mapping. To balance potential overestimation by a larger distance and underestimation by a smaller distance, and to reduce potential outliers, for the final mapping we considered a cluster to be a group of two or more occurrences (minimum cluster size = 2), and an outlier to be a single occurrence that is at least 5° away from a cluster (outlier distance = 5°) (see results for details).

### Producing species distribution maps and lineage richness maps, and evaluation

We selected the best-performing parameter combination as the optimal combination. For species with > 20 occurrences, we obtained the final distribution map by determining the overlap between the polygon map and the SDM map. For species with fewer than 20 occurrences, the final distribution map was equal to the polygon map. Finally, we generated lineage richness maps by stacking the final species distribution maps.

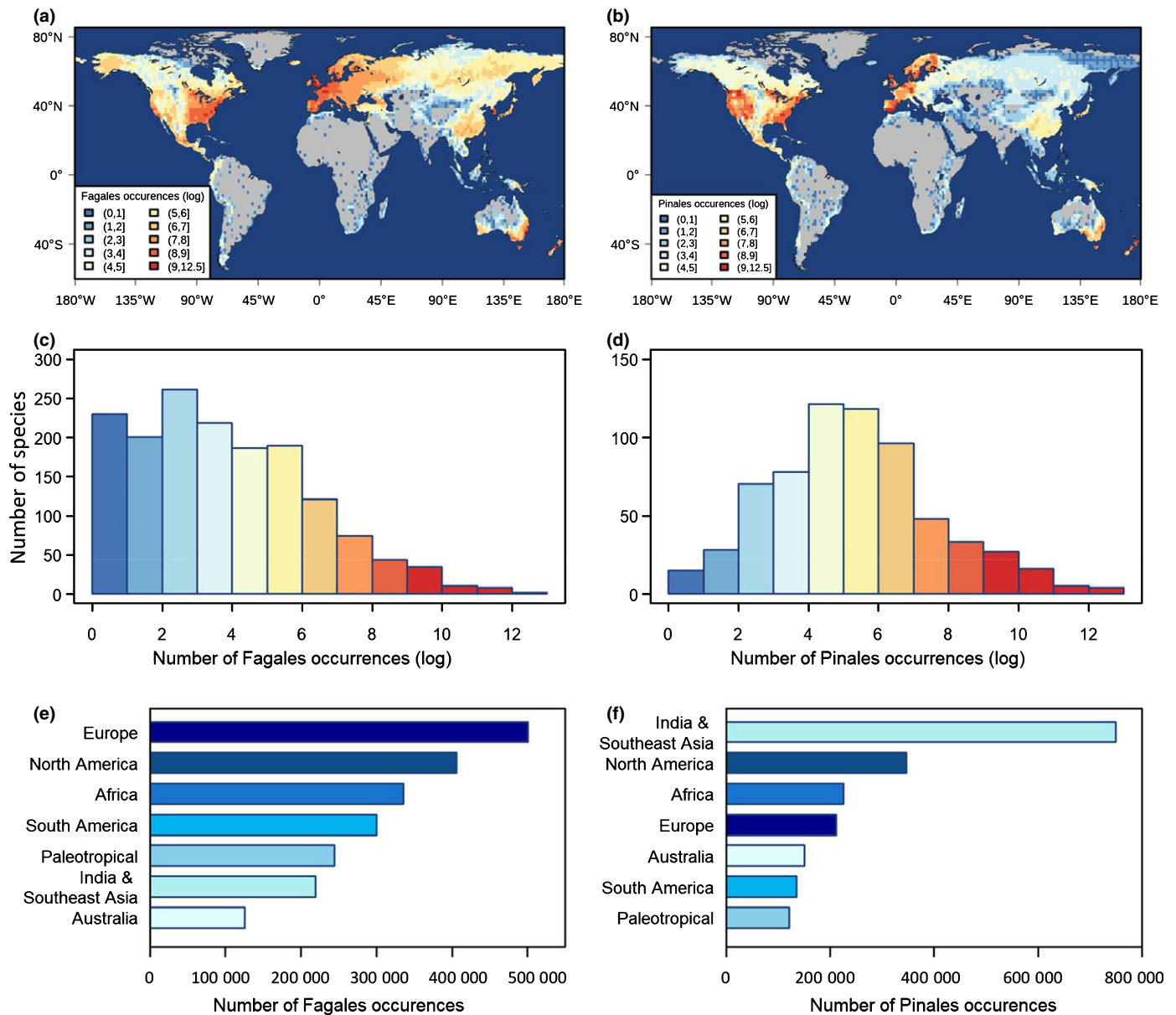
We evaluated species distribution maps and lineage richness maps separately. We manually checked all distribution maps using resources including Flora of China (eFloras, 2020), The PLANTS Database of the US Department of Agriculture (USDA & NRCS, 2020), PlantZAfrica (<http://pza.sanbi.org/>), Flora Malesiana (<http://portal.cybertaxonomy.org/flora-malesiana/>) and Plants of the World Online of Kew (POWO, 2019). We used four levels to assess the consensus: (1) total mismatch; (2) similarity between our maps and the references but with a large area missing or additional in either; (3) similarity between our maps and the references but with a small area missing or additional in either; and (4) complete match (see Notes S3 for details).

As there are available regional distribution maps for *Quercus* (Xu *et al.*, 2019) and coarse-resolution maps for *Pinus* (Critchfield & Little, 1966), we evaluated the richness maps of the two genera by comparing our high-resolution richness maps of *Quercus* (431 species) and *Pinus* (110 species) with previously published richness maps separately. We determined the similarity by calculating the correlation using Spearman's  $\rho$ .

## Results

### Occurrence data collection

We collected 5934 880 valid occurrence records from 48 databases (Table S1) for the 15 families in the two lineages, including 6065 different species names. After correcting species names using the Catalogue of Life (2021) and data cleaning, we retained 1932 species, covering all families and genera of Fagales and Pinales (Table S2). There were 1318 species with > 20 records, 84 species (e.g. *Quercus robur*, *Pinus sylvestris*, *Juniperus communis*) of which had more than 10 000 records (*Pinus halepensis* had the largest number of records: 242 561). There were 402 species with 4–20 records, which was insufficient for SDM; and 208 rare species with fewer than four records, which was insufficient for polygon mapping (Fig. 2; Table S3). For 84 species, > 50% of the records were removed during data cleaning (Table S3). By manually checking these species, we determined that most of them represent widely cultivated or rare species with only few occurrence records available, implying that the applied cleaning procedures were justified. All occurrence records of three species, *Quercus bawanglingensis*, *Quercus obconicus*, and *Betula glandulosa*, were mistakenly removed and we manually retained these records. Specifically, for Fagales, we originally collected 3134 710 records. Correcting for species names and occurrence data cleaning left us a dataset encompassing 2372 272 records,



**Fig. 2** Density maps of occurrence records (log-transformed) collected for Fagales (a) and Pinales (b). Sampling bias for Fagales (c) and Pinales species (d), where each block shows the number of species with specific numbers (log-transformed) of validated records collected. Sampling bias between different bioregions for Fagales (e) and Pinales (f), where each block shows the number of species with specific numbers of validated records collected.

