





# Global models and predictions of plant diversity based on advanced machine learning techniques

Lirong Cai<sup>1</sup> , Holger Kreft<sup>1,2</sup> , Amanda Taylor<sup>1</sup> , Pierre Denelle<sup>1</sup> , Julian Schrader<sup>1,3</sup> , Franz Essl<sup>4</sup> , Mark van Kleunen<sup>5,6</sup> , Jan Pergl<sup>7</sup> , Petr Pyšek<sup>7,8</sup> , Anke Stein<sup>5</sup>, Marten Winter<sup>9</sup> , Julie F. Barcelona<sup>10</sup> , Nicol Fuentes<sup>11</sup> , Inderjit<sup>12</sup> , Dirk Nikolaus Karger<sup>13</sup> , John Kartesz<sup>14</sup>, Andreij Kuprijanov<sup>15</sup>, Misako Nishino<sup>14</sup>, Daniel Nickrent<sup>16</sup> , Arkadiusz Nowak<sup>17,18</sup> , Annette Patzelt<sup>19</sup> , Pieter B. Pelser<sup>10</sup> , Paramjit Singh<sup>20</sup> , Jan J. Wieringa<sup>21</sup>  and Patrick Weigelt<sup>1,2,22</sup> 

<sup>1</sup>Biodiversity, Macroecology and Biogeography, University of Göttingen, 37077 Göttingen, Germany; <sup>2</sup>Centre of Biodiversity and Sustainable Land Use, University of Göttingen, 37077 Göttingen, Germany; <sup>3</sup>School of Natural Sciences, Macquarie University, 2109 Sydney, NSW, Australia; <sup>4</sup>Bioinvasions, Global Change, Macroecology-Group, University of Vienna, 1030 Vienna, Austria; <sup>5</sup>Ecology, Department of Biology, University of Konstanz, 78464 Konstanz, Germany; <sup>6</sup>Zhejiang Provincial Key Laboratory of Plant Evolutionary Ecology and Conservation, Taizhou University, 318000 Taizhou, China; <sup>7</sup>Department of Invasion Ecology, Czech Academy of Sciences, Institute of Botany, 25243, Průhonice, Czech Republic; <sup>8</sup>Department of Ecology, Faculty of Science, Charles University, 12844 Prague, Czech Republic; <sup>9</sup>German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany; <sup>10</sup>School of Biological Sciences, University of Canterbury, 8140 Christchurch, New Zealand; <sup>11</sup>Departamento de Botánica, Facultad de Ciencias Naturales y Oceanográficas, Universidad de Concepción, 4030000 Concepción, Chile; <sup>12</sup>Department of Environmental Studies and Centre for Environmental Management of Degraded Ecosystems (CEMDE), University of Delhi, 110007 Delhi, India; <sup>13</sup>Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 8903 Birmensdorf, Switzerland; <sup>14</sup>Biota of North America Program (BONAP), Chapel Hill, NC 27516, USA; <sup>15</sup>650065 Kemerovo, Russia; <sup>16</sup>Plant Biology Section, School of Integrative Plant Science, College of Agriculture and Life Science, Cornell University, Ithaca, NY 14853, USA; <sup>17</sup>Department of Botany and Nature Protection, University of Warmia and Mazury in Olsztyn, 10-728 Olsztyn, Poland; <sup>18</sup>PAS Botanical Garden, 02-973 Warszawa, Poland; <sup>19</sup>Hochschule Weihenstephan-Triesdorf, University of Applied Sciences, Vegetation Ecology, 85354 Freising, Germany; <sup>20</sup>Mohali, 140301 Punjab, India; <sup>21</sup>Naturalis Biodiversity Center, 2333 CR Leiden, the Netherlands; <sup>22</sup>Campus-Institut Data Science, 37077 Göttingen, Germany

## Summary

Authors for correspondence:  
Lirong Cai  
Email: lcai@uni-goettingen.de

Patrick Weigelt  
Email: pweigelt@uni-goettingen.de

Received: 13 July 2022  
Accepted: 29 September 2022

New Phytologist (2023) 237: 1432–1445  
doi: 10.1111/nph.18533

**Key words:** biodiversity, diversity–environment models, phylogenetic diversity, species richness, vascular plants.

- Despite the paramount role of plant diversity for ecosystem functioning, biogeochemical cycles, and human welfare, knowledge of its global distribution is still incomplete, hampering basic research and biodiversity conservation.
- Here, we used machine learning (random forests, extreme gradient boosting, and neural networks) and conventional statistical methods (generalized linear models and generalized additive models) to test environment-related hypotheses of broad-scale vascular plant diversity gradients and to model and predict species richness and phylogenetic richness worldwide. To this end, we used 830 regional plant inventories including c. 300 000 species and predictors of past and present environmental conditions.
- Machine learning showed a superior performance, explaining up to 80.9% of species richness and 83.3% of phylogenetic richness, illustrating the great potential of such techniques for disentangling complex and interacting associations between the environment and plant diversity. Current climate and environmental heterogeneity emerged as the primary drivers, while past environmental conditions left only small but detectable imprints on plant diversity.
- Finally, we combined predictions from multiple modeling techniques (ensemble predictions) to reveal global patterns and centers of plant diversity at multiple resolutions down to 7774 km<sup>2</sup>. Our predictive maps provide accurate estimates of global plant diversity available at grain sizes relevant for conservation and macroecology.

## Introduction

Vascular plants comprise well over 340 000 species (Govaerts *et al.*, 2021) and are fundamental to terrestrial ecosystems maintaining ecosystem functioning (Tilman *et al.*, 2014) and providing ecosystem services (Isbell *et al.*, 2011; Cardinale *et al.*, 2012).

To preserve and manage this important part of global biodiversity, knowledge of its spatial distribution and location of biodiversity centers is critical. Mapping plant distributions and diversity has a long and rich tradition starting in the 19<sup>th</sup> century, with the collation of regional plant species numbers and expert-drawn isolines of species richness (Wulff, 1935; Barthlott

Although it is widely accepted that plant diversity reflects the complex interplay of evolutionary, geological, and ecological processes, disentangling the drivers of global plant diversity remains an important topic of modern macroecology (Kreft & Jetz, 2007; Tietje *et al.*, 2022). Several hypotheses related to geography, past and present climate, and environmental heterogeneity of a region have been proposed to explain plant diversity patterns (Currie *et al.*, 2004; Mittelbach *et al.*, 2007; Fine, 2015; Supporting Information Table S1). Large and heterogeneous areas, for example, are hypothesized to support more species by offering a greater diversity of resources and habitats, thus promoting species coexistence (Connor & McCoy, 1979) and offering refugia during environmental fluctuations (Stein *et al.*, 2014). Also, areas with warm, wet, and relatively stable climates such as humid tropical forests should support more species owing to high speciation (Rohde, 1992; Mittelbach *et al.*, 2007; Brown, 2014) and low extinction rates (Gillooly & Allen, 2007; Eiserhardt *et al.*, 2015). Geographic isolation could simultaneously promote extinction (Brown & Kodric-Brown, 1977; Ouborg, 1993) and speciation (Kisel & Barraclough, 2010), by making populations less well-connected. Finally, historical processes like past plate tectonics and climatic change have influenced diversity patterns through altered biotic isolation and exchange or species range shifts (Dynesius & Jansson, 2000; Svenning *et al.*, 2015;

Couvreur *et al.*, 2021). However, past environmental conditions remain underrepresented in global models of plant diversity and their legacies in modern plant distributions are still poorly understood (Kissling *et al.*, 2012; Hagen *et al.*, 2021).

Diversity–environment relationships are often complex, nonlinear, and scale-dependent (Francis & Currie, 2003; Keil & Chase, 2019). Many environmental predictors interact and show high levels of collinearity, thus presenting major challenges for conventional statistical models such as generalized linear models (GLMs) and generalized additive models (GAMs). Machine learning approaches represent powerful modeling tools that can effectively deal with multidimensional and correlated data and can reveal nonlinear relationships and interactions of predictors without *a priori* specification (Olden *et al.*, 2008; Crisci *et al.*, 2012). Therefore, machine learning has become a promising alternative to conventional techniques in ecology (Hengl *et al.*, 2017; Park *et al.*, 2020; Sabatini *et al.*, 2022). However, its performance in modeling global plant diversity has yet to be explored. In addition to relying on one particular model type, combining predictions based on multiple modeling techniques (i.e. ensemble predictions) might decrease prediction uncertainties (Araújo & New, 2007) and can thereby further improve predictions of global plant diversity patterns.

Here, we present improved models and predictions of two key facets of vascular plant diversity, that is, species richness and phylogenetic richness, at a global extent using advanced statistical modeling techniques. In addition to nonspatial and spatial GLMs and GAMs, we systematically assess the predictive performance of machine learning methods, including random forests, extreme gradient boosting (XGBoost), and neural networks. Specifically, our aims are as follows: to compare the performance of different modeling techniques in revealing complex diversity–environment relationships and to improve global geo-statistical plant diversity models; to test hypotheses on plant diversity gradients related to geography, environmental heterogeneity, current climate, and past environmental conditions, and to quantify their relative importance for plant species and phylogenetic richness; and, to predict both facets of plant diversity at multiple grain sizes across the globe. Our study is based on *c.* 300 000 species from checklists and floras for 830 regions across the globe (Fig. S1) collated in the Global Inventory of Floras and Traits (Weigelt *et al.*, 2020; GIFT; Notes S1), and a large, dated megaphylogeny of vascular plants (Jin & Qian, 2019).

