Forest Ecology and Management 504 (2022) 119838

Contents lists available at ScienceDirect

Forest Ecology and Management

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

Impact of spatial configuration of training data on the performance of Amazonian tree species distribution models

Pablo Pérez Chaves [a,*], Kalle Ruokolainen [a], Jasper Van doninck [a,b], Hanna Tuomisto [a]

[a] University of Turku, Department of Biology, 20014 Turku, Finland
[b] Michigan State University, Department of Integrative Biology, 48824 Michigan, United States of America

ARTICLE INFO

Keywords:
Amazonia
Species distribution models
Remote sensing
Landsat
Trees
Tropical forest management
Peru

ABSTRACT

Remote sensing can provide useful explanatory variables for tree species distribution modeling, but only a few studies have explored this potential in Amazonia at local scales. Particularly for tropical forest management it would be useful to be able to predict the potential distribution of important tree taxa in areas where field data is as yet missing. Forest concessions produce valuable census data that cover large areas with high sampling effort and can be used as occurrence data in species distribution models (SDM). Nevertheless, these tree records are often spatially clumped and possibly only provide accurate predictions over areas close to where the training occurrence records are located. Here, we aim at investigating to what degree SDM performance and spatial predictions differ between models that have different spatial configurations of the occurrence data. For this, we divided the available occurrence data from a forest concession census in Peruvian Amazonia into different spatial configurations (narrow, elongated and compact), each of which contained approximately 20% of the full dataset. We then modelled the distributions of five tree taxa using Landsat data and elevation. More elongated configurations of the training data were more representative of the available environmental space, and also produced more robust SDMs. Average model performance (expressed as AUC) was 5% higher and variation in model performance 50% lower when elongated rather than compact configurations of training area were used. This confirms that covering only a small fraction of the environmental variability in the area of interest may lead to misleading SDM predictions, which needs to be taken into account when forest management decisions are based on SDMs.

1. Introduction

Remote sensing data provide spatially and temporally continuous information for species distribution models (Bradley et al. 2012; Cord et al. 2013; Rocchini 2013; He et al. 2015; Chaves et al. 2018), which are, in turn, useful for biodiversity assessments (van Ewijk et al. 2014; Turner 2014) and for contributing to the understanding of variation in species composition (He et al. 2015; Van doninck & Tuomisto 2018; Chaves, 2021; Chaves et al., 2020; Tuomisto et al., 2019). Knowing where a species occurs and where not is a basic ingredient in successful planning of its management and conservation. In many cases, obtaining direct observations of the presence and absence of the target species in the entire area of interest is not possible, but a good species distribution model (SDM) can serve as a surrogate (Guisan & Zimmermann 2000; Guisan & Thuiller 2005; Pearson 2010; Franklin & Miller 2010; Soberón 2010; Peterson et al. 2011; Franklin 2013; Guisan et al. 2017; Araújo et al. 2019). Species distribution models have been used mainly at regional to global scales at coarse spatial resolution, typically using as predictors climatic variables (Guisan & Zimmermann 2000; Guisan & Thuiller 2005; Franklin & Miller 2010; Peterson et al. 2011; Guisan et al. 2017; Gomes et al. 2018; Raghavan et al. 2019) or low-resolution remote sensing data (Prates-Clark et al. 2008; Buermann et al. 2008; Van doninck et al. 2020). However, in many cases where SDMs can be of practical use, the operational decisions are made at a much finer spatial resolution (hundreds of meters or at most a few kilometers). For example, planning of sustainable use of timber resources at the face of climate change requires understanding of both temporally slowly changing climatic drivers and more local practically permanent edaphic determinants of tree species distribution (Rehfeldt et al. 2015). Deriving reliable predictions of timber species distributions at local scales in Amazonia could facilitate forest management and planning activities, such as estimating the forest potential of areas that have not been censed

* Corresponding author.
E-mail address: papech@utu.fi (P.P. Chaves).

https://doi.org/10.1016/j.foreco.2021.119838
Received 21 June 2021; Received in revised form 22 October 2021; Accepted 31 October 2021
Available online 8 November 2021

yet.

Landsat data (with spatial resolution of 30 m) are among the potentially most interesting sources of information here because of appropriate spatial and spectral resolution, good temporal coverage and free access. Landsat has potential especially in the tropics, where accurate maps of any kind are scarce. Remote sensing data such as Landsat data and topographic variables have been recently used to model the distribution of tree taxa at local scales in Amazonia (Figueiredo et al. 2015; Figueiredo et al. 2016; Chaves et al. 2018). Furthermore, in Amazonian rain forests, it has been particularly observed that differences in canopy reflectance, as measured by Landsat, can predict spatial patterns both in soil properties and in the species composition of plants (Tuomisto, Poulsen, et al. 2003; Tuomisto, Ruokolainen, et al. 2003; Salovaara et al. 2005; Higgins et al. 2011; Higgins et al. 2012; Sirén et al. 2013; Muro et al. 2016; Van doninck & Tuomisto 2018; Tuomisto et al. 2019; Chaves et al. 2020) and are also useful predictors for modeling the distribution of understory plants (Van doninck et al. 2020). Elevation, another variable that can be measured remotely, has been related to geological substrates and soil nutrients in Amazonia (Vormisto et al. 2004; Costa et al. 2005; Higgins et al. 2011) and topography has also been found to covary with tree species distributions (Fortunel et al. 2018; Zuleta et al. 2018). Therefore, average reflectance values derived from Landsat data and elevation can be considered as ecologically informative remote sensing layers (Bradley et al. 2012; He et al. 2015; Leitão & Santos 2019) that are potentially useful for predicting the distribution of trees at local scales.

The building or training of a distribution model necessarily requires occurrence records whose geographical accuracy corresponds to the spatial resolution of the environmental variables and the aimed resolution of model predictions (Figueiredo et al. 2016; Connor et al. 2018). Forest censuses can have precise geographical coordinates and, therefore, they seem promising as a source of training data for SDMs at fine resolutions. In some tropical countries, such as Peru, legislation requires that forest concessions produce census data of timber trees, and these often cover larger areas with higher sampling effort than national forest inventory schemes do. In spite of higher sampling effort, census data from forest concessions might be highly concentrated to just one part of the area of interest, hence possibly limiting the reliability of model interpretations to areas close to the locations of the occurrence records. Earlier studies have documented that both the data partitioning scheme and the number of occurrences used for training SDMs can have an effect on SDM predictions (Stockwell & Peterson 2002; Elith et al. 2006; Hernandez et al. 2006; Wisz et al. 2008; Veloz 2009; Mateo et al. 2010; Gonzalez et al. 2011; Hijmans 2012; Radosavljevic Aleksandar et al. 2013; Boria et al. 2014; Fourcade et al. 2014; van Proosdij et al. 2016; Boria & Blois 2018). In particular, this effect becomes visible if different spatial configurations of the training data cover different parts of the relevant environmental space. Here, we aim to quantify–with real forest concession data–differences in model performance and spatial predictions when the training data is divided into different spatial configurations and shapes.

We recently modelled the distribution of five tree taxa in Southern Peruvian Amazonia using Landsat satellite imagery and elevation at local scales (Chaves et al., 2018). Since then, the number of available species occurrence records has increased five-fold, and it is of both methodological and practical interest to test how robust the SDM results are to spatial biases in this particular training data of a poorly known ecosystem. To reach this aim, we divide the now available data into 24 different spatial configurations of equal area as the original study but different shapes (from more compact to more elongated) and compare their results both with each other and with results obtained with the full data. We then assess how well each of the spatial configurations represents the available environmental space in the study area and to what degree this is related to the performance of the corresponding models. Our aim is to provide such information about the performance of SDMs that can be used when assessing their applicability and reliability for forest management decisions.

2. Materials and methods

2.1. Study area and species data

The study area covers about 6,300 km^2 in Iberia and Tahuamanu districts, Madre de Dios region, Southern Peruvian Amazonia. Climate in the area is tropical and humid with mean monthly temperature ranging from 24,3 to 25,1 °C and total annual rainfall ranging from 1470 to 2225 mm, as extracted from CHELSA (Karger et al. 2017). The terrain is flat to undulating, with elevation ranging between 260 and 410 m above sea level.

