Naturalis Biodiversity Center

Naturalis Repository

Ecological niche models and point distribution data reveal a differential coverage of the cacao relatives (Malvaceae) in South American protected areas

Matheus Colli-Silva, José Rubens Pirani, Alexander Zizka

DOI:
https://doi.org/10.1016/j.ecoinf.2022.101668

Downloaded from
Naturalis Repository

Article 25fa Dutch Copyright Act (DCA) - End User Rights
This publication is distributed under the terms of Article 25fa of the Dutch Copyright Act (Auteurswet) with consent from the author. Dutch law entitles the maker of a short scientific work funded either wholly or partially by Dutch public funds to make that work publicly available following a reasonable period after the work was first published, provided that reference is made to the source of the first publication of the work.

This publication is distributed under the Naturalis Biodiversity Center 'Taverne implementation' programme. In this programme, research output of Naturalis researchers and collection managers that complies with the legal requirements of Article 25fa of the Dutch Copyright Act is distributed online and free of barriers in the Naturalis institutional repository. Research output is distributed six months after its first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and copyrights owner(s) of this work. Any use of the publication other than authorized under this license or copyright law is prohibited.

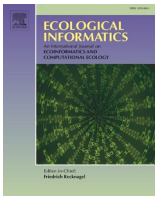
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the department of Collection Information know, stating your reasons. In case of a legitimate complaint, Collection Information will make the material inaccessible. Please contact us through email: collectie.informatie@naturalis.nl. We will contact you as soon as possible.

Ecological Informatics 69 (2022) 101668

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Ecological niche models and point distribution data reveal a differential coverage of the cacao relatives (Malvaceae) in South American protected areas

Matheus Colli-Silva [a,*], José Rubens Pirani [a], Alexander Zizka [b,c,d]

[a] Department of Botany, Institute of Biosciences, University of São Paulo, Rua do Matão 277, 05508-090 São Paulo, Brazil
[b] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstr. 4, D-04103 Leipzig, Germany
[c] Naturalis Biodiversity Center, P.O. Box 9517, 2300RA Leiden, Netherlands
[d] Department of Biology, Philipps-University Marburg, Karl-von-Frisch Str. 8, D-35043 Marburg, Germany

ARTICLE INFO

Keywords:
Biodiversity documentation
Biological conservation
Flowering plants
Malvaceae
Neotropics
Species distribution modeling

ABSTRACT

For many regions, such as in South America, it is unclear how well the existent protected areas network (PAs) covers different taxonomic groups and if there is a coverage bias of PAs towards certain biomes or species. Publicly available occurrence data along with ecological niche models might help to overcome this gap and to quantify the coverage of taxa by PAs ensuring an unbiased distribution of conservation effort. Here, we use a published occurrence database of 271 species from the cacao family (Malvaceae) to explore how South American PAs cover species with different distribution, abundance, and threat status. Furthermore, we compared the performance of online databases, expert knowledge, and modelled species distributions in estimating species coverage in PAs. We found 79 species from our survey (29% of the total) lack any record inside South American PAs and that 20 out of 23 species potentially threatened with extinction are not covered by PAs. The area covered by South American PAs was low across biomes, except for Amazonia, which had a relative high PA coverage, but little information on species distribution within PA available. Also, raw georeferenced occurrence data were underestimating the number of species in PAs, and projections from ecological niche models were more prone to overestimating the number of species represented within PAs. We discuss that the protection of South American flora in heterogeneous environments demand for specific strategies tailored to particular biomes, including collecting specimens inside PAs in less collected areas, and the delimitation of areas for protection in more known areas.

1. Introduction

Identifying spatial biodiversity patterns is crucial for conservation, given accelerating land use change and habitat destruction (Diniz-Filho et al., 2013; Gaston et al., 2008). For vascular plants, data collection of relevant areas is often challenging, especially in regions with high levels of species richness and endemism with collection biases, such as South America (Oliveira et al., 2017; Ulloa-Ulloa et al., 2017). It is against this background that large-scale databases of species distributions compiled from natural history collections, observation networks and, increasingly, citizen science initiatives have become primary sources for assessing biodiversity patterns and are important resources for conservation planning (Greve et al., 2016; Robertson et al., 2014; Williams and Crouch, 2017).

Despite recent endeavors to digitize herbarium collections and making them available (see Robertson et al., 2014), little occurrence information is accessible for large parts of South America, namely in Amazonia (Feeley, 2015). As for most areas of Earth, biodiversity documentation is heavily biased towards accessible regions, e.g., near cities, roads and waterways (Daru et al., 2019; Oliveira et al., 2017; Zizka et al., 2020a). Consequently, the South American protected areas network (henceforth "South American PAs") may only represent a biased subset of continental plant diversity, putatively missing taxa with a narrow distribution. This bias can be problematic for conservation planning, because species threatened with extinction may not be covered by PAs (Oliveira et al., 2017), and current efforts to evaluate species' extinction risk in the first place—i.e., Red Listing—primarily rely on occurrence records from PAs (Bridgewater, 2016; Le Saout et al.,

* Corresponding author.
E-mail address: matheus.colli.silva@alumni.usp.br (M. Colli-Silva).

https://doi.org/10.1016/j.ecoinf.2022.101668
Received 1 July 2021; Received in revised form 24 April 2022; Accepted 2 May 2022
Available online 7 May 2022
1574-9541/© 2022 Elsevier B.V. All rights reserved.

2013). Currently, PAs cover around 24% of South American land surface with large differences among countries, provinces, and biomes for numerous reasons (Baldi et al., 2019). In fact, despite significant efforts to use the protection of endangered species or other biological factors as justification on PA creation (Rylands and Brandon, 2005; Salvio and Gomes, 2018), political and economic reasons were often at least equally important for PA designations (Jim and Xu, 2004; Joppa et al., 2008; Rydén et al., 2020).

One attempt to overcome low collection efforts and biased sampling in key areas are ecological niche models (ENMs) based on occurrence records and environmental variables (Hopkins, 2007; Kramer-Schadt et al., 2013). ENMs can provide putative species distributions for areas with little data, and therefore can give an estimate of the biodiversity in each area and point to potential diversity hotspots in need of further field sampling. Many different ENM algorithms of varying complexity exist (Miller, 2010); nevertheless, for many South American species little physiological data and only few occurrence records are available. Thus, in practice, the use of ENMs for cross-taxonomic studies is limited to algorithms that can work with such limited data, for instance Maxent (Anderson and Gonzalez, 2011), regression or random forest-based algorithms (Evans et al., 2010) or the bioclim algorithm (Palkar et al., 2020).

Hence, addressing the proportion of biodiversity covered by PAs in South America under different mechanistic approaches seems to be a necessary task for conservation planning. Yet, typical challenges for large-scale approaches are a heterogeneous sampling effort thorough particular PAs on the continent (Colli-Silva et al., 2016; Daru et al., 2019), a poor documentation of the flora of many PAs, and geographic errors and misidentifications in existing botanical collections (Maldonado et al., 2015; Yesson et al., 2007; Zizka et al., 2020b). Here, we selected a published dataset from species of the cacao family (Malvaceae) as model group. Malvaceae is widespread throughout the tropics, with species richness centers in different regions, including different environments of Tropical Americas (Colli-Silva and Pirani, 2020). Specifically, in this study, we aimed to explore how ecologically different species (i.e., species from different biomes, and species potentially threatened and not threatened) are represented within PAs, and how species of these groups are covered by South American PAs considering different spatial approaches (i.e., ENMs vs. point occurrence data).