1326 (98.1%) of the 1351 (excluding hybrid species) Fagales species of the Catalogue of Life (Tables S2–S4). For Pinales, we originally collected 2800 170 records. After the final data cleaning, the dataset encompassed 2246 672 records with all 606 Pinales species (except hybrid species) of the Catalogue of Life (Tables S2–S4). For the missing 25 species, we searched the literature for their occurrence information and added them into the database (Table S5).

### Performance of species distribution models

From the seven pseudo-absence generating strategies, the geographically stratified strategy scored the highest mean TSS

(0.580), followed by the random strategy (0.579), the environmentally stratified strategy (0.578), the environmentally semi-stratified strategy (0.571) and the target-group (0.567 with all species as the target group, and 0.546 with the family as the target group). The geographic strategy and the density strategy yielded low performance, with mean TSS values of 0.445 and 0.362, respectively (Table S6). The difference among the five strategies with the highest TSS values was not significant ( $P$ -value = 0.99). To avoid the uncertainties potentially introduced by random sampling, we selected the geographically stratified strategy as the pseudo-absence generating strategy for all SDM. The assessment of the different SDM methods and complexity levels yielded high performance for low complexity GLMs (polynomial degree = 1,

2, or 3) and GAMs (degrees of freedom = 2, 3, or 5), and for high complexity GBMs (maximum number of trees = 500, 1000, or 10 000) and RF models (minimum node size = 10, 3, or 1) (Table S6). All selected models except those from medium and complex GLMs (with degree of the polynomial = 2 and 3) had mean TSS scores above 0.6. The mean TSS value of the 12 models under the geographically stratified strategy was 0.641 (Table S6).

### Parameter optimization of polygon mapping

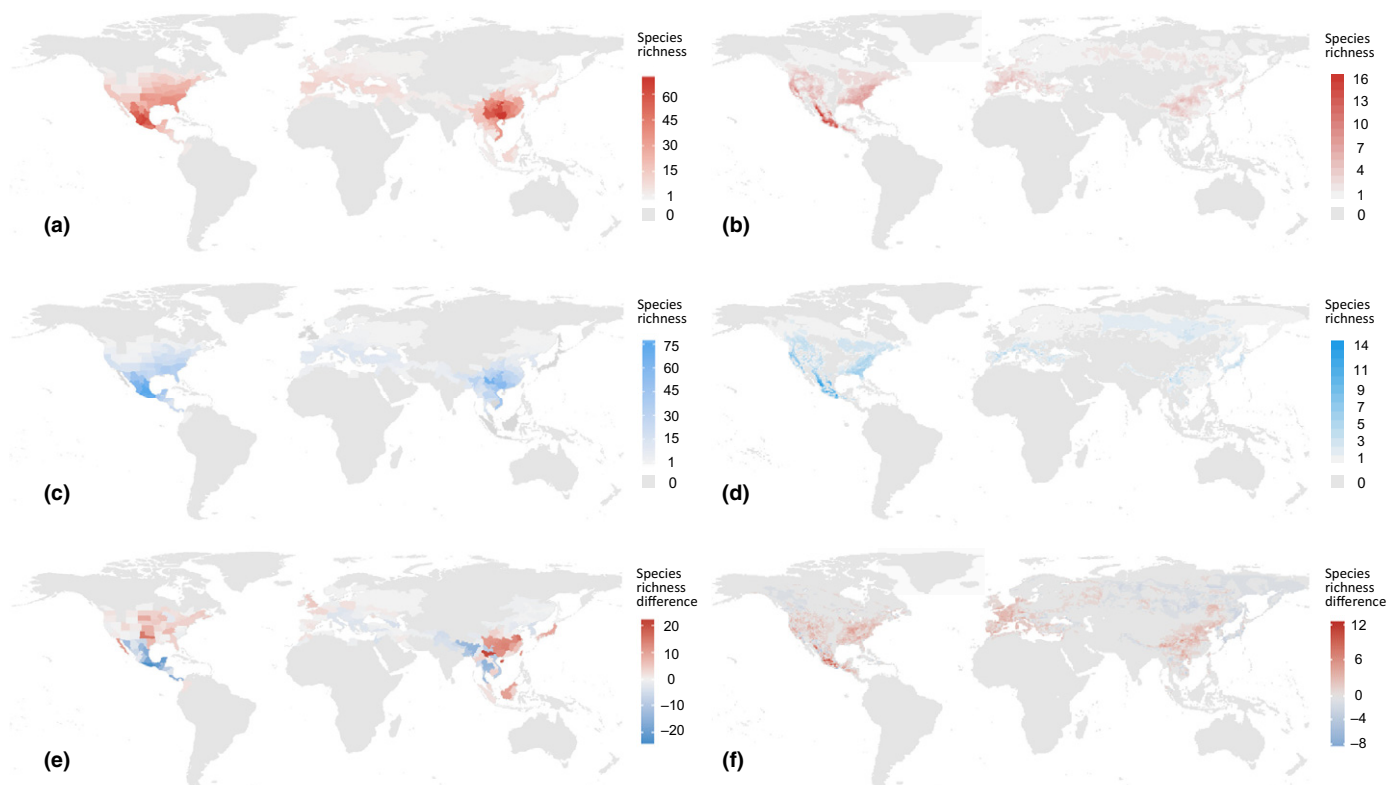
In the polygon parameter exploration, the 30 possible parameter combinations resulted in different numbers of species for which polygon maps could be generated and a different fraction of occurrences falling within the combined map (Table S7). The parameter combinations of outlier distance and minimum cluster size yielding the best performances regarding the number of species for which polygon maps could be generated were: (7°, 1), (5°, 1), (7°, 2), (3°, 1), (5°, 2), and (2°, 1) (the number of maps generated ranges from 198 to 188 for 200 species, Table S7). Among these parameter combinations, the fraction of occurrences falling inside the combined area did not differ significantly (0.87 on average,  $P$ -value = 0.969, Table S7).