We used forest census data produced by Consolidado Otorongo forestry concession in approximately 15,000 ha of its management areas. Commercial trees with diameter at breast height (DBH) at least 30 cm were registered along linear transects and their local name, height, DBH, volume and location were recorded. The beginning and end of each evaluation transect were georeferenced using GPS devices. Each individual tree was georeferenced in the field using tape-measured X and Y coordinates in relation to the starting point of the inventory line. in the field using a relative system of X and Y coordinates. The relative positions of the trees in the inventory line were obtained with a measuring tape. Geographical coordinates were later assigned to each tree by combining their field-measured within-transect locations with the transect GPS coordinates.

To ensure comparability of results, we focus on the same tree taxa (Table 1) that were modelled in a previous study (Chaves et al. 2018). These taxa were selected on the basis of their abundance and taxonomic consistency (Chaves et al. 2018). Since the censuses were performed for commercial rather than scientific purposes and voucher specimens were not collected, the species identifications may not be entirely accurate. Therefore, the focus is on such genera that only contain one species within the inventory dataset and also external sources, such as the Global Biodiversity Information Facility (GBIF), indicate that they have few species in the area in general. Nevertheless, we will be referring to them with their generic names and using the term “taxon” instead of “species”.

We calculated the average number of trees of the same species within a radius of 150 m to address the possible effect of conspecific canopy trees on reflectance values. If many individuals of the same taxon are found together, they may affect the overall surface reflectance to the degree that it indicates actual occurrence of the target taxon to a larger

Table 1
Number of occurrences in the forest census data per taxon as used in the species distribution modeling process for trees in Peruvian Amazonia. Different training datasets were used in different SDMs: the already published occurrence data from Chaves et al. (2018) (“original”), subsets of the same surface area as the original (20% of the total available training data) combined into 24 different spatial configurations (“small”; see Figure S1 for details), and all training data together (“large”). All SDMs used the same test dataset (50% of total census area). An abbreviation (Abb.) to be used in some of the figures is given after each species name.

Species	Abb.	Training dataset			Test dataset
		Original	Small*	Large	
Amburana cearensis	AMB	210	163 (91–267)	1055	809
Apuleia leiocarpa	APU	198	230 (126–342)	1290	1381
Crepidospermum goudotianum	CRE	71	67 (38–102)	469	498
Dipteryx odorata	DIP	491	424 (290–539)	2620	2502
Manilkara bidentata	MAN	168	220 (159–298)	1285	1170

(*) Average number of occurrences. Minimum and maximum in brackets.

degree and is less accurate as a model of general canopy properties.

2.2. Remotely sensed predictors

Given that our study area is relatively small (6,300 km^2) and entirely situated in lowland tropical rainforest, climatic variability can be expected to have only a minimal effect on species distributions, if any. Soil variability is more likely to be important, but given the accuracy problems of digital soil maps in these poorly sampled areas (Moulatlet et al. 2017) as well as their coarse spatial resolution, we chose to base our analyses on Landsat satellite imagery and elevation (referred together as remote sensing data).

We obtained reflectance data from a Landsat TM/ETM + image composite over the Amazon rainforest biome (Van doninck & Tuomisto 2018). The composite was based on all image acquisitions of the dry season months July–September in the years 2000–2009 The composite was subsequently cropped to our study area, which was within Landsat paths 11–12 and rows 67–68. We used four of the seven Landsat TM/ETM + bands as predictors for the SDMs: red (band 3), near-infrared (band 4), and shortwave infrared (bands 5 and 7). Blue and green (bands 1 and 2, respectively) were excluded because of extensive residual atmospheric contamination. Normalized difference vegetation index (NDVI) was also used as a predictor variable. As an additional predictor variable in the models, we used elevation above sea level from ASTER GDEM Version 003 (Aster Global Digital Elevation Model) granules S11W071, S11W070, S12W071 and S12W070, which have the same 30-m spatial resolution as the Landsat data. Earlier studies in Amazonia have found average filtering with a moderate-size window to increase the accuracy of forest classification, floristic modeling and species distribution models (Rajaniemi et al. 2005; Salovaara et al. 2005; Chaves et al. 2018) as well as predicting soil characteristics and floristic composition (Van doninck & Tuomisto 2018). Filtering reduces the amount of local noise while preserving the broader-scale patterns. Therefore, we applied an averaging filter based on a 5x5 pixel window (150 × 150 m) to the remote sensing data before further analyses.

2.3. Modeling process

To obtain fair comparisons among models based on different training data configurations and to avoid overfitting of the models (Elith et al. 2006; Veloz 2009; Radosavljevic Aleksandar et al. 2013), we divided the available occurrence data into two halves. One of these was used as the test data for all models, and the other was divided in different ways to obtain different training data configurations (Fig. 1). Three sets of models were built for each taxon. The first set was based on the original training dataset used in a previous study (Chaves et al. 2018) (referred to as "original" model sets). This covered about 20% of the now available occurrence data, so the new data were divided into 24 different spatial configurations of similar area as the original but different shapes (narrow, elongated and compact; Figure S1). The length–width ratios of the narrow, elongated and compact spatial configurations were approximately 20:1, 20:2 and 20:12, respectively. The occurrence data within each of these was used to build another set of models (referred to as "small" model set). Finally, all available occurrences within the training area were used for building complete species distribution models (referred to as "large" model sets). Numbers of occurrence records

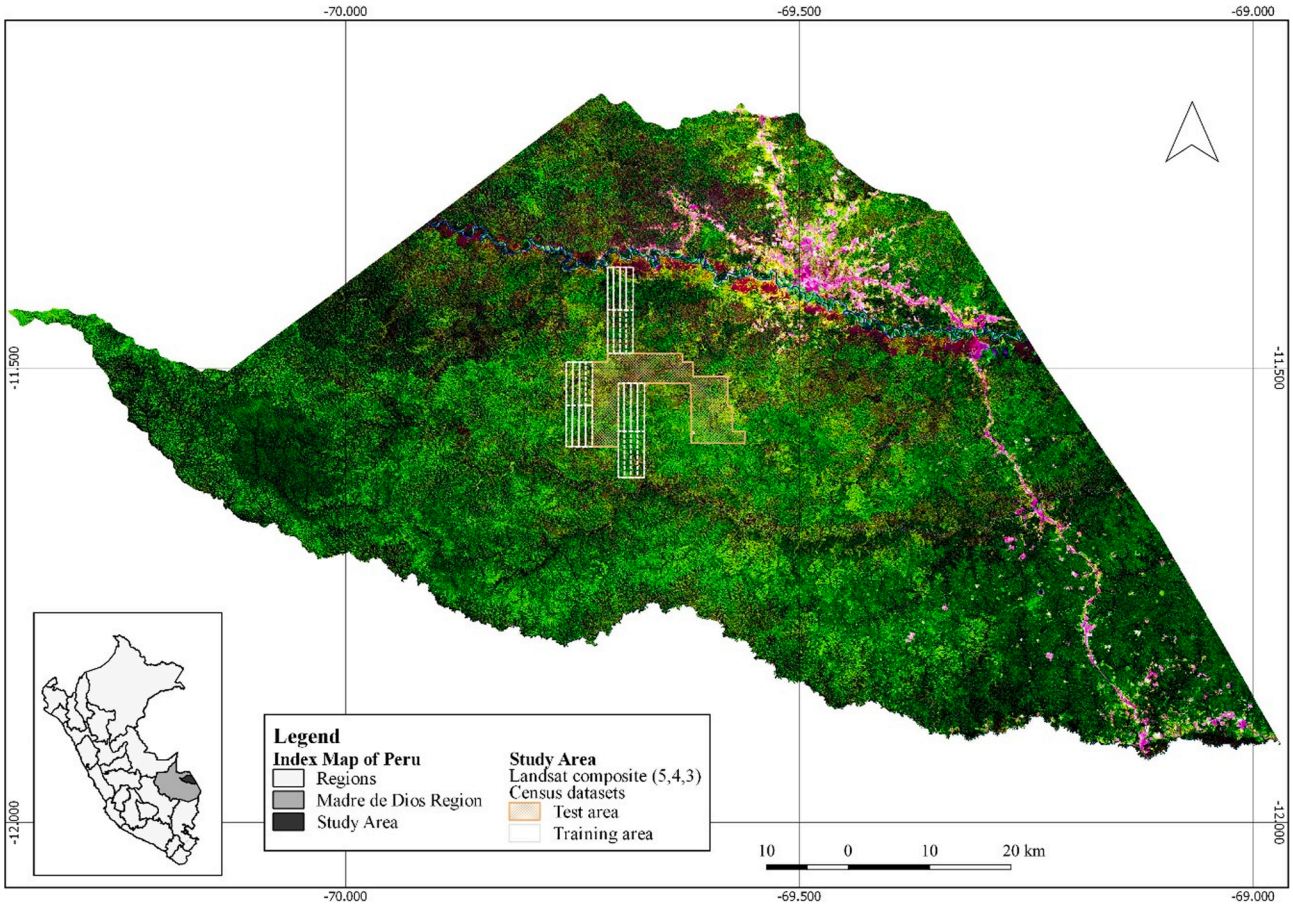

Fig. 1. Satellite image of the study area in Southern Peruvian Amazonia with polygons indicating the extents of the forest census datasets. Trees occurring within the orange dashed polygons were used for model validation (test data) and trees coming from the white polygons were used in different combinations for model calibration (training data; see Figure S1 for details). In the satellite image, pink colors correspond to deforested areas (mainly due to agricultural expansion), red to inundated areas, blue to the Tahuamanu river and different shades of green to different forest types.

available for each taxon are shown in Table 1.

