PLOS ONE

RESEARCH ARTICLE

Using worldwide edaphic data to model plant species niches: An assessment at a continental extent

Santiago José Elías Velazco[1]*, Franklin Galvão[1], Fabricio Villalobos[2], Paulo De Marco Júnior[3]

1 Laboratório de Ecologia Florestal, Departamento de Ciências Agrarias, Universidade Federal do Paraná, Curitiba, Paraná, PR, Brasil, **2** Laboratorio de Macroecología Evolutiva, Red de Biología Evolutiva, Instituto de Ecología, AC, Xalapa, Veracruz, México, **3** Laboratório de Teoria, Metacomunidades e Ecologia de Paisagens, Departamento de Ecologia, ICB, Universidade Federal de Goiás, Goiânia, GO, Brasil

* sjevelazco@gmail.com

OPEN ACCESS

Citation: Velazco SJE, Galvão F, Villalobos F, De Marco Júnior P (2017) Using worldwide edaphic data to model plant species niches: An assessment at a continental extent. PLoS ONE 12(10): e0186025. https://doi.org/10.1371/journal.pone.0186025

Editor: Manuel Joaquín Reigosa, University of Vigo, SPAIN

Received: May 16, 2017

Accepted: September 22, 2017

Published: October 19, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: SJEV is supported by a doctoral research grant from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. PDMJ and FG are supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico. FV is supported by Consejo Nacional de Ciencia y Tecnología. The authors received no specific funding for this work.

Abstract

Ecological niche modeling (ENM) is a broadly used tool in different fields of plant ecology. Despite the importance of edaphic conditions in determining the niche of terrestrial plant species, edaphic data have rarely been included in ENMs of plant species perhaps because such data are not available for many regions. Recently, edaphic data has been made available at a global scale allowing its potential inclusion and evaluation on ENM performance for plant species. Here, we take advantage of such data and address the following main questions: What is the influence of distinct predictor variables (e.g. climatic vs edaphic) on different ENM algorithms? and what is the relationship between the performance of different predictors and geographic characteristics of species? We used 125 plant species distributed over the Neotropical region to explore the effect on ENMs of using edaphic data available from the SoilGrids database and its combination with climatic data from the CHELSA database. In addition, we related these different predictor variables to geographic characteristics of the target species and different ENM algorithms. The use of different predictors (climatic, edaphic, and both) significantly affected model performance and spatial complexity of the predictions. We showed that the use of global edaphic plus climatic variables generates ENMs with similar or better accuracy compared to those constructed only with climate variables. Moreover, the performance of models considering these different predictors, separately or jointly, was related to geographic properties of species records, such as number and distribution range. The large geographic extent, the variability of environments and the different species' geographical characteristics considered here allowed us to demonstrate that global edaphic data adds useful information for plant ENMs. This is particularly valuable for studies of species that are distributed in regions where more detailed information on soil properties is poor or does not even exist.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Ecological niche and species distribution modeling (ENM and SDM, respectively) are widely-used tool in different fields of plant ecology, including the prediction of new populations of rare species [1]; or potential distribution of invasive species [2]; informing conservation practices for threatened taxa [3,4]; estimating the effect of climatic change on species distributions [5,6]; describing macroecological patterns [7] and studying past species distributions under a paleobiological approach [8]. Despite the broad application of ENMs in botanical studies, debates surrounding several aspects of ENMs continue to date. One of the most important of these aspects relates to the actual data used in ENMs [9].

Data used for conducting ENMs can be grouped in two sets: biogeographical data about the distribution (or presence/absence) of species (i.e. occurrence records) and environmental data (i.e. predictor variables) used to predict those distributions [10]. ENM performance is sensitive to several characteristics of these two datasets [11–21]. For example, regarding biogeographical data, ENMs can be affected by different aspects of the species' distributional patterns and their sampling such as: prevalence (considered here as the ratio between the quantity of presence and absences), range size and spatial autocorrelation [11,12,22]; which in turn are related to the available sample size [14], data biases along road networks or cities [20,23], geographical accuracy [15] and environmental variability captured by the records [13]. All of these aspects can interact with the environmental data selected to fit the ENMs, and affect model accuracy [21,24] and resultant suitability [19]. Even if occurrence data were bias-free, environmental data can still severely affect ENM performance, especially if inappropriate environmental variables are used as predictors [25].

Choosing a particular environmental variable for ENM depends on the modeling purpose and its biological significance to the species under study [26]. Obviously, different species may have particular constraints related to their dependency on environmental factors and no single variable is expected to be equally meaningful for all species. For instance, variables related to soil properties are considered to be particularly important in determining the distribution of plant species, but have little direct effect on the distribution of the majority of animal species [27]. Considering this plant-soil relationship, predictors can be grouped, following [28], in: (i) resource, matter and energy consumed by an organism, such as oxygen, water, macronutrients and micronutrients; (ii) variables that have direct physiological importance, such as pH, cation exchange capacity, aluminum concentration, hydromorphic condition; and (iii) indirect variables that do not have important physiological effects, such as porosity, bulk density, texture (clay, silt and sand fraction) and soil depth.

In contrast with climatic variables commonly used for ENMs, which describe environmental variation at regional scales (a.k.a. "macroclimatic" variables; e.g. CHELSA; [29]), edaphic variables vary at local scales and with great complexity [30]. For example, within the same landscape, climatic conditions can be very homogenous throughout while soil properties can vary widely according to different parental material [31], topographic position [32] or land-use [33]. Indeed, there are several examples in the literature where soil properties control the distribution of plant species or the structure, composition, and physiognomy of a community within an otherwise climatically homogeneous geographical extent. For instance, mangrove distribution is strongly influenced by soil properties such as salinity, acidity, hydromorphy and nutrient supply [34]. Swamp forests, like the *Caxeitais* (dominated by *Tabebuia cassinoides* (Lam.) DC.) of the Brazilian coast, are mainly distributed over organic and hydromorphic soils [35]. The halophyte vegetation from Chile and Europe is restricted to continental salines [36,37]. Furthermore, soil scarcity can also determine natural plant formations such as those inhabiting rocky outcrops [38]. Narrow plant endemics are also frequently associated with

specific types of soil, rock, and bedrock [39]. Even certain soil nutrients can determine the distributional transition from one vegetation type to another, such as that between Neotropical seasonal forests and savannas where the concentration of aluminum or potassium define the structure of these vegetation types [40,41].

Consequently, it is clear that edaphic conditions play an important role in determining the niche of terrestrial plant species [25,42]. Accordingly, several studies have tested the effect of including edaphic variables in ENMs for plant species such as the importance of soil nutritional variables for predicting plant distribution [43]; the improvement of plant ENMs performance when using physical and chemical soil data [21,24]; and the effect of both landscape and edaphic data in predicting future plant distributions under climate change scenarios [9,44]. All of these studies reinforced the idea that plant ENMs could be improved by using a single or a group of edaphic variables. Unfortunately, edaphic variables are still not frequently used as predictors in plant ENMs, which continue to be limited to climatic variables [42]. One reason for this lack of consideration of edaphic variables in plant ENMs may be related to the geographical extent for which these data are available. Such availability has been usually restricted to certain countries or regions (e.g. USA, China or the European Union), whereas in many other regions, as in many Latin America countries, these data are simply not available. Recently, however, the ISRIC World Soil Information with the SoilGrids database has provided data related to physical, chemical and taxonomical characteristics of soils across the globe [45]. Therefore, this database allows the construction of ENMs for plant species inhabiting large regions of the world or species occurring in countries that differ in the quantity and quality of the available edaphic data.

Indeed, despite including detailed soil data, most plant ENM studies have been conducted on extents that are usually smaller than the complete geographic distribution of plant species. Such ENMs may not comprise the full environmental variability that characterizes a species distribution and thus may affect model performance [46,47]. Here, we evaluate the potential effect of using the SoilGrids global dataset in improving ENMs for plant species. We used 125 species distributed along the Neotropical region of the Americas, where many countries do not have detailed soil data, to explore the effect of using global edaphic data and its combination with climatic data in the prediction of models constructed under commonly used ENM algorithms. In addition, we related the different variable sets to certain geographical characteristics of target species (e.g. occurrence area, number of records and density of records) and different algorithms.

Methods

Overview

To evaluate the effect of adding global edaphic data into ENMs and its relationship with different modeling algorithms, we adopted a factorial experimental design with two factors: Predictor and Algorithm, with three and four levels respectively, totalizing 12 combinations of factor levels. The first factor, Predictor, comprised three different models based on different predictor sets (edaphic, climatic, or both). For the second factor, Algorithm, we used four types of ENM algorithms (Fig 1). The 12 treatments were applied to 125 plant species, our experimental units, thus 1500 models were fitted (see below).

Study area

Our study area extended from the south of the United States of America to the austral extremes of Chile and Argentina. This area covers a wide variety of climatic conditions, geological formations, and soil types, but many of its constituent countries lack edaphic data.