## Materials and Methods

## Species distribution data and species richness

To calculate species and phylogenetic richness, we used the species composition of native vascular plants in regional checklists and floras from GIFT (Weigelt *et al.*, 2020; v.2.1: <http://gift.uni-goettingen.de>). In GIFT, all nonhybrid species names are standardized and validated based on taxonomic information provided by The Plant List (v.1.1, <http://www.theplantlist.org>) and additional resources available via iPlant's Taxonomic Name

Resolution Service (TNRS; Boyle *et al.*, 2013; Weigelt *et al.*, 2020). The original database contains > 3000 geographic regions representing islands, protected areas, biogeographical regions, and administrative units (e.g. countries and provinces). We excluded regions with incomplete native vascular plant checklists, incomplete data for predictor variables, or an area of < 100 km<sup>2</sup>. Furthermore, we coped with overlapping regions in two steps. First, for overlapping regions from one individual literature source, we kept only nonoverlapping regions preferring smaller over larger regions (e.g. the individual states of Brazil instead of the country). Second, for overlapping regions from different literature sources, we retained smaller and larger regions if smaller regions covered only parts of the larger regions. Otherwise, we removed the larger regions. A total of 298 087 vascular plant species from 775 mainland regions and 55 islands or island groups were used to proceed with the calculation of species richness (i.e. taxonomic richness) and phylogenetic richness. The geographic regions in the dataset were distributed representatively across the entire globe, covering all major biomes (Fig. S1).

### Phylogeny reconstruction and phylogenetic richness

We used a large, dated megatree of vascular plants, GBOTB\_extended (Jin & Qian, 2019), as a backbone to generate a phylogeny for all species in the dataset. The megatree was derived from the GBOTB tree for seed plants by Smith & Brown (2018) and the phylogeny for pteridophytes in Zanne *et al.* (2014). We excluded taxa not identified to the species level for calculating phylogenetic richness, leading to a dataset including 295 417 species in 466 families of vascular plants. All families and 10 128 out of 14 962 genera (67.7%) in the dataset were included in the megatree. We bound the remaining genera and species into their respective families and genera using 'Scenario 3' in the R package V.PHYLOMAKER (Jin & Qian, 2019). In 'Scenario 3', the weighted positioning of the additional taxa depends on the length and amount of already existing tips per taxon. 91.95% out of the 295 417 species in the dataset were from genera already present in the backbone. It is suggested that patterns of phylogenetic richness are similar regardless of whether the phylogeny used is resolved at the genus or species level (Qian & Jin, 2021). To test for the effect of adding missing genera to the phylogeny on phylogenetic richness, we carried out a sensitivity analysis and found consistent patterns, indicating that our method is robust (see Methods S1 for details).

Several indices exist for capturing different dimensions of phylogenetic diversity including richness, divergence, and regularity (Tucker *et al.*, 2017). Here, we focus on phylogenetic richness, which represents the amount of unique phylogenetic history present in an assemblage (Tucker *et al.*, 2017). We chose Faith's PD, a common measure of phylogenetic richness, calculated as the sum of the branch lengths of all species coexisting in a region (Faith, 1992), which is directly comparable to species richness. Even though highly correlated with species richness (Pearson's  $r = 0.98$ ), we did not standardize phylogenetic richness (i.e. assessing the deviation of phylogenetic richness from expectations

based on species richness) in our main analyses as we were not interested in whether the phylogenetic structure of a region is overdispersed or clustered, but rather aimed to capture both taxonomic and phylogenetic aspects of plant diversity. However, we present an analysis on the drivers of deviations in phylogenetic richness from species richness in Table S2.

### Predictor variables

We identified a set of candidate predictor variables hypothesized to affect plant distributions and diversity and classified them into four categories: geography, current climate, environmental heterogeneity, and past environmental conditions. Twenty-five predictors were considered in the original dataset (Table S1). These have been shown or hypothesized to contribute to geographic patterns of plant diversity in previous studies (Kreft & Jetz, 2007; Kissling *et al.*, 2012; Stein *et al.*, 2014; Keil & Chase, 2019). Geographic variables were region area (km<sup>2</sup>) and the summed proportion of landmass area in the surrounding area of the target region within buffer distances of 100, 1000, and 10 000 km, serving as a measure of geographic isolation (Weigelt & Kreft, 2013). Current climatic variables included 13 biologically relevant temperature and precipitation variables. These variables represent annual averages, seasonality, and limiting climatic factors (e.g. length of the growing season), capturing the main aspects of climate important for plant diversity (Karger *et al.*, 2017). Furthermore, gross primary productivity (Zhao & Running, 2010) was included as a measure of potential plant productivity based on available solar energy and water. Climatic variables were extracted as mean values across the input raster layers per region. The number of soil types (Hengl *et al.*, 2017) and elevational range (Danielson & Gesch, 2011) were calculated for each region as proxies for environmental heterogeneity within regions.

To determine the contribution of past environmental conditions to modern diversity patterns, we calculated biome area variation since the Pliocene and the Middle Miocene, temperature anomaly since the mid-Pliocene warm period, temperature stability since the last glacial maximum (LGM), and velocity of temperature change since the LGM. Terrestrial biomes are affected by multiple drivers containing atmospheric circulation, precipitation, and temperature patterns, and thus, changes in biome distributions represent major environmental changes through geological time. To calculate biome area variation, we used biome distribution maps at present (Olson *et al.*, 2001), the LGM (*c.* 25–15 ka; Ray & Adams, 2001), the mid-Pliocene warm period (mid-Piacenzian, *c.* 3.264–3.025 Ma; Dowsett *et al.*, 2016), and the Middle Miocene (*c.* 17–15 Ma; Henrot *et al.*, 2010). The three paleo-time periods represented particularly different climates compared with present-day conditions and showed distinct biome distributions, which are hypothesized to have left imprints on current plant diversity (Svenning *et al.*, 2015; Sandel *et al.*, 2020). As biome definitions differed across the four datasets, we regrouped biomes to match across datasets and then calculated biome area changes (see Methods S2 for details; Table S3). We acknowledge the potential drawbacks



of this approach due to the coarse resolution and uncertainty of the original past biome maps. Because of the coarse resolution of the Middle Miocene map and absent data for some geographic regions, we used biome area variation only since the Pliocene and excluded Miocene biome variation from further analyses.

In addition, we calculated temperature stability from two paleo-time periods until present, that is, the LGM and the mid-Pliocene warm period, representing cooler and warmer climates than the current climate, respectively. Temperature stability since the LGM was calculated using the `CLIMATESTABILITY` R package (Owens & Guralnick, 2019). It takes temperature differences between 1000 yr time slices expressed as standard deviation and averages the results across all time slices. The stability is then calculated as the inverse of the mean standard deviation rescaled to (0,1). Temperature anomaly since the mid-Pliocene was calculated as the difference in mean annual temperature between the mid-Pliocene warm period and present day. The velocity of temperature change since the LGM was calculated as the ratio between temporal change and contemporary spatial change in temperature, representing the speed with which a species would have to move its range to track analogous climatic conditions (Sandel *et al.*, 2011). For details on paleoclimate estimates, see Methods S2.

An alternative way to evaluate the effects of biogeographic history on plant diversity is to account for predefined discrete geographic regions influencing diversity via differences in diversification history and dispersal barriers. We therefore included floristic kingdoms (Takhtajan, 1986) as an additional categorical variable in the models and compared the performance of models with and without floristic kingdoms to assess whether we managed to model the effect of biogeographic history properly by only including the variables that directly quantify past environmental change.

## Statistical models

**Predictor variable selection** To quantify diversity–environment relationships, we fitted five different types of models with species richness and phylogenetic richness as response variables: GLMs, GAMs, random forests, XGBoost, and neural networks. To compare model performance across model types, we used the same set of predictors across models. As there was significant collinearity between the 23 predictors in the initial dataset, we removed variables with low contribution to predictions until the variance inflation factors (VIFs) of all remaining variables were below a threshold of five. It has been suggested that a VIF value that exceeds five indicates a problematic amount of collinearity (James *et al.*, 2013). The contribution to predictions was based on a preliminary ranking of predictor variables using random forests and a stepwise forward strategy for variable introduction (Genuer *et al.*, 2015). Along these lines, we selected a subset of 15 predictor variables minimizing redundancy and maximizing model performance to fit models (bold in Table S1; Fig. S2). The predictors retained represented all aspects (geography, current climate, environmental heterogeneity, and past environment) that are hypothesized to affect plant diversity patterns.

**Modeling** To perform GLMs and GAMs, we used a negative binomial error distribution with a log link function for species richness to cope with the overdispersion of the response variable and a Gaussian error distribution with a log link function for phylogenetic richness. For the GLMs, some predictors were log-transformed owing to their skewed distribution (i.e. area, temperature seasonality, number of wet days, precipitation seasonality, precipitation of warmest quarter, gross primary productivity, elevational range, number of soil types, and velocity in temperature since the LGM). After log transformation, all continuous predictor variables were standardized to zero mean and unit variance to aid model fitting and make their parameter estimates comparable. Although fitting GLMs with 15 predictors might seem excessive, it is suggested not to exclude predictors hypothesized to be important when collinearity is minimized and not a hindrance to analysis (Morrissey & Ruxton, 2018). Thus, in our GLMs, we built the full model including 15 predictors and then simplified the model using Akaike's information criterion (AIC). Predictors were tested in turn, and removed if AIC values were larger in the complex models than in the simpler ones (Phillips *et al.*, 2019; Table S4). To account for the interactive effects of environmental predictors on diversity patterns, we fitted GLMs including energy–water, energy–environmental heterogeneity, and area–environment interactions, as suggested by previous studies (Kreft & Jetz, 2007; Stein *et al.*, 2014; Keil & Chase, 2019). Models including interactions were simplified based on AIC values. First, all interactions were tested, and then, any main effects (i.e. individual predictors) that were not included in the retained interactions were tested (Phillips *et al.*, 2019). In GAMs, we used penalized regression smoothers (with nine spline bases for species richness and 10 spline bases for phylogenetic richness) for each predictor to estimate the smooth terms. The number of spline bases was selected from values between two and 10 using random cross-validation to optimize model performance (i.e. minimizing the root-mean-square error (RMSE)). Additionally, we used a gamma value of 1.4 to reduce overfitting without compromising model fit (Wood, 2006) and also included a double penalty to variable coefficients. We used the R packages `MASS` (Venables & Ripley, 2002) to fit negative binomial GLMs and `MGCV` (Wood, 2006) to fit GAMs.