### Species range maps and evaluation

In total, we generated 1957 species range maps, including 1141 maps based on combining polygon and SDM maps, 549 maps

generated only from polygons, and 267 maps using point buffers. The maps based on combinations of polygons and SDM covered 78% (median value) of the original polygon maps (Table S8), indicating that SDM results generally modified the polygon maps by removing regions of unsuitable habitat. We manually inspected the 1690 combined or polygon maps by comparing them to the references, and 93% of the maps showed a good match (with a rating of level 3 or level 4, Table S9).

We compared the resulting richness maps with previously published richness maps and found they matched similar patterns (Figs 3, S4). We compared *Quercus* richness maps against those presented by Xu *et al.* (2019) and found a similar pattern (Spearman's  $\rho = 0.83$ ) with no significant difference ( $t$ -value = 0.19,  $P$ -values = 0.84), and in 135 out of 180 regions the difference in richness between our maps and those of Xu *et al.* was < 6. This generally indicates a high level of agreement between the two sets of maps. Our approach produced more species in eastern North America, western Europe, and eastern and south-eastern Asia, but fewer species in western North America, Central America, eastern Europe, north-eastern Asia, and Himalayan regions. We found that the regions with the highest species diversity are located in south-central China, eastern North America, and Central America, where the difference between our richness map and that of Xu *et al.* (2019) was also largest (Fig. 3a,c,e). Comparing our richness maps with those generated by the maps of Critchfield & Little (1966), we found a similar pattern (Spearman's  $\rho = 0.73$ ), with the maps produced by our pipeline generating significantly more species ( $t$ -value = 16.28,  $P$ -values < 0.01). In



**Fig. 3** Biodiversity distribution maps of the species richness of *Quercus* (a) and *Pinus* (b). Region-wise species richness for *Quercus* (c) and *Pinus* (d) and the difference between the species richness and a previously published regional-level distribution database (e, f).

most regions of the world, our richness map had more species, and the largest richness difference was found in Central America, where there is a hotspot of *Pinus*. In some regions in eastern and north-eastern Asia, our richness map generated fewer species (Fig. 3b,d,f).

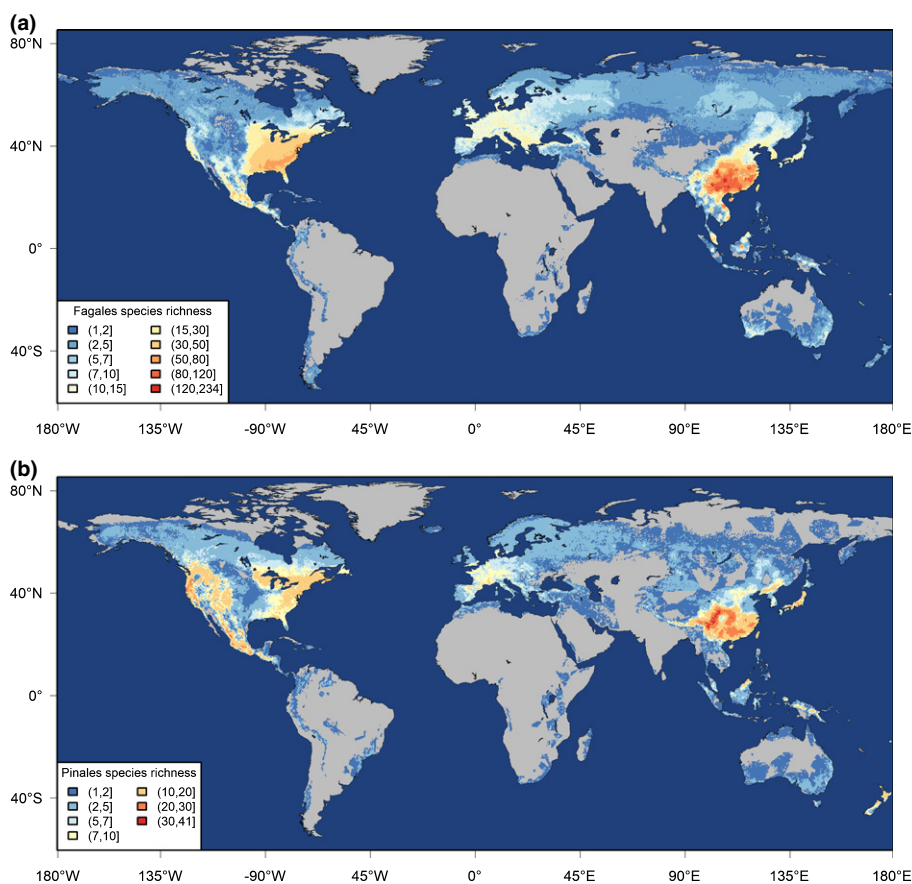
### Species richness patterns

We found that Fagales and Pinales have similar distribution patterns (Spearman's  $\rho = 0.77$ ). They are both distributed globally and have their main biodiversity centre in southern China (Fig. 4). For Fagales, secondary biodiversity centres are located in south-eastern North America, Central America, and Borneo. For Pinales, additional biodiversity centres are located along the west coast of North America, and in Central America, central Japan and New Caledonia (Fig. 4). Both Fagales and Pinales follow a latitudinal diversity gradient, with a peak in richness at around 30°N. Family-level distribution maps are available in Notes S4; Fig. S5.

### Discussion

Species range maps are central for fundamental research in macroecology and biogeography (Rocchini *et al.*, 2011), as well as for conservation and restoration programmes (Ferrier, 2002; Miller *et al.*, 2012). However, detailed distribution maps are still

lacking for most vascular plant species at a global scale. Previous efforts to map plant species ranges and diversity properties were limited to a few taxa (e.g. *Pinus*; Critchfield & Little, 1966), a specific geographic extent (e.g. China; Wang *et al.*, 2010), or used often simple polygon mapping at coarser resolution when applied globally (e.g. 1°, Guo *et al.*, 2020). Integration of different datasets including the increasing online databases (as listed in Table S1; Serra-Diaz *et al.*, 2017) opens possibilities for novel data science approaches to map plants globally. Here, we demonstrate that based on global database compilations, our comprehensive pipeline can transform the scattered distribution information into global distribution maps in batch-mode processing (Fig. 1). Notably, comparing to previous approaches, our algorithm can automatically identify and map ranges of different populations for species with disjunct distribution, by appropriate parameter settings, as well as optional filters such as bioregion partitions and environmental associations. Moreover, we propose a comprehensive SDM mapping algorithm composed of four modelling methods of differing complexity and seven pseudo-absence sampling strategies. Our open pipeline helps to map distribution ranges more accurately and allows to define settings that are specific to the characteristics of the target clades. It can thus be expected that our pipeline will boost the availability of species range maps for future research and conservation planning. Later, we discuss elements of the pipeline that may help users of the pipeline to optimize their applications, including data



**Fig. 4** Biodiversity distribution maps of the species richness of Fagales (a) and Pinales (b) mapped as 1 km × 1 km grid cells.



preparation and cleaning steps, and the core process of mapping species distributions.