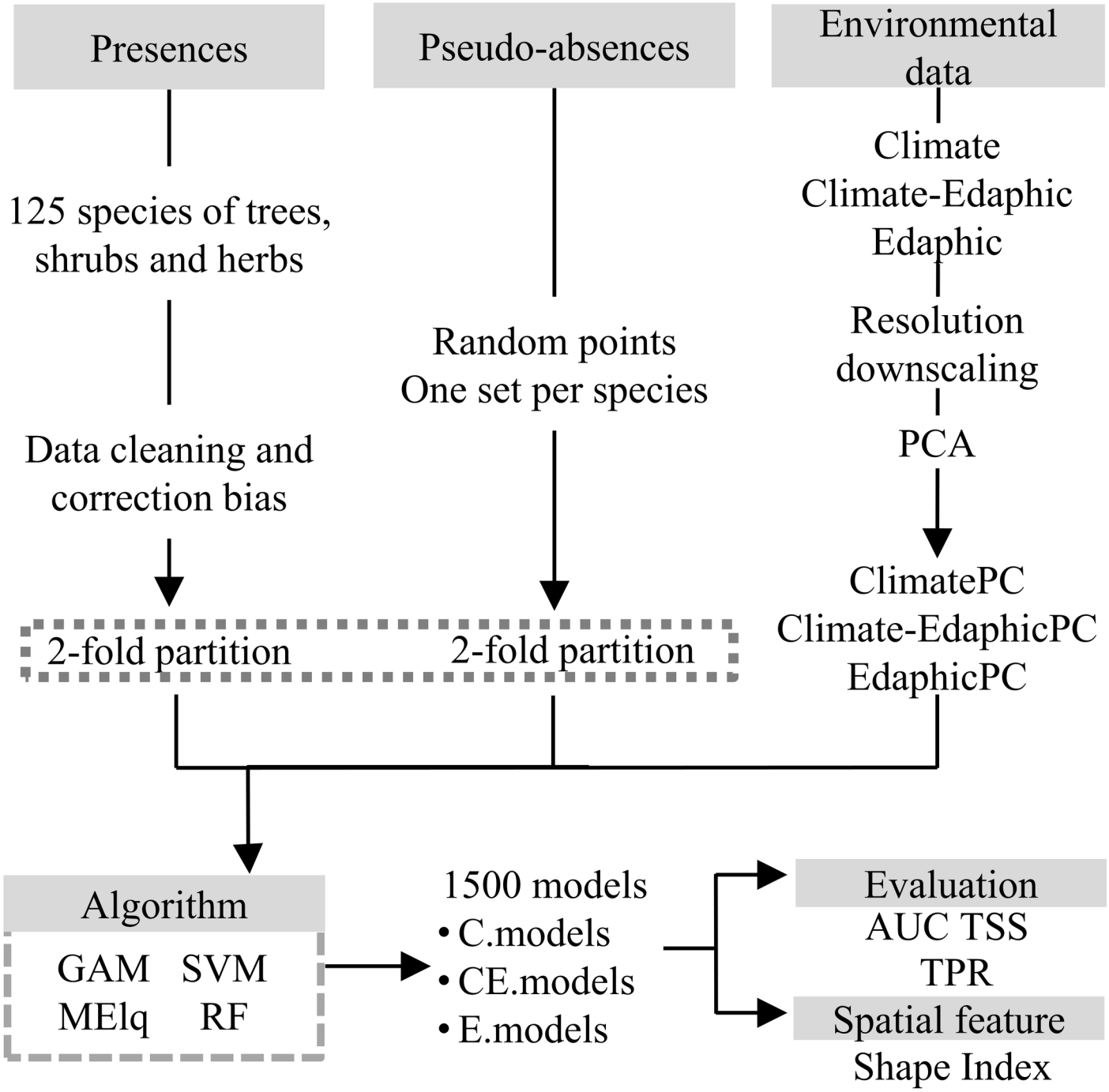

Fig 1. Experimental design for testing the effect of using edaphic variables in ENMs for plant species.

https://doi.org/10.1371/journal.pone.0186025.g001

Consequently, our selected plant species (see below) occur in different biomes, from arid regions such as the *Chihuahuan* and *Caatinga* steppe and warm-humid biomes such as the Brazilian Atlantic and *Chaco-Darién* moist forests to the cold regions of the *Nothofagus* forests and *Andean páramos* (see S1 Table for the complete species list).

Environmental data

We used three sets of environmental variables for building ENMs: climate-only, edaphic-only and both climatic and edaphic variables together [21,43], hereafter called C.models, E.models and CE.models, respectively. Note that all of these predictors were continuous variables. For the C.models, we employed the 19 bioclimatic variables from the recently developed CHELSA v1.1 online database [29]. These variables were built based on monthly averages of climate data, mainly temperature and precipitation as collected from meteorological stations, for the 1979–2013 period and interpolated to the global surface [29]. E.models were built with 56 variables related to physical and chemical soil properties obtained from the SoilGrids database available from ISRIC-World Soil Information [45], the data were downloaded in June of 2015 (Table 1). The SoilGrids database provides global maps of soil classes and some edaphic variables (see Table 1). In addition, this database has an automated updating system that progressively increases its accuracy when new input data becomes available in the international soil profile databases [45]. The CE.models were built combining the climate and edaphic datasets, summing up to 75 variables. Both climatic and edaphic datasets were acquired with a spatial resolution of 30 arc-seconds ($\approx$ 1 km^2 cell size) and upscaled to 5 arc-minutes ($\approx$ 10 km^2 cell size). This upscaling (resolution change) was based on the aggregation, by taking their average value, of lower resolution cells into higher resolution cells. Later, these datasets were cropped to the extent of the study region ranging from -120˚ to -30˚ in longitude and -60˚ to 35˚ in latitude.

Different modeling approaches present different sensitivity to collinearity of predictor variables [48]. However, no single methodological procedure has been considered ideal for solving or handling collinearity [48]. Here, we opted to conduct a principal component analysis (PCA) on the original environmental dataset and use the scores of each derived principal components (PCs) as new predictors variables [16,49]. The PCA is a multivariate technique that produces uncorrelated components from the original data sorted according to the amount of total variance that it explains. We selected a number of PCs that explained more than 95% of

Table 1. Climate and edaphic variables (names and units) used as predictors in plant ecological niche models.

Climate (Source: CHELSA)	Unit	Edaphic (Source: SoilGrids)	Unit
Annual Mean Temperature	˚C	Depth to bedrock up to maximum 240 cm	cm
Mean Diurnal Range	˚C	Predicted probability of occurrence of R horizon	%
Isothermality	˚C	Mean of bulk density*	kg/m3
Temperature Seasonality	˚C	Mean of coarse fragments volumetric*	%
Max temperature of warmest week	˚C	Mean of soil texture fraction clay*	%
Min temperature of coldest week	˚C	Mean of soil texture fraction silt*	%
Temperature annual range	˚C	Mean of soil texture fraction sand*	%
Mean temperature of wettest quarter	˚C	Mean of cation exchange capacity*	cmolc/kg
Mean temperature of driest quarter	˚C	Mean of soil organic carbon stock*	Tn/ha
Mean temperature of warmest quarter	˚C	Mean of soil organic carbon content*	‰
Mean temperature of coldest quarter	˚C	Mean of soil pH in H2O*	
Annual precipitation	Mm		
Precipitation of wettest week	Mm		
Precipitation of driest week	Mm		
Precipitation seasonality	C of V		

*Data for six depths

https://doi.org/10.1371/journal.pone.0186025.t001

the total variance in the original dataset [46]. The major advantages of this procedure are the correction of multicollinearity among the original variables, the use of almost all information contained in a large dataset that is captured in the PCs, and the reduction of the number of variables used in the models. Accordingly, C.models and E.models were built with the first six PCs and the first 11 for the CE.models (see S2 Table and S3 Table for more information about variance explained and variables' coefficients for the selected PCs). The reduced number of new variables (PCs) reveals the high collinearity in our original variable set. In fact, the first two PCs of the PCAs conducted for each variable set explained more than 50% of the variance (in the S1 Fig the relationships of original variables and the first two PCs of each variable set are depicted).

Plant species data and cleaning

We selected 125 terrestrial plant species distributed within the Neotropical region with the purpose of considering the wide variety of environmental conditions in our study region. Data for these taxa was restricted to the species level, thus infraspecific taxa were not considered. Our final species dataset comprised trees (82), shrubs (27), herbs (8) and palm (8) species. We considered only species with more than 20 checked records (described below; see S1 Table). This dataset comprised species inhabiting extreme latitudes such as *Atriplex canescens*, *Prosopis glandulosa* or *Parthenium incanum* in the north, and *Nothofagus antarctica*, *N. pumillo*, and *Mulguraea tridens* in the south. These species also differ in regard to their geographic range sizes, from those with narrow distributions such as *Juglans australis* to those considered as cosmopolites such as *Trema micranta*, *Ipomoea carnea* and *Inga vera*.

We conducted a taxonomic revision for these taxa verifying their accepted names and synonymy using The Plant List Version 1.1. (http://www.theplantlist.org/) and Tropicos (http://www.tropicos.org), checked by the Taxonomic Name Resolution Service v3.2 [50] based on APG III [51]. After confirming accepted names and synonymy, we used these names to search occurrence records for these species in the Global Biodiversity Information Facility (http://www.gbif.org/) and the speciesLink database (http://splink.cria.org.br/).

Occurrence records available in those databases may contain some taxonomic and geographic coordinate errors [52]. Our first step for data cleaning was the elimination of all records allocated outside the study area and those with repeated geographic coordinates. We also removed those species' records corresponding to invasive or cultivated distributions, thus leaving only those records that pertain to the natural distribution of species. This last step was conducted by using information about species distributions available in the Catalogue of Life (http://www.catalogueoflife.org/), Flora del Conosur (www.darwin.edu.ar/Proyectos/FloraArgentina/fa.htm), List of Species of the Brazilian Flora (http://floradobrasil.jbrj.gov.br/), Smithsonian Tropical Research Institute-Scientific Databases (http://biogeodb.stri.si.edu/bioinformatics/en/), PLANTS Database (http://plants.usda.gov/java/) and Tropicos national species list from Bolivia, Panamá, Paraguay, Peru and Ecuador (http://www.tropicos.org/). In order to clean records temporally, we only considered records that were collected between 1979 and 2013, thus corresponding to the temporal span of our climate variables.

It is common that species' occurrence records are biased towards roads, cities or countries [23,53,54]. Therefore, these records are not the result of random and homogeneous sampling along the geographic distribution of a species, compromising the accuracy of ENMs [20]. We used a systematic sampling given its suggested effectiveness to correct geographic bias [55] by creating a grid with a resolution of 10 arc min (≈ 20 km^2 cell size) and then selecting one occurrence per cell. The number of cleaned records for species ranged from 20 to 1227 (see S1 Table).

Modeling procedures

The diverse algorithms usually employed to build ENMs have different input requirements [56], degrees of complexity [57], stability [58] and predictive abilities [22,59]. For these reasons, we also explored how different algorithms respond to distinct sets of environmental variables. We used four methods which are commonly used in ENM and highlighted for their performance: Generalized Additive Models (GAM), Maximum Entropy (ME), Random Forest (RF) and Support Vector Machine (SVM).

GAM is a non-parametric extension of GLM (Generalized Linear Model) that replaces the linear relationship between the dependent and independent variables by the sum of a smooth function [60]. Owing to its combination of a link function and a smooth function, the GAM method has the ability to deal with highly non-linear and non-monotonic relationships between the response and explanatory variables [61]. These GAMs were fitted using a binomial distribution with all single predictor variables, i.e. without backward or forward selection and interaction. The Newton method was used to optimize the estimation of the smoothing parameter.

ME is a machine learning method based on the principle of maximum entropy [62,63]. This principle is based on minimizing the relative entropy between two probability densities defined in feature space [64]. This is a high performance technique [59] and is less sensitive to spatial errors than others algorithms [65]. This method can be tuned with different features such as linear, quadratic, product, threshold, hinge and binary; the default use of all these features can cause overfitting and affect the models performance [66]. Thus we used linear and quadratic features ([67], hereafter MElq), both of these constrain the approximation of the probability distribution in a way that the variables' mean and variance should be close to its observed values [62]. We also used 1000 maximum iterations, default regularization values, logistic output format and 10000 maximum background points.