2. Methods

2.1. Database and spatial delimitations

In this study, we used a revised occurrence database of Colli-Silva and Pirani (2020). This database focuses on three of the nine subfamilies of Malvaceae sensu Alverson et al. (1998)—Byttnerioideae, Helicteroideae and Sterculioideae—which are species-rich in South America and include multiple taxa of economic importance such as cacao (*Theobroma cacao* L.), the West Indian elm (*Guazuma ulmifolia* Lam.) and the tropical chestnuts (*Sterculia* spp.). This database was built considering an extensive survey of the taxonomic literature and an expert review of the materials deposited in relevant herbarium collections for the group (including the largest American, European and South American herbaria, namely the MO, NY, RB, MBM, HUEFS, CEN, CEPEC, LIL, SPF, INPA and IPA collections (herbarium acronyms follows Thiers, 2022)). It also included an extensive review of the geographic distribution of all records, attributing or correcting geographic coordinates based on the location information described in the voucher labels found at herbaria.

The database comprises c. 15,000 occurrence records for 271 species in 11 genera for South America. Species occur in different habitats, encompassing either narrowly distributed or widespread species in genera prevalent in open, seasonally dry areas (*Ayenia* L., *Helicteres* L., *Melochia* L., *Rayleya* Cristóbal, *Waltheria* L.), forested formations (*Herrania* Goudot, *Theobroma* L., *Pterygota* Schott. & Endl., *Sterculia* L.), as well as genera with species occurring in multiple biomes (*Byttneria* Loefl. and *Guazuma* Adans.). For more details on the database, we refer to Colli-Silva and Pirani (2020).

For PA delimitations, we downloaded the World Database on Protected Areas (UNEP-WCMC, 2021), for a homogeneous and standard nomenclature of PAs across different countries in South America. We adopted the biome delimitations of Olson et al. (2001), except that we considered the Atlantic Forest and Amazonia as two biomes in their own right, because these two major blocks of South American "Tropical and Subtropical Moist Broadleaf Forests" sensu Olson et al. (2001) differ in their biogeographic history, accessibility, and conservation status for our study group.

2.2. Ecological niche modeling

All downstream analyses were run in the R Environment (R Core Team, 2021). Firstly, we modelled species ranges based on Colli-Silva and Pirani (2020) occurrence database (see section above) and climatic variables, using the "sdm" package v. 1.0–67 (Naimi and Araújo, 2016). To do so, we downloaded 10 min resolution bioclimatic variables from the WorldClim project v. 2, which are derived from monthly temperature and rainfall values often used in ecological modeling techniques (Fick and Hijmans, 2017; Karger et al., 2017).

We reduced geographically sampling bias of nested point occurrences by a spatial thinning of species occurrence records with the "spThin" package v. 0.2.0 (Aiello-Lammens et al., 2015) with a 25 km buffer. For the cases where we developed ENMs (i.e. species with 10 or more occurrence records), background points were extracted from the area delimited by a minimum convex hull around the occurrence points. We used the thinned occurrences to model species distributions using the following workflow: (1) for species with more than 15 records, we did a Principal Component Analysis (using the "rasterPCA" function of the RStoolbox v. 0.2.6 package; Leutner et al., 2019) of the bioclimatic variables and considered the first three principal components as predictors in our ENM; (2) for taxa with 10 to 14 records, we did a similar approach but considered only the first two principal components; (3) for species with 3 to 9 records, we approximated the range by a convex hull polygon around the occurrence points; finally, (4) for species with less than 3 records, we adopted a buffer of 25 km radius around the occurrence as species range. The rationale for the choice of the number of specimen records to develop ENMs or to approximate species' distribution range by convex hull polygons is that less than three records are prone to develop inaccurate ENMs (Proosdij et al., 2015).

We used an ensemble/consensus approach of obtaining an unweighted averaging mean of random forest, Maxent and bioclim models, with five replicates and cross-validations each. We then generated species presence/absence distributions using a threshold of equal sensitivity and specificity (see Liu et al., 2005). For model evaluation, we reported the mean Area Under the Curve (AUC) and the True Skill Statistic (TSS; check Allouche et al., 2006 for more info on these metrics).

2.3. Obtaining threat status

As only less than 20% of our study species were officially Red Listed by the IUCN, the International Union for Conservation of Nature (www.iucn.org), we estimated species' threat levels using a preliminary conservation assessment via the "IUCNN" package v. 0.9.3 (Zizka et al., 2020c) based on Colli-Silva and Pirani (2020) point occurrence database (see Section 2.1.), classifying species either as "threatened" or "non-threatened", according to "IUCNN" threat status suggestions.

IUCNN suggests preliminary threat categories by approximating species' Red List assessments according to the scheme of the IUCN (Zizka et al., 2020c). Briefly, IUCNN training models are based on three neural network algorithms that predict the conservation status of a species as a regression task. Each model incorporated by the package is conceptually different but have individual advantages and limitations. Detailed performances of every model, as well as their implementation in R

Environment are detailed by Zizka et al. (2020c). Approximations were based on geographic species occurrences, environmental data as prediction features, and existing IUCN Red List assessments as training data. While these assessments are only an approximation of IUCN full assessments, the accuracy of the approach in approximating assessments on the threatened vs. non-threatened level has been shown to be above 80% in other taxa (Zizka et al., 2020c). Hence, in the absence of a full assessment, "IUCNN" provides a rapid proxy to discuss potential conservation measures when only occurrence data is available.

As training dataset for the IUCNN models, we downloaded a second alternate database of preserved specimens obtained from GBIF, the Global Biodiversity Information Facility repository (GBIF.org, 2021). It should be noted that, prior to generating the training set, we performed standard data cleaning procedures of the GBIF database abovementioned, removing invalid and inaccurate coordinates (e.g., points in the sea, country or state centroids, incomplete coordinates), using the "CoordinateCleaner" v. 2.0–18 package (Zizka et al., 2019).

2.4. Assessing the coverage of species in PAs

To evaluate the coverage of species in PAs considering their occurrence in different biomes (see Section 2.1.) and their threat status (see Section 2.3.), we calculated the empirical representation index F as the mean fraction of records within PA across all species s in the dataset. We then compared F with a simulated null model F_r. To do so, we sampled occurrences for s species across South America, with the same distribution of occurrence numbers but with randomized locations and calculated F_r based on this simulated data. We replicated this calculation 1000 times ($r = 1000$) and obtained an F_r fraction for all species, according to the following equation:

$$F_r = \frac{1}{s} * \left(\sum_s \frac{n_s}{N_s} \right)$$

Where n_s is the number of records that the species s has inside South American PAs and N_s is the total number of records of the species s for a given replicate r. Data for all s species were obtained, and we used a two-side Kolmogorov-Smirnov Test, which also considers the variance of data, to test if the empirical distribution and the simulated replicates differed at a significance level of 0.05. We assumed as null hypothesis that empirical and simulated distributions would not be different.

Finally, to study the coverage in different biomes, we used the same approach as described above, but restricted the sampling of random points to the respective biome of interest. To assess which species were underrepresented in PAs, we separately examined potentially threatened species (see Section 2.3.) and species from different biomes (assigning species to a biome if more than 70% of its records were within that biome).