### Biodiversity patterns of Fagales and Pinales

In this study, for the first time, we present high-resolution species richness maps for Fagales and Pinales, as well as for subordinate families (Figs 4, S5). Our tests against independent, regional distribution maps indicate that the species richness distributions for *Pinus* and *Quercus* were consistent with previous coarser mapping approaches (Fig. 3), demonstrating the power of our approach for future mapping of plant families at a comparably high spatial resolution. With the maps generated in this study, we found a congruent pattern between the two clades, especially the biodiversity in south-(west)ern China and Central America (Fig. 4). Shelter and cradle theory (López-Pujol *et al.*, 2011; Hipp *et al.*, 2018; Sosa *et al.*, 2018; Sundaram *et al.*, 2019), climate and environment heterogeneity (Qian *et al.*, 2007; Noss *et al.*, 2015; Dakhil *et al.*, 2021; also see Notes S5; Table S10 for a primary analysis), and deep-history tectonic events (Manos & Stanford, 2001; Svenning, 2003; Bouchal *et al.*, 2014; Xing *et al.*, 2014; Xing & Ree, 2017; Zheng *et al.*, 2018; Zhang *et al.*, 2021) have been proposed to explain the biodiversity hotspots and related patterns. However, due to the different evolutionary history of the two clades, further explorations are still needed to reveal the common or unique mechanisms behind these congruent patterns. Nevertheless, our visualization and mapping of species richness patterns provides insights for further studies of those clades.

### Data quality and mapping validation

We have acquired a unique collection of occurrence data from various data sources (Fig. 2; Table S1) to produce a compilation of species range maps for two important temperate tree clades. Among the compiled occurrences, about 9% were removed due to invalid species names and afterwards about 8% were removed due to incorrect distributions (Tables S3, S4), indicating the importance of data cleaning. The R-package COORDINATE-CLEANER helped us to clean about 20% of the invalid occurrences, removing dubious records without the need of a distribution reference checklist (e.g. species in Juglandaceae and Myricaceae), providing a useful tool for data cleaning. After the data cleaning and mapping steps, the validation with *Quercus* and *Pinus* indicated overall good performance of our mapping approach at a large scale. At smaller scales, more discrepancies were observed; for instance, the single data source of Critchfield & Little (1966) led to a smaller range area in most regions (Fig. 3b,d,f). Meanwhile, mapping bias may also introduce differences between the two sets of maps at small scales; for example, in northern Asia, the *Pinus* species in our maps have smaller distribution ranges (Fig. 3b,d,f), represented primarily by two species, *Pinus pumila* and *Pinus sibirica*. These two species are widely distributed across north-central Asia and north-eastern Asia, respectively. However, their ranges created by polygon mapping are small and scattered due to unevenness and a deficiency of occurrences (Fig. 2; Meyer *et al.*, 2016). In such cases of data deficiency, adjusting

parameters manually could help reduce this problem to some extent. Given the lack of high-resolution mapping of the two clades, it is impossible to quantitatively evaluate each of the species ranges at a fine scale, but the evaluation of the species range maps demonstrates that overall robust patterns are recovered (Fig. 3; Table S9).

### Parameter optimization

The development of mapping pipelines requires an optimization procedure to increase precision (Burgman & Fox, 2003; Chefaoui & Lobo, 2008; Barbet-Massin *et al.*, 2012; Li & Wang, 2013; Merow *et al.*, 2014; Meyer *et al.*, 2017), which guides the selection of optimal parameters. In particular, the complexity of SDM algorithms may strongly influence the results of suitability maps (Iturbide *et al.*, 2015; Merow *et al.*, 2017; Brun *et al.*, 2020). Brun *et al.*, (2020) found that intermediate parameterization complexity performed best, and model performance peaked at 10–11 variables. In our study, we observed that intermediate parameterization complexity in GLM (polynomial degree = 2) and GAM (degree of freedom = 5) had higher TSS values, while complex parameterization in GBM (maximum number of trees = 10 000) and RF (minimum node size = 1) scored higher (Table S6). Furthermore, the output of SDM is influenced by selected pseudo-absences (Iturbide *et al.*, 2015), as absences provide a contrast to presence data to indicate potentially unsuitable conditions (VanDerWal *et al.*, 2009). Different strategies were shown to result in different model performances (Table S6; Senay *et al.*, 2013). In our parameter optimization, we found that the three environmentally or geographically-stratified strategies and the random strategy all generally performed well at the global scale (Table S6). However, in a study on oak distribution in Europe, random sampling underestimated areas of high suitability because false absences introduced uncertainty, especially when occurrences failed to represent the realized niche (Chefaoui & Lobo, 2008; Iturbide *et al.*, 2015). To reduce false absences, environmentally or geographically weighted strategies have been proposed to generate pseudo-absences, which have proved to have better performance in classification and machine learning algorithms (Barbet-Massin *et al.*, 2012), as was applied in this study.

In contrast to the large number of studies on SDM, polygon (hull) mapping methods are generally applied in simple form and their optimization is not explored. A common parameter is to determine if the polygons around the occurrence points are connected (Bivand & Rundel, 2017). In the  $\alpha$ -hull method, an additional parameter  $\alpha$  is used to determine the disk radius (Rodríguez-Casal & López-Pateiro, 2010). Since hull methods are regularly criticized for their tendency to overestimate (Burgman & Fox, 2003; Graham & Hijmans, 2006; Meyer *et al.*, 2017) due to their simple parameterization, exploration of new parameters and their optimization are useful, especially for species with too few occurrences to create a SDM map. Here, we used a more complex approach than previously done in hull mapping methods by introducing two parameters: outlier distance and minimum size of a cluster. By exploring these parameters, we conclude that a large outlier distance may result in overestimated

ranges, as outliers are erroneously included in the range or a corridor is formed connecting the clusters, while a small outlier distance might generate more separated polygons that lead to a scattered distribution pattern, underestimating the real distribution range of a species. Further, a large value for the minimum size of a cluster may mistakenly remove the occurrences from a disjunct small population, while a small value may fail to remove outliers. Therefore, the optimization of these parameters was important in generating accurate maps. Since the majority of our species have relatively few occurrences (Fig. 2; Table S3) and the datasets were cleaned thoroughly, we ended up using an outlier distance of 5° and a small minimum cluster size of two, thereby keeping a large number of occurrence points (Table S7). However, for species with a large number of occurrences, this parameter set sometimes failed to remove outliers. For example, for *Abies balsamea*, a minimum cluster size set to < 7 would produce a polygon that includes the outliers in the western coast outside its natural range in eastern North America (POWO, 2019). Filtering based on environmental layers (e.g. elevation or temperature) can help remove unsuitable area in polygons, and especially in this study, polygon deficiency was generally solved after overlaying the SDM maps, which illustrates that combining the two mapping approaches enabled us to map species ranges more accurately compared with using individual approaches only.