SVM uses linear models to find a decision function, which is a hyperplane determined by non-linear decision boundaries that split samples in different classes within a higher-dimensional space [68,69]. The optimal hyperplane is the one that maximizes the buffer between the boundary (i.e. support vectors) and the data [70]. Mapping of the input data in a high-dimensional feature space is defined by a kernel function [71]. These models were built based on probability classes, performed with a radial basis kernel (RBF) and with a constant cost value ($C = 1$).

RF comprises a family of algorithms that perform classification and regression analyses. RF is a modification of bagging trees, which build a model based on the average of a large collection of non-correlated trees [72]. In each node of these trees, a random sample of m predictors is chosen as split candidates from the full set of predictors [73]. These algorithms have the advantage of not overfitting the data [74] and use the out-of-bag (OOB) sample to construct different variable importance measures [72]. To determine the optimal number of variables randomly sampled at each split the RF algorithm was tuned automatically. 500 trees were used at the tuning step, with default values of the step factor and the improvement in OOB error parameter. We considered those models with the minimum OOB error as our final RF models.

Given that we did not have real absences of our species, we created pseudo-absences to fit GAM, SVM and RF models. The prevalence and the method of pseudo-absence allocation can affect ENM performance, which can vary for distinct algorithms [75–77]. To reduce potential noise, we used a prevalence of 1, thus the number of pseudo-absences for each species was equal to its presences. These pseudo-absences were allocated across the study area, which constitutes the biogeographic domain that the modeled species could have used as an accessible

area over relevant periods of time [78,79]. We used one soil layer as a raster mask for creating the pseudo-absences given that some cells with climate data may have no soil data (i.e. "empty cells"), such as lakes and some mountain regions.

Model evaluation

Models were evaluated by a 2-fold cross-validation where the presences of each species and its respective pseudo-absences were partitioned into 50–50% training-testing sets. To control for spatial autocorrelation between training and testing records, we used a checkerboard partitioning method similar to [80]. This method generates checkerboard grids that partition the records into bins by subdividing the geographic extent equally. For this, a particular grid resolution (i.e. cell size) must be chosen a priori, which does not guarantee a balanced number of records in each bin [80]. Therefore, we adapted the method to select the grid resolution that optimizes representation and balance of records within bins. To do so, we created 30 grids with resolutions varying from 0.5 to 15 degrees, with a gradual increase of 0.5. The optimum grid resolution was the one which (i) represented both training and testing records and (ii) minimized the difference between the number of training and testing records. Finally, to maintain a prevalence of 1, we randomly allocated pseudo-absences within each partition group.

Model performance was assessed by dependent and independent threshold metrics [81]. We used the True Positive Rate (TPR) and the True Skill Statistic (TSS) [82] as threshold-dependent indices and the Receiver Operating Curve (AUC) as a threshold-independent evaluation. The threshold was the value that maximized the sum of sensitivity and specificity that produced the most accurate predictions [83]. The complexity of the different spatial patterns of binary predictions (ENM outputs) was evaluated using the shape index (SI). This index measures the complexity of the predicted patches of pixels (i.e. potentially suitable cells) by considering the relationship between the sums of each patch perimeter (p_i) divided by the square root of patch area (a_i), $\mathrm{SI} = \sum_1^i (0.25\, p_i/\sqrt{a_i})$ [84].

Data analysis

We used Repeated Measures ANOVAs to test the effect of the Predictor (C.models, E.models and CE.models), Algorithm (e.g. GAM, SVM, etc.) and their interaction on TPR, TSS, AUC and SI indices. We assumed that the Predictor and Algorithm as within-subject factors. To perform this analysis correctly and avoid a high Type II error rate, it was necessary that the data met the sphericity condition: the variances of the differences between combinations of levels do not differ. We used Mauchly's Sphericity Test at 95% confidence to validate the sphericity condition of the covariance matrix. When this condition was rejected, the degrees of freedom were corrected by the Greenhouse-Geisser method and used Type III sums of squares. We performed a post-hoc test using linear contrasts based on linear mixed effect models, considering the Predictor and Algorithm as fixed factors and the species as random factor. These models were used to perform pairwise comparisons of means between different predictors for a single algorithm at 95% confidence level. The p-values were corrected using the false discovery rate procedure.

After evaluating the models, their predicted suitabilities were projected onto the geographical space. For each species, we conducted pair-wise comparisons between the suitabilities of different kinds of models and algorithms by calculating the Kendall rank correlation coefficient (τ) with cells of the entire study area. Values of this coefficient range from -1 (perfect disagreement) to 1 (perfect agreement), with values near zero representing independence between the compared ranks.

We used Pearson correlation (r) to explore the relationship between variation captured by records for different predictor sets and species' geographic characteristics, which were, for each species: (i) geographical extent, based on the number of cells within a minimum convex polygon comprising all of a species' records; (ii) number of records and (iii) density of records, which is the ratio between a species' number of records and its geographical extent. For each ENM algorithm, we explored the effect of such species' characteristics and predictors on TSS [12] by fitting linear mixed-effect models. These characteristics were considered as fixed effects within the mixed-effect models, along with the models with different predictors (C.models, E. models and CE.models), whereas the species were considered as random effects. TSS values were arcsine transformed. We used the variance inflation factor (VIF) to test for collinearity among predictors (species geographic characteristics), their significances were determined by a likelihood ratio test.

Construction of ENMs and statistical analyses were conducted in the R environment v. 3.3.2 [85]. The *dismo* v. 1.1.1 package [86] was used to create pseudo-absences, model prediction and validation, and to fit MElq using Maxent v. 3.3.3. The GAMs, SVMs and RF models were fitted using the *gbm* v. 2.1.1 [87], *kernlab* v. 0.9.25 [88] and *randomForest* v. 4.6.12 [89] packages, respectively. We used the packages *raster* v. 2.5.8 [90], *SDMTools* v. 1.1.221 [91], and *pcaPP* v. 1.9.61 [92] to handle raster, calculate the shape index, and the Kendall rank correlation coefficient. To fit the linear mixed effect models, repeated measures ANOVAs and the pairwise mean contrasts, we used the packages *nlme* v. 3.1.128 [93], *lsmeans* 2.26.3 [94] and *car* v. 2.1–5 [95], respectively.

Results

The use of different predictors (climatic, edaphic, and both) significantly affected model performance, as measured by the TSS, TPR and AUC indices. They also affected the spatial complexity of the geographic predictions (SI). Moreover, TSS, TPR, AUC and SI showed different responses regarding the use of the distinct ENM algorithms. The interactions between predictors and algorithms were significant for TSS, AUC and SI (Table 2).

According to TSS, TPR, and AUC, C.models and CE.models performed better than E.models, regardless of the ENM algorithm used. Nonetheless, MElq performed better for the CE. models, regarding TSS, whereas SVM, GAM and RF did not show differences between C.models and CE.models. These results were different for the sensitivity, given that the CE.models showed the best values for SVM. Moreover, no algorithm differed significantly regarding only the C.models (Fig 2A). Regardless of the predictor set, the SVM and RF algorithms had the highest values of TSS and AUC, followed by MElq and GAM. Regarding the spatial complexity of predictions, C.models showed the most aggregated and continuous prediction, whereas the E.models had the most spread and complex patterns. The CE.models had an intermediate shape complexity. Independent of the predictor set, RF created the most complex spatial patterns, whereas SVM showed the lowest SI (Fig 2A).

Mean values of Kendal rank correlation of suitabilities always showed positive values for the pair-wise comparison of models with different predictors (Fig 3A). The highest values of suitability correlation were for C.models-CE.models and E.models-CE.models for all algorithms. MElq had the most similar suitability for these paired comparisons. The lowest correlation was between the suitability of C.models and E.models, with mean values smaller than 0.4 for all algorithms. For the C.models-E.models comparison, the highest correlations were for MElq and RF, whereas in the C.models-CE.models comparison, highest correlation was for MElq and GAM, and for the E.models-CE.models comparison highest correlations were for MElq, RF and SVM (Fig 3A). These comparisons of suitability between algorithms showed

Table 2. Results of the repeated measures ANOVA for the TSS, TPR, AUC and SI, considering the algorithm (GAM, MEql, SVM and RF) and predictor (climate, climate-edaphic, edaphic) factors.

Index	Factors	Sum of Squares	Df	Mean Square	F
TSS	Algorithm	0.577	2.501	0.294	79.429***
	Predictor	1.344	1.463	0.858	76.316***
	Algorithm * Predictor	0.027	4.220	0.006	4.657***
TPR	Algorithm	0.043	2.588	0.021	13.024***
	Predictor	0.242	1.560	0.184	44.654***
	Algorithm * Predictor	0.008	5.119	0.001	1.505^{ns}
AUC	Algorithm	0.261	2.035	0.107	76.929***
	Predictor	0.379	1.457	0.249	70.262***
	Algorithm * Predictor	0.014	3.541	0.004	6.013***
SI	Algorithm	17965.610	1.705	10534.815	176.159***
	Predictor	77863.060	1.487	52370.852	690.234***
	Algorithm * Predictor	2286.069	4.684	488.070	47.693***

Degrees of freedom were corrected using Greenhouse-Geisser estimate of sphericity. TSS: true skill statistic; TPR: true positive rate, AUC: area under curve; SI: shape index.

Significance

*** $P < 0.001$

** $P < 0.01$

* $P < 0.05$

^{ns} $P > 0.05$

https://doi.org/10.1371/journal.pone.0186025.t002

that GAM-MElq were the most similar, followed by RF-SVM, for any predictor set. The lowest correlation was found between GAM-SVM and between MElq-SVM (Fig 3B).

The relationship between different geographic characteristics of species revealed that species with wider distributions had also more records sampled ($r = 0.790$, $p < 0.001$) but they showed lower record density ($r = -0.640$, $p < 0.001$). However, the relationship between number of records and their density was weak ($r = -0.090$, $p < 0.334$) (S2 Fig). Widely distributed species presented higher standard deviation for the first principal component of the climate predictors ($r = 0.670$, $p < 0.001$). These patterns were weaker for the climate-edaphic ($r = 0.240$, $p = 0.007$) and edaphic predictors set ($r = 0.130$, $p = 0.146$) (S3 Fig; S4 Table).