2.5. Contrasting species coverage of ENMs and point occurrence data

To contrast the performance of ENMs and point occurrence data, we used the Colli-Silva and Pirani (2020) database as gold standard for comparison with two other databases: (1) point occurrence data for the same species of Colli-Silva and Pirani (2020), but whose occurrence points were automatedly downloaded from GBIF repository (GBIF.org, 2021) with no further revision, except for standard data cleaning procedures (see Section 2.3.), and (2) predicted range from the distribution modeling binaries (see Section 2.2).

For ENM and point occurrence data, we counted the species richness and number of endemic species in each country. Furthermore, the presence or absence of every species in South American PAs according to ENM or point occurrence data results were summarized to create a confusion matrix. Then, we calculated omission (false negatives) and commission (false positives) errors as well as other accuracy metrics using the "caret" package v. 6.0–84 (Kuhn et al., 2019).

3. Results

3.1. Ecological niche models (ENMs)

We obtained AUC scores higher than 0.8 for most of the distribution models (Appendix S1 in Supplementary Material), indicating good model fit in general. Our ENMs suggested moderate to high species richness in sparsely sampled areas in Central and Northwestern South America (Fig. 1). Modelled richness also suggested western Amazonia, the Caatinga in Northeastern Brazil and the dry deciduous forest formations of central-western Brazil as centers of species richness in the study group.

ENMs suggested higher species numbers than found by raw occurrences for most countries, for instance in Peru and Uruguay (Table 1). However, for some countries, such as in Argentina, Paraguay and Guyana, species richness decreased compared to point occurrences. Brazil, Colombia, Paraguay and Bolivia were also the countries projected to harbor more species and endemic species (Table 1). Compared to raw occurrences, the distribution models suggested a lower country-level endemism, especially in Bolivia, Paraguay and Venezuela (Table 1). Model outputs are available as Supplementary Material (Appendix S2).

3.2. Species coverage in PAs

The coverage of occurrence records in South American PAs varied among species, biomes, and threat levels. In total, 79 of the 271 study species (29%) lacked known occurrences in any PAs; of these, 20 were indicated as potentially threatened by our preliminary assessment (i.e., as "Vulnerable", "Endangered" or "Critically Endangered"; see Table S2 for the full occurrence data per species). Our preliminary conservation assessment suggested 23 species as potentially threatened and only three of these, *Byttneria abutiloides* A.St.-Hil. and Naudin, *Byttneria obtusata* Benth. And *Helicteres urupensis* Leane had at least one record inside any PA. Interestingly, these three species are all restricted to forested environments: *Byttneria abutiloides* from the Brazilian Atlantic Forest, and *B. obtusata* and *H. urupensis* from different sites of Amazonia (see also Colli-Silva and Pirani, 2020).

In general, species were underrepresented in PAs compared to a random expectation (Fig. 2a) and threatened species were less covered in PAs than non-threatened species (Fig. 2b and c). Noteworthy exceptions from the Amazonian and the Atlantic Forest were *Pterygota, Sterculia* and *Theobroma* species, which are well represented in PAs from forested biomes. In contrast, examples of underrepresented genera in PAs were *Ayenia* and *Byttneria*, mostly from open seasonally dry formations, and *Herrania* from less collected areas of Amazonia, especially Western Brazil, Eastern Colombia and Northern Peru.

Within the individual South American biomes, Amazonian and Atlantic Forest species were best represented (Fig. 3). For the other biomes with less delimited PAs, i.e., Deserts and Xeric Woodlands and Tropical and Subtropical Grasslands and Savannas, species coverage was as underrepresented given the PAs network for these biomes (Fig. 3). Amazonia exhibited, on median, higher percentages of species inside protected areas (27.3%), whilst other biomes exhibited 7 to 11% of median records per species inside PAs, a smaller fraction due to less protected-areas delimited.

3.3. Accuracy of ENMs and point-occurrence approaches

The GBIF point occurrence database revealed high levels of sensitivity and specificity, while ENMs had a low sensitivity scores, reflecting errors of commission, with where there is an incorrect account of the species in PAs (Table 2). Conversely, lower levels of sensitivity reported in our results (Table 2) reflect that ENM predictions failed in predicting the occurrence of species inside PAs. In other words, despite the use of spatial filters to reduce over-prediction of species richness, ENM predictions were more prone to errors of commission (overestimating the

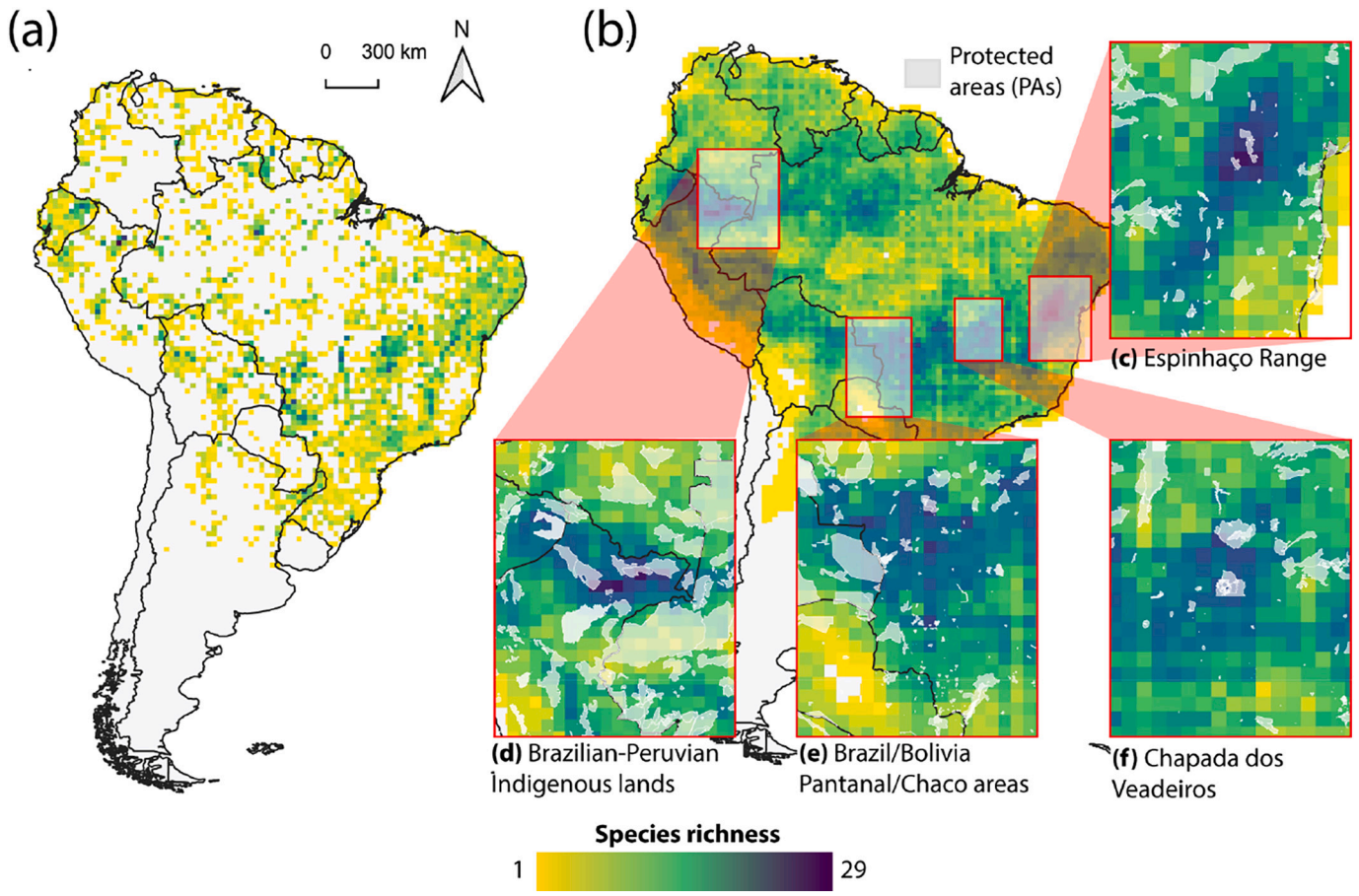


Fig. 1. Species richness maps based on (a) manual revisited point occurrences and (b) ENM predictions at 0.5° gridcells. From the map (b), we selected four major areas (c-f), with highest levels of predicted species richness, intersecting it with the South American PAs: (c) the Espinhaço Range, (d) Brazilian-Peruvian Amazonian indigenous lands, (e) Brazil/Bolivia Pantanal/Chaco areas and (f) Chapada dos Veadeiros. The shade areas represent PAs.