In addition, we applied machine learning techniques, that is, random forests, XGBoost, and neural networks, to fit global models of plant diversity. Random forests are an ensemble learning method that builds a large collection of decision trees and outputs average predictions of the individual regression trees, while XGBoost is an ensemble model of decision trees trained sequentially fitting the residual errors in each iteration. Several innovations make XGBoost highly effective, including a novel tree learning algorithm for handling sparse data and a theoretically justified weighted quantile sketch procedure enabling handling instance weights in approximate tree learning (Chen & Guestrin, 2016). Neural networks are a machine learning method that comprises a collection of connected units (neurons) and their connections (edges). For these machine learning methods, species and phylogenetic richness were log-transformed before modeling to reduce the skewness of their distributions. A set of tuning parameters (i.e. hyperparameters), which cannot directly be

estimated from the data, needs to be set beforehand. These hyperparameters determine the training strategy and related efficiency of the algorithms. It is commonly suggested to tune hyperparameters to maximize model performance before running models for a certain problem (Bergstra & Bengio, 2012). We used the *train* function from the R package *CARET* to optimize the model tuning parameters for the three machine learning models used here (Kuhn, 2008). We used repeated random cross-validation and selected the hyperparameters that produced the lowest RMSE. We then refitted the final models using these optimal hyperparameters. The R package *RANGER* was used to fit random forests (Wright & Ziegler, 2017), *XGBOOST* to fit XGBoost (Chen & Guestrin, 2016), and *NEURALNET* to fit neural networks (Günther & Fritsch, 2010). Unlike GLMs and GAMs, machine learning can detect and model interactions of predictors without *a priori* specification, and we visualized interactions in machine learning models using partial dependence plots. For details on tuning parameters, model fitting using machine learning techniques, and visualization of interactions, see Methods S3.

**Spatial terms** Species distribution data and environmental predictors are often spatially autocorrelated. On the one hand, this might lead to biased parameter estimates, which need to be accounted for (Dormann *et al.*, 2007). On the other hand, including spatial information in models could increase their predictive power (Keil & Chase, 2019). Because of this, we generated spatial models using different modeling techniques. To account for spatial autocorrelation in GLM residuals, we used simultaneous autoregressive (SAR) models of the spatial error type, which is recommended for use when dealing with spatially autocorrelated species distribution data (Kissling & Carl, 2008). We evaluated SAR models with different neighborhood structures and spatial weights (lag distances between 200 and 3000 km, weighted and binary coding). As the final SAR model, we chose a model with weighted neighborhood structure and 800 km lag distance both for species and for phylogenetic richness, which had the minimal AIC and the best reduction in spatial autocorrelation in the residuals. Species and phylogenetic richness were log-transformed before modeling. In GAMs, we added a two-dimensional spline on geographical coordinates, which accounts for spatial autocorrelation in model residuals (Dormann *et al.*, 2007; Keil & Chase, 2019). To cope with spatial autocorrelation in machine learning models, we included cubic polynomial trend surfaces (i.e. latitude (Y), centered longitude (X) as well as  $X^2$ ,  $XY$ ,  $Y^2$ ,  $X^3$ ,  $X^2Y$ ,  $XY^2$ , and  $Y^3$ ; Bjorholm *et al.*, 2005; Li, 2019). Overall, the spatial models successfully removed spatial autocorrelation from model residuals (Fig. S3).

**Comparison with established models** To compare our models to published global models of plant species richness, we rebuilt these models for the dataset analyzed here. First, we fitted the best model as in Krefl & Jetz (2007), a combined six-predictor model using GLMs; and second, we built a GAM using the same model structure as Keil and Chase's smooth model (Keil & Chase, 2019), which contained a two-dimensional spline on geographical coordinates, 15 single predictors, and interactions

between each individual predictor and area. We ran models including the same 15 predictor variables and floristic kingdom using random forests and XGBoost and compared them with the models without floristic kingdom. Adding floristic kingdom increased collinearity between predictors. However, the two tree-based models are able to handle multicollinearity when they are used for prediction. Random forests in the *RANGER* R package can handle categorical variables automatically; however, XGBoost works only with numeric vectors. We therefore converted all other forms of data into numeric vectors. Here, we used one-hot encoding (0,1) to convert the floristic kingdom into dummy variables for the XGBoost model.

**Variable importance** To estimate the relative importance of each environmental predictor, we used a consistent method across model types. We randomly reshuffled values of the predictor of interest in the dataset, predicted the response variables based on the modified dataset, and calculated the Spearman rank correlation coefficient between those predictions and the predictions using the original dataset. The relative importance of the predictor of interest was calculated as one minus the correlation coefficient divided by the sum of one minus the correlation coefficients of all predictors (Thuiller *et al.*, 2009). Likewise, to compare the relative importance of different categories of predictor variables (categories in Table S1), we permuted values of a subset of predictors belonging to one category, correlated the predictions of the model using the modified dataset and predictions using the original dataset, and estimated the importance of each category as one minus the Spearman rank correlation coefficient divided by the sum of one minus the correlation coefficients of all predictor categories. Relationships between diversity metrics and predictor variables were visualized as partial dependence plots (see Methods S3 for details).

## Cross-validation

To assess the accuracy of model predictions across all different model types, we used random 10-fold cross-validation and spatial 68-fold cross-validation following Ploton *et al.* (2020; for details, see Methods S4). To quantify model predictive performance, we summarized the cross-validation results using the RMSE and two different pseudocoefficients of determination to quantify the amount of variation explained by the model based on out-of-bag samples.  $R^2\_CORR$  is the coefficient of determination of a linear model of the predicted and observed values from all repetitions of the cross-validation.  $R^2\_Accuracy$  is the amount of variation explained by the model, calculated as  $R^2\_Accuracy = (1 - SSE/SST)$  (Hengl *et al.*, 2017), where SSE is the sum of the squared error between observation and prediction and SST is the total sum of squares. The model with the lowest RMSE and highest  $R^2\_CORR/R^2\_Accuracy$  was identified as the best predictive model. For all models, we calculated cross-validation results for log-transformed observed and predicted species and phylogenetic richness, because species and phylogenetic richness were log-transformed before modeling for machine learning models and fitted with log link functions in GLMs and GAMs.

Variation explained according to spatial cross-validation was consistently lower than variation explained according to random cross-validation, likely because the former offers biased and pessimistic estimates (Wadoux *et al.*, 2021). Spatial cross-validation excludes entire portions of regions with specific combinations of environmental characteristics and biogeographic histories from the training data and is therefore less representative of the globe and its environmental spectrum, likely causing predictions outside the covariate space within the models. By contrast, random cross-validation is almost unbiased when the sampling design is systematic or random (Wadoux *et al.*, 2021). Because the geographic regions in our dataset were distributed representatively across the entire globe, covering all major biomes (Fig. S1), we argue that random cross-validation offers relatively unbiased assessments of model performance.

## Predictions

We used the resulting models to predict vascular plant species and phylogenetic richness across global grids of four different resolutions (i.e. 7774, 23 322, 69 967, and 209 903 km<sup>2</sup> hexagon size). We used the DGGRIDR R package (Barnes & Sahr, 2017) to produce a grid of equal-area and equidistant hexagons across the Earth's surface clipped for global coastlines. Islands smaller than 1.5 times the grid cell size were treated as entire entities instead of subdividing them into several partial grid cells. For each hexagon, we calculated the same predictor variables as for the geographic regions used for fitting the models. We then used the models to predict vascular plant species and phylogenetic richness and mapped the predictions across the hexagon grid. Due to missing values in some predictor variables, a few values had to be interpolated for predictions (see Methods S5 for details).

Besides predictions based on individual models, we used an ensemble prediction procedure, which averages the predictions based on the models fitted by different techniques weighted by model accuracy (the inverse of the model squared error) from the random cross-validation process (Marmion *et al.*, 2009). Because spatial cross-validation was likely biased (Wadoux *et al.*, 2021), we used model accuracy from random cross-validation. In addition to the hexagon grids, we generated plant diversity maps in raster format at a resolution of 30 arc seconds based on predictions for the 7774 km<sup>2</sup> hexagons (see Methods S5; Fig. S4). As centers of plant diversity based on the ensemble predictions, we defined regions with predicted richness values higher than the 90<sup>th</sup> quantile, that is, containing at least 1765 plant species and 41 866 Ma of phylogenetic richness at a resolution of 7774 km<sup>2</sup>.