The linear mixed-effect models revealed, for all algorithms, that the species geographical extent negatively affected the TSS, the number of records affected SVM, whereas for this algorithm number of records affected the TSS positively, implying better model performance (S4 Fig). In addition, we found that different predictors affected model accuracy but the interaction among predictors and species geographic characteristic differed among the ENM algorithms. Interaction between predictor sets and geographical extent were significant for GAM, MElq and RF, but not for SVM, which had significant interaction between the number of records and predictor sets. (Table 3; S4 Fig). Species geographic characteristics, predictor sets, and their interaction explained between 54 and 68% of model performance (TSS) variability (Tabxle 3).

As expected, the response of the ENMs to different predictors varied individually for each species. Thus, for some species, the use of edaphic data (E.models and CE.models) considerably improved model accuracy in comparison with those models constructed with climate-only predictors (e.g. *Astronium graveolens*, *Cedrela odorata*, *Ficus insipida*, *Genipa Americana*, *Guarea glabra* and *Salix humboldtiana*). Conversely, edaphic-only predictors notably decreased model performance for other species (e.g. *Casearia decandra*, *Phytolacca dioica*,

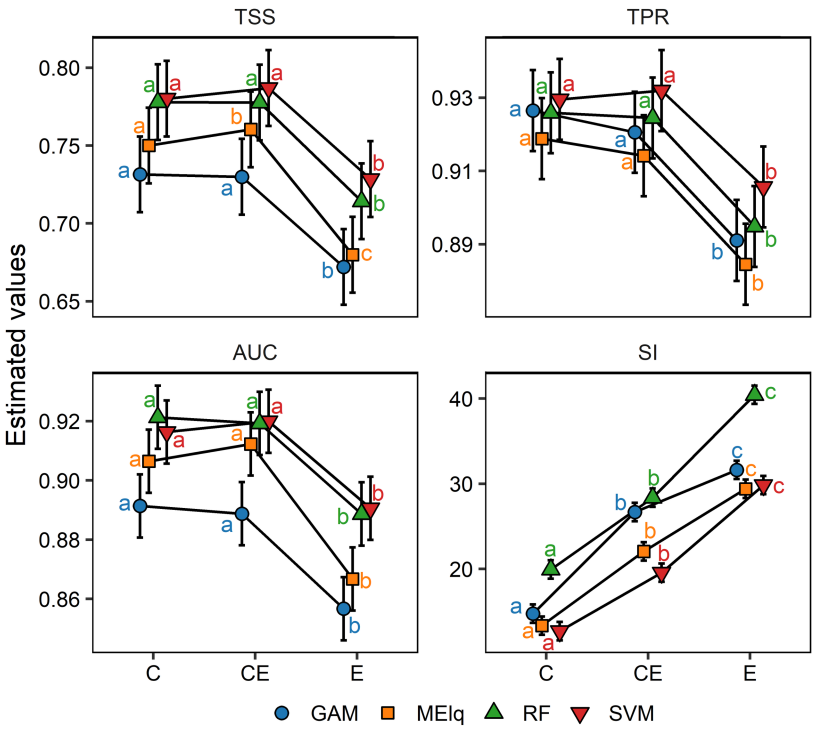

Fig 2. Estimated values and confidence interval (bars) for TSS, AUC, TPR and SI of models fitted with three set of predictors combined with four algorithms. Means with same letter for different predictor and same algorithm denote significant difference using the linear contrast ($P < 0.05$). TSS: true skill statistic, TPR: true positive rate, AUC: area under curve; SI: shape index, C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.

https://doi.org/10.1371/journal.pone.0186025.g002

Hevea brasiliensis, *Matayba eleagnoides*, *Schinus molle* and *Baccharis crispa*). In addition, there were species that presented similar outputs irrespective of the kind of predictors used for modeling (e.g. *Chuquiraga avellanedae*, *Nothofagus pumilio* and *Persea schiedeana*; Fig 4). The effect of different predictor variables on the species' suitability pattern varied among species. For example, for species such as *Salix chilensis* and *Guarea glabra* that have broad distributions, CE.models and E.models showed suitable areas that were more constrained compared to those from C.models. Conversely, *Bulnesia sarmientoy* showed an expansion of the suitable areas for those models that used edaphic data compared to those that did not include these data (Fig 5).

Based on SVM models, better performance when using the CE.models was observed for 54 species, whereas for 53 species this was true when using the C.models and five species showed the best model performance when using E.models. Also, there were species whose models had the same maximum accuracy independent of the predictors set used. For example, models with climatic-only or climatic-edaphic predictors performed equally well for 13 species, whereas edaphic-only or climatic-edaphic predictors did the same for three species. Finally, models for only three species showed the same TSS irrespective of the considered predictor variables (Fig 5).

Discussion

We have shown here the advantages of using worldwide edaphic data as predictors in ecological niche models for plant species. More specifically, we showed that ENMs constructed with commonly used climatic variables plus edaphic variables did not affect negatively the

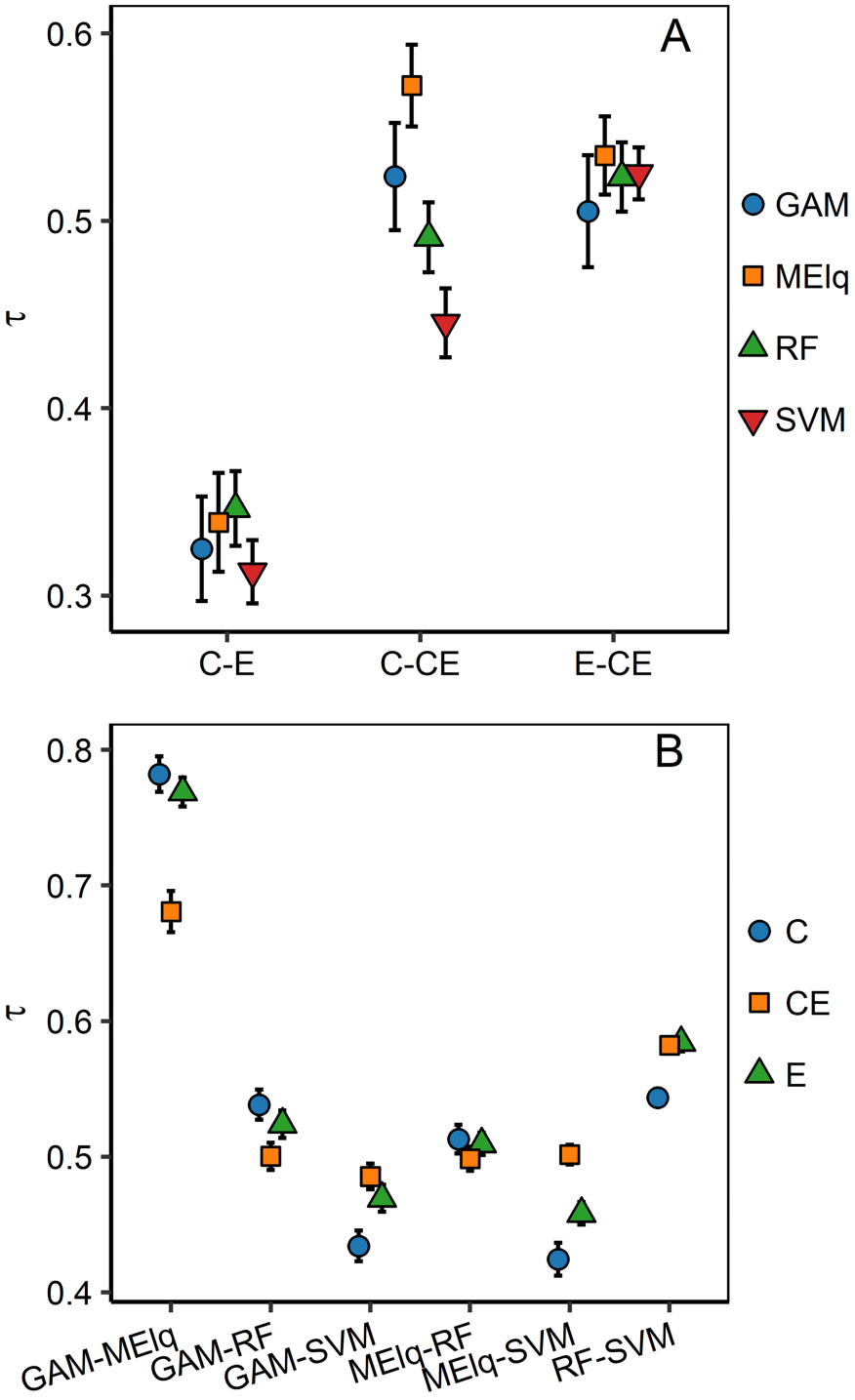

Fig 3. Mean and confidence interval of Kendall rank correlation coefficient (τ) of pair comparison of suitability. (A) Suitability comparison between models with predictors sets for different algorithm. (B) Suitability comparison between algorithms for different predictors sets. C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.

https://doi.org/10.1371/journal.pone.0186025.g003

performance of ENMs but instead improved the accuracy for some algorithms. This happened even when the ENMs based only on edaphic variables did not provide accurate predictions for

Table 3. Summary of linear mixed effect models for four algorithms and the significance of covariates. The model selection was based on the likelihood ratio test.

Algorithm	Covariates	LRT χ^2	Df	p-value	R^2
GAM	GE	110.204	1	<0.001	
	NR	0.081	1	0.777	
	DR	0.001	1	0.969	
	Predictor	63.803	2	<0.001	
	GE*Predictor	21.418	2	<0.001	
	NR*Predictor	5.704	2	0.058	
	DR*Predictor	0.628	2	0.628	54.255
MElq	GE	110.405	1	<0.001	
	NR	2.104	1	0.147	
	DR	0.003	1	0.957	
	Predictor	12.642	2	0.002	
	GE*Predictor	10.736	2	0.005	
	NR*Predictor	4.654	2	0.097	
	DR*Predictor	0.191	2	0.909	68.686
RF	GE	138.093	1	<0.001	
	NR	0.424	1	0.515	
	DR	0.387	1	0.534	
	Predictor	139.853	2	<0.001	
	GE*Predictor	2.678	2	0.262	
	NR*Predictor	3.558	2	0.169	
	DR*Predictor	1.876	2	0.391	64.059
SVM	GE	59.592	1	<0.001	
	NR	19.461	1	<0.001	
	DR	0.281	1	0.596	
	Predictor	133.931	2	<0.001	
	GE*Predictor	4.389	2	0.111	
	NR*Predictor	19.322	2	<0.001	
	DR*Predictor	0.233	2	0.890	68.614

GE: geographical extent; NR: Number of records; DR: density of records; Predictor: models constructed with climate, climate-edaphic or edaphic variables; Df: degree of freedom; LRT χ^2: Chi square for the likelihood radio test. R^2: marginal determination coefficient calculated for the final models with significance values < 0.05 of their covariates.

https://doi.org/10.1371/journal.pone.0186025.t003

any algorithm. Owing to the particular spatial patterning of these different predictor sets, climatic and edaphic, their use also affected the shape and spatial complexity of ENM outputs. In addition, the performance of models considering these predictor sets, separately or jointly, was strongly related to geographic properties of species records, irrespective of the algorithm. Our findings highlight the feasibility and advantages of including global soil data, along with climatic variables, into ENMs to achieve accurate predictions of plant species distributions.