Table 1
Species richness and endemism levels of South American countries based on known (point occurrences) and projected (species distribution modeling, ENM) species occurrences. The first three columns show the total number of species for each country, and the latter three shows the number of endemic species to each country. For each case, besides the number of species considering known and projected occurrences, there are also the "change" column indicating the percent change from one approach to another. For "endemic species to the country", the percent from all species are in parenthesis.

Country	Total no. of species			Endemic species to the country		
	Total point occurrences	Total ENM	Change from ENM to point occurences	Point occurrences (% Total)	ENM (% Total)	Change from ENM to point occurences
Argentina	44	38	−14%	4 (9%)	5 (10%)	20%
Bolivia	69	73	6%	4 (6%)	2 (0%)	−100%
Brazil	179	195	9%	80 (45%)	59 (30%)	−26%
Chile	0	2	–	–	–	–
Colombia	67	74	10%	13 (19%)	6 (10%)	−54%
Ecuador	54	57	6%	6 (11%)	1 (0%)	−83%
French Guyana	18	19	6%	–	–	–
Guyana	34	32	−6%	1 (3%)	1 (0%)	0%
Peru	17	80	371%	2 (12%)	5 (10%)	150%
Paraguay	73	31	−58%	9 (12%)	0 (0%)	−100%
Suriname	16	22	38%	–	–	–
Uruguay	2	7	250%	–	–	–
Venezuela	48	51	6%	3 (6%)	1 (0%)	−67%

number of species represented within PAs). Nevertheless, there were also cases of errors of omission (underestimation of the presence of species), and projected species richness in some countries decreased when compared to the point occurrence database.

4. Discussion

Our results showed that different species are unequally covered by South American PAs, as almost a third of all species had no occurrence records in any PA. Secondly, the coverage in PAs was also different among threat levels and biomes, as we found most of the potentially

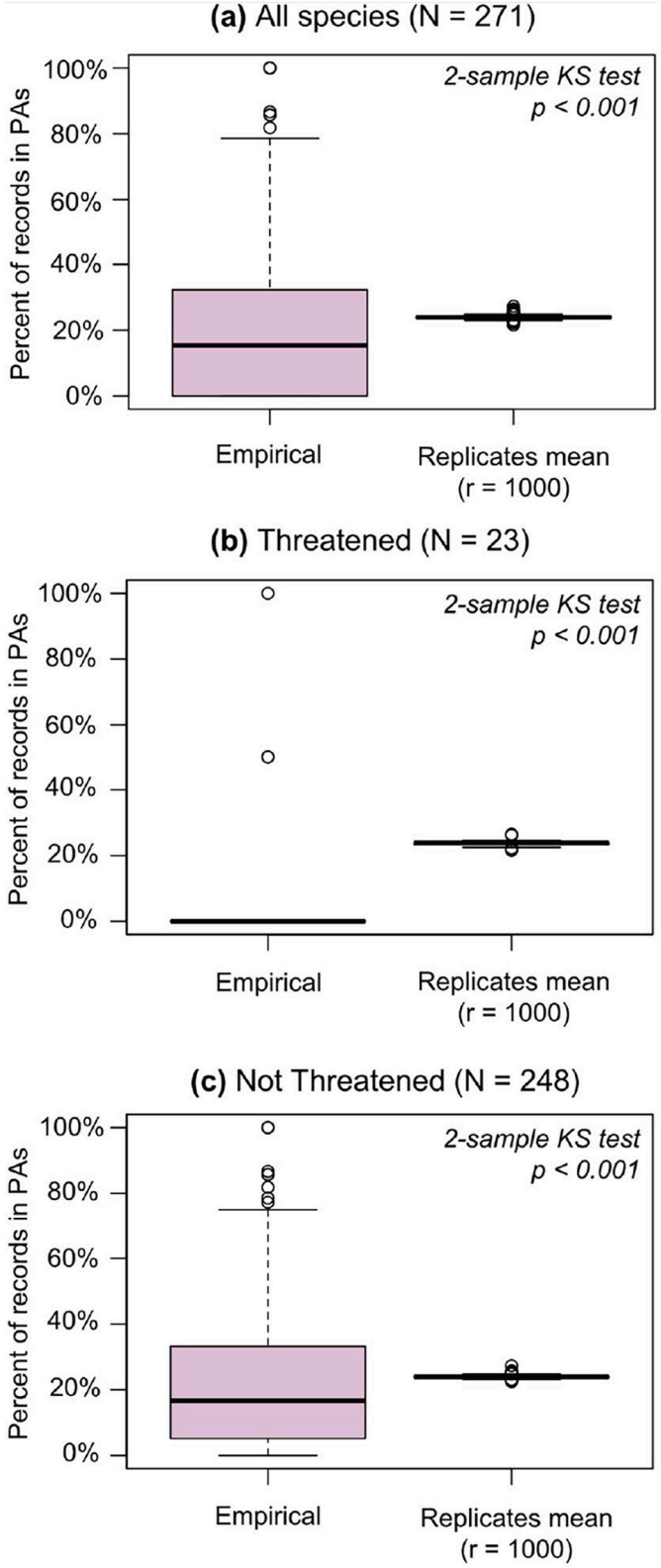


Fig. 2. Compared boxplots of the percent of records in PAs for species given empirical and simulated distributions. Distribution data consider all species, then only threatened, and only not threatened considering our preliminary conservation assessment. All data used to build these graphs are available as Supplementary Material (Table S3).

threatened species only outside PAs. Thirdly, we found that point occurrence records without any further data curation underestimated species coverage in areas where sampling effort is low (namely in Amazonia), whereas ENMs overestimated the occurrence of species in PAs in general.

Our results can be interpreted under a number of perspectives: 1) considering the impact of different approaches to obtain species occurrences inside PAs (either considering point occurrences or their projected area); 2) the heterogeneity of their presence or absence inside PAs (which species, from which biomes or threat status, are better represented in PAs); 3) in which levels these species are more, less or as expected to occur inside PAs when comparing to a simulated scenario under null conditions. In the next sections, we shall discuss each interpretation in detail.

4.1. *Impacts of ENMs and point-occurrence approaches*

The reported over-prediction of species (Table 2) present in PAs by ENMs is expected, because rarely all environmentally suitable areas are effectively occupied by a species (Aranda and Lobo, 2011; Mendes et al., 2020). The common approach to account for this limitation by using spatial buffers as filters proved to be insufficient in the case of our study, cautioning against the use of simple ENMs to predict the species richness of specific PAs. Nevertheless, we see the major use of correlative ENMs in detecting potential conservation gaps in unprotected or undocumented areas and promising locations for field sampling (e.g., Graham and Hijmans, 2006; Hopkins, 2005; Oliveira et al., 2019). In this sense, despite their limitations, ENMs are useful when little information is available.