The census dataset provided information about the presence of tree individuals larger than 30 cm of diameter, but no absence data. Therefore, we used the complementary log–log (cloglog) link function (Phillips et al. 2017) of MaxEnt algorithm, which uses presence-only data and background information to model species distributions (Phillips et al. 2017). MaxEnt has performed equally well or better than other modeling algorithms (Giovanelli et al., 2010; Phillips et al., 2006; Valavi, 2021; Elith et al., 2006; Hernandez et al., 2006; Merckx et al., 2011; Merow et al., 2014; Wisz et al., 2008), and it has been found to derive consistent predictions across different calibration areas (Giovanelli et al. 2010) and to be less sensitive to configuration settings (Hallgren et al. 2019). To avoid excessive model complexity, we used only the linear and quadratic features of MaxEnt (Syfert et al. 2013). We used the same features for all the models to facilitate comparisons of model predictions among the different tree taxa and models sets ("original", "small" and "large" training data sets).

Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), using the same test dataset for each model. From the test dataset we extracted all the suitability values derived from the SDM predictions and we assessed differences in the mean suitability values between the "original", "small" and "large" models of each taxon using a Tukey's test. We compared the spatial predictions of each taxon and model set using a Pearson correlation. For the "small" model set, since it consisted of 24 models per taxon (each of the 24 spatial configurations of the training area), we calculated the average and standard deviation of the spatial predictions of each configuration ("compact", "elongated" and "narrow"). All analyses were carried out in R version 3.5.1 (R Core Team 2020) using the packages "raster" (Hijmans 2017), "rgdal" (Bivand et al. 2016) and "dismo" (Hijmans et al. 2015).

2.4. Environmental space and model performance

We quantified a 2-dimensional environmental space where elevation was one of the dimensions. The other dimension was the first axis of a principal components analysis (PCA) of the 4 Landsat bands that were used in the modeling exercise (hereafter referred to as "Landsat-PCA"). This choice was made because elevation was the most important contributing variable in all SDMs while the contributions of the Landsat bands varied among the taxa (Supplementary material - Table S2). NDVI was not included in the PCA since the bands it is based on were already included. We extracted different Landsat-PCA and elevation values in two ways: (i) from the geographic areas (e.g: values from the entire study area or from the training areas – Fig. 1) and (ii) from the occurrence points (tree location) of the training data. For the latter (ii), we extracted the Landsat-PCA and elevation values for the coordinates that correspond to the occurrence data points within each of the 24 spatial configurations ("training occurrence") and calculated their median and range. For the former (i), we extracted Landsat-PCA and elevation values from: (a) the entire study area, (b) the entire training area in the "large" model and (c) each of the 24 spatial configurations of training data in the "small" model set ("training area") and calculated different environmental space metrics.

Based on the values extracted from the geographic training areas, we plotted the extracted values from each of the 24 spatial configurations of the training area in a bidimensional environmental space (Landsat-PCA and elevation) and defined a convex hull around them. We then quantified the coverage of each of the 24 spatial configurations in terms of the percentage of area they covered in the environmental space (A%), and the density of extracted values they contained (D%). For each configuration, we defined the centroid, range and standard deviation of the extracted values in both environmental dimensions. Additionally, we extracted the Landsat-PCA and elevation values for the coordinates that correspond to the taxon occurrence data points within each of the 24 spatial configurations ("training occurrence") and calculated their median and range.

We use the term "environmental space metrics" to refer to all the statistics mentioned above based on the training occurrence data points (median and range) and based on the training areas (A%, D%, centroid, range and standard deviation). We used Pearson's coefficient of correlation to quantify to what degree model performance (AUC) was related to each of the environmental space metrics and the number of taxon occurrences used for model training.

3. Results

3.1. Model performance and suitability estimates

For all tree taxa, average model performance (expressed as AUC) was lowest when compact training data configurations were used, and higher when elongated or narrow configurations were used (by 3% and 5%, respectively; Table 2). The highest increase in model performance between compact and narrow configuration was seen for *Crepidospermum* (more than 6%) and the smallest for *Apuleia* (3.5%). Additionally, the standard deviation of the AUC values obtained for models that used different narrow training data configurations was less than half of the standard deviation for models that used different compact configurations (Table 2).

Model performance (AUC values) based on the test dataset varied among the taxa and was between 0.56 and 0.83 when using the "original" dataset for model training and between 0.42 and 0.87 when using the "small" sets of the training data (Fig. 2a). When all of the training data were used ("large" in Fig. 2a), AUC values ranged between 0.75 and 0.85 and were invariably higher (by 10–33%) than in models for the same taxon that used smaller training data sets. When comparing the "original" and "small" model sets, average performance was similar for *Crepidospermum* and *Amburana*, but better when using the "small" sets than the "original" for *Manilkara*, *Dipteryx* and *Apuleia* (Fig. 2a). The biggest differences in model performance between training data sets were seen for *Apuleia* and the smallest for *Amburana* and *Crepidospermum* (Fig. 2a).

The models gave, on average, significantly higher suitability values when using the "large" training dataset compared to the "original" and "small" ones. We found statistically significant ($p < 0.05$, Tukey's test) differences in the average suitability values between each pair of model sets ("original", "small" and "large") per taxon.

All tree taxa had, on average, less than two conspecific neighbours within a radius of 150-m, except for *Dipteryx*. The average number of conspecific neighbours per taxon within a 150-m radius were 1.2, 1.6, 0.7, 2.4 and 1.7 for AMB, APU, CRE, DIP and MAN respectively.

3.2. Environmental space, number of occurrences and model performance

Each of the 24 spatial configurations of the "small" training data sets covered partly different segments of the environmental space represented in the study area (Fig. 3). On average, narrow configurations of

Table 2
Average and standard deviation values of MaxEnt model performance (AUC) for five tree taxa when using different spatial configurations of the training data (compact, elongated, and narrow) in Peruvian Amazonia. AUC Standard deviation values are shown in parentheses. AMB = *Amburana*, APU = *Apuleia*, CRE = *Crepidospermum*, DIP = *Dipteryx*, MAN = *Manilkara*.