Soils are the consequence of different forming factors such as climate, organisms, topography, parent material, time [96], among other local factors [97]. Any particular combination of these factors will give rise to particular processes that can be extremely complex [97,98], involving disintegration, integration, weathering, decomposition, neoformation and transformation [99]. These factors and processes acting in soil genesis define the chemical and physical characteristics of soils, which will ultimately determine the underground environment for terrestrial plants. We want to reinforce here the widely accepted idea that soil is one of the most

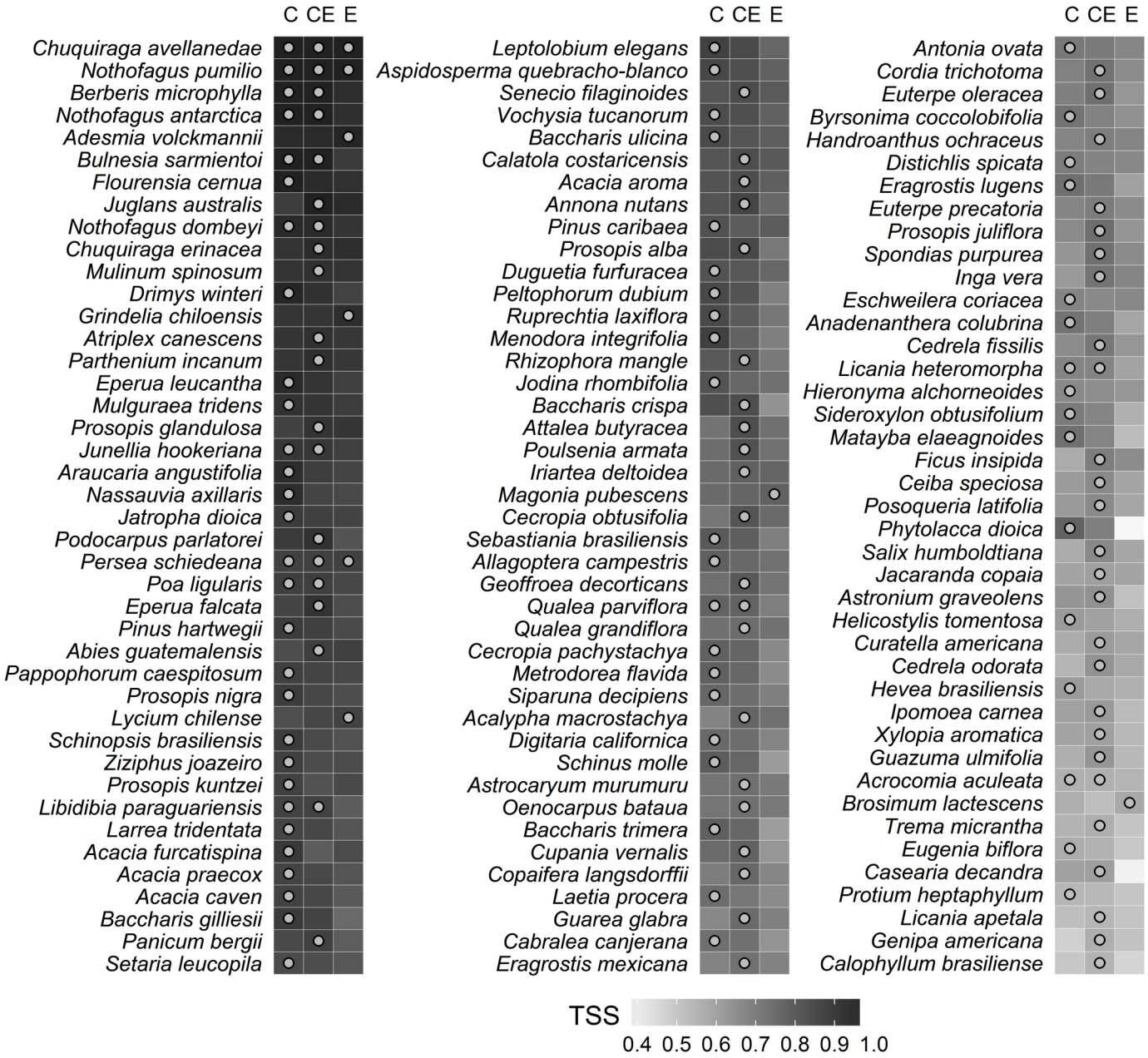

Fig 4. List of modelled species and model accuracy performed with SVM and three predictors sets. Species are sorted by their overall TSS mean, the points represent the highest model accuracy for a particular species. TSS: true skill statistic, C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.

https://doi.org/10.1371/journal.pone.0186025.g004

important factors affecting plant ecology and that the common practice of only using climatic predictors represents a weak conceptual basis for the application of ENMs for plant species (see also [25]). Despite potential disadvantages of our testing framework, such as the lack of real species absences and the extension of the edaphic data used, we demonstrated that the combination of climatic and global edaphic predictor variables increases the accuracy of ENMs for several of our studied plant species. In fact, our findings are consistent with previous studies that found similar results for plants at smaller extents and finer resolutions [9,21,43,44].

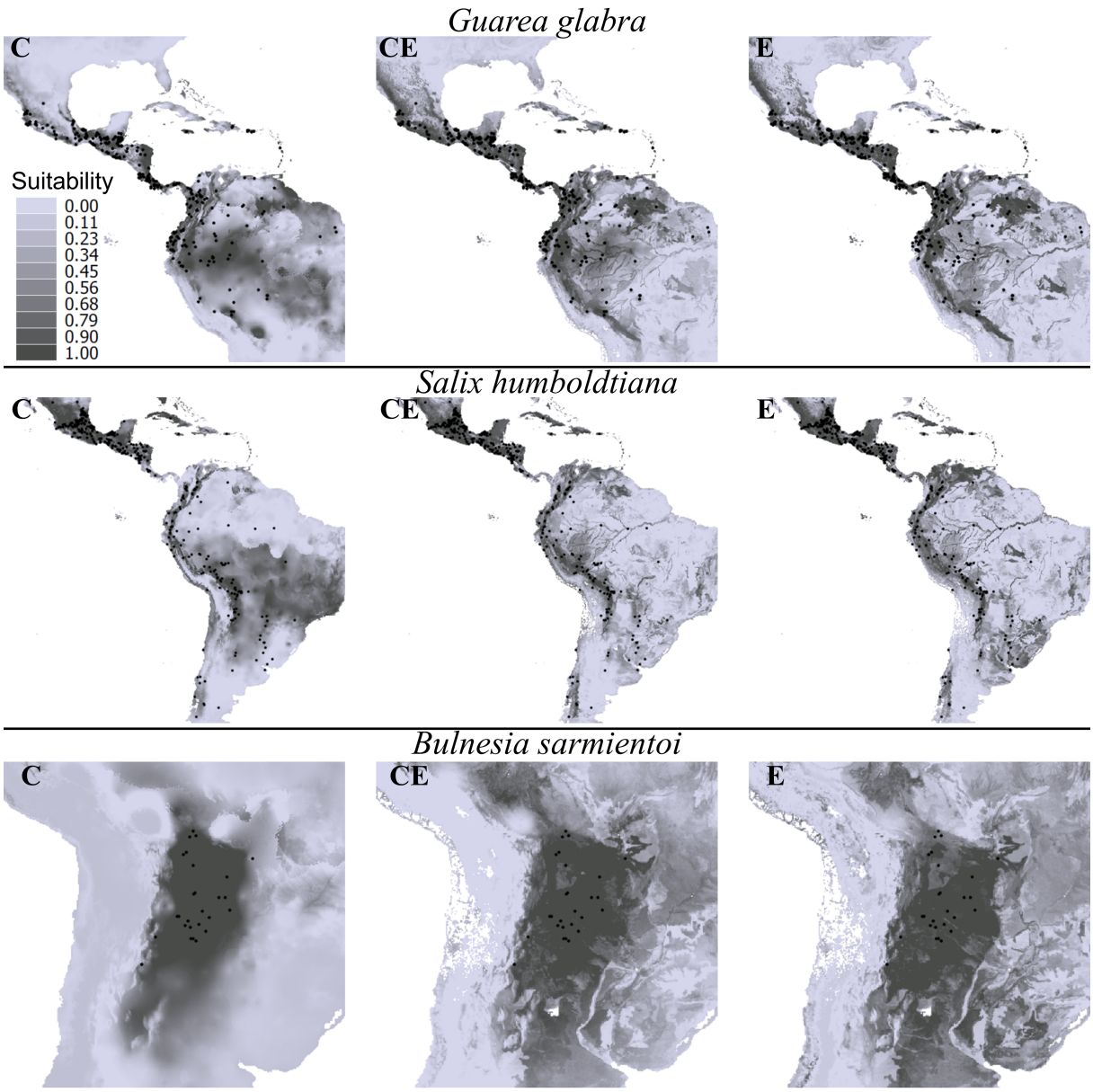

Fig 5. Examples of change in suitability predicted by SVM models for three species derived from the use of three predictors set. C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.

https://doi.org/10.1371/journal.pone.0186025.g005

The type of organisms under study must guide the selection of predictor variables for ENMs and SDMs. Accordingly, soil properties should be considered when applying ENMs for plant species, whereas these properties may be neglected when modeling animal species [27]. However, few studies doing ENM or SDM with plant species have used variables related to soil or, for that matter, variables different than climatic ones [42,100]. Adding or excluding variables when describing the species environmental niches can affect the form of the resulting hypervolume in multivariate space (S5 Fig) [101]. Owing to the reciprocity between the environmental and geographic space [102], changes in the multivariate space can affect ENM predictions on geographic space. In our case, this fact may be responsible for the low consensus

between the suitability patterns (Fig 3A) and the geometry of predicted maps (Fig 2) of our modeled species. Regardless of the ENM algorithm used (e.g. GAM vs MElq), using edaphic predictors alone or jointly with climatic variables (E.models and CE.models, respectively) augmented the spatial complexity of the predicted plant distributions. This results from the complex spatial variation of soil properties [30] in comparison with climatic variables. Indeed, the spatial complexity of edaphic variables could be responsible for the lower TSS and AUC values predicted by our models based only on edaphic predictors, which are consistent with results obtained for several plant species in Canada with models constructed with the same type of variable and more detailed edaphic data [21]. Given that different algorithms or variables could produce models with the same accuracy but with different spatial predictions (see Fig 5; [103]), we highlight the importance of acknowledging that model evaluation must be based both on an accuracy metric (e.g. TSS, AUC, etc.) and a preliminary visual examination based on the ecological knowledge of the studied species and their relationships with the selected predictors [26].