It should be noted that the models we used in our study—random forest, Maxent and bioclim, which are the widest applied algorithms in ENM studies (Mi et al., 2017; Phillips et al., 2006)—were ensembled to create the outcomes, which resulted in high AUC and TSS scores (Table 1; Appendix S1). We attribute this characteristic to the fact that while logistic regression models are highly interpretable (but less accurate), because it is easy to extract individual coefficients, they end up being less accurate. Conversely, random-forest models are highly accurate but less interpretable. Both accuracy and interpretability are important characteristics in ENMs, and ensembling the information provided by both algorithms resulted in a good niche characterization of the species in our group. Moreover, our results reinforce what other studies have indicated on the utility to use ENMs to predict species distribution (Costa et al., 2009; García, 2006). Particularly in the study of Costa et al. (2009), the authors highlight the advantages of ENMs in projecting squamate distribution in areas far apart from known localities or inaccessible PAs of the Brazilian Cerrado. Namely, the "Parque Nacional da Chapada das Mesas", in Northeastern Brazil, known for being little disturbed but poorly collected for their study group. In their work, predictions are better reached using specific models, such as Maxent, and maximized the chance of tackling areas for further sampling. In the case of our study, both the park and the model were included, and we found similar results, which shows the use of ENMs as a remedy to avoid sampling bias in poorly collected areas.

Another remedy is the mobilization of existing data. Recent efforts in this direction include for instance the documentation of the flora of Brazilian PAs in and openly accessible online Floral Catalogue of Brazilian Protected Areas coordinated by the Rio de Janeiro Botanical Garden ("Catálogo de Plantas das Unidades de Conservação do Brasil", https://catalogo-ucs-brasil.jbrj.gov.br/), and the Distributed Information System for Biological Collections: Integrating Species Analyst and SinBiota ("speciesLink network") initiative to make available occurrence records from species of the Brazilian fauna and flora derived from museums (Canhos et al., 2015). In South America, a similar initiative exists for the flora of the Guianas as well (http://portal.cybertaxonomy.org/flora-guianas/node/1). Our results stress the importance of such initiatives together with targeted floristic and taxonomical surveys

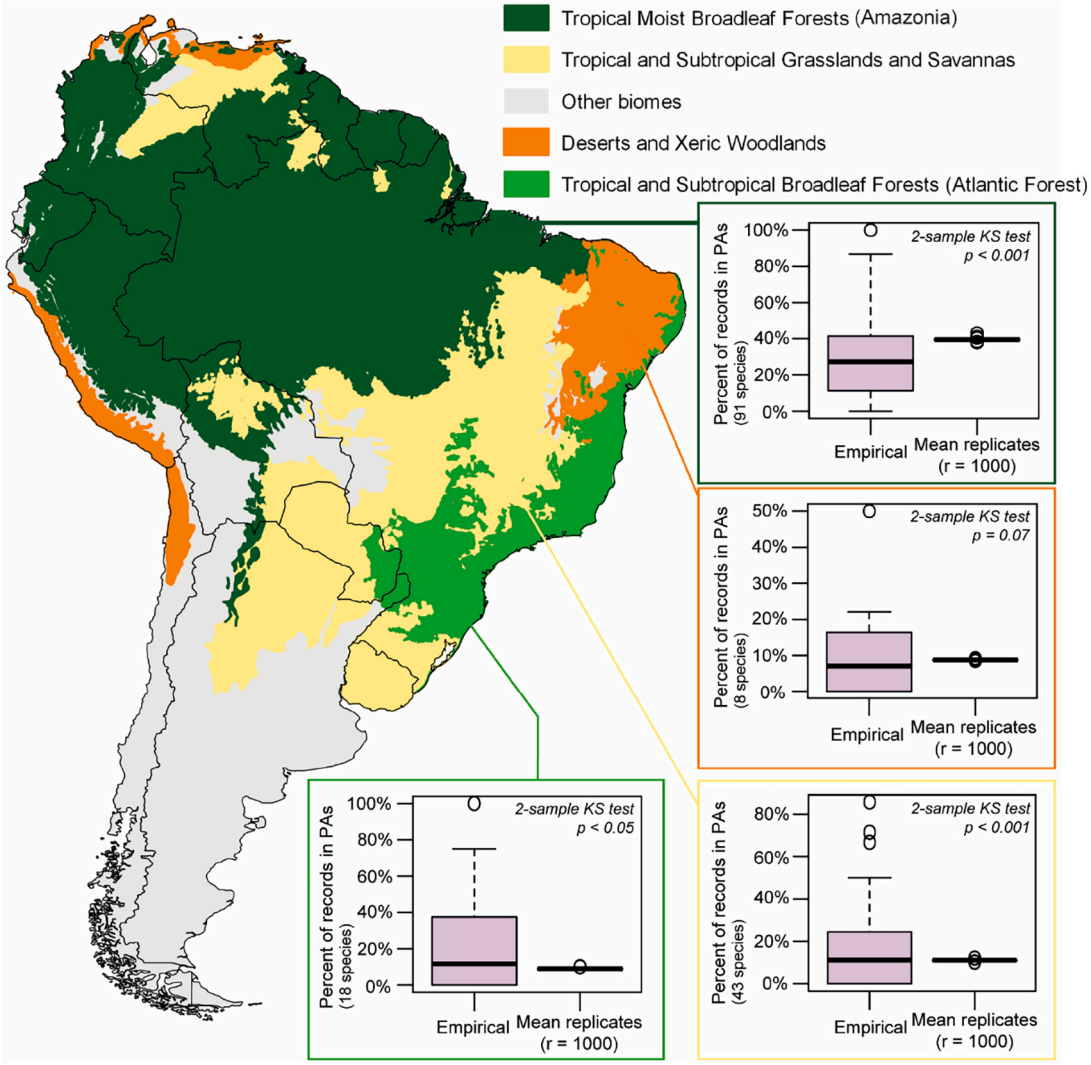


Fig. 3. Compared boxplots of the percent of records in PAs given empirical and simulated distributions, but per major biomes. This figure consider species with >70% of their records within the main biome delimitations sensu Olson et al. (2001), and random points were generated within each biome delimitation (see Section 2.4 in Methods). All data used to build these graphs are available as Supplementary Material (Table S3).

Table 2
Goodness of fit when comparing (1) GBIF automated cleaned occurrence database (GBIF.org, 2021) and modelled (ENM framework) with a manually verified occurrence database (Colli-Silva and Pirani, 2020) as gold standard. See Table S1 in Supplementary Material for a full dataset used in this analysis.

Metrics	GBIF database	ENM predictions
Sensitivity (reflects commission errors)	0.78	0.06
Specificity (reflects omission errors)	0.85	0.97
Positive Predicted Value	0.68	0.50
Negative Predicted Value	0.90	0.71
Prevalence	0.29	0.29
Detection Rate	0.23	0.02
Detection Prevalence	0.33	0.04

within particular PAs (Hopkins, 2005; Ribeiro et al., 2018; Zappi et al., 2019), to achieve global commitments such as the Global Strategy for Plant Conservation goals (CBD-Convention on Biological Diversity, 2016), which requires that 75% of a country's threatened species should be covered within its PAs network. In addition, targeted collections are indispensable for updating the taxonomical knowledge on individual genera and families and to update herbarium collections (Hopkins, 2007).