Spatial configuration	AUC Model performance (AUC Standard deviation)				
	AMB	APU	CRE	DIP	MAN
Compact	0.765 (0.109)	0.678 (0.131)	0.801 (0.083)	0.747 (0.111)	0.714 (0.123)
Elongated	0.789 (0.051)	0.682 (0.096)	0.833 (0.026)	0.768 (0.055)	0.733 (0.060)
Narrow	0.809 (0.028)	0.702 (0.061)	0.851 (0.018)	0.787 (0.052)	0.751 (0.051)

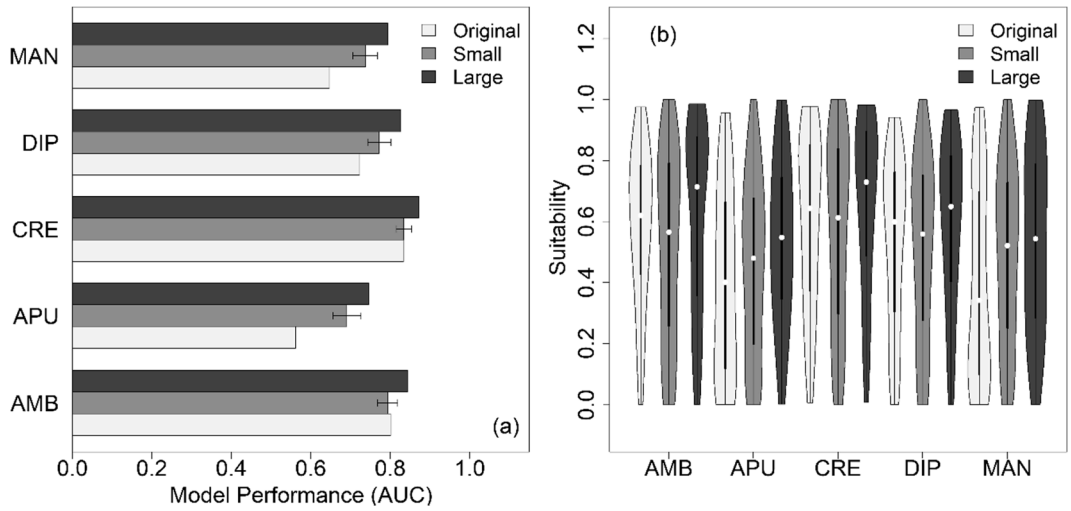

Fig. 2. Model performance and suitability values of SDMs derived for 5 tree taxa in Peruvian Amazonia. (**a**) Model performance (AUC) using the test data; "original" sets used similar occurrence data as Chaves et al. (2018), "small" sets have the same surface area as the "original" data but represent 24 different spatial configurations (see Supplementary Figure - S1 for details), and "large" sets used all occurrences within the training area (**b**) Distribution of predicted suitability values per taxon and per model set as obtained using the independent test dataset. AMB = *Amburana*, APU = *Apuleia*, CRE = *Crepidospermum*, DIP = *Dipteryx*, MAN = *Manilkara*.

the training data represented the environmental space better (20% of the environmental space, defined by two environmental variables, which contained 70% of total occurrence point density) than elongated configurations (16% of space and 56% of density) and compact configurations (13% of space and 46% of density). The position of each training area configuration in the environmental space varied mainly along the elevation axis, as reflected in the centroid positions in Fig. 3.

Model performance was strongly correlated with the centroid position along the elevation dimension for most of the taxa (Table 3). Only for *Crepidospermum*, model performance was correlated with the proportion of environmental space covered within each training area (Table 3). Model performance was not correlated with either elevation or Landsat-PCA range, nor with the centroid position in the Landsat-PCA axis for any taxon, but it was strongly correlated with both the median values of Landsat-PCA and elevation when extracted from the training occurrences within each spatial configuration (Table 3). Finally, model performance was only correlated with the number of occurrences used in the model for *Crepidospermum* (Table 3). The relationship between AUC and all the environmental space metrics is also shown in the supplementary material (Supplementary materials - Figure S3)

3.3. Model projection and predicted suitability

The predicted suitability patterns differed for each taxon depending on which training data set was used, and this was related to general model performance (Fig. 4, SM3). For *Crepidospermum* and *Amburana*, whose models invariably obtained high AUC values, the predicted suitability patterns were relatively stable, i.e. they varied little between model sets ("original", "small" and "large"). For *Dipteryx*, *Manilkara* and especially *Apuleia*, whose models obtained lower AUC values, the results were more sensitive to changes in training data configuration. Models built using the "large" sets always had higher AUC values than models built using the other sets ("original" and "small"). When comparing the "large" and "original" models of the same taxon, the spatial suitability patterns remained similar, except for *Apuleia*, whose suitability predictions differed spatially (Fig. 4).

Comparison of the results based on different "small" training data sets revealed that using compact spatial configurations of the training data led to lower performance of the models (AUC) than when using elongated or narrow configurations (Fig. 5). Additionally, the standard deviation of the predicted suitability values was more than twice as high for models trained using compact configurations than for models trained with narrow configurations (Fig. 6), indicating that the predictions of the former may be less robust.

4. Discussion

4.1. Model performance and the spatial configuration of the training area

We found that using elongated spatial configurations of the training data for model building increased average SDM model performance (AUC) by approximately 5% and reduced variation in performance by at least 50% when compared to using compact configurations. This is in agreement with previous studies, carried out at broader extents, which have shown how data partitioning and the number of occurrences might lead to different modeling outcomes (Stockwell & Peterson 2002; Elith et al. 2006; Hernandez et al. 2006; Wisz et al. 2008; Veloz 2009; Mateo et al. 2010; Gonzalez et al. 2011; Hijmans 2012; Radosavljevic Aleksandar et al. 2013; Boria et al. 2014; Fourcade et al. 2014; van Proosdij et al. 2016; Boria & Blois 2018). Our results provide a quantification of the phenomenon at the local extent in a generally poorly known forest area.

The improved performance and robustness of SDMs based on elongated training areas can be assumed to emerge because elongated training areas generally occupied a larger proportion of the environmental space than compact ones (20% vs 13%), which made the training data representative for a larger proportion of the occurrence points (70% vs 45%). How representative the training set is of the environmental variation within the test set and within the geographical area to which the model is applied is rarely explicitly quantified in species distribution modeling. We addressed this question graphically, and treated mismatch by excluding from modeling geographical areas with environmental conditions beyond the range encountered in the training set. Another option for the geographical delimitation of the study area is to include only pixels that are environmentally sufficiently similar to those of the training set (Meyer & Pebesma 2021)

Overall, the standard deviation values of the suitability predictions were higher when compact spatial configurations of the training data were used to build the models than when elongated configurations were used. This highlights the importance of taking into consideration the spatial configuration of the training area when using geographically highly concentrated occurrence data for modeling the distribution of trees at local extents. Furthermore, in all model predictions, the high suitability values were concentrated mainly around the training areas. It