### Challenges and future improvements

Besides methodological limitations, our mapping approach is impacted by incompleteness and uncertainties in the occurrence data. A first uncertainty is associated with the cleaning of presence data, which might not entirely remove problematic records (Zizka *et al.*, 2020). In particular, the records of nonnative species, especially those close to native ranges, may not always be successfully cleaned, which could cause overestimation of the species range. For instance, *Larix decidua* is native to central Europe and surrounding regions (POWO, 2019), but due to its widespread cultivation, the surrounding North Sea regions are also included in our reconstructed map.

A second uncertainty is associated with a lack of records. For instance, in this study, the accuracy of species distribution in Borneo should be further improved. Though it is widely accepted that Indonesia is a hotspot of plant biodiversity, the number of occurrence records collected there was relatively low, even after searching for datasets in several languages (Fig. 2; this data deficiency is also described by Collen *et al.* (2008); Raes *et al.* (2009) and Cahyaningsih *et al.* (2021)), which may have led to underestimated ranges and species numbers in this region.

Third, our approach is dependent on the quality of the independent reference checklist used for data cleaning. In Kew's regional-level distribution database, mainland China is only divided into nine regions: China North-Central, China South-Central, China Southeast, Hainan, Inner Mongolia, Manchuria, Qinghai, Tibet, and Xinjiang. North-Central China, Inner Mongolia, and South-Central China in particular cover extensive areas with high intra-regional environmental variability (Ren *et al.*, 2007). In contrast, Kew's database divides North and Central

America into much smaller regions, leading to more accurate validation of distribution maps in these regions. Therefore, a good reference checklist for data cleaning is important for enhancing map quality at finer scales. When working at smaller spatial scales, finer regional checklists are recommended to remove outliers.

While most maps are accurate, we identified some artefacts, particularly in maps of tropical and subtropical species, where ranges are overestimated or underestimated, or where artefactual linear range borders are observed, especially in *Lithocarpus* and *Quercus*. We expect that the main reason for these artefacts is insufficient data for rare and narrow-ranged species. Since our pipeline is automated and therefore easily applicable to data updates, future versions of the presented maps will increase in accuracy as data coverage increases. Solutions for supplementing and completing datasets might come from national forest inventories (Serra-Diaz *et al.*, 2017) and citizen science projects (e.g. iNaturalist and eBird; Bradter *et al.*, 2018), and a data merging workflow, such as the one developed in this study, could be used to add these data to already existing datasets. Furthermore, although our pipeline could reduce uncertainty, an improvement in public databases is still necessary and requires the support of taxonomists and improved AI identification technology.

### Conclusion

In conclusion, our study highlights the power of combining multiple occurrence and range datasets, as well as the crucial importance of improved data cleaning methods and the collection of additional data through innovative approaches in biodiversity science (e.g. citizen science projects and online observation reporting), for the global mapping of species distribution ranges. The maps generated here are provided to the scientific community open access, and future efforts will include expanding the mapping to more families and regularly updating existing maps as more data become available. The mapping approach developed here will further the field of macroecology and the study of global distribution patterns and may significantly aid future conservation efforts.

### Acknowledgements












The authors thank Benjamin Flück and Alexander Skeels from ETH Zürich, Yunyi Shen from the University of Wisconsin–Madison, Tong Lyu from Peking University, and Xiaoting Xu from Sichuan University for technical support; and Melissa Dawes for feedback and proofing the manuscript. This work was supported by a China Scholarship Council grant awarded to LL, a Swiss National Science Foundation grant awarded to LP (no. 310030\_188550) and an ETH postdoctoral fellowship granted to LMB.

### Author contributions

LL, LP, NEZ, FL and OH designed the research. LL, FL, JSS, ZW and DNK collected the data. LL, FL, FF, CA and PB

performed the data analysis. LL, LMB and LP wrote the manuscript with contributions from all authors.

## ORCID

Camille Albouy  <https://orcid.org/0000-0003-1629-2389>  
 Lydian M. Boschman  <https://orcid.org/0000-0002-1802-0187>  
 Philipp Brun  <https://orcid.org/0000-0002-2750-9793>  
 Fabian Fopp  <https://orcid.org/0000-0003-0648-8484>  
 Oskar Hagen  <https://orcid.org/0000-0002-7931-6571>  
 Dirk N. Karger  <https://orcid.org/0000-0001-7770-6229>  
 Flurin Leugger  <https://orcid.org/0000-0001-9027-6892>  
 Lisha Lyu  <https://orcid.org/0000-0001-7855-8109>  
 Loïc Pellissier  <https://orcid.org/0000-0002-2289-8259>  
 Joeri Sergej Strijk  <https://orcid.org/0000-0003-1109-7015>  
 Niklaus E. Zimmermann  <https://orcid.org/0000-0003-3099-9604>

## Data availability

The codes for data cleaning, parameter optimization, and mapping described in this study are openly available in GitLab at <https://gitlab.ethz.ch/gdplants/gdplants>. The modified environmental layers used for SDM mapping are openly available in EnviDat at <https://www.envidat.ch/dataset/sdm-env-layers-gdplants> (doi: 10.16904/envidat.309). The species distribution maps generated in this study are openly available in EnviDat at <https://www.envidat.ch/dataset/species-distribution-maps-gdplants> (doi: 10.16904/envidat.308).