## Uncertainty

To assess variation of the predictions across models, we calculated the coefficient of variation of predicted values for each hexagon grid cell. The coefficient of variation is defined as the ratio of the standard deviation to the mean, which accounts for the differences in diversity between regions and thereby avoids artificially high uncertainty of high-diversity regions. Additionally, we calculated standard errors of predictions for GLMs, GAMs, and

random forests. For XGBoost and neural networks, we modeled the relationship between model residuals and environmental predictors from the raw data and used this model to predict uncertainty across the hexagon grids.

## Results and Discussion

### Performance of plant diversity models

Our results reveal a great potential of machine learning, particularly decision tree methods, for modeling plant diversity–environment relationships and for accurately predicting plant diversity across various scales. Overall, the predictive power of the models was high (Table 1). Machine learning models and GAMs outperformed GLMs, and spatial models (i.e. models containing spatial terms to account for the spatial nonindependence of regions; Dormann *et al.*, 2007) showed an overall better performance than nonspatial models (except GLMs for species richness). Extreme gradient boosting, an ensemble of sequentially trained decision trees, produced the most accurate predictions both for species richness (70.3% variation explained based on spatial cross-validation and 80.9% based on random cross-validation) and for phylogenetic richness (73.7% and 83.3%, respectively), which was consistent across spatial and nonspatial models.

The good predictive performance of machine learning models can be attributed to their ability to uncover complex, nonlinear diversity–environment relationships (Figs S5, S6) and interactive effects (Figs S7–S18). We found strong interactions between spatial terms and environmental variables (Figs S7–S18). This indicates regional differences in plant diversity and diversity–environment relationships and shows that different combinations of environmental variables are important when predicting diversity across geographic regions (Keil & Chase, 2019). Moreover, machine learning models revealed strong interactions between energy and water availability, energy and environmental heterogeneity, as well as area and environmental variables (Figs S7–S18). Also, the accuracy of GLMs increased when including the interactions that turned out to be important in machine learning models (70.4% vs 63.6% in species richness based on random cross-validation; 63.5% vs 45.2% in phylogenetic richness), highlighting the role of complex interactive effects among biotic and abiotic factors in shaping global plant diversity patterns (Francis & Currie, 2003; Kreft & Jetz, 2007; Keil & Chase, 2019). By implicitly accounting for grain dependence and complex interactions among spatial and environmental variables, our machine learning models outperform previous models of plant diversity (Kreft & Jetz, 2007; Keil & Chase, 2019; Table S4), improving our understanding of diversity–environment relationships and allowing for improved predictions of plant diversity across scales.

### Drivers of global patterns of vascular plant diversity

Current climatic variables emerged as the most important drivers of plant diversity, accounting for 34.4–48.1% of the variation in species richness and 39.7–58.2% in phylogenetic richness across

**Table 1** Performance of global models of vascular plant diversity based on cross-validation.

Models	Species richness				Phylogenetic richness (Faith's PD)			
	Random cross-validation		Spatial cross-validation		Random cross-validation		Spatial cross-validation	
	RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
Nonspatial models								
Full GLM	0.525	0.636	0.582	0.561	0.514	0.452	0.552	0.359
Minimum adequate GLM	0.520	0.643	0.548	0.608	0.513	0.454	0.548	0.369
GLM with interaction terms	0.471	0.704	0.502	0.664	0.412	0.635	0.453	0.559
GAM	0.437	0.742	0.507	0.658	0.359	0.723	0.430	0.604
Random forests	0.415	0.761	0.511	0.639	0.317	0.784	0.395	0.667
Extreme gradient boosting	0.389	0.791	0.487	0.673	0.295	0.813	0.384	0.685
Neural networks	0.451	0.718	0.604	0.496	0.328	0.769	0.419	0.628
Spatial models								
SAR	0.537	0.600	0.548	0.584	0.416	0.629	0.426	0.611
GAM	0.413	0.769	0.499	0.667	0.340	0.751	0.416	0.633
Random forests	0.398	0.780	0.502	0.653	0.303	0.803	0.379	0.694
Extreme gradient boosting	0.371	0.809	0.463	0.703	0.279	0.833	0.351	0.737
Neural networks	0.422	0.753	0.587	0.522	0.314	0.789	0.433	0.597

Each model was evaluated for its predictive performance using both random 10-fold and spatial 68-fold cross-validation. Nonspatial models were fitted with 15 predictors representing geography, current climate, environmental heterogeneity, and past environment conditions (Supporting Information Table S1) except for the minimum adequate generalized linear model (GLM) and the GLM with interaction terms. Spatial models in addition contained spatial terms (i.e. simultaneous autoregressive (SAR) models, generalized additive models (GAMs) including splines of geographic coordinates, and machine learning methods including cubic polynomial trend surfaces). The minimum adequate GLM was obtained by simplifying the full GLM based on Akaike's information criterion (AIC). The GLM with interaction terms was fitted including all predictors of the full GLM and interactions of energy-water, energy-heterogeneity, and area-environment-related variables and was then simplified based on AIC. Because the response variables (i.e. species and phylogenetic richness) were log-transformed in models, the accuracy statistics are provided on a log scale. Based on all out-of-bag samples, values shown are root-mean-square error (RMSE); the amount of variation explained by the model calculated as one minus the ratio of the sum of the squared error between observation and prediction to the total sum of squares ( $R^2$ ). For more detailed cross-validation results, see Table S4.

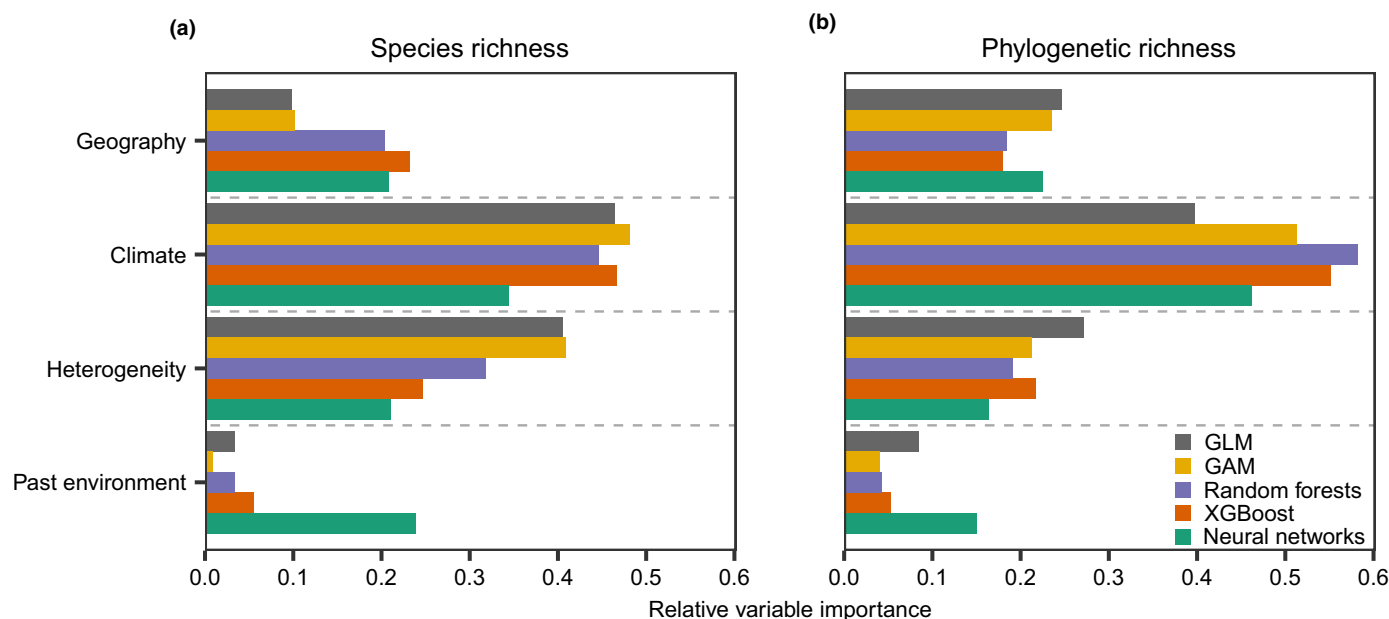
models (Fig. 1; Table S1). High energy and water availability and low seasonality promoted species and phylogenetic richness (Figs S5, S6), supporting other large-scale studies that report strong effects of the current climate on plant diversity (Francis & Currie, 2003; Hawkins *et al.*, 2003; Kreft & Jetz, 2007). Environmental heterogeneity (measured here as elevational range and number of soil types within a region) explained 21.0–40.9% of the variation in species richness and 16.3–27.2% in phylogenetic richness, with increasing heterogeneity leading to higher plant diversity as expected (Stein *et al.*, 2014). Even though species and phylogenetic richness were highly correlated (Pearson's  $r = 0.98$ ), some differences emerged in diversity–environment relationships. For example, environmental heterogeneity explained less variation in phylogenetic richness than in species richness. This potentially reflects a signal of *in situ* speciation that is promoted by high environmental heterogeneity, creating clusters of closely related species resulting in relatively low phylogenetic richness compared with species richness (Forest *et al.*, 2007). This notion was also supported by a negative effect of the number of soil types on the residual variation of phylogenetic richness after accounting for species richness (Table S2).