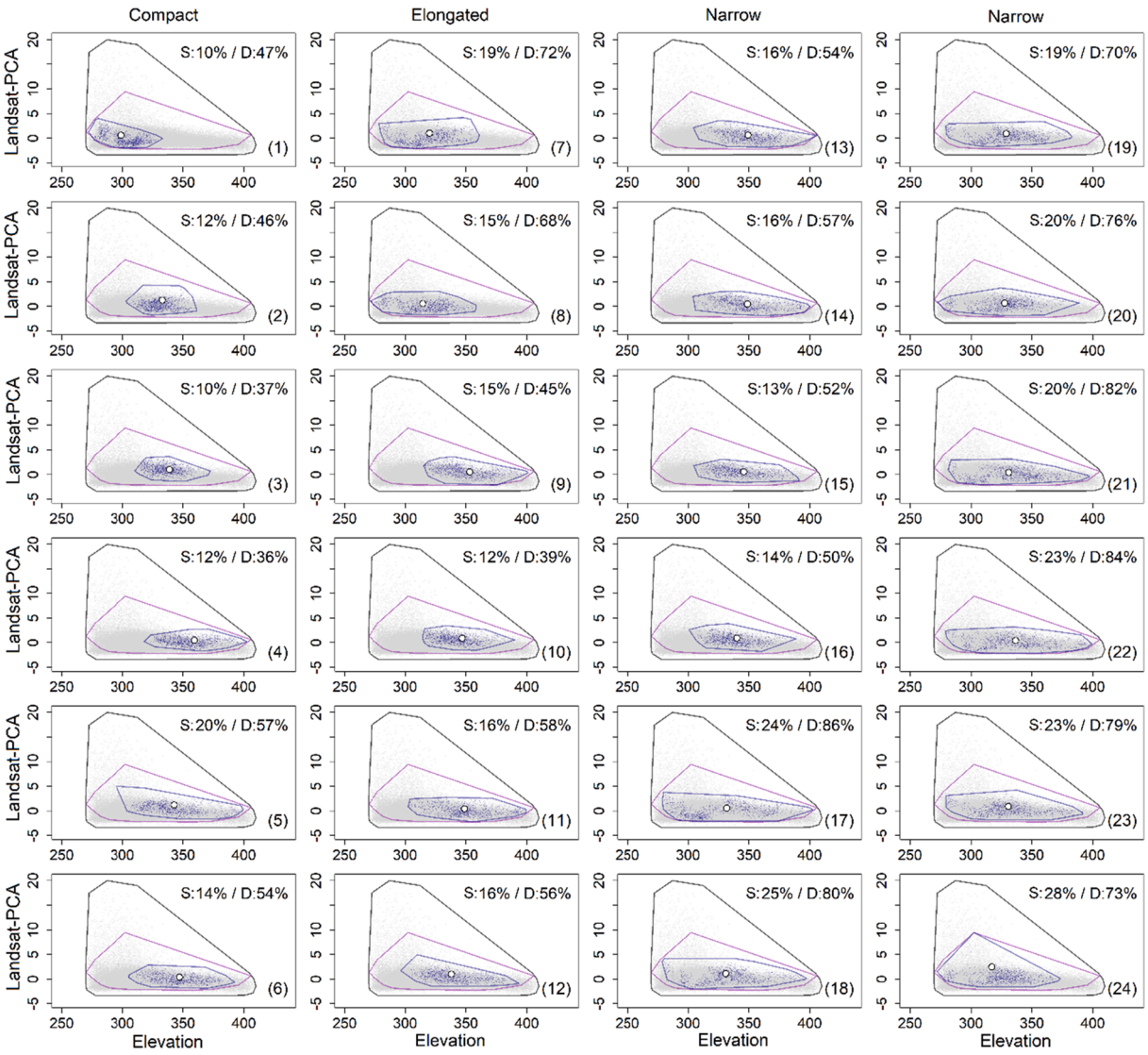

Fig. 3. Environmental space covered in the different spatial configurations of training data used for species distribution model building for trees in Peruvian Amazonia (identifying number of each configuration is shown in parentheses; see Supplementary materials - Figure S1). The environmental space is represented by elevation and by the first axis of a principal component analysis of the Landsat bands 3, 4, 5 and 7 (Landsat-PCA). Grey polygons are convex hulls depicting the available environmental space within the whole study area. Purple convex hulls show the environmental space covered within the entire training area. Blue convex hulls show the environmental space covered within each of the 24 spatial configurations of the "small" training data sets used for model building. Grey and blue points represent pixels in the environmental space of the whole study area and each of the 24 "small" training data sets, respectively. The white circles depict the centroids of each blue convex hull. The proportion of the environmental space (S%) and point density (D%) covered within each blue convex hull is shown in the graph. Compact, elongated and narrow in the column title refer to the different types of spatial configurations of the training area.

is, therefore, necessary to assess the spatial uncertainty in species distribution models in more detail (Rocchini et al. 2011), especially for those predictions that are far from the training areas. Standard deviation maps of the predicted suitability values help to identify locations where the predictions are more robust, but they need to be interpreted with care. In general, the predicted suitability values in our models were low in areas far from the training areas, and this can be expected to be a general trend. As a consequence, the standard deviation in those locations will be low as well. Therefore, although low standard deviation can be thought to indicate model stability, this may mostly apply to locations close to the training area.

When we divided the training area into 24 spatial configurations of similar area, the number of occurrences per taxon within each configuration varied but was always above the minimum number recommended for deriving accurate models (Stockwell & Peterson 2002; Wisz et al. 2008; van Proosdij et al. 2016). Even though we did not aim at controlling for the number of occurrences, having 24 model sets of different numbers of occurrences enabled us to assess the relationship between model performance and the number of occurrences. We found that a larger number of occurrences did not improve model performance. This appears to disagree with the general findings that model performance increases with increasing samples size (Stockwell & Peterson 2002; Hernandez et al. 2006; Wisz et al. 2008; Mateo et al. 2010; Merckx et al. 2011; van Proosdij et al. 2016), but may simply reflect the fact that the variation in the number of occurrences per spatial configuration was small (Table 1). Within this range, the number of occurrences appeared less important than their spatial distribution, because the latter had a bigger impact on how representative the occurrence points were of the entire area of interest. In addition, model performance was explained by the environmental space metrics used in our research (Landsat reflectance values and elevation). In Amazonian forests, canopy reflectance and elevation have been related to soil nutrients (Costa et al. 2005; Higgins et al. 2011; Sirén et al. 2013; Van doninck & Tuomisto 2018; Tuomisto et al. 2019), so the median

Table 3
Pearson correlation coefficients between model performance (AUC) and training data set properties for five tree taxa in Peruvian Amazonia. The environmental space was characterized based on the training areas of each of the 24 "small" model sets and their training occurrences. Landsat-PCA is the first axis of a principal component analysis based on 4 Landsat bands (4–6 and 7) used in the modeling procedure. AMB = *Amburana*, APU = *Apuleia*, CRE = *Crepidospermum*, DIP = *Dipteryx*, MAN = *Manilkara*. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Env. Space	Env. Metric	Pearson correlation coefficient				
		AMB	APU	CRE	DIP	MAN
Training area	Space (%)	0.38	0.15	0.47*	0.27	0.20
	Point density (%)	0.10	0.28	0.26	0.03	0.14
	Centroid (Elevation)	0.46*	0.65***	0.33	0.43*	0.55**
	Centroid (Landsat-PCA)	0.37	0.35	0.36	0.38	0.34
	St. Deviation (Elevation)	0.17	0.10	0.32	0.03	0.04
	St- Deviation (Landsat-PCA)	0.27	0.63***	0.18	0.48**	0.55**
	Range (Landsat-PCA)	0.23	0.24	0.21	0.24	0.22
	Range (Elevation)	0.24	0.03	0.38	0.12	0.07
Training occurrences	Median (Elevation)	0.41*	0.75***	0.24	0.60**	0.61**
	Median (Landsat-PCA)	0.77***	0.83***	0.60**	0.76***	0.72***
	Range (Landsat-PCA)	0.26	0.25	0	0.47*	0.17
	Range (Elevation)	0.15	0.05	0.22	0.16	0.05
	Number of occurrences	0.08	0.13	0.58**	0.09	0.13

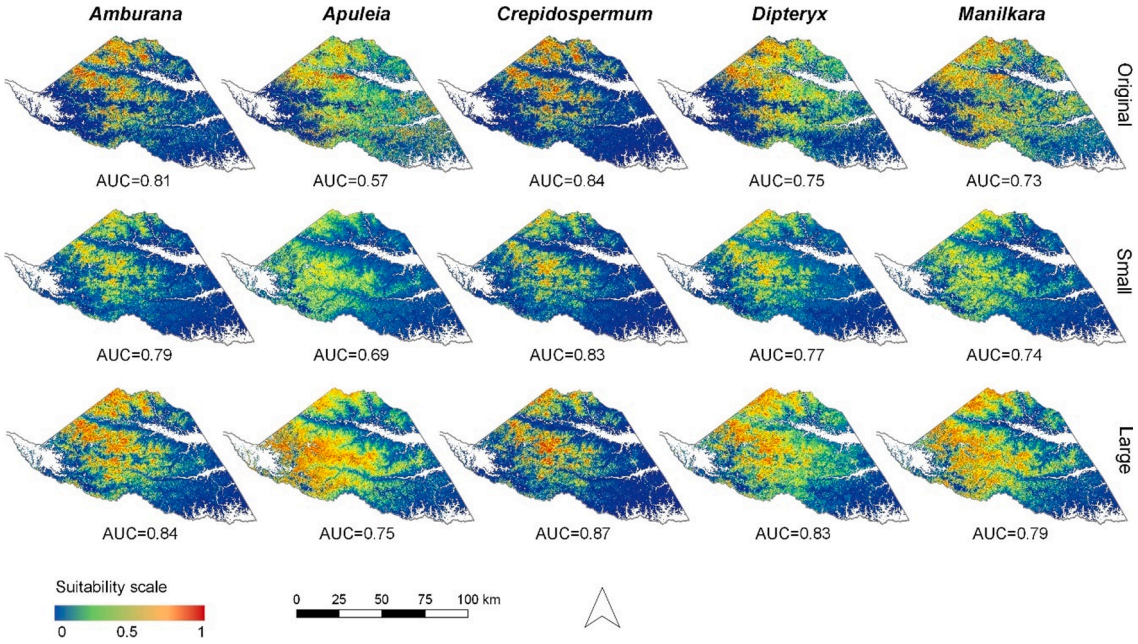

Fig. 4. Predicted suitability values for five tree taxa obtained from models based on different sets of training data ("original", "small", and "large"; see S1) in Peruvian Amazonia. The second row ("small") shows the average of the suitability predictions from 24 different spatial configurations of the training data. Values of predictor variables outside the range covered by the occurrence data were masked out and excluded from the analyses (shown in white).

environmental conditions extracted from the occurrence data locations is likely to be related to the average habitat conditions where each tree taxon occurs.