## References

- Allouche O, Tsoar A, Kadmon R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223–1232.
- Araujo MB, Guisan A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33: 1677–1688.
- Araújo MB, Williams PH. 2000. Selecting areas for species persistence using occurrence data. *Biological Conservation* 96: 331–345.
- Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* 3: 327–338.
- Beck J, Böller M, Erhardt A, Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* 19: 10–15.
- Bellard C, Bertelsmeier C, Leadley P, Thuiller W, Courchamp F. 2012. Impacts of climate change on the future of biodiversity. *Ecology Letters* 15: 365–377.
- Bivand R, Rundel C. 2017. *RGEOS: interface to geometry engine-open source (GEOS)*. R package v.0.5-2. [WWW document] URL <https://CRAN.R-project.org/package=rgeos> [accessed 10 October 2019].
- Bland LM, Collen B, Orme CDL, Bielby J. 2015. Predicting the conservation status of data-deficient species. *Conservation Biology* 29: 250–259.
- Bouchal J, Zetter R, Grímsson F, Denk T. 2014. Evolutionary trends and ecological differentiation in early Cenozoic Fagaceae of western North America. *American Journal of Botany* 101: 1332–1349.
- Bradter U, Mair L, Jönsson M, Knappe J, Singer A, Snäll T. 2018. Can opportunistically collected Citizen Science data fill a data gap for habitat suitability models of less common species? *Methods in Ecology and Evolution* 9: 1667–1678.
- Breiman L. 2001. Random forests. *Machine Learning* 45: 5–32.
- Brodribb TJ, Pittermann J, Coomes DA. 2012. Elegance versus speed: examining the competition between conifer and angiosperm trees. *International Journal of Plant Sciences* 173: 673–694.
- Brun P, Thuiller W, Chauvier Y, Pellissier L, Wüest RO, Wang Z, Zimmermann NE. 2020. Model complexity affects species distribution projections under climate change. *Journal of Biogeography* 47: 130–142.
- Burgman MA, Fox JC. 2003. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation* 6: 19–28.
- Butchart SHM, Stattersfield AJ, Baillie J, Bennun LA, Stuart SN, Akçakaya HR, Hilton-Taylor C, Mace GM. 2005. Using red list indices to measure progress towards the 2010 target and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360: 255–268.
- Cahyaningsih R, Brehm JM, Maxted N. 2021. Gap analysis of Indonesian priority medicinal plant species as part of their conservation planning. *Global Ecology and Conservation* 26: e01459.
- Chamberlain S, Ram K, Barve V, Mcglinn D, Chamberlain MS. 2017. *RGBIF: interface to the global biodiversity information facility API*. R package v.1.2.0. [WWW document] URL <https://CRAN.R-project.org/package=rgbif> [accessed 20 February 2019].
- Cheek M, Nic Lughadha E, Kirk P, Lindon H, Carretero J, Looney B, Douglas B, Haelewaters D, Gaya E, Llewellyn T *et al.* 2020. New scientific discoveries: plants and fungi. *Plants, People, Planet* 2: 371–388.
- Chefaoui RM, Lobo JM. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling* 210: 478–486.
- Collen B, Ram M, Zamin T, McRae L. 2008. The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science* 1: 75–88.
- Critchfield WB, Little EL. 1966. *Geographic distribution of the pines of the world*. Washington, DC, USA: USDA Forest Service.
- Dakhil MA, Li J, Pandey B, Pan K, Liao Z, Olatunji OA, Zhang L, Eid EM, Abdelaal M. 2021. Richness patterns of endemic and threatened conifers in south-west China: topographic-soil fertility explanation. *Environmental Research Letters* 16: 34017.
- De Vos JM, Joppa LN, Gittleman JL, Stephens PR, Pimm SL. 2015. Estimating the normal background rate of species extinction. *Conservation Biology* 29: 452–462.
- Di Cola V, Broennimann O, Petitpierre B, Breiner FT, D'Amen M, Randin C, Engler R, Pottier J, Pio D, Dubuis A *et al.* 2017. *ECOSPAT: an R package to support spatial analyses and modeling of species niches and distributions*. *Ecography* 40: 774–787.
- Duputié A, Zimmermann NE, Chuine I. 2014. Where are the wild things? Why we need better data on species distribution. *Global Ecology and Biogeography* 23: 457–467.
- Di Febbraro M, Sallustio L, Vizzarri M, De Rosa D, De Lisio L, Loy A, Eichelberger BA, Marchetti M. 2018. Expert-based and correlative models to map habitat quality: which gives better support to conservation planning? *Global Ecology and Conservation* 16: e00513.
- eFloras. 2020. *eFloras*. St Louis, MO, USA and Cambridge, MA, USA: Missouri Botanical Garden and Harvard University Herbaria.
- Ferrier S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* 51: 331–363.
- Fraginière Y, Bétrisey S, Cardinaux L, Stoffel M, Kozłowski G. 2015. Fighting their last stand? A global analysis of the distribution and conservation status of gymnosperms. *Journal of Biogeography* 42: 809–820.
- Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29: 1189–1232.
- Govaerts R, Frodin DG. 1998. *World checklist and bibliography of Fagales*. Richmond, UK: Royal Botanic Gardens, Kew.
- Graham CH, Hijmans RJ. 2006. A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography* 15: 578–587.
- Greenwell B, Boehmke B, Cunningham J. 2018. *GBM: generalized boosted regression models*. R package v.2.1.5. [WWW document] URL <https://CRAN.R-project.org/package=gbm> [accessed 23 January 2019].