It is well known that the performance of an ENM algorithm can vary according to the characteristics of the species niche [104], the training dataset [16,77,105], and how the algorithm is tuned [66,106]. One reason that could explain the discrepancy on the accuracy of our different algorithms is the prevalence between presence and absences. For instance, ENMs built under SVM and RF produce better models when the presences/absences ratio is 1, whereas GAM models achieve higher performances with lower presences/absences ratio values [107]. In our case, the fact that we used a number of pseudo-absences equal to that of presences for all algorithms can have negatively affected GAM compared to the other algorithms. Nevertheless, the high positive correlation between GAM and MElq suitabilities is noteworthy. Such correlation may result from the ME tuning with linear and quadratic terms, a feature that reduces the complexity of the ME model making it more similar to GAM. Despite ME being characterized as a high-performance algorithm [59], here this method was outperformed by SVM and RF even when it was conducted with 10,000 background points instead of using the pseudo-absences as in the other algorithms [106]. This finding may be particular for ENMs of plant species, for which SVM and RF have been referred as two of the most accurate algorithms for modeling Neotropical plants [108]. In addition, we found that SVM and RF were the most accurate methods, although producing different predictions, a fact that reflects their ability to represent complex non-linear relationships. Moreover, each one of these algorithms has important advantages; RF constructs models that avoid overfitting [74] whereas SVM has the ability to construct stable models even with a large set of covariates [108]. In fact, both algorithms were the best classifiers with the UC Irvine Machine Learning Repository [109].

The geographic range size of species can influence the performance of ENMs [11,12,22]. For instance, small-ranged species could have a limited variability of environmental conditions captured by its presences [110] and results from their ENMs may be more marginal (i.e. the difference between the mean environmental condition of the species and the mean of the study areas [111]) in comparison to other widely distributed species within a particular study region. Our study supports this interpretation given that narrowly distributed species had lower climatic variation represented by their occurrence records. This pattern is more evident in models that considered climatic variables only whereas models including edaphic variables showed wide variability and low correlation between environmental variation and species' range-size (S3 Fig). This could explain why some species with the largest geographic ranges showed increased accuracy with the use of the edaphic predictors, whereas for many of the restricted species similar accuracy was observed when using climatic variables only or climatic and edaphic variables together. This finding stresses the dissimilar nature of climatic and

edaphic variables and the different way in which those predictors interact with the geographic characteristics of species records.

Our results also showed that species with wide geographic distributions and large numbers of occurrences produce models with lower accuracy, a tendency that is consistent with previous findings [11,12,22,105], whereas species with a denser aggregation of records showed the opposite trend (S4 Fig). On this point we agree with [112], in that the relationship between model accuracy and the geographic range of a species is strongly affected by the extent used to construct the models, which is consequently related with the relative occurrence area and marginality of species. In addition, different extents and resolutions can influence the relative importance of predictors and the predicted suitability [113]. As the extent increases for an individual model, the environmental difference between predicted and unpredicted cells may increase simply by the broader environmental variability captured by larger extents. This may inflate the metrics designed to estimate model accuracy [114], especially for species distributed in marginal areas of the environmental space. Other causes for inflating the estimated model accuracy is the relationship between the number of occurrence records and accuracy. In our case, widely distributed species turned out to be more sampled (i.e. had more records) but, as mentioned above, species with more records also showed lower accuracies (S2 Fig). Again, one simple explanation for this is that a low discrimination of the environment between predicted and unpredicted cells is expected for species with large distributional area [11].

To the extent of our knowledge, this is the first attempt to evaluate the reliability of using global edaphic information to perform ENMs of plant species over large regions. Nevertheless, ENMs/SDMs experiments have some practical limitations because these approaches are sensible by several factors such as the extent of the area used to construct the models [115], the covariates selected and their grain [21,113], the geographical characteristics and quality of the species records [12,22], the pseudo-absences allocation methods [77], the algorithms used and their tuning [106] or even the species selected to perform the experiments [110], we acknowledge that there is no "silver bullet" approach that is capable of dealing with all those potential situations [104]. Therefore, we suggest that all comparative modeling studies such as ours need to be extrapolated with care. Of course, we are aware that the edaphic data we used may have some deficiencies related to the information on soil sampling and covariates used to generate these data [45] and that it is difficult to make generalizations to other regions of the world. However, the large geographic extent, the variability of environments and the different species geographic characteristics considered here allowed us to show that such global edaphic data adds useful information for plant distribution modeling. This is particularly valuable for studies of species that are distributed in regions where more detailed information on soil properties is poor or does not even exist. Importantly, we do not imply that these global edaphic data must be used in all future studies applying ENMs for plant species, but we do encourage modelers to test some of these edaphic variables and evaluate their model outputs against those conducted with climatic variables only. Recently the SoilGrids was improved by using more accurate technics and with finer-resolution data [116], thus we suggest that future studies consider the effect of different resolutions of soil data when applied to plant ENMs.

Supporting information

S1 Fig. Ordination diagram for the first two axis of three PCAs conducted with three variable set. C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.
(TIF)

S2 Fig. Relationship between geographical extent, number of records and density of records for the 125 target species.
(TIFF)

S3 Fig. Relationship between the standard deviation for the first principal component, of three predictors set, captured by the records of each species and their relationship with geographical extent, number of records and density of records. C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.
(TIFF)

S4 Fig. Effect of geographical extent, number of records and density of records on the TSS for GAM, MElq, RF and SVM conducted with three predictors sets. TSS: this index was transformed to arcsine; C: models with climate predictors; CE: models with climate and edaphic predictors; E: models with edaphic predictors.
(TIFF)

S5 Fig. Predicted suitability by the SVM method for *Cedrella odorata* and its relationship between the geographical and environmental space for three predictors sets. The right panel shows the first two principal components of the PCA conducted for each variable set. C: models with climate predictors, CE: models with climate and edaphic predictors, E: models with edaphic predictors.
(TIFF)

S1 Table. List of species modeled, families, habit and number of cleaned record (NR).
(PDF)

S2 Table. Principal components selected from the PCAs, their eigenvalues, variance explained and cumulative variance explained for each variable set.
(PDF)

S3 Table. Coefficients of the principal components selected from the PCAs performed for each dataset.
(PDF)

S4 Table. List of species modeled, number of records (NR), geographical extent (GR), density of records (DP) and standard deviation of the first principal component of climatic variables only (C.SD), climatic and edaphic variables (CE.SD) and edaphic variables only (E.SD).
(PDF)

S1 File. Principal components of climate variables used to perform the climate-only models. The data are in GeoTIFF format.
(RAR)

S2 File. Principal components of edaphic variables used to perform edaphic-only models. The data are in GeoTIFF format compressed.
(RAR)

S3 File. Principal components of climatic and edaphic variables used to perform climatic and edaphic models. The data are in GeoTIFF format compressed.
(RAR)

Acknowledgments

SJEV was supported by a doctoral research grant from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). PDMJ and FG thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for productivity grants. FV thanks A. Lira-Noriega, A.T. Peterson, O. Rojas-Soto and J. Soberón for several discussions on ENMs, and CONACyT for financial support.

Author Contributions

Conceptualization: Santiago José Elías Velazco, Paulo De Marco Júnior.

Formal analysis: Santiago José Elías Velazco.

Investigation: Santiago José Elías Velazco.

Methodology: Santiago José Elías Velazco.

Resources: Santiago José Elías Velazco.

Software: Santiago José Elías Velazco.

Supervision: Franklin Galvão, Fabricio Villalobos, Paulo De Marco Júnior.

Visualization: Santiago José Elías Velazco, Franklin Galvão, Fabricio Villalobos, Paulo De Marco Júnior.

Writing – original draft: Santiago José Elías Velazco, Fabricio Villalobos, Paulo De Marco Júnior.

Writing – review & editing: Santiago José Elías Velazco, Franklin Galvão, Fabricio Villalobos, Paulo De Marco Júnior.

References

1. Williams JN, Seo C, Thorne J, Nelson JK, Erwin S, O'Brien JM, et al. Using species distribution models to predict new occurrences for rare plants. Divers Distrib. 2009; 15: 565–576. https://doi.org/10.1111/j.1472-4642.2009.00567.x
2. Václavík T, Meentemeyer RK. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecol Model. 2009; 220: 3248–3258. https://doi.org/10.1016/j.ecolmodel.2009.08.013
3. Sousa-Silva R, Alves P, Honrado J, Lomba A. Improving the assessment and reporting on rare and endangered species through species distribution models. Glob Ecol Conserv. 2014; 2: 226–237. https://doi.org/10.1016/j.gecco.2014.09.011
4. Wan J, Wang C, Yu J, Nie S, Han S, Liu J, et al. Developing conservation strategies for Pinus koraiensis and Eleutherococcus senticosus by using model-based geographic distributions. J For Res. 2016; 27: 389–400. https://doi.org/10.1007/s11676-015-0170-5
5. Priti H, Aravind NA, Uma Shaanker R, Ravikanth G. Modeling impacts of future climate on the distribution of Myristicaceae species in the Western Ghats, India. Ecol Eng. 2016; 89: 14–23. https://doi.org/10.1016/j.ecoleng.2016.01.006
6. Still SM, Frances AL, Treher AC, Oliver L. Using Two Climate Change Vulnerability Assessment Methods to Prioritize and Manage Rare Plants: A Case Study. Nat Areas J. 2015; 35: 106–121. https://doi.org/10.3375/043.035.0115
7. Dubuis A, Pottier J, Rion V, Pellissier L, Theurillat J-P, Guisan A. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches: Predicting plant species richness. Divers Distrib. 2011; 17: 1122–1131. https://doi.org/10.1111/j.1472-4642.2011.00792.x
8. Svenning J-C, Fløjgaard C, Marske KA, Nógues-Bravo D, Normand S. Applications of species distribution modeling to paleobiology. Quat Sci Rev. 2011; 30: 2930–2947. https://doi.org/10.1016/j.quascirev.2011.06.012