Our results showed that increasing the training area by a factor of four improved model performance by 20%, on average. For *Amburana* and *Crepidospermum*, quadrupling the training area increased average model performance by 10%, for *Dipteryx* by 13% and for *Apuleia* and *Manilkara*, by more than 30%. This probably reflects the fact that a larger proportion of the environmental space was sampled, which allowed deriving better species distribution models.

4.2. Implications and recommendations for tropical forest managment

Combining forest census data from forest concessions and freely available remote sensing data, such as Landsat satellite imagery and elevation, offer a unique opportunity to predict the distribution of important tree taxa in nearby areas where field data is still missing. In Peru, forest concessions are granted for up to 40 years and their area can be up to 40,000 ha. Forest concessions are divided into units that are yearly managed (so called "annual cutting units"). All commercial trees with more than 30 cm in diameter within the annual cutting unit need to be measured and censused before any harvesting activities are allowed. With the forest census data of an annual cutting unit, it is possible to predict the potential distribution of important tree taxa for the next annual cutting units in order to estimate the forest potential and plan forest management activities. However, robust predictions can only be derived with adequate spatial configuration of the training areas. The smaller the part of the environmental variability in the study area that is covered by the training data, the bigger the risk that misleading SDM predictions are derived.

Our findings emphasize the importance of taking into consideration the spatial configuration of the training area when using geographically highly concentrated occurrence data for modeling the distribution of trees at finer scales. In such cases, we recommend dividing the training

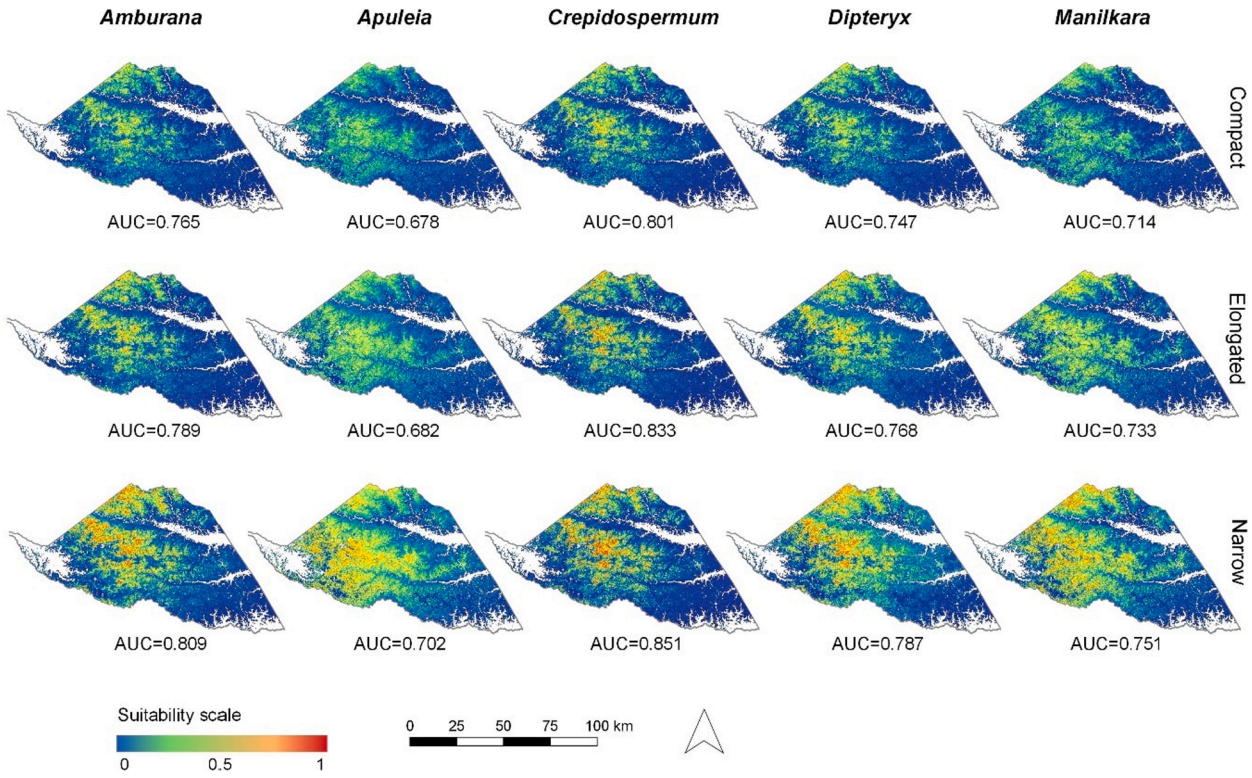

Fig. 5. Predicted suitability values for five tree taxa in Peruvian Amazonia obtained from species distribution models based on different spatial configurations of the training data ("compact", "elongated", and "narrow"; see S1). The figure shows average suitability predictions of each spatial configuration type. Values of predictor variables outside the range covered by the occurrence data were masked out and excluded from the analyses (shown in white).

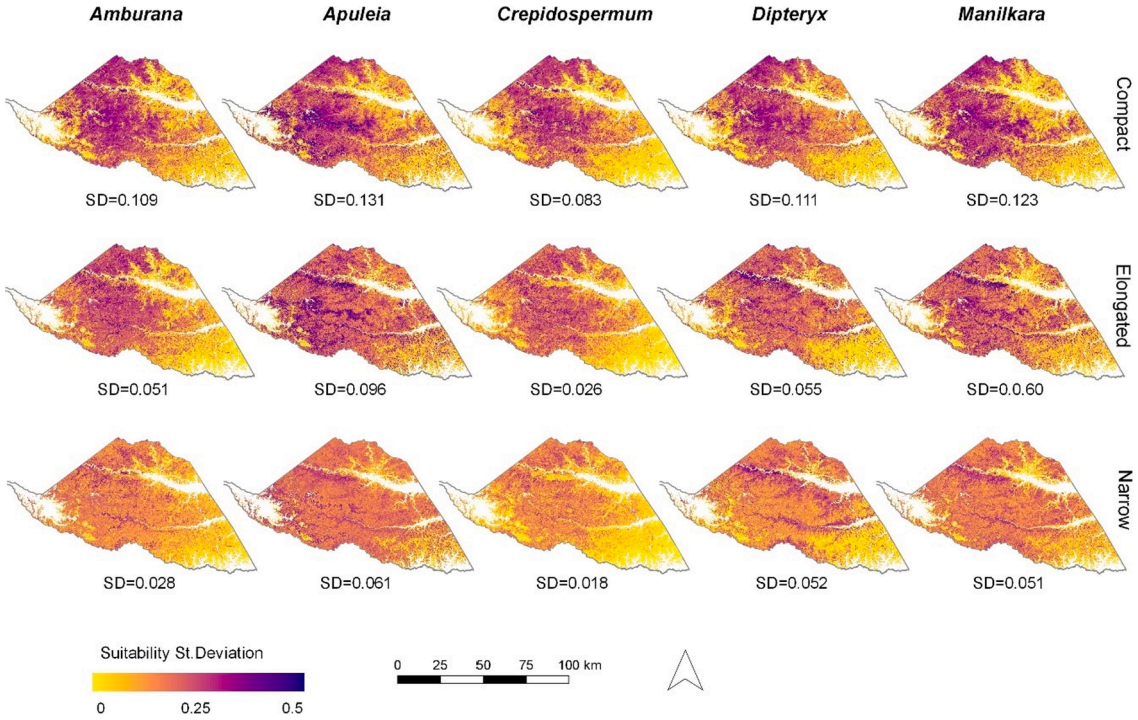

Fig. 6. Standard deviation of the predicted suitability values for five tree taxa as calculated with the models shown in Fig. 5.

occurrence data in elongated configurations. Our approach is particularly relevant in cases where tropical forest management decisions are made based on SDMs, since both model performance and robustness can be considerably improved if spatial configurations are considered in model building. If training data are biased to just one part of the area of interest, the risk is big that areas far away will be predicted to have low suitability for the species of interest just because there is no training data in matching conditions.

We acknowledge the possibility of misidentifications and species lumping in our data, and therefore we referred to them with generic names and the term "taxa" rather than "species". If species with different habitat preferences get lumped in the analyses, their combined habitat preferences will be broader and probably more difficult to model than those of the individual species. On the other hand, lumping increases the number of occurrence records and may thereby improve model performance. Encouragingly, earlier studies have shown that there is ecological signal in genus-level and even family-level data, which gives confidence in results based on data with possible lumping (Higgins & Ruokolainen 2004; Emilio et al. 2010; Cayuela et al. 2011).