- Guisan A, Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8: 993–1009.
- Guo W-Y, Serra-Diaz JM, Schrödt F, Eiserhardt WL, Maitner BS, Merow C, Violle C, Anand M, Belluau M, Bruun HH *et al.* 2020. Half of the world's tree biodiversity is unprotected and is increasingly threatened by human activities. *bioRxiv* doi: 10.1101/2020.04.21.052464.
- Hagen O, Vaterlaus L, Albouy C, Brown A, Leugger F, Onstein RE, de Santana CN, Scotese CR, Pellissier L. 2019. Mountain building, climate cooling and the richness of cold-adapted plants in the Northern Hemisphere. *Journal of Biogeography* 46: 1792–1807.
- Harrell FE Jr, Lee KL, Mark DB. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15: 361–387.
- Hastie T. 2018. *Package 'GAM'*. R package v.1.16.1. [WWW document] URL <https://CRAN.R-project.org/package=gam> [accessed January 2019].
- Hastie TJ, Tibshirani RJ. 1990. *Generalized additive models*. Boca Raton, FL, USA: CRC Press.
- Heller NE, Zavaleta ES. 2009. Biodiversity management in the face of climate change: a review of 22 years of recommendations. *Biological Conservation* 142: 14–32.
- Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, Ribeiro E, Samuel-Rosa A, Kempen B, Leenaars JGB, Walsh MG *et al.* 2014. SoilGrids1km—Global soil information based on automated mapping. *PLoS ONE* 9: e105992.
- Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B *et al.* 2017. SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE* 12: e0169748.
- Hipp AL, Manos PS, González-Rodríguez A, Hahn M, Kaproth M, McVay JD, Avalos SV, Cavender-Bares J. 2018. Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. *New Phytologist* 217: 439–452.
- Huamán Z, Hoekstra R, Bamberg JB. 2000. The inter-genebank potato database and the dimensions of available wild potato germplasm. *American Journal of Potato Research* 77: 353–362.
- Hurlbert AH, Jetz W. 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences, USA* 104: 13384–13389.
- IPCC. 2019. *IPCC special report on the ocean and cryosphere in a changing climate*. [WWW document] URL <https://www.ipcc.ch/srocc/> [accessed 1 January 2020].
- Iturbide M, Bedia J, Herrera S, del Hierro O, Pinto M, Gutiérrez JM. 2015. A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling* 312: 166–174.
- Kadmon R, Farber O, Danin A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* 14: 401–413.
- Karger DN, Conrad O, Böhrer J, Kawohl T, Kreft H, Soria-Auza RW, Zimmermann NE, Linder HP, Kessler M. 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific Data* 4: 170122.
- Kier G, Mutke J, Dinerstein E, Ricketts TH, Küper W, Kreft H, Barthlott W. 2005. Global patterns of plant diversity and floristic knowledge. *Journal of Biogeography* 32: 1107–1116.
- Kreft H, Jetz W. 2007. Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences, USA* 104: 5925–5930.
- Li X, Wang Y. 2013. Applying various algorithms for species distribution modelling. *Integrative Zoology* 8: 124–135.
- Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* 2: 18–22.
- López-Pujol J, Zhang F-MM, Sun H-QQ, Ying T-SS, Ge S. 2011. Centres of plant endemism in China: places for survival or for speciation? *Journal of Biogeography* 38: 1267–1280.
- Maitner BS, Boyle B, Casler N, Condit R, Donoghue J, Durán SM, Guaderrama D, Hinchliff CE, Jørgensen PM, Kraft NJB *et al.* 2018. The BIEN R package: a tool to access the botanical information and ecology network (BIEN) database. *Methods in Ecology and Evolution* 9: 373–379.
- Manos PS, Stanford AM. 2001. The historical biogeography of Fagaceae: tracking the tertiary history of temperate and subtropical forests of the Northern Hemisphere. *International Journal of Plant Sciences* 162: S77–S93.
- Merow C, Smith MJ, Edwards TC, Guisan A, McMahon SM, Normand S, Thuiller W, Wüest RO, Zimmermann NE, Elith J. 2014. What do we gain from simplicity versus complexity in species distribution models? *Ecography* 37: 1267–1281.
- Merow C, Wilson AM, Jetz W. 2017. Integrating occurrence data and expert maps for improved species range predictions. *Global Ecology and Biogeography* 26: 243–258.
- Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19: 992–1006.
- Meyer L, Diniz-Filho JAF, Lohmann LG. 2017. A comparison of hull methods for estimating species ranges and richness maps. *Plant Ecology and Diversity* 10: 389–401.
- Meyer WB, Meyer WB, Turner BL II. 1994. *Changes in land use and land cover: a global perspective*. Cambridge, UK: Cambridge University Press.
- Miller JS, Porter-Morgan HA, Stevens H, Boom B, Krupnick GA, Acevedo-Rodríguez P, Fleming J, Gensler M. 2012. Addressing target two of the global strategy for plant conservation by rapidly identifying plants at risk. *Biodiversity and Conservation* 21: 1877–1887.
- Morueta-Holme N, Enquist BJ, McGill BJ, Boyle B, Jørgensen PM, Ott JE, Peet RK, Šimová I, Sloat LL, Thiers B *et al.* 2013. Habitat area and climate stability determine geographical variation in plant species range sizes. *Ecology Letters* 16: 1446–1454.
- Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A: General* 135: 370–384.
- Noss RF, Platt WJ, Sorrie BA, Weakley AS, Means DB, Costanza J, Peet RK. 2015. How global biodiversity hotspots may go unrecognized: lessons from the North American Coastal Plain. *Diversity and Distributions* 21: 236–244.
- Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 344: 1246752.
- Pimm SL, Joppa LN. 2015. How many plant species are there, where are they, and at what rate are they going extinct? *Annals of the Missouri Botanical Garden* 100: 170–176.
- Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, Vesk PA, McCarthy MA. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution* 5: 397–406.
- POWO. 2019. *Plants of the world online*. Facilitated by the Royal Botanic Gardens, Kew. [WWW document] URL <http://www.plantsoftheworldonline.org/> [accessed 1 February 2019].
- Qian H, White PS, Song J-S. 2007. Effects of regional vs. ecological factors on plant species richness: an intercontinental analysis. *Ecology* 88: 1440–1453.
- R Core Team. 2013. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [WWW document] URL <https://www.R-project.org/> [accessed 1 October 2018].
- Raes N, Roos MC, Slik JWF, Van Loon EE, Ter Steege H. 2009. Botanical richness and endemism patterns of Borneo derived from species distribution models. *Ecography* 32: 180–192.
- Rahbek C, Graves GR. 2001. Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences, USA* 98: 4534–4539.
- Ren H, Shen W-J, Lu H-F, Wen X-Y, Jian S-G. 2007. Degraded ecosystems in China: status, causes, and restoration efforts. *Landscape and Ecological Engineering* 3: 1–13.
- Rocchini D, Hortal J, Lengyel S, Lobo JM, Jiménez-Valverde A, Ricotta C, Bacaro G, Chiarucci A, Jimenez-Valverde A, Ricotta C *et al.* 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography* 35: 211–226.
- Rodríguez-Casal A, López-Pateiro B. 2010. Generalizing the convex hull of a sample: the R package ALPHAHULL. *Journal of Statistical Software* 34: 1–28.
- Senay SD, Wörner SP, Ikeda T. 2013. Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE* 8: e71218.
- Serra-Diaz JM, Enquist BJ, Maitner B, Merow C, Svenning JC. 2017. Big data of tree species distributions: how big and how good? *Forest Ecosystems* 4: 1–12.