4.2. *Different coverage in South American PAs*

For our study group, species are less covered by PAs in open/dry biomes when compared to their occurrence in PAs from forested environments (Fig. 3). This can be a result of the reduced proportion of these biomes protected by PAs due to the historical neglect of open/dry biomes compared to forests in conservation decisions (Baldi et al., 2019; Pennington et al., 2006). Such neglect has various causes, as discussed in Pennington et al. (2006) including (1) the failure to consider different, naturally fragmented open/dry biomes as a distinct biome; (2) the disagreement on the nomenclature of each biome (e.g., adopting the nomenclature "Cerrado" vs. "savannas"; "Caatinga" vs. "dry deciduous forest" in areas of Brazil); or (3) the ongoing fragmentation of already naturally fragmented biomes due to human land use.

Conversely, larger PAs, such as those found in Amazonia, may be effective in protecting their biodiversity and are under less human pressure (Jones et al., 2018), although this is a matter of debate (see Fahrig, 2001). These large PAs seem to be essential for protecting our study species, particularly in Northern Brazil, the Guianas, Colombia, Peru, Ecuador, and Venezuela (Table 1). Our findings of a lower-than-expected coverage of species in Amazonia (Fig. 3) may be an artifact of the difficult accesses to these PAs and the resulting incomplete

knowledge of their flora and fauna as suggested in other studies (Feeley, 2015; Hopkins, 2007; Oliveira et al., 2016), including for our study group (Colli-Silva and Pirani, 2020).

In contrast to the PAs of Amazonia, most remnants of intact habitat in the coastal regions of South America or in the center remain unprotected (Baldi et al., 2019; Watson et al., 2014). Examples are areas of open/dry habitats encompassing the Brazilian Caatinga or the Chaco. Most of these areas are also failing to reach some of the Aichi Biodiversity Targets (Baldi et al., 2019)—a set of global targets established at the Convention on Biological Diversity (CBD), held in Nagoya, Japan, in 2010—indicating that improved strategical planning towards protecting species and centers of plant endemism in these areas is necessary.

Lastly, our results show a good representation of our study group in the PAs of the Atlantic Forest where plant species are both relatively well sampled and where the floras of PAs are relatively better known (Oliveira et al., 2017; Ribeiro et al., 2018; Rodrigues et al., 2004). This contrasts with findings from the open/dry biomes—where the coverage was as expected based on the area protected but with few species protected in absolute numbers, and Amazonia, where we found a lower fraction of species covered by large areas of PA. Increasing initiatives observed in conserving and restoring the Atlantic Forest over the past decade (Joly et al., 2014; Rezende et al., 2018; Tabarelli et al., 2010), and ongoing efforts in documenting biodiversity (e.g., Colli-Silva et al., 2020) may reflect the better performance of the PAs in the Atlantic Forest. Still, some authors have considered the existing efforts as insufficient to guarantee the conservation of Atlantic Forest biodiversity (Tabarelli et al., 2005).

In conclusion, despite the progress made in the past decades in understanding Neotropical diversity (see Antonelli et al., 2018) and defining PAs (including national parks, reserves, and indigenous territories), our results indicate that, by using species of the Malvaceae family as model organisms, a large fraction of South American plant diversity remains unprotected or poorly known (as revealed by the contrast of ENMs and point-occurrence approaches). This may be related to the limited knowledge diversity, for instance in Amazonia which is mostly based on collections from spatially restricted historical expeditions in Venezuela (Huber and Wurdack, 1984), Colombia (Baker et al., 1954), Ecuador (Renner, 1993) and Brazil (Hopkins, 2005; Prance, 1971), or the long-term neglect of certain habitats, as in the case of open/dry biomes. Nonetheless, the issue of limited and biased data availability, in particular from Amazonia, will likely persist, since the number of expeditions has dropped in the last two decades (ter Steege et al., 2016), and the focus of on-going collection on accessible areas (Oliveira et al., 2017, 2019), due to high cost of field expeditions in remote PAs.

Declaration of Competing Interest

None.

Acknowledgments

We thank the São Paulo Research Foundation (FAPESP) for funding (Grant IDs 2017/19295–1 and 2019/04530–0). We also thank the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil for funding the post-graduation in which the first author was enrolled and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig for hosting MCS for the time of this study. JRP is grateful to the National Council for Scientific and Technological Research (CNPq), Brazil for financial support. AZ is thankful for funding by iDiv via the German Research Foundation (DFG FZT 118), through sDiv, the Synthesis Centre of iDiv.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2022.101668.

References

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., Anderson, R.P., 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography 38, 541–545. https://doi.org/10.1111/ecog.01132.

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

Alverson, W.S., Karol, K.G., Baum, D.A., Chase, M.W., Swensen, S.M., Mccourt, R., Sytsma, K., 1998. Circumscription of the Malvales and relationships to other Rosidae: evidence from rbcL sequence data. Am. J. Bot. 85, 876–887. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

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

Antonelli, A., Ariza, M., Albert, J., Andermann, T., Azevedo, J., Bacon, C., Faurby, S., Guedes, T., Hoorn, C., Lohmann, L.G., Matos-Maraví, P., Ritter, C.D., Sanmartín, I., Silvestro, D., Tejedor, M., ter Steege, H., Tuomisto, H., Werneck, F.P., Zizka, A., Edwards, S.V., 2018. Conceptual and empirical advances in Neotropical biodiversity research. PeerJ 6, e5644. https://doi.org/10.7717/peerj.5644.

Aranda, S.C., Lobo, J.M., 2011. How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. Ecography 34, 31–38. https://doi.org/10.1111/j.1600-0587.2010.06134.

Baker, R.E.D., Cope, F.W., Holliday, P.C., Bartley, B.G., Taylor, D.J., 1954. The Anglo-Colombian Cacao Collecting Expedition. Report on Cacao Research, 1953. Imperial College of Tropical Agriculture, St Augustine, Trinidad, pp. 8–29.

Baldi, G., Schauman, S., Texeira, M., Marinaro, S., Martin, O.A., Gandini, P., Jobbágy, E. G., 2019. Nature representation in South American protected areas: country contrasts and conservation priorities. PeerJ 7, e7155. https://doi.org/10.7717/peerj.7155.

Bridgewater, P., 2016. The Anthropocene biosphere: do threatened species, Red Lists and protected areas have a future role in nature conservation? Biodivers. Conserv. 25, 603–607. https://doi.org/10.1007/s10531-016-1062-5.

Canhos, D.A.L., Sousa-Baena, M.S., de Souza, S., Maia, L.C., Stehmann, J.R., Canhos, V. P., De Giovanni, R., Bonacelli, M.B.M., Los, W., Peterson, A.T., 2015. The importance of biodiversity E-infrastructures for megadiverse countries. PLoS Biol. 13, e1002204 https://doi.org/10.1371/journal.pbio.1002204.

CBD-Convention on Biological Diversity, 2016. The Targets 2011–2020. Available at. https://www.cbd.int/gspc/targets.shtml.

Colli-Silva, M., Pirani, J.R., 2020. Estimating bioregions and undercollected areas in South America by revisiting Byttnerioideae, Helicteroideae and Sterculioideae (Malvaceae) occurrence data. Flora 271, 151688. https://doi.org/10.1016/j.flora.2020.151688.