Geographic variables (area and geographic isolation) explained 9.8–23.1% of the variation in species richness and 18.0–24.6% in phylogenetic richness. Larger regions tend to have higher *in situ* speciation rates owing to more opportunities for geographic isolation within a region and lower extinction rates due

to larger populations (Terborgh, 1973; Kisel & Barraclough, 2010). These effects should be most pronounced in self-contained, isolated regions like islands, mountains, or other isolated habitats and less so in regions that are similar to their surroundings (Rosenzweig, 2003; Testolin *et al.*, 2021). Additionally, larger regions often provide a greater variety of habitats, offering more environmental niches to be occupied by species. Geographic isolation, measured here as the proportion of surrounding landmass, did not explain much variation (0.0–3.9% in species richness; 0.5–3.5% in phylogenetic richness; Fig. S19) for both diversity facets, possibly because our dataset consisted mainly of mainland regions (93.4% of all regions). While geographic isolation is a main driver of insular plant diversity (Weigelt & Kreft, 2013), isolation and peninsular effects seem to play only a minor role on the mainland, where geographic isolation can be expected to be more important for compositional uniqueness of regions and endemism, rather than for richness (Sandel *et al.*, 2020).

We hypothesized that higher plant diversity would accumulate in regions with long-term climate stability because of low extinction and high speciation rates (Fine, 2015; Svenning *et al.*, 2015). We therefore assessed the effects of temperature stability and biome variation as proxies for past climatic change for two paleo-time periods, that is, the LGM and the mid-Pliocene warm period. In contrast to the expected legacy effects of historical variables on modern plant diversity, past





**Fig. 1** Relative importance of environmental variable categories for explaining global patterns of vascular plant diversity across five nonspatial models. (a) Species richness; (b) phylogenetic richness (Faith's PD). Relative importance for different variable categories (scaled to sum up to one) was calculated as one minus the Spearman rank correlation coefficient between predictions of the model using a dataset where the values of the predictors of interest were randomly reshuffled and predictions using the original dataset. Environmental variables falling into each category are shown in Supporting Information Table S1. For the importance of individual environmental variables, see Fig. S19. GAM, generalized additive model; GLM, generalized linear model; XGBoost, extreme gradient boosting.

environmental conditions contributed only 0.8–5.5% to explaining species richness in most of our models, but up to 23.8% in neural networks. Likewise, past environmental conditions showed higher explanatory power (15.0%) for phylogenetic richness in neural networks than in other models (4.0–8.5%). Models including spatial trend surfaces or discrete biogeographic regions (i.e. floristic kingdoms) to account for regional idiosyncrasies (after statistically controlling for current and past environments) further improved model fits (Tables 1, S4). This suggests that in addition to climate stability since the LGM or mid-Pliocene warm period, biogeographic history predating the Pliocene or regional idiosyncrasies other than climatic changes affected modern plant diversity. These historical regional effects are possibly due to dispersal barriers and idiosyncratic colonization and diversification histories (Qian & Ricklefs, 2004; Ricklefs & He, 2016).

### Improved global plant diversity maps

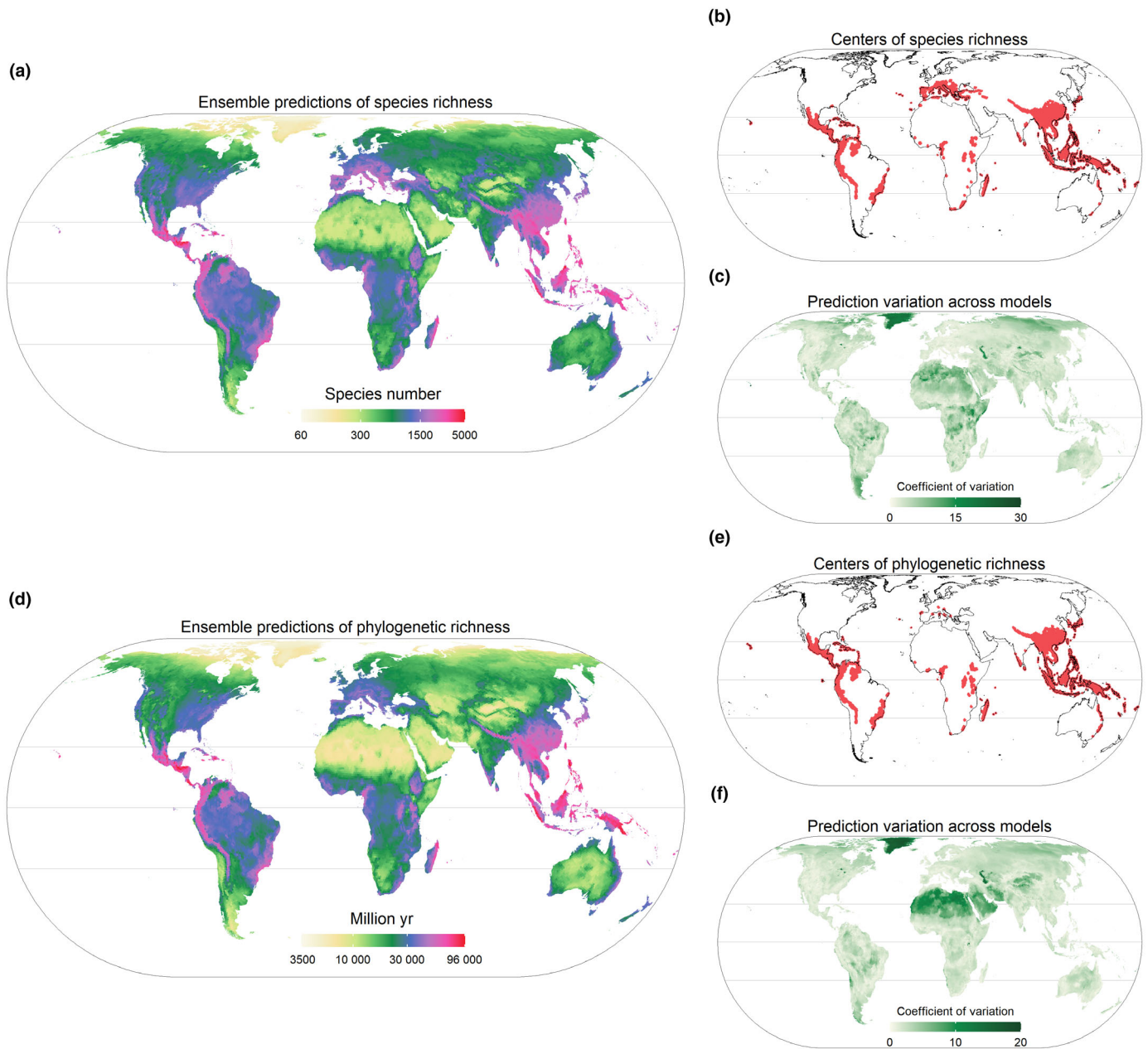
We produced global diversity maps for species and phylogenetic richness of vascular plants, based on individual well-performing models and model ensembles. Because of its outstanding predictive power and ability to handle missing data, we consider XGBoost (including geographic coordinates) the most powerful single model for predicting plant diversity (Figs S20d, S21d). In addition, we present ensemble predictions, which reduce the uncertainty introduced by the choice of one particular modeling technique and therefore improve prediction accuracy (Marmion *et al.*, 2009). Including region area and its interactions with other predictor variables allowed us to predict plant diversity across

global grids of equal-area and equidistant hexagons of different grain sizes (i.e. 7774, 23 322, 69 967, and 209 903 km<sup>2</sup>; Figs S22, S23). All model predictions and their uncertainties are accessible at <https://gift.uni-goettingen.de/shiny/predictions/>.

Our ensemble predictions (Fig. 2a,d) describe the global patterns of species and phylogenetic richness with unprecedented detail and accuracy. The maps capture how diversity varies along environmental gradients and identify global centers of plant diversity (Fig. 2b,e). The highest concentrations of plant species and phylogenetic richness are predicted in Central America, southern Mexico, Andes–Amazonia, the Caribbean, southeastern Brazil, the Cape region of Southern Africa, Madagascar, Malay Archipelago, Indochina, and southern China (Fig. 2b,e), which is in line with empirical observations and previous studies (Myers *et al.*, 2000; Barthlott *et al.*, 2005; Kreft & Jetz, 2007). While patterns of phylogenetic richness closely resembled species richness (Pearson's  $r = 0.97$ ), discrepancies occurred, for example, around the Mediterranean, in Central America, the Caucasus, and the Himalayas (Fig. S24). Differences might result from unequal taxonomic efforts (e.g. many closely related species described separately in Europe) or the uneven distribution of evolutionarily old or young clades across the globe (Thorne, 1999; Endress, 2001). The former suggests that predictions of phylogenetic diversity provide a taxonomically less biased representation of global plant diversity patterns.