- Sosa V, De-Nova JA, Vásquez-Cruz M. 2018. Evolutionary history of the flora of Mexico: dry forests cradles and museums of endemism. *Journal of Systematics and Evolution* 56: 523–536.
- Sundaram M, Donoghue MJ, Farjon A, Filer D, Mathews S, Jetz W, Leslie AB, Gardens RB, Tw S, Farjon A *et al.* 2019. Accumulation over evolutionary time as a major cause of biodiversity hotspots in conifers. *Proceedings of the Royal Society B: Biological Sciences* 286: 1–8.
- Svenning JC. 2003. Deterministic Plio-Pleistocene extinctions in the European cool-temperate tree flora. *Ecology Letters* 6: 646–653.
- The Nature Conservancy. 2009. *Global ecoregions, major habitat types, biogeographical realms and the nature conservancy terrestrial assessment units*. [WWW document] URL <https://geospatial.tnc.org/datasets/b1636d640ede4d6ca8f5e369f2dc368b/about> [accessed 1 October 2018].
- Tittensor DP, Walpole M, Hill SLL, Boyce DG, Britten GL, Burgess ND, Butchart SHM, Leadley PW, Regan EC, Alkemade R *et al.* 2014. A mid-term analysis of progress toward international biodiversity targets. *Science* 346: 241–244.
- USDA, NRCS. 2020. *The PLANTS database*. Greensboro, NC, USA: National Plant Data Team.
- VanDerWal J, Shoo LP, Graham C, Williams SE. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling* 220: 589–594.
- Vasconcelos TS, Rodríguez MÁ, Hawkins BA. 2012. Species distribution modelling as a macroecological tool: a case study using New World amphibians. *Ecography* 35: 539–548.
- Walther G-R, Post E, Convey P, Menzel A, Parmesan C, Beebee TJC, Fromentin J-M, Hoegh-Guldberg O, Bairlein F. 2002. Ecological responses to recent climate change. *Nature* 416: 389–395.
- Wang XQ, Ran JH. 2014. Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution* 75: 24–40.
- Wang Z, Fang J, Tang Z, Lin X. 2010. Patterns, determinants and models of woody plant diversity in China. *Proceedings of the Royal Society B: Biological Sciences* 278: 2122–2132.
- Weigelt P, König C, Kreft H. 2020. GIFT – a global inventory of floras and traits for macroecology and biogeography. *Journal of Biogeography* 47: 16–43.
- Wisn MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A; NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14: 763–773.
- Wüest RO, Zimmermann NE, Zurell D, Alexander JM, Fritz SA, Hof C, Kreft H, Normand S, Cabral JS, Szekely E *et al.* 2020. Macroecology in the age of big data – where to go from here? *Journal of Biogeography* 47: 1–12.
- Xing Y, Onstein RE, Carter RJ, Stadler T, Peter LH. 2014. Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity, and turnover rates are disconnected. *Evolution* 68: 2821–2832.
- Xing Y, Ree RH. 2017. Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. *Proceedings of the National Academy of Sciences, USA* 114: E3444–E3451.
- Xu W-B, Guo W-Y, Serra-Diaz JM, Schrodt F, Eiserhardt WL, Enquist BJ, Maitner BS, Merow C, Violle C, Anand M *et al.* 2020. Quaternary climate change explains global patterns of tree beta-diversity. *bioRxiv* doi: 10.1101/2020.11.14.382846.
- Xu X, Dimitrov D, Shrestha N, Rahbek C, Wang Z. 2019. A consistent species richness–climate relationship for oaks across the Northern Hemisphere. *Global Ecology and Biogeography* 28: 1051–1066.
- Yang Y, Wang Z, Xu X. 2017. *Taxonomy and distribution of global gymnosperms*. Shanghai, China: Shanghai Scientific and Technical Publishers.
- Zhang Q, Ree RH, Salamin N, Xing Y, Silvestro D. 2021. Fossil-informed models reveal a boreotropical origin and divergent evolutionary trajectories in the walnut family (Juglandaceae). *Systematic Biology* 71: 242–258.
- Zheng W, Zeng W, Tang Y, Shi W, Cao K. 2018. Species diversity and biogeographical patterns of Lauraceae and Fagaceae in northern tropical and subtropical regions of China. *Shengtai Xuebao/Acta Ecologica Sinica* 38: 8676–8687.
- Zizka A, Antunes Carvalho F, Calvente A, Rocio Baez-Lizarazo M, Cabral A, Coelho JFR, Colli-Silva M, Fantinati MR, Fernandes MF, Ferreira-Araújo T *et al.* 2020. No one-size-fits-all solution to clean GBIF. *PeerJ* 8: e9916.
- Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R *et al.* 2019. COORDINATECLEANER: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 10: 744–751.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

**Fig. S1** Generalized additive model (GAM) fits of correlations between the number of occurrences and the true skill statistic (TSS) values of seven pseudo-absence strategies (random, target-group, geographic, density, geographically stratified, environmentally stratified, environmentally semi-stratified).

**Fig. S2** True skill statistic (TSS) rankings of the seven pseudo-absence sampling strategies (random, target-group, geographic, density, geographically stratified, environmentally stratified, environmentally semi-stratified).

**Fig. S3** Bioregion partitioning by The Nature Conservancy (2009).

**Fig. S4** Region-wise species richness for Fagales and Pinales and the difference between the species richness and the Kew's regional-level distribution database.

**Fig. S5** Biodiversity maps of Betulaceae, Casuarinaceae, Fagaceae, Juglandaceae, Myricaceae, Nothofagaceae, and Tico-dendraceae (Fagales), and of Araucariaceae, Cephalotaxaceae, Cupressaceae, Phyllocladaceae, Pinaceae, Podocarpaceae, Sciadopytiaceae, Taxaceae (Pinales).

**Notes S1** Pseudo-absence strategies.

**Notes S2** The effect of occurrence numbers on pseudo-absence strategies.

**Notes S3** Details of map validation.

**Notes S4** Family-level distribution patterns.

**Notes S5** A primary analysis between species richness and environmental factors.

**Table S1** Databases and operations used to access occurrences, and numbers of accessible occurrences of Fagales and Pinales from each database.

**Table S2** Summary of species name processing and maps generated of Betulaceae, Casuarinaceae, Fagaceae, Juglandaceae, Myricaceae, Nothofagaceae, and Tico-dendraceae (Fagales) and of Araucariaceae, Cupressaceae, Pinaceae, Podocarpaceae, Sciadopytiaceae, and Taxaceae (Pinales).

**Table S3** Summary of occurrence processing of species of Fagales and Pinales.

**Table S4** Summary of occurrence record processing of Betulaceae, Casuarinaceae, Fagaceae, Juglandaceae, Myricaceae, Nothofagaceae, and Ticodendraceae (Fagales) and of Araucariaceae, Cupressaceae, Pinaceae, Podocarpaceae, Sciadopityaceae, and Taxaceae (Pinales).

**Table S5** Missing species and accessible supplementary references.

**Table S6** True skill statistic (TSS) values of seven different pseudo-absence strategies (random, target-group, geographic, density, geographically stratified, environmentally stratified, environmentally semi-stratified) and five model complexity levels for each of the generalized linear models (GLMs), generalized additive models (GAMs), generalized boosting machines (GBMs), and random forest (RF) models.

**Table S7** Polygon mapping parameter optimization based on the number of species for which polygon maps can be generated and

the number of occurrences falling within the overlap between range polygons and species distribution modelling (SDM) maps.

**Table S8** Coherence between polygons and combined maps of Fagales and Pinales.

**Table S9** Grades of species range mapping quality.

**Table S10** Correlation between the species richness of Fagales and Pinales and environmental variables (average annual temperature, aridity, frost change frequency, precipitation in the driest quarter, mean diurnal temperature range, and precipitation seasonality; soil organic carbon content, pH, and clay content; temperature change since the Last Glacial Maximum; topographic heterogeneity).

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.