Colli-Silva, M., Bezerra, T.L., Franco, G.A.D.C., Ivanauskas, N.M., Souza, F.M., 2016. Registros de espécies vasculares em unidades de conservação e implicações para a lista da flora ameaçada de extinção no estado de São Paulo. Rodriguésia 67, 405–425. https://doi.org/10.1590/2175-7860201667212.

Colli-Silva, M., Reginato, M., Cabral, A., Forzza, R.C., Pirani, J.R., Vasconcelos, T.N.C., 2020. Evaluating shortfalls and spatial accuracy of biodiversity documentation in the Atlantic Forest, the most diverse and threatened Brazilian phytogeographic domain. TAXON 69, 567–577. https://doi.org/10.1002/tax.12239.

Costa, G.C., Nogueira, C., Machado, R.B., Colli, G.R., 2009. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. Biodivers. Conserv. 19, 883–899. https://doi.org/10.1007/s10531-009-9746-8.

Daru, B.H., le Roux, P.C., Gopalraj, J., Park, D.S., Holt, B.G., Greve, M., 2019. Spatial overlaps between the global protected areas network and terrestrial hotspots of evolutionary diversity. Glob. Ecol. Biogeogr. 28, 757–766. https://doi.org/10.1111/geb.12888.

Diniz-Filho, J.A.F., Loyola, R.D., Raia, P., Mooers, A.O., Bini, L.M., 2013. Darwinian shortfalls in biodiversity conservation. Trends Ecol. Evol. 28, 689–695. https://doi.org/10.1016/j.tree.2013.09.003.

Evans, J.S., Murphy, M.A., Holden, Z.A., Cushman, S.A., 2010. Modeling Species Distribution and Change Using Random Forest. Predictive Species and Habitat Modeling in Landscape Ecology, pp. 139–159. https://doi.org/10.1007/978-1-4419-7390-0_8.

Fahrig, L., 2001. How much habitat is enough? Biol. Conserv. 100, 65–74. https://doi.org/10.1016/s0006-3207(00)00208-1.

Feeley, K., 2015. Are we filling the data void? An assessment of the amount and extent of plant collection records and census data available for tropical South America. PLoS One 10, e0125629. https://doi.org/10.1371/journal.pone.0125629.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

García, A., 2006. Using ecological niche modelling to identify diversity hotspots for the herpetofauna of Pacific lowlands and adjacent interior valleys of Mexico. Biol. Conserv. 130, 25–46. https://doi.org/10.1016/j.biocon.2005.11.030.

Gaston, K.J., Jackson, S.F., Cantú-Salazar, L., Cruz-Piñón, G., 2008. The ecological performance of protected areas. Ann. Rev. Ecol. Evol. Syast. 39, 93–113. https://doi.org/10.1146/annurev.ecolsys.39.110707.173529.

GBIF.org, 13 February 2021. GBIF Occurrence Download. https://doi.org/10.15468/dl.4yvvx3.

Graham, C.H., Hijmans, R.J., 2006. A comparison of methods for mapping species ranges and species richness. Glob. Ecol. Biogeogr. 15, 578–587. https://doi.org/10.1111/j.1466-8238.2006.00257.x.

Greve, M., Lykke, A.M., Fagg, C.W., Gereau, R.E., Lewis, G.P., Marchant, R., Marshall, A. R., Ndayishimiye, J., Bogaert, J., Svenning, J.-C., 2016. Realising the potential of herbarium records for conservation biology. S. Afr. J. Bot. 105, 317–323. https://doi.org/10.1016/j.sajb.2016.03.017.

Hopkins, M.J.G., 2005. Flora da Reserva Ducke, Amazonas, Brasil. Rodriguésia 56, 9–25. https://doi.org/10.1590/2175-78602005568602.

Hopkins, M.J.G., 2007. Modelling the known and unknown plant biodiversity of the Amazon Basin. J. Biogeogr. 34, 1400–1411. https://doi.org/10.1111/j.1365-2699.2007.01737.x.

Huber, O., Wurdack, J.J., 1984. A history of botanical exploration in Territorio Federal Amazonas, Venezuela. Smithson. Contrib. Bot. 56, 1–83. https://doi.org/10.5479/si.0081024X.56.

Jim, C.Y., Xu, S.S.W., 2004. Recent protected-area designation in China: an evaluation of administrative and statutory procedures. Geogr. J. 170, 39–50. https://doi.org/10.1111/j.0016-7398.2004.05038.x.

Joly, C.A., Metzger, J.P., Tabarelli, M., 2014. Experiences from the Brazilian Atlantic Forest: ecological findings and conservation initiatives. New Phytol. 204, 459–473. https://doi.org/10.1111/nph.12989.

Jones, K.R., Venter, O., Fuller, R.A., Allan, J.R., Maxwell, S.L., Negret, P.J., Watson, J.E. M., 2018. One-third of global protected land is under intense human pressure. Science 360, 788–791. https://doi.org/10.1126/science.aap9565.

Joppa, L.N., Loarie, S.R., Pimm, S.L., 2008. On the protection of "protected areas.". Proc. Natl. Acad. Sci. 105, 6673–6678. https://doi.org/10.1073/pnas.0802471105.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Sci. Data 4. https://doi.org/10.1038/sdata.2017.122.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19, 1366–1379. https://doi.org/10.1111/ddi.12096.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Strucca, L., Tang, Y., Candan, C., Hunt, T., 2019. "caret": Classification and Regression Training. R Package Version 6.0–88. https://CRAN.R-project.org/package=caret.

Le Saout, S., Hoffmann, M., Shi, Y., Hughes, A., Bernard, C., Brooks, T.M., Rodrigues, A. S., 2013. Protected areas and effective biodiversity conservation. Science 342 (6160), 803–805. https://doi.org/10.1126/science.1239268.

Leutner, N., Horning, N., Schwalb-Willmann, J., 2019. "RStoolbox": Tools for Remote Sensing Data Analysis. R Package Version 0.2.6. https://CRAN.R-project.org/package=RStoolbox.

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

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., Chilquillo, E., Rønsted, N., Antonelli, A., 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob. Ecol. Biogeogr. 24, 973–984. https://doi.org/10.1111/geb.12326.

Mendes, P., Velazco, S.J.E., de Andrade, A.F.A., De Marco, P., 2020. Dealing with overprediction in species distribution models: how adding distance constraints can improve model accuracy. Ecol. Model. 431, 109180 https://doi.org/10.1016/j.ecolmodel.2020.109180.

Mi, C., Huettmann, F., Guo, Y., et al., 2017. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ 5, e2849. https://doi.org/10.7717/peerj.2849.

Miller, J., 2010. Species distribution modeling. Geogr. Compass 4, 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x.

Naimi, B., Araújo, M.B., 2016. Sdm: a reproducible and extensible R platform for species distribution modelling. Ecography 39, 368–375. https://doi.org/10.1111/ecog.01881.

Oliveira, U., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J.B., Silva, D.P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., Ascher, J.S., de Vasconcelos, M.F., De Marco, P., Löwenberg-Neto, P., Dias, P.G., Ferro, V.G., Santos, A.J., 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Divers. Distrib. 22, 1232–1244. https://doi.org/10.1111/ddi.12489.

Oliveira, U., Soares-Filho, B.S., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J.B., Silva, D. P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., Ascher, J.S., de Vasconcelos, M.F., De Marco, P., Löwenberg-Neto, P., Ferro, V.G., Santos, A.J., 2017. Biodiversity conservation gaps in the Brazilian protected areas. Sci. Rep. 7 https://doi.org/10.1038/s41598-017-08707-2.