9. Austin MP, Van Niel KP. Impact of landscape predictors on climate change modelling of species distributions: a case study with Eucalyptus fastigata in southern New South Wales, Australia: Impact of landscape predictors on climate change modelling. J Biogeogr. 2011; 38: 9–19. https://doi.org/10.1111/j.1365-2699.2010.02415.x
10. Franklin J. Mapping Species Distributions: spatial Inference and prediction. United States of America: Cambridge University Press; 2009.
11. McPherson J, Jetz W, Rogers DJ. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J Appl Ecol. 2004; 41: 811–823. https://doi.org/10.1111/j.0021-8901.2004.00943.x
12. Luoto M, Pöyry J, Heikkinen RK, Saarinen K. Uncertainty of bioclimate envelope models based on the geographical distribution of species: Uncertainty of bioclimate envelope models. Glob Ecol Biogeogr. 2005; 14: 575–584. https://doi.org/10.1111/j.1466-822X.2005.00186.x
13. Hortal J, Jiménez-Valverde A, Gómez JF, Lobo JM, Baselga A. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos. 2008; 117: 847–858. https://doi.org/10.1111/j.0030-1299.2008.16434.x
14. Jiménez-Valverde A, Lobo J, Hortal J. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecol. 2009; 10: 196–205. https://doi.org/10.1556/ComEc.10.2009.2.9
15. Newbold T. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Prog Phys Geogr. 2010; 34: 3–22. https://doi.org/10.1177/0309133309355630
16. Dupin M, Reynaud P, Jarošík V, Baker R, Brunel S, Eyre D, et al. Effects of the Training Dataset Characteristics on the Performance of Nine Species Distribution Models: Application to Diabrotica virgifera virgifera. Thrush S, editor. PLoS ONE. 2011; 6: e20957. https://doi.org/10.1371/journal.pone.0020957 PMID: 21701579
17. Beale CM, Lennon JJ. Incorporating uncertainty in predictive species distribution modelling. Philos Trans R Soc B Biol Sci. 2012; 367: 247–258. https://doi.org/10.1098/rstb.2011.0178 PMID: 22144387
18. Fernández M, Hamilton H, Kueppers LM. Characterizing uncertainty in species distribution models derived from interpolated weather station data. Ecosphere. 2013; 4: 1–17. https://doi.org/10.1890/ES13-00049.1
19. Harris RMB, Porfirio LL, Hugh S, Lee G, Bindoff NL, Mackey B, et al. To be or not to be? Variable selection can change the projected fate of a threatened species under future climate. Ecol Manag Restor. 2013; 1–5. https://doi.org/10.1111/emr.12055
20. Beck J, Böller M, Erhardt A, Schwanghart W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol Inform. 2014; 19: 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002
21. Beauregard F, Blois S. Beyond a climate-centric view of plant distribution: edaphic variables add value to distribution models. Hérault B, editor. PLoS ONE. 2014; 9: e92642. https://doi.org/10.1371/journal.pone.0092642 PMID: 24658097
22. Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R. A comparative evaluation of presence-only methods for modelling species distribution: A comparative evaluation of presence-only methods for modelling species distribution. Divers Distrib. 2007; 13: 397–405. https://doi.org/10.1111/j.1472-4642.2007.00346.x
23. Meyer C, Kreft H, Guralnick R, Jetz W. Global priorities for an effective information basis of biodiversity distributions. Nat Commun. 2015; 6: 8221. https://doi.org/10.1038/ncomms9221 PMID: 26348291
24. Dubuis A, Giovanettina S, Pellissier L, Pottier J, Vittoz P, Guisan A. Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. Rocchini D, editor. J Veg Sci. 2013; 24: 593–606. https://doi.org/10.1111/jvs.12002
25. Mod HK, Scherrer D, Luoto M, Guisan A. What we use is not what we know: environmental predictors in plant distribution models. Scheiner S, editor. J Veg Sci. 2016; 27: 1308–1322 https://doi.org/10.1111/jvs.12444
26. Austin M. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecol Model. 2007; 200: 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005
27. Austin MP. Case studies of the use of environmental gradients in vegetation and fauna modelling: theory and practice in Australia and New Zealand. In: Scott JM, Heglund PJ, Samson F, Haufler J, Morrison M, Raphael M, et al., editors. Predicting Species Occurrences: Issues of Accuracy and Scale. Covelo, California: Island Press; 2002. pp. 73–82.
28. Austin MP, Smith TM. A new model for the continuum concept. Vegetatio. 1989; 83: 35–47. https://doi.org/10.1007/BF00031679

29. Karger DN, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza RW, et al. Climatologies at high resolution for the earth's land surface areas. ArXiv Prepr ArXiv160700217. 2016; Available: https://arxiv.org/abs/1607.00217

30. Heuvelink GB., Webster R. Modelling soil variation: past, present, and future. Geoderma. 2001; 100: 269–301. https://doi.org/10.1016/S0016-7061(01)00025-8

31. Anderson DW. The effect of parent material and soil development on nutrient cycling in temperate ecosystems. Biogeochemistry. 1988; 5: 71–97.

32. Ceddia MB, Vieira SR, Villela ALO, Mota L dos S, Anjos LHC dos, Carvalho DF de. Topography and spatial variability of soil physical properties. Sci Agric. 2009; 66: 338–352.

33. Mwanjalolo Jackson-Gilbert M, Makooma Moses T, Rao KPC, Musana B, Bernard F, Leblanc B, et al. Soil Fertility in relation to Landscape Position and Land Use/Cover Types: A Case Study of the Lake Kivu Pilot Learning Site. Adv Agric. 2015; 2015: 1–8. https://doi.org/10.1155/2015/752936

34. Reef R, Feller IC, Lovelock CE. Nutrition of mangroves. Tree Physiol. 2010; 30: 1148–1160. https://doi.org/10.1093/treephys/tpq048 PMID: 20566581

35. Rachwal MFG, Curcio GR. Atributos pedológicos e ocorrência de caixeta no litoral paranaense, Brasil. Sci For. 2001; 59: 156–163.

36. Teillier S, Becerra P. Flora y vegetacion del salar de Ascotan, andes del norte de chile. Gayana Botánica. 2003; 60: 114–122.

37. Melečková Z, Dítě D, jun PE, Píš V, Galvánek D. Succession of Saline Vegetation in Slovakia after a Large-Scale Disturbance. Ann Bot Fenn. 2014; 51: 285–296. https://doi.org/10.5735/085.051.0504

38. Gröger A, Huber O. Rock outcrop habitats in the Venezuelan Guayana lowlands: their main vegetation types and floristic components. Braz J Bot. 2007; 30: 599–609.

39. Bárcenas-Argüello ML, Gutiérrez-Castorena M del C, Terrazas T. The Role of Soil Properties in Plant Endemism–A Revision of Conservation Strategies. In: Soriano MCMCH, editor. Soil Processes and Current Trends in Quality Assessment. Rijeka, Croatia: InTech; 2013.

40. Ruggiero PGC, Batalha MA, Pivello VR, Meirelles ST. Soil-vegetation relationships in cerrado (Brazilian savanna) and semideciduous forest, Southeastern Brazil. Plant Ecol. 2002; 160: 1–16.

41. Lloyd J, Domingues TF, Schrodt F, Ishida FY, Feldpausch TR, Saiz G, et al. Edaphic, structural and physiological contrasts across Amazon Basin forest–savanna ecotones suggest a role for potassium as a key modulator of tropical woody vegetation structure and function. Biogeosciences. 2015; 12: 6529–6571. https://doi.org/10.5194/bg-12-6529-2015

42. Thuiller W. On the importance of edaphic variables to predict plant species distributions—limits and prospects. J Veg Sci. 2013; 24: 591–592. https://doi.org/10.1111/jvs.12076 PMID: 26819539

43. Coudun C, Gégout J-C, Piedallu C, Rameau J-C. Soil nutritional factors improve models of plant species distribution: an illustration with Acer campestre (L.) in France. J Biogeogr. 2006; 33: 1750–1763. https://doi.org/10.1111/j.1365-2699.2005.01443.x

44. Bertrand R, Perez V, Gégout J-C. Disregarding the edaphic dimension in species distribution models leads to the omission of crucial spatial information under climate change: the case of Quercus pubescens in France. Glob Change Biol. 2012; 18: 2648–2660. https://doi.org/10.1111/j.1365-2486.2012.02679.x

45. Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, Ribeiro E, et al. SoilGrids1km—Global Soil Information Based on Automated Mapping. Bond-Lamberty B, editor. PLoS ONE. 2014; 9: e105992. https://doi.org/10.1371/journal.pone.0105992 PMID: 25171179

46. Sánchez-Fernández D, Lobo JM, Hernández-Manrique OL. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. Divers Distrib. 2010; 17: 163–171. https://doi.org/10.1111/j.1472-4642.2010.00716.x

47. Carretero MA, Sillero N. Evaluating how species niche modelling is affected by partial distributions with an empirical case. Acta Oecologica. 2016; 77: 207–216. https://doi.org/10.1016/j.actao.2016.08.014

48. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013; 36: 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

49. Cruz-Cárdenas G, López-Mata L, Villaseñor JL, Ortiz E. Potential species distribution modeling and the use of principal component analysis as predictor variables. Rev Mex Biodivers. 2014; 85: 189–199. https://doi.org/10.7550/rmb.36723

50. Boyle B, Hopkins N, Lu Z, Garay JAR, Mozzherin D, Rees T, et al. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics. 2013; 14: 16. https://doi.org/10.1186/1471-2105-14-16 PMID: 23324024

51. Anagiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. 2009; 161: 105–121. https://doi.org/10.1111/j.1095-8339.2009.00996.x

52. Goodwin ZA, Harris DJ, Filer D, Wood JR, Scotland RW. Widespread mistaken identity in tropical plant collections. Curr Biol. 2015; 25: R1066–R1067. Available: http://www.sciencedirect.com/science/article/pii/S0960982215012282 https://doi.org/10.1016/j.cub.2015.10.002 PMID: 26583892

53. Mccarthy KP, Fletcher RJ Jr, Rota CT, Hutto RL. Predicting Species Distributions from Samples Collected along Roadsides. Conserv Biol. 2012; 26: 68–77. https://doi.org/10.1111/j.1523-1739.2011.01754.x PMID: 22010858

54. Reddy S, Dávalos LM. Geographical sampling bias and its implications for conservation priorities in Africa. J Biogeogr. 2003; 30: 1719–1727.