Two potential problems in our models emerge from the fact that the occurrence data only contained canopy trees with DBH $\geq$ 30 cm. Firstly, it is impossible to distinguish between true and false absences of a taxon, because the presences of individuals smaller than the threshold value were not registered in the field census. We mitigated this problem by using a presence-only modeling method. Secondly, pixel values in Landsat images depend on the reflectance characteristics of the forest canopy, which in turn are largely determined by the canopy trees. Therefore, it is possible that the SDM of a given taxon is affected by its own spectral characteristics, which might lead to the SDMs modeling where the big trees of the taxon are growing at the moment rather than more generally where the environmental characteristics are suitable for it. We reduced this risk by filtering the Landsat data and carrying out the SDMs using 150-m pixels. Exceedingly few 150-m pixels had more than two individuals of the same taxon, as these forests have a very high species diversity. Therefore, the contribution of any individual tree crown to the pixel's overall reflectance is small. Obviously, in less species-rich forests the situation may be different and the issue needs to be given more attention.

5. Conclusions

More elongated configurations of the training data were more representative of the available environmental space (as measured by remote sensing data at the local scale in Peruvian Amazonia) and produced better and more robust SDMs for five canopy tree taxa than compact configurations did. Using elongated rather than compact training data configuration increased model performance by up to 5% and reduced variance in model performance between models by 50%. Quadrupling the area covered by the training data increased model performance even more, by 20%. These results are in agreement with earlier observations and provide a quantitative estimate of the importance of both the amount and the spatial configuration of the training area on SDM performance.

Funding

PPC was supported by funding from University of Turku Graduate School and JVd by funding from the Academy of Finland (grant 273,737 to HT and grant 296,406 to Risto Kalliola).

CRediT authorship contribution statement

Pablo Pérez Chaves: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Kalle Ruokolainen:** Conceptualization, Writing – review & editing. **Jasper Van doninck:** Resources, Writing – review & editing. **Hanna Tuomisto:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pablo Perez Chaves reports financial support was provided by University of Turku.

Acknowledgments

Special thanks are due to Consolidado Otorongo Forest Concession for providing the forest census databases for this research.

Data Availability

The forest census data used in this study is property of Consolidado Otorongo Forest Concession and access was given for this study only. Those interested in requesting access to the data may contact the company (see: www.bozovich.com) at peru@bozovich.com. The Amazonian Landsat TM/ETM + composite for July-September 2000-2009 can be accessed through Fairdata (http://urn.fi/urn:nbn:fi:att:71ba2590-7112-4669-a4b3-a427c85c7a86).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foreco.2021.119838.

References

Araújo, M.B., Anderson, R.P., Barbosa, A.M., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. *Science*. Advances 5, eaat4858.

Bivand, R., Keitt, T., & Rowlingson, B. 2016. *rgdal: Bindings for the Geospatial Data Abstraction Library*.

Boria, R.A., Blois, J.L., 2018. The effect of large sample sizes on ecological niche models: Analysis using a North American rodent, Peromyscus maniculatus. Ecological Modelling 386, 83–88.

Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling 275, 73–77.

Bradley, B.A., Olsson, A.D., Wang, O., Dickson, B.G., Pelech, L., Sesnie, S.E., Zachmann, L.J., 2012. Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data? Ecological Modelling 244, 57–64.

Buermann, W., Saatchi, S., Smith, T.B., Zutta, B.R., Chaves, J.A., Milá, B., & Graham, C. H. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography* 35: 1160–1176.

Cayuela, L., de la Cruz, M., & Ruokolainen, K. 2011. A method to incorporate the effect of taxonomic uncertainty on multivariate analyses of ecological data. *Ecography* 34: 94–102.

Chaves, P.P., Ruokolainen, K., & Tuomisto, H. 2018. Using remote sensing to model tree species distribution in Peruvian lowland Amazonia. *Biotropica* 50: 758–767.

Chaves, P.P, et al., 2021. Using forestry inventories and satellite imagery to assess floristic variation in bamboo-dominated forests in Peruvian Amazonia. Journal of Vegetation Science. https://doi.org/10.1111/jvs.12938.

Chaves, P., Zuquim, G., Ruokolainen, K., Van doninck, J., Kalliola, R., Gómez Rivero, E., Tuomisto, H., 2020. Mapping Floristic Patterns of Trees in Peruvian Amazonia Using Remote Sensing and Machine Learning. Remote Sensing 12 (9), 1523. https://doi.org/10.3390/rs12091523.

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F., Liu, J., 2018. Effects of grain size and niche breadth on species distribution modeling. Ecography 41 (8), 1270–1282.

Cord, A.F., Meentemeyer, R.K., Leitão, P.J., Václavík, T., Whittaker, R., 2013. Modelling species distributions with remote sensing data: bridging disciplinary perspectives (R. Whittaker, Ed.). Journal of Biogeography 40 (12), 2226–2227.

Costa, F.R.C., Magnusson, W.E., Luizao, R.C., 2005. Mesoscale distribution patterns of Amazonian understorey herbs in relation to topography, soil and watersheds. Journal of Ecology 93, 863–878.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M., & E. Zimmermann, N. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.

Emilio, T., Nelson, B.W., Schietti, J., Desmoulière, S.-M., Espírito Santo, H.M.V., Costa, F. R.C., 2010. Assessing the relationship between forest types and canopy tree beta diversity in Amazonia. Ecography 33 (4), 738–747.

van Ewijk, K.Y., Randin, C.F., Treitz, P.M., Scott, N.A., 2014. Predicting fine-scale tree species abundance patterns using biotic variables derived from LiDAR and high spatial resolution imagery. Remote Sensing of Environment 150, 120–131.

Figueiredo, S.M. de M., Venticinque, E.M., Figueiredo, E.O., & Ferreira, E.J.L. 2015. Predicting the distribution of forest tree species using topographic variables and vegetation index in eastern Acre, Brazil. *Acta Amazonica* 45: 167–174.

Figueiredo, S.M. de M., Venticinque, E.M., Figueiredo, E.O., Figueiredo, S.M. de M., Venticinque, E.M., & Figueiredo, E.O. 2016. Spatial scale effects of sampling on the interpolation of species distribution models in the southwestern amazon. *Revista Árvore* 40: 617–625.

Fortunel, C., Lasky, J.R., Uriarte, M., Valencia, R., Wright, S.J., Garwood, N.C., Kraft, N. J.B., 2018. Topography and neighborhood crowding can interact to shape species growth and distribution in a diverse Amazonian forest. Ecology 99 (10), 2272–2283.

Fourcade, Y., Engler, J.O., Rödder, D., Secondi, J., Valentine, J.F., 2014. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. PLOS ONE 9 (5), e97122.

Franklin, J., 2013. Species distribution models in conservation biogeography: developments and challenges. Diversity and Distributions 19 (10), 1217–1223.

Franklin, J., Miller, J.A., 2010. Mapping species distributions: Spatial inference and prediction. Cambridge University Press.

Giovanelli, J.G.R., de Siqueira, M.F., Haddad, C.F.B., Alexandrino, J., 2010. Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods. Ecological Modelling 221 (2), 215–224.