Thanks to the high-resolution environmental data and modeling techniques that account for complex interactions, regions with steep elevational gradients show finer-tuned variation in predicted effects presented here than in previous studies (Barthlott *et al.*, 2005; Kreft & Jetz, 2007). For example, the eastern slopes





**Fig. 2** Global patterns of vascular plant diversity predicted across an equal-area hexagon grid of 7774 km<sup>2</sup> resolution. Species richness (a) and phylogenetic richness (Faith's PD, d) are based on ensemble predictions of five different models (i.e. three spatial models using machine learning methods, a spatial generalized additive model, and a nonspatial generalized linear model with interactions) weighted by model accuracy; species richness (b) and phylogenetic richness (e) centers are defined as regions with predicted richness values higher than the 90<sup>th</sup> quantile of the predictions (i.e. containing at least 1765 plant species and 41 866 Ma of phylogenetic richness per 7774 km<sup>2</sup>). Variation of predictions across models used for the ensemble predictions is calculated as coefficient of variation of predicted values for species richness (c) and phylogenetic richness (f). Horizontal lines depict the equator and borders of the tropics. In (a, d), log<sub>10</sub> scale is used and all maps use Eckert IV projection. For maps of all different models and resolutions and data download, see <https://gift.uni-goettingen.de/shiny/predictions/>.

of the Andes, southern Himalayan slopes, and the northern Kunlun Mountains in China show a finer differentiation from adjacent dryer and less diverse regions than in Kreft & Jetz (2007). At the same time, our ensemble predictions show relatively high values in species-poor regions like nonglaciated parts of Greenland and the Sahara. Here, and in other regions with extreme values of plant diversity, individual models perform better than the

ensemble model (Figs S20, S21), which tends to attenuate extreme values. Besides the important differences just outlined, the ensemble predictions presented here were strongly correlated with model predictions in Kreft & Jetz (2007; Pearson's  $r = 0.872$ ; Fig. S25). Aside from the different modeling techniques used and how they account for complex and interactive diversity–environment relationships, differences with previous

maps could derive from the accumulation of knowledge on plant diversity worldwide and the continuously updated species distribution data in GIFT used for modeling.

Regions with high species and phylogenetic richness were found to be distributed mostly in mountainous regions (Fig. S26). Specifically, tropical mountain ranges, including the tropical Andes, eastern African highlands, and various Asian mountains (e.g. in southern China and the Malay Archipelago), are global centers of plant diversity. The high diversity of tropical mountain ranges, as also found in previous studies (Testolin *et al.*, 2021), is linked to warm and wet climates and heterogeneous environments (Antonelli *et al.*, 2018). Multiple biogeographical and evolutionary processes, including speciation, dispersal, and persistence that are driven by long-term orogenic and climatic dynamics in mountains, have led to outstanding regional plant diversity (Antonelli *et al.*, 2018; Rahbek *et al.*, 2019). Orogenic processes constantly change soil composition, nutrient levels, and local climate of mountainous regions, thus creating novel and heterogeneous habitats where plant lineages diversify and colonize from neighboring areas (Antonelli *et al.*, 2018). Moreover, climatic fluctuations stimulate diversification by driving dynamic shifts in habitat connectivity within mountains (Rahbek *et al.*, 2019). Due to their steep environmental gradients and heterogeneous nature, mountain regions provide refugia in times of unfavorable climate (Bennett *et al.*, 1991; Rahbek *et al.*, 2019).

Differences among models (measured as the coefficient of variation) were greatest in regions with extreme environments, such as deserts and Arctic regions (Fig. 2c,f). Arctic regions also consistently showed the highest prediction uncertainty across models (Figs S27, S28). The uncertainties in regions with extreme environments probably stem from two sources. First, extremely species-poor regions might be less well represented in published diversity data. Regions with extreme environments are often part of artificially delimited regions instead of being sampled individually (e.g. Chad and Libya sampled instead of the Sahara). Those artificially delimited regions are more environmentally heterogeneous, which attenuates the extreme values of environmental factors as well as plant diversity. Machine learning models are known to not extrapolate well under such conditions (Elith *et al.*, 2010). Second, even for regions with relatively homogeneous environments, checklists and floras do not only include information on predominant but also azonal vegetation, making them richer than expected from their prevailing conditions and observed at a more local scale (compared to alpha diversity predictions in Sabatini *et al.*, 2022).

## Conclusions

We present the most accurate and comprehensive predictive global maps of regional vascular plant species and phylogenetic richness available to date. They are based on significantly improved global models using comprehensive global inventory-based plant distribution data, high-resolution past and current environmental information, and advanced machine learning models. Our findings illustrate that machine learning methods applied to large

distribution and environmental datasets help to disentangle underlying complex and interacting associations between the environment and plant diversity. Machine learning methods therefore help to improve both fundamental understanding and quantitative knowledge in biogeography and macroecology. The updated global diversity maps of vascular plant diversity at multiple grain sizes (available at <https://gift.uni-goettingen.de/shiny/predictions/>) provide a solid foundation for large-scale biodiversity monitoring and research on the origin of plant diversity and subsequently support future global biodiversity assessments and environmental policies.

## Acknowledgements

LC was supported by China Scholarship Council (CSC) Grant (no. 201808330443). PP and JP were supported by EXPRO grant no. 19-28807X (Czech Science Foundation) and long-term research development project RVO 67985939 (Czech Academy of Sciences). MW acknowledges DFG funding via iDiv (DFG FZT 118, 202548816). FE appreciates funding by the Austrian Science Foundation FWF (Global Plant Invasions-project, grant I 5825-B). We thank Alexandr Ebel and Christian König for contributing data and discussions about the manuscript. Open Access funding enabled and organized by Projekt DEAL.

## Author contributions

LC, HK and PW conceived the idea and developed the conceptual framework of the study. LC, HK, AT, PD, JS, FE, MvK, JP, PP, AS, MW, JFB, NF, I, DNK, JK, AK, MN, DN, AN, AP, PBP, PS, JJW and PW were involved in collecting the data. LC performed the statistical analyses. LC wrote the first draft of the manuscript with input from HK, AT and PW. LC, HK, AT, PD, JS, FE, MvK, JP, PP, AS, MW, JFB, NF, I, DNK, JK, AK, MN, DN, AN, AP, PBP, PS, JJW and PW contributed to the writing and interpretation of the results.

## ORCID

Julie F. Barcelona  <https://orcid.org/0000-0001-5087-8637>

Lirong Cai  <https://orcid.org/0000-0001-9432-2024>

Pierre Denelle  <https://orcid.org/0000-0002-4729-3774>

Franz Essl  <https://orcid.org/0000-0001-8253-2112>

Nicol Fuentes  <https://orcid.org/0000-0002-3773-9832>

Inderjit  <https://orcid.org/0000-0002-4142-1392>

Dirk Nikolaus Karger  <https://orcid.org/0000-0001-7770-6229>

Mark van Kleunen  <https://orcid.org/0000-0002-2861-3701>

Holger Kreft  <https://orcid.org/0000-0003-4471-8236>

Daniel Nickrent  <https://orcid.org/0000-0001-8519-0517>

Arkadiusz Nowak  <https://orcid.org/0000-0001-8638-0208>

Annette Patzelt  <https://orcid.org/0000-0003-3510-4582>

Pieter B. Pelsers  <https://orcid.org/0000-0002-6990-1419>

Jan Pergl  <https://orcid.org/0000-0002-0045-1974>

Petr Pyšek  <https://orcid.org/0000-0001-8500-442X>

Julian Schrader  <https://orcid.org/0000-0002-8392-211X>  
 Paramjit Singh  <https://orcid.org/0000-0001-7909-6284>  
 Amanda Taylor  <https://orcid.org/0000-0002-0420-2203>  
 Patrick Weigelt  <https://orcid.org/0000-0002-2485-3708>  
 Jan J. Wieringa  <https://orcid.org/0000-0003-0566-372X>  
 Marten Winter  <https://orcid.org/0000-0002-9593-7300>

## Data availability

Predictions of vascular plant species and phylogenetic richness and model uncertainties based on the various statistical models applied here are available at <https://gift.uni-goettingen.de/shiny/predictions/>. In addition, all predictions and uncertainties as well as the data and R codes needed to run the analyses are available at doi: [10.6084/m9.figshare.19539085](https://doi.org/10.6084/m9.figshare.19539085). The list of original plant checklists and floras from the Global Inventory of Floras and Traits (GIFT) can be found in Notes S1.