55. Fourcade Y, Engler JO, Rödder D, Secondi J. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. PLoS ONE. 2014; 9: e97122. https://doi.org/10.1371/journal.pone.0097122 PMID: 24818607

56. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, et al. Ecological niches and geographic distributions. Princeton, N.J: Princeton University Press; 2011.

57. Rangel TF, Loyola RD. Labeling Ecological Niche Models. Nat Conserv. 2012; 10: 119–126. https://doi.org/10.4322/natcon.2012.030

58. Duan R-Y, Kong X-Q, Huang M-Y, Fan W-Y, Wang Z-G. The Predictive Performance and Stability of Six Species Distribution Models. Hernandez-Lemus E, editor. PLoS ONE. 2014; 9: e112764. https://doi.org/10.1371/journal.pone.0112764 PMID: 25383906

59. Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 2006; 29: 129–151. https://doi.org/10.1111/j.2006.0906–7590.04596.x

60. Hastie T, Tibshirani R. Generalized Aditive Models. Stat Sci. 1986; 1: 297–318.

61. Guisan A, Edwards TC, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model. 2002; 157: 89–100.

62. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006; 190: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

63. Phillips SJ, Dudík M, Schapire RE. A maximum entropy approach to species distribution modeling. Proceedings of the twenty-first international conference on Machine learning. ACM; 2004. p. 83. Available: http://dl.acm.org/citation.cfm?id=1015412

64. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists: Statistical explanation of MaxEnt. Divers Distrib. 2011; 17: 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

65. Graham CH, Elith J, Hijmans RJ, Guisan A, Peterson AT, Loiselle BA, et al. The influence of spatial errors in species occurrence data used in distribution models: Spatial error in occurrence data for predictive modelling. J Appl Ecol. 2007; 45: 239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

66. Anderson RP, Gonzalez I. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. Ecol Model. 2011; 222: 2796–2811. https://doi.org/10.1016/j.ecolmodel.2011.04.011

67. Souza RA, De Marco P. The use of species distribution models to predict the spatial distribution of deforestation in the western Brazilian Amazon. Ecol Model. 2014; 291: 250–259. https://doi.org/10.1016/j.ecolmodel.2014.07.007

68. Kamath C. Scientific data mining: a practical perspective. Philadelphia: Society for Industrial and Applied Mathematics; 2009.

69. Salcedo-Sanz S, Rojo-Álvarez JL, Martínez-Ramón M, Camps-Valls G. Support vector machines in engineering: an overview. Wiley Interdiscip Rev Data Min Knowl Discov. 2014; 4: 234–267. https://doi.org/10.1002/widm.1125

70. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer New York; 2013.

71. Hornik K, Meyer D, Karatzoglou A. Support vector machines in R. J Stat Softw. 2006; 15: 1–28.

72. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag New York; 2009.

73. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York, NY: Springer New York; 2013.

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

75. Lobo JM, Jiménez-Valverde A, Hortal J. The uncertain nature of absences and their importance in species distribution modelling. Ecography. 2010; 33: 103–114. https://doi.org/10.1111/j.1600-0587.2009.06039.x

76. Hanberry BB, He HS, Palik BJ. Pseudoabsence generation strategies for species distribution models. PLoS ONE. 2012; 7: e44486. https://doi.org/10.1371/journal.pone.0044486 PMID: 22952985

77. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. Selecting pseudo-absences for species distribution models: how, where and how many?: How to use pseudo-absences in niche modelling? Methods Ecol Evol. 2012; 3: 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

78. Soberón J, Peterson AT. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers Inform. 2005; 2: 1–10. https://doi.org/10.17161/bi.v2i0.4

79. Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol Model. 2011; 222: 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

80. Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, et al. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. McPherson J, editor. Methods Ecol Evol. 2014; 5: 1198–1205. https://doi.org/10.1111/2041-210X.12261

81. Liu C, White M, Newell G. Measuring and comparing the accuracy of species distribution models with presence-absence data. Ecography. 2011; 34: 232–243. https://doi.org/10.1111/j.1600-0587.2010.06354.x

82. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. J Appl Ecol. 2006; 43: 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

83. Jiménez-Valverde A, Lobo JM. Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica. 2007; 31: 361–369. https://doi.org/10.1016/j.actao.2007.02.001

84. McGarigal K, Cushman SA, Ene E. FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps [Internet]. Amherst: University of Massachusetts; 2012. Available: http://www.umass.edu/landeco/research/fragstats/fragstats.html

85. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. 2017. Available: https://www.R-project.org/

86. Hijmans RJ, Phillips S, Leathwick J, Elith J. dismo: Species Distribution Modeling. R package version 1.0–15 [Internet]. 2016. Available: https://CRAN.R-project.org/package=dismo

87. Ridgeway G. gbm: Generalized Boosted Regression Models. R package version 2.1.1 [Internet]. 2015. Available: https://CRAN.R-project.org/package=gbm

88. Karatzoglou A, Alex Smola, Hornik K, Zeileis A. kernlab—An S4 Package for Kernel Methods in R. J Stat Softw. 2004; 11: 1–20. Available: http://www.jstatsoft.org/v11/i09/

89. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2: 18–22.

90. Hijmans RJ. raster: Geographic Data Analysis and Modeling. R package version 2.5–2 [Internet]. 2015. Available: https://CRAN.R-project.org/package=raster

91. VanDerWal J, Falconi L, Januchowski S, Luke Shoo, Storlie C. SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. R package version 1.1–221. [Internet]. 2014. Available: https://CRAN.Rproject.org/package=SDMTools

92. Peter Filzmoser, Fritz H, Kalcher K. pcaPP: Robust PCA by Projection Pursuit. R package version 1.9–60 [Internet]. 2014. Available: https://CRAN.R-project.org/package=pcaPP

93. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Non linear Mixed Effects Models. R package version 3.1–125 [Internet]. 2016. Available: http://CRAN.R-project.org/package=nlme

94. Lenth RV. Least-Squares Means: The R Package lsmeans. J Stat Softw. 2016; 69. https://doi.org/10.18637/jss.v069.i01

95. Fox J, Weisberg S. An R Companion to Applied Regression. [Internet]. Second edition. Thousand Oaks, California, USA; 2011. Available: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

96. Jenny H. Factors of soil formation: a system of quantitative pedology. New York: Dover; 1994.

97. Schaetzl RJ, Anderson S. Soils genesis and geomorphology [Internet]. Cambridge; New York: Cambridge University Press; 2005. Available: http://dx.doi.org/10.1017/CBO9780511815560

98. Breemen N van, Buurman P. Soil formation [Internet]. Dordrecht; Boston: Kluwer Academic; 2002. Available: http://site.ebrary.com/id/10067342

99. Arnold RW. Pedology and pedogenesis. Encyclopedia of Soil Science Arnold. Dordrecht, Netherlands: Springer; 2008. p. 902.

100. Diekmann M, Michaelis J, Pannek A. Know your limits–The need for better data on species responses to soil variables. Basic Appl Ecol. 2015; 16: 563–572. https://doi.org/10.1016/j.baae.2015.08.010

101. Hutchinson GE. An Introduction to Population Ecology. New Haven, CT: Yale University Press; 1978.

102. Colwell RK, Rangel TF. Hutchinson's duality: the once and future niche. Proc Natl Acad Sci. 2009; 106: 19651–19658. Available: http://www.pnas.org/content/106/Supplement_2/19651.short https://doi.org/10.1073/pnas.0901650106 PMID: 19805163

103. Pliscoff P, Luebert F, Hilger HH, Guisan A. Effects of alternative sets of climatic predictors on species distribution models and associated estimates of extinction risk: A test with plants in an arid environment. Ecol Model. 2014; 288: 166–177. https://doi.org/10.1016/j.ecolmodel.2014.06.003

104. Qiao H, Soberón J, Peterson AT. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Kriticos D, editor. Methods Ecol Evol. 2015; 6: 1126–1136. https://doi.org/10.1111/2041-210X.12397

105. Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, van Loon EE, Raes N, Reemer M, et al. Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria–Dutch Hoverflies as a Case Study. Chapman MG, editor. PLoS ONE. 2013; 8: e63708. https://doi.org/10.1371/journal.pone.0063708 PMID: 23691089

106. Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. Maxent is not a presence-absence method: a comment on Thibaud et al. O'Hara RB, editor. Methods Ecol Evol. 2014; 5: 1192–1197. https://doi.org/10.1111/2041-210X.12252

107. Barbet-Massin M, Thuiller W, Jiguet F. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? Ecography. 2010; 33: 878–886. https://doi.org/10.1111/j.1600-0587.2010.06181.x

108. Lorena AC, Jacintho LFO, Siqueira MF, Giovanni RD, Lohmann LG, de Carvalho ACPLF, et al. Comparing machine learning classifiers in potential distribution modelling. Expert Syst Appl. 2011; 38: 5268–5275. https://doi.org/10.1016/j.eswa.2010.10.031

109. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014; 15: 3133–3181.

110. Syphard AD, Franklin J. Species traits affect the performance of species distribution models for plants in southern California. J Veg Sci. 2010; 21: 177–189. https://doi.org/10.1111/j.1654-1103.2009.01133.x

111. Hirzel AH, Hausser J, Chessel D, Perrin N. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology. 2002; 83: 2027–2036.

112. Jiménez-Valverde A, Lobo JM, Hortal J. Not as good as they seem: the importance of concepts in species distribution modelling. Divers Distrib. 2008; 14: 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x

113. Vale CG, Tarroso P, Brito JC. Predicting species distribution at range margins: testing the effects of study area extent, resolution and threshold selection in the Sahara-Sahel transition zone. Robertson M, editor. Divers Distrib. 2014; 20: 20–33. https://doi.org/10.1111/ddi.12115

114. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr. 2008; 17: 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

115. Anderson RP, Raza A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela: Effect of study region on models of distributions. J Biogeogr. 2010; 37: 1378–1393. https://doi.org/10.1111/j.1365-2699.2010.02290.x

116. Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, et al. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS ONE. 2015; 10: e0125814. https://doi.org/10.1371/journal.pone.0125814 PMID: 26110833