Oliveira, U., Soares-Filho, B.S., Santos, A.J., Paglia, A.P., Brescovit, A.D., de Carvalho, C. J.B., Silva, D.P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., Ascher, J.S., Vasconcelos, M.F., Marco, P.D., Löwenberg-Neto, P., Ferro, V.G., 2019. Modelling highly biodiverse areas in Brazil. Sci. Rep. 9 https://doi.org/10.1038/s41598-019-42881-9.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial ecoregions of the world: a new map of life on earth. BioScience 51, 933. https://doi.org/10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2.

Palkar, R.S., Janarthanam, M.K., Sellappan, K., 2020. Prediction of potential distribution and climatic factors influencing *Garcinia indica* in the Western Ghats of India using ecological niche modeling. Nat. Acad. Sci. Lett. 43, 585–591. https://doi.org/10.1007/s40009-020-00918-y.

Pennington, R.T., Lewis, G.P., Ratter, J.A., 2006. An overview of the plant diversity, biogeography and conservation of Neotropical savannas and seasonally dry forests. In: Pennington, R.T., R.T., Lewis, G.P., Ratter, J.A. (Eds.), Neotropical Savannas and Seasonally Dry Forests: Plant Diversity, Biogeography and Conservation. CRC Press, Boca Raton, pp. 1–31.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Prance, G.T., 1971. An index of plant collectors in Brazilian Amazonia. Acta Amazon. 1, 25–65. https://doi.org/10.1590/1809-43921971011025.

Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., Raes, N., 2015. Minimum required number of specimen records to develop accurate species distribution models. Ecography 39, 542–552. https://doi.org/10.1111/ecog.01509.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Renner, S.S., 1993. A history of botanical exploration in Amazonian Ecuador, 1739-1988. Smithson. Contrib. Bot. 82, 1–35.

Rezende, C.L., Scarano, F.R., Assad, E.D., Joly, C.A., Metzger, J.P., Strassburg, B.B.N., Tabarelli, M., Fonseca, G.A., Mittermeier, R.A., 2018. From hotspot to hopespot: an opportunity for the Brazilian Atlantic Forest. Perspect. Ecol. Conserv. 16, 208–214. https://doi.org/10.1016/j.pecon.2018.10.002.

Ribeiro, B.R., Martins, E., Martinelli, G., Loyola, R., 2018. The effectiveness of protected areas and indigenous lands in representing threatened plant species in Brazil. Rodriguésia 69, 1539–1546. https://doi.org/10.1590/2175-7860201869404.

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L., Desmet, P., 2014. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PLoS One 9, e102623. https://doi.org/10.1371/journal.pone.0102623.

Rodrigues, A.S.L., Andelman, S.J., Bakarr, M.I., Boitani, L., Brooks, T.M., Cowling, R.M., Fishpool, L.D.C., da Fonseca, G.A.B., Gaston, K.J., Hoffmann, M., Long, J.S., Marquet, P.A., Pilgrim, J.D., Pressey, R.L., Schipper, J., Sechrest, W., Stuart, S.N., Underhill, L.G., Waller, R.W., Watts, M.E.J., Yan, X., 2004. Effectiveness of the global protected area network in representing species diversity. Nature 428, 640–643. https://doi.org/10.1038/nature02422.

Rydén, O., Zizka, A., Jagers, S.C., Lindberg, S.I., Antonelli, A., 2020. Linking democracy and biodiversity conservation: empirical evidence and research gaps. Ambio 49, 419–433. https://doi.org/10.1007/s13280-019-01210-0.

Rylands, A.B., Brandon, K., 2005. Brazilian protected areas. Conserv. Biol. 19, 612–618. https://doi.org/10.1111/j.1523-1739.2005.00711.x.

Salvio, G.M.M., Gomes, C.R., 2018. Protected area systems in South American countries. Floresta e Ambiente 25. https://doi.org/10.1590/2179-8087.113417.

Tabarelli, M., Pinto, L.P., Silva, J.M.C., Hirota, M., Bede, L., 2005. Challenges and opportunities for biodiversity conservation in the Brazilian Atlantic Forest. Conserv. Biol. 19, 695–700. https://doi.org/10.1111/j.1523-1739.2005.00694.x.

Tabarelli, M., Aguiar, A.V., Ribeiro, M.C., Metzger, J.P., Peres, C.A., 2010. Prospects for biodiversity conservation in the Atlantic Forest: lessons from aging human-modified landscapes. Biol. Conserv. 143, 2328–2340. https://doi.org/10.1016/j.biocon.2010.02.005.

ter Steege, H., Vaessen, R.W., Cárdenas-López, D., Sabatier, D., Antonelli, A., de Oliveira, S.M., Pitman, N.C.A., Jørgensen, P.M., Salomão, R.P., 2016. The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. Sci. Rep. 6 https://doi.org/10.1038/srep29549.

Thiers, B., 2022. Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/ (accessed 12 January 2022).

Ulloa-Ulloa, C., Acevedo-Rodríguez, P., Beck, S., Belgrano, M.J., Bernal, R., Berry, P.E., Brako, L., Celis, M., Davidse, G., Forzza, R.C., Gradstein, S.R., Hokche, O., León, B., León-Yánez, S., Magill, R.E., Neill, D.A., Nee, M., Raven, P.H., Stimmel, H., Strong, M.T., 2017. An integrated assessment of the vascular plant species of the Americas. Science 358, 1614–1617. https://doi.org/10.1126/science.aao0398.

UNEP-WCMC, 2021. Protected Area Profile for Latin America and Caribbean from the World Database of Protected Areas. March 2021. Available at. www.protectedplanet.net (accessed on 12 March 2021).

Watson, J.E.M., Dudley, N., Segan, D.B., Hockings, M., 2014. The performance and potential of protected areas. Nature 515, 67–73. https://doi.org/10.1038/nature13947.

Williams, V.L., Crouch, N.R., 2017. Locating sufficient plant distribution data for accurate estimation of geographic range: the relative value of herbaria and other sources. S. Afr. J. Bot. 109, 116–127. https://doi.org/10.1016/j.sajb.2016.12.015.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A., Culham, A., 2007. How global is the global biodiversity information facility? PLoS One 2, e1124. https://doi.org/10.1371/journal.pone.0001124.

Zappi, D.C., Moro, M.F., Walker, B., Meagher, T., Viana, P.L., Mota, N.F.O., Watanabe, M. T.C., Nic Lughadha, E., 2019. Plotting a future for Amazonian *canga vegetation* in a *campo rupestre* context. PLoS One 14, e0219753. https://doi.org/10.1371/journal.pone.0219753.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N.,

Zizka, V., Antonelli, A., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744–751. https://doi.org/10.1111/2041-210x.13152.

Zizka, A., Antonelli, A., Silvestro, D., 2020a. Sampbias, a method for quantifying geographic sampling biases in species distribution data. Ecography 44, 25–32. https://doi.org/10.1111/ecog.05102.

Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N.M.C., Santos, T.A.B., dos Santos-Costa, R.C., Serrano, F.C., Alves da Silva, A.P., de Souza Soares, A., Cavalcante de Souza, P.G., Calisto Tomaz, E., Vale, V.F., Vieira, T.L., Antonelli, A., 2020b. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916. https://doi.org/10.7717/peerj.9916.

Zizka, A., Silvestro, D., Vitt, P., Knight, T.M., 2020c. Automated conservation assessment of the orchid family with deep learning. Conserv. Biol. https://doi.org/10.1111/cobi.13616.