## References

- Antonelli A, Kissling WD, Flantua SGA, Bermúdez MA, Mulch A, Muellner-Riehl AN, Kreft H, Linder HP, Badgley C, Fjeldså J *et al.* 2018. Geological and climatic influences on mountain biodiversity. *Nature Geoscience* 11: 718–725.
- Araújo MB, New M. 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* 22: 42–47.
- Barnes R, Sahr K. 2017. DGGDR: discrete global grids for R. doi: [10.5281/zenodo.1322866](https://doi.org/10.5281/zenodo.1322866).
- Barthlott W, Mutke J, Rafiqpoor D, Kier G, Kreft H. 2005. Global centers of vascular plant diversity. *Nova Acta Leopoldina NF* 92: 61–83.
- Bennett KD, Tzedakis PC, Willis KJ. 1991. Quaternary refugia of north European trees. *Journal of Biogeography* 18: 103–115.
- Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13: 281–305.
- Bjorholm S, Svenning J-C, Skov F, Balslev H. 2005. Environmental and spatial controls of palm (Arecaceae) species richness across the Americas. *Global Ecology and Biogeography* 14: 423–429.
- Boyle B, Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, McKay SJ *et al.* 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* 14: 16.
- Brown JH. 2014. Why are there so many species in the tropics? *Journal of Biogeography* 41: 8–22.
- Brown JH, Kodric-Brown A. 1977. Turnover rates in insular biogeography: effect of immigration on extinction. *Ecology* 58: 445–449.
- Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, Narwani A, Mace GM, Tilman D, Wardle DA *et al.* 2012. Biodiversity loss and its impact on humanity. *Nature* 486: 59–67.
- Chen T, Guestrin C. 2016. XGBOOST: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM, 785–794.
- Connor EF, McCoy ED. 1979. The statistics and biology of the species-area relationship. *The American Naturalist* 113: 791–833.
- Couvreux TLP, Dauby G, Blach-Overgaard A, Deblauwe V, Dessein S, Droissart V, Hardy OJ, Harris DJ, Janssens SB, Ley AC *et al.* 2021. Tectonics, climate and the diversification of the tropical African terrestrial flora and fauna. *Biological Reviews* 96: 16–51.
- Crisci C, Ghattas B, Perera G. 2012. A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling* 240: 113–122.
- Currie DJ, Mittelbach GG, Cornell HV, Field R, Guegan J-F, Hawkins BA, Kaufman DM, Kerr JT, Oberdorff T, O'Brien E *et al.* 2004. Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters* 7: 1121–1134.
- Danielson JJ, Gesch DB. 2011. *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. [WWW document] URL <http://pubs.er.usgs.gov/publication/ofr20111073> [accessed 27 February 2018].
- Dormann C, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Daniel Kissling W *et al.* 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30: 609–628.
- Dowsett H, Dolan A, Rowley D, Moucha R, Forte AM, Mitrovica JX, Pound M, Salzmann U, Robinson M, Chandler M *et al.* 2016. The PRISM4 (mid-Piacenzian) paleoenvironmental reconstruction. *Climate of the Past* 12: 1519–1538.
- Dynesius M, Jansson R. 2000. Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. *Proceedings of the National Academy of Sciences, USA* 97: 9115–9120.
- Eiserhardt WL, Borchsenius F, Sandel B, Kissling WD, Svenning J-C. 2015. Late Cenozoic climate and the phylogenetic structure of regional conifer floras world-wide. *Global Ecology and Biogeography* 24: 1136–1148.
- Elith J, Kearney M, Phillips S. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* 1: 330–342.
- Endress PK. 2001. The flowers in extant basal angiosperms and inferences on ancestral flowers. *International Journal of Plant Sciences* 162: 1111–1140.
- Enquist BJ, Condit R, Peet RK, Schildhauer M, Thiers BM. 2016. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints*. doi: [10.7287/peerj.preprints.2615v2](https://doi.org/10.7287/peerj.preprints.2615v2).
- Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1–10.
- Fine PVA. 2015. Ecological and evolutionary drivers of geographic variation in species diversity. *Annual Review of Ecology, Evolution, and Systematics* 46: 369–392.
- Forest F, Grenyer R, Rouget M, Davies TJ, Cowling RM, Faith DP, Balmford A, Manning JC, Procheş S, van der Bank M *et al.* 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* 445: 757–760.
- Francis AP, Currie DJ. 2003. A globally consistent richness-climate relationship for angiosperms. *The American Naturalist* 161: 523–536.
- GBIF. 2020. *GBIF: the global biodiversity information facility (year) what is GBIF?* [WWW document] URL <https://www.gbif.org/what-is-gbif> [accessed 13 January 2020].
- Genuer R, Poggi J-M, Tuleau-Malot C. 2015. vsurf: an R package for variable selection using random forests. *The R Journal* 7: 19–33.
- Gillooly JF, Allen AP. 2007. Linking global patterns in biodiversity to evolutionary dynamics using metabolic theory. *Ecology* 88: 1890–1894.
- Govaerts R, Nic Lughadha E, Black N, Turner R, Paton A. 2021. The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity. *Scientific Data* 8: 215.
- Günther F, Fritsch S. 2010. NEURALNET: training of neural networks. *The R Journal* 2: 30–38.
- Hagen O, Skeels A, Onstein RE, Jetz W, Pellissier L. 2021. Earth history events shaped the evolution of uneven biodiversity across tropical moist forests. *Proceedings of the National Academy of Sciences, USA* 118: e2026347118.
- Hawkins BA, Field R, Cornell HV, Currie DJ, Guégan J-F, Kaufman DM, Kerr JT, Mittelbach GG, Oberdorff T, O'Brien EM *et al.* 2003. Energy, water, and broad-scale geographic patterns of species richness. *Ecology* 84: 3105–3117.
- 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.
- Henrot A-J, François L, Favre E, Butzin M, Ouberdous M, Munhoven G. 2010. Effects of CO<sub>2</sub>, continental distribution, topography and vegetation changes on the climate at the Middle Miocene: a model study. *Climate of the Past* 6: 675–694.
- Isbell F, Calcagno V, Hector A, Connolly J, Harpole WS, Reich PB, Scherer-Lorezen M, Schmid B, Tilman D, van Ruijven J *et al.* 2011. High plant diversity is needed to maintain ecosystem services. *Nature* 477: 199–202.
- James G, Witten D, Hastie T, Tibshirani R. 2013. *An introduction to statistical learning: with applications in R*. New York, NY, USA: Springer.
- Jin Y, Qian H. 2019. V.PHYLOMAKER: an R package that can generate very large phylogenies for vascular plants. *Ecography* 42: 1353–1359.



- Karger DN, Conrad O, Böhner 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.
- Keil P, Chase JM. 2019. Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution* 3: 390–399.
- Kisel Y, Barraclough TG. 2010. Speciation has a spatial scale that depends on levels of gene flow. *The American Naturalist* 175: 316–334.
- Kissling WD, Baker WJ, Balslev H, Barfod AS, Borchsenius F, Dransfield J, Govaerts R, Svenning J-C. 2012. Quaternary and pre-Quaternary historical legacies in the global distribution of a major tropical plant lineage. *Global Ecology and Biogeography* 21: 909–921.
- Kissling WD, Carl G. 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography* 17: 59–71.
- König C, Weigelt P, Schrader J, Taylor A, Kattge J, Kreft H. 2019. Biodiversity data integration—the significance of data resolution and domain. *PLoS Biology* 17: e3000183.
- Kreft H, Jetz W. 2007. Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences, USA* 104: 5925–5930.
- Kuhn M. 2008. Building predictive models in R using the CARET package. *Journal of Statistical Software* 28: 1–26.
- Li L. 2019. Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. *Remote Sensing* 11: 1378.
- Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W. 2009. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* 15: 59–69.
- Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19: 992–1006.
- Mittelbach GG, Schemske DW, Cornell HV, Allen AP, Brown JM, Bush MB, Harrison SP, Hurlbert AH, Knowlton N, Lessios HA *et al.* 2007. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecology Letters* 10: 315–331.
- Morrissey MB, Ruxton GD. 2018. Multiple regression is not multiple regressions: the meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology* 10. doi: [10.3998/ptpbio.16039257.0010.003](https://doi.org/10.3998/ptpbio.16039257.0010.003).
- Mutke J, Barthlott W. 2005. Patterns of vascular plant diversity at continental to global scales. *Biologische Skriften* 55: 521–538.
- Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853–858.
- Olden JD, Lawler JJ, Poff NL. 2008. Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology* 83: 171–193.
- Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D'Amico JA, Itoua I, Strand HE, Morrison JC *et al.* 2001. Terrestrial ecoregions of the world: a new map of life on earth. *Bioscience* 51: 933–938.
- Ouborg NJ. 1993. Isolation, population size and extinction: the classical and metapopulation approaches applied to vascular plants along the dutch rhine-system. *Oikos* 66: 298–308.
- Owens HL, Guralnick R. 2019. CLIMATESTABILITY: an R package to estimate climate stability from time-slice climatologies. *Biodiversity Informatics* 14: 8–13.
- Park DS, Willis CG, Xi Z, Kartesz JT, Davis CC, Worthington S. 2020. Machine learning predicts large scale declines in native plant phylogenetic diversity. *New Phytologist* 227: 1544–1556.
- Phillips HRP, Guerra CA, Bartz MLC, Briones MJI, Brown G, Crowther TW, Ferlian O, Gongalsky KB, van den Hoogen J, Krebs J *et al.* 2019. Global distribution of earthworm diversity. *Science* 366: 480–485.
- Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N *et al.* 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* 11: 4540.
- Qian H, Jin Y. 2021. Are phylogenies resolved at the genus level appropriate for studies on phylogenetic structure of species assemblages? *Plant Diversity* 43: 255–263.
- Qian H, Ricklefs RE. 2004. Taxon richness and climate in angiosperms: is there a globally consistent relationship that precludes region effects? *The American Naturalist* 163: 773–779.
- Qian H, Zhang J, Jiang M-C. 2022. Global patterns of fern species diversity: an evaluation of fern data in GBIF. *Plant Diversity* 44: 135–140.
- Rahbek C, Borregaard MK, Antonelli A, Colwell RK, Holt BG, Nogues-Bravo D, Rasmussen CMØ, Richardson K, Rosing MT, Whittaker RJ *et al.* 2019. Building mountain biodiversity: geological and evolutionary processes. *Science* 365: 1114–1119.
- Ray N, Adams JM. 2001. A GIS-based vegetation map of the world at the last glacial maximum (25,000–15,000 BP). *Internet Archaeology* 11. doi: [10.1111/ia.11.2](https://doi.org/10.1111/ia.11.2).
- Ricklefs RE, He F. 2016. Region effects influence local tree species diversity. *Proceedings of the National Academy of Sciences, USA* 113: 674–679.
- Rohde K. 1992. Latitudinal gradients in species diversity: the search for the primary cause. *Oikos* 65: 514–527.
- Rosenzweig ML. 2003. Reconciliation ecology and the future of species diversity. *Oryx* 37: 194–205.
- Sabatini FM, Jiménez-Alfaro B, Jandt U, Chytrý M, Field R, Kessler M, Lenoir J, Schrodt F, Wiser SK, Arfin Khan MAS *et al.* 2022. Global patterns of vascular plant alpha diversity. *Nature Communications* 13: 4683.
- Sabatini FM, Lenoir J, Hattab T, Arnst EA, Chytrý M, Dengler J, Ruffray PD, Hennekens SM, Jandt U, Jansen F *et al.* 2021. sPlotOpen—an environmentally balanced, open-access, global dataset of vegetation plots. *Global Ecology and Biogeography* 30: 1740–1764.
- Sandel B, Arge L, Dalsgaard B, Davies RG, Gaston KJ, Sutherland WJ, Svenning J-C. 2011. The influence of late quaternary climate-change velocity on species endemism. *Science* 334: 660–664.
- Sandel B, Weigelt P, Kreft H, Keppel G, van der Sande MT, Levin S, Smith S, Craven D, Knight TM. 2020. Current climate, isolation and history drive global patterns of tree phylogenetic endemism. *Global Ecology and Biogeography* 29: 4–15.
- Smith SA, Brown JW. 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* 105: 302–314.
- Stein A, Gerstner K, Kreft H. 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecology Letters* 17: 866–880.
- Svenning J-C, Eiserhardt WL, Normand S, Ordonez A, Sandel B. 2015. The influence of paleoclimate on present-day patterns in biodiversity and ecosystems. *Annual Review of Ecology, Evolution, and Systematics* 46: 551–572.
- Takhtajan AL. 1986. *Floristic regions of the world*. Oakland, CA, USA: University of California Press.
- Terborgh J. 1973. On the notion of favorableness in plant ecology. *The American Naturalist* 107: 481–501.
- Testolin R, Attorre F, Borchardt P, Brand RF, Bruehlheide H, Chytrý M, De Sanctis M, Dolezal J, Finckh M, Haider S *et al.* 2021. Global patterns and drivers of alpine plant species richness. *Global Ecology and Biogeography* 30: 1218–1231.
- Thorne RF. 1999. Eastern Asia as a living museum for archaic angiosperms and other seed plants. *Taiwania* 44: 413–422.
- Thuiller W, Lafourcade B, Engler R, Araújo MB. 2009. BIOMOD—a platform for ensemble forecasting of species distributions. *Ecography* 32: 369–373.
- Tietje M, Antonelli A, Baker WJ, Govaerts R, Smith SA, Eiserhardt WL. 2022. Global variation in diversification rate and species richness are unlinked in plants. *Proceedings of the National Academy of Sciences, USA* 119: e2120662119.
- Tilman D, Isbell F, Cowles JM. 2014. Biodiversity and ecosystem functioning. *Annual Review of Ecology, Evolution, and Systematics* 45: 471–493.
- Tucker CM, Cadotte MW, Carvalho SB, Davies TJ, Ferrier S, Fritz SA, Grenyer R, Helmus MR, Jin LS, Mooers AO *et al.* 2017. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews* 92: 698–715.
- Venables WN, Ripley BD. 2002. *Modern applied statistics with s*. New York, NY, USA: Springer.
- Wadoux AMJ-C, Heuvelink GBM, de Bruin S, Brus DJ. 2021. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling* 457: 109692.



- 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.
- Weigelt P, Kreft H. 2013. Quantifying island isolation—insights from global patterns of insular plant species richness. *Ecography* 36: 417–429.
- Wood SN. 2006. *Generalized additive models: an introduction with R*. Boca Raton, FL, USA: Chapman & Hall/CRC.
- Wright MN, Ziegler A. 2017. RANGER: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77: 1–17.
- Wulff EW. 1935. Versuch einer Einteilung der Vegetation der Erde in pflanzengeographische Gebiete auf Grund der Artenzahl. *Repertorium Specierum Novarum Regni Vegetabilis* 12: 57–83.
- Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlenn DJ, O'Meara BC, Moles AT, Reich PB *et al.* 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506: 89–92.
- Zhao M, Running SW. 2010. Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. *Science* 329: 940–943.

## Supporting Information

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

**Fig. S1** Observed species and phylogenetic richness of vascular plants for 830 geographic regions used to train the models.

**Fig. S2** Correlations among all predictors and their density distributions.

**Fig. S3** Spatial correlograms of raw diversity data, and residuals from nonspatial and spatial models, respectively, fitted for species richness and phylogenetic richness.

**Fig. S4** Comparison between ensemble predictions of vascular plant diversity across an equal-area grid of 7774 km<sup>2</sup> hexagons and a raster layer of resampled ensemble predictions at 30-arc-second resolution.

**Fig. S5** Estimated effects of predictor variables on species richness of vascular plants across five nonspatial models (partial dependence plots).

**Fig. S6** Estimated effects of predictor variables on phylogenetic richness of vascular plants across five nonspatial models (partial dependence plots).

**Fig. S7** Interaction strength of each predictor variable for explaining species richness (overall) in the spatial random forest model and two-way interaction strengths between the nine top-ranked covariates and all other covariates.

**Fig. S8** Estimated effects of the nine two-way interactions (two-predictor partial dependence plots) in the spatial random forest model for species richness.

**Fig. S9** Interaction strength of each predictor variable for explaining species richness (overall) in the spatial extreme

gradient boosting model and two-way interaction strengths between the nine top-ranked covariates and all other covariates.

**Fig. S10** Estimated effects of the nine two-way interactions (two-predictor partial dependence plots) in the spatial extreme gradient boosting model for species richness.

**Fig. S11** Interaction strength of each predictor variable for explaining species richness (overall) in the spatial neural network model and two-way interaction strengths between the nine top-ranked covariates and all other covariates.

**Fig. S12** Estimated effects of the nine two-way interactions (two-predictor partial dependence plots) in the spatial neural network model for species richness.

**Fig. S13** Interaction strength of each predictor variable for explaining phylogenetic richness (overall) in the spatial random forest model and two-way interaction strengths between the nine top-ranked covariates and all other covariates.

**Fig. S14** Estimated effects of the nine two-way interactions (two-predictor partial dependence plots) in the spatial random forest model for phylogenetic richness.

**Fig. S15** Interaction strength of each predictor variable for explaining phylogenetic richness (overall) in the spatial extreme gradient boosting model and two-way interaction strengths between the nine top-ranked covariates and all other covariates.

**Fig. S16** Estimated effects of the nine two-way interactions (two-predictor partial dependence plots) in the spatial extreme gradient boosting model for phylogenetic richness.

**Fig. S17** Interaction strength of each predictor variable for explaining phylogenetic richness (overall) in the spatial neural network model and two-way interaction strengths between the nine top-ranked covariates and all other covariates.

**Fig. S18** Estimated effects of the nine two-way interactions (two-predictor partial dependence plots) in the spatial neural network model for phylogenetic richness.

**Fig. S19** Relative importance of environmental variables explaining the global pattern of vascular plant diversity across five nonspatial models.

**Fig. S20** Species richness of vascular plants predicted across an equal-area grid of 7774 km<sup>2</sup> hexagons based on different models (i.e. spatial models using machine learning methods and generalized additive models and a nonspatial generalized linear model with interactions).

**Fig. S21** Phylogenetic richness of vascular plants predicted across an equal-area grid of 7774 km<sup>2</sup> hexagons based on different models (i.e. spatial models using machine learning methods and

generalized additive models and a nonspatial generalized linear model with interactions).

**Fig. S22** Species richness of vascular plants based on ensemble predictions across different grid sizes (i.e. spatial models using machine learning methods and generalized additive models and a nonspatial generalized linear model with interactions).

**Fig. S23** Phylogenetic richness of vascular plants based on ensemble predictions across different grid sizes (i.e. spatial models using machine learning methods and generalized additive models and a nonspatial generalized linear model with interactions).

**Fig. S24** Residuals (deviation) from the linear regression between species richness and phylogenetic richness based on ensemble predictions ( $\text{phylogenetic richness} = 22.1 \times \text{species richness}$ ,  $R^2 = 0.947$ ,  $P < 0.0001$ ).

**Fig. S25** Comparison between vascular plant species richness based on ensemble predictions produced in the scope of this paper (SR.Ensemble) and species richness extracted from Kreft and Jetz's predictions (Kreft & Jetz, 2007) (SR.Kreft) (a,  $\text{SR.Kreft} = 1.01 \times \text{SR.Ensemble}$ ,  $R^2 = 0.76$ ,  $P < 0.0001$ ).

**Fig. S26** Vascular plant diversity based on ensemble predictions across an equal-area grid of 7774 km<sup>2</sup> hexagons and mountain regions.

**Fig. S27** Uncertainty in predicted species richness from the five models used for the ensemble predictions (i.e. spatial models using machine learning methods and generalized additive models and a nonspatial generalized linear model with interactions).

**Fig. S28** Uncertainty in predicted phylogenetic richness from the five models used for the ensemble predictions (i.e. spatial models

using machine learning methods and generalized additive models and a nonspatial generalized linear model with interactions).

**Methods S1** Sensitivity analyses of phylogenetic richness.

**Methods S2** Past environmental variables.

**Methods S3** Statistical models.

**Methods S4** Cross-validation.

**Methods S5** Handling of missing values in predictor variables for predicting and calculating predictions in raster format.

**Notes S1** References of checklists and floras from the Global Inventory of Floras and Traits (GIFT) used to compile the regional species composition data.

**Table S1** List of environmental predictor variables hypothesized to affect plant diversity patterns.

**Table S2** Coefficients of a linear model between the residuals (deviation) from the linear regression between species richness and phylogenetic richness, and the 15 predictor variables identified to best explain plant diversity.

**Table S3** Homogenization of biome classifications for current maps, last glacial maximum, Pliocene (mid-Piacenzian), and Middle Miocene.

**Table S4** Model assessment results.

Please note: Wiley is 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.

See also the Commentary on this article by Puglielli & Pärtel, 237: 1074–1077.