Gomes, V.H.F., IJff, S.D., Raes, N., Amaral, I.L., Salomão, R.P., Coelho, L.S., Matos, F.D. A., Castilho, C.V., Filho, D.A.L., López, D.C., Guevara, J.E., Magnusson, W.E., Phillips, O.L., Wittmann, F., Carim, M.J.V., Martins, M.P., Irume, M.V., Sabatier, D., Molino, J.-F., Bánki, O.S., Guimarães, J.R.S., Pitman, N.C.A., Piedade, M.T.F., Mendoza, A.M., Luize, B.G., Venticinque, E.M., Novo, E.M.M.L., Vargas, P.N., Silva, T.S.F., Manzatto, A.G., Terborgh, J., Reis, N.F.C., Montero, J.C., Casula, K.R., Marimon, B.S., Marimon, B.-H., Coronado, E.N.H., Feldpausch, T.R., Duque, A., Zartman, C.E., Arboleda, N.C., Killeen, T.J., Mostacedo, B., Vasquez, R., Schöngart, J., Assis, R.L., Medeiros, M.B., Simon, M.F., Andrade, A., Laurance, W.F., Camargo, J. L., Demarchi, L.O., Laurance, S.G.W., Farias, E.S., Nascimento, H.E.M., Revilla, J.D. C., Quaresma, A., Costa, F.R.C., Vieira, I.C.G., Cintra, B.B.L., Castellanos, H., Brienen, R., Stevenson, P.R., Feitosa, Y., Duivenvoorden, J.F., C, G.A.A., Mogollón, H.F., Targhetta, N., Comiskey, J.A., Vicentini, A., Lopes, A., Damasco, G., Dávila, N., García-Villacorta, R., Levis, C., Schietti, J., Souza, P., Emilio, T., Alonso, A., Neill, D., Dallmeier, F., Ferreira, L.V., Araujo-Murakami, A., Praia, D., Amaral, D.D., Carvalho, F.A., Souza, F.C., Feeley, K., Arroyo, L., Pansonato, M.P., Gribel, R., Villa, B., Licona, J.C., Fine, P.V.A., Cerón, C., Baraloto, C., Jimenez, E.M., Stropp, J., Engel, J., Silveira, M., Mora, M.C.P., Petronelli, P., Maas, P., Thomas-Caesar, R., Henkel, T.W., Daly, D., Paredes, M.R., Baker, T.R., Fuentes, A., Peres, C.A., Chave, J., Pena, J.L.M., Dexter, K.G., Silman, M.R., Jørgensen, P.M., Pennington, T., Fiore, A., Valverde, F.C., Phillips, J.F., Rivas-Torres, G., Hildebrand, P., Andel, T.R., Ruschel, A.R., Prieto, A., Rudas, A., Hoffman, B., Vela, C.I.A., Barbosa, E.M., Zent, E.L., Gonzales, G.P.G., Doza, H.P.D., Miranda, I.P.A., Guillaumet, J.-L., Pinto, L.F.M., Bonates, L.C.M., Silva, N., Gómez, R.Z., Zent, S., Gonzales, T., Vos, V.A., Malhi, Y., Oliveira, A.A., Cano, A., Albuquerque, B.W., Vriesendorp, C., Correa, D.F., Torre, E.V., Heijden, G., Ramirez-Angulo, H., Ramos, J.F., Young, K.R., Rocha, M., Nascimento, M.T., Medina, M.N.U., Tirado, M., Wang, O., Sierra, R., Torres-Lezama, A., Mendoza, C., Ferreira, C., Baider, C., Villarroel, D., Balslev, H., Mesones, I., Giraldo, L.E.U., Casas, L.F., Reategui, M.A.A., Linares-Palomino, R., Zagt, R., Cárdenas, S., Farfan-Rios, W., Sampaio, A.F., Pauletto, D., Sandoval, E.H.V., Arevalo, F.R., Huamantupa-Chuquimaco, I., Garcia-Cabrera, K., Hernandez, L., Gamarra, L.V., Alexiades, M.N., Pansini, S., Cuenca, W.P., Milliken, W., Ricardo, J., Lopez-Gonzalez, G., Pos, E., & ter Steege, H. 2018. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. *Scientific Reports* 8: 1003.

Gonzalez, S.C., Soto-Centeno, J.A., Reed, D.L., 2011. Population distribution models: species distributions are better modeled using biologically relevant data partitions. BMC Ecology 11, 20.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8 (9), 993–1009.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press, Cambridge.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135 (2-3), 147–186.

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., Mackey, B., 2019. Species distribution models can be highly sensitive to algorithm configuration. Ecological Modelling 408, 108719. https://doi.org/10.1016/j.ecolmodel.2019.108719.

He, K.S., Bradley, B.A., Cord, A.F., Rocchini, D., Tuanmu, M.-N., Schmidtlein, S., Turner, W., Wegmann, M., Pettorelli, N., Nagendra, H., Horning, N., 2015. Will remote sensing shape the next generation of species distribution models? Remote Sensing in Ecology and Conservation 1 (1), 4–18.

Hernandez, P.A., Graham, C.H., Master, L.L., & Albert, D.L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29: 773–785.

Higgins, M.A., Asner, G.P., Perez, E., Elespuru, N., Tuomisto, H., Ruokolainen, K., & Alonso, A. 2012. Use of Landsat and SRTM Data to Detect Broad-Scale Biodiversity Patterns in Northwestern Amazonia. *Remote Sensing* 4: 2401–2418.

Higgins, M.A., & Ruokolainen, K. 2004. Rapid Tropical Forest Inventory: a Comparison of Techniques Based on Inventory Data from Western Amazonia. *Conservation Biology* 18: 799–811.

Higgins, M.A., Ruokolainen, K., Tuomisto, H., Llerena, N., Cardenas, G., Phillips, O.L., Vásquez, R., Räsänen, M., 2011. Geological control of floristic composition in Amazonian forests. Journal of Biogeography 38 (11), 2136–2149.

Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93 (3), 679–688.

Hijmans, R.J. 2017. *raster: Geographic Data Analysis and Modeling.*

Hijmans, R.J., Phillips, S., & Elith, J.L. and J. 2015. *dismo: Species Distribution Modeling.*

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. *Scientific Data* 4: 170122.

Leitão, P.J., & Santos, M.J. 2019. Improving Models of Species Ecological Niches: A Remote Sensing Overview. *Frontiers in Ecology and Evolution* 7:.

Mateo, R.G., Felicísimo, Á.M., & Muñoz, J. 2010. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science* 21: 908–922.

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., & Vanaverbeke, J. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *ECOLOGICAL MODELLING* 222: 588–597.

Merow Cory, Silander John A., & Warton David. 2014. A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution* 5: 215–225.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution 12 (9), 1620–1633.

Moulatlet, G.M., Zuquim, G., Figueiredo, F.O.G., Lehtonen, S., Emilio, T., Ruokolainen, K., Tuomisto, H., 2017. Using digital soil maps to infer edaphic affinities of plant species in Amazonia: Problems and prospects. Ecology and Evolution 7 (20), 8463–8477.

Muro, J., doninck, J.V., Tuomisto, H., Higgins, M.A., Moulatlet, G.M., Ruokolainen, K., 2016. Floristic composition and across-track reflectance gradient in Landsat images over Amazonian forests. ISPRS Journal of Photogrammetry and Remote Sensing 119, 361–372.

Pearson, R.G., 2010. Species' distribution modeling for conservation educators and practitioners. Lessons in conservation 3, 54–89.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B. (Eds.), 2011. Ecological Niches and Geographic Distributions (MPB-49)Ecological Niches and Geographic Distributions (MPB-49). Princeton University Press.

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography 40 (7), 887–893.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3-4), 231–259.

Prates-Clark, C.D.C., Saatchi, S.S., Agosti, D., 2008. Predicting geographical distribution models of high-value timber trees in the Amazon Basin using remotely sensed data. Ecological Modelling 211 (3-4), 309–323.

Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography 39 (6), 542–552.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Aleksandar, R., Anderson, R.P., Miguel, A., 2013. Making better Maxent models of species distributions: complexity, overfitting and evaluation. Journal of Biogeography 41, 629–643.

Raghavan, R.K., Barker, S.C., Cobos, M.E., Barker, D., Teo, E.J.M., Foley, D.H., Nakao, R., Lawrence, K., Heath, A.C.G., Peterson, A.T., 2019. Potential Spatial Distribution of the Newly Introduced Long-horned Tick, Haemaphysalis longicornis in North America. Scientific Reports 9, 498.

Rajaniemi, S., Tomppo, E., Ruokolainen, K., Tuomisto, H., 2005. Estimating and mapping pteridophyte and Melastomataceae species richness in western Amazonian rainforests. International Journal of Remote Sensing 26 (3), 475–793.

Rehfeldt, G.E., Worrall, J.J., Marchetti, S.B., Crookston, N.L., 2015. Adapting forest management to climate change using bioclimate models with topographic drivers. Forestry: An International Journal of Forest Research 88 (5), 528–539.

Rocchini, D., 2013. Seeing the unseen by remote sensing: satellite imagery applied to species distribution modelling. Journal of Vegetation Science 24 (2), 209–210.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography: Earth and Environment 35 (2), 211–226.

Salovaara, K.J., Thessler, S., Malik, R.N., Tuomisto, H., 2005. Classification of Amazonian primary rain forest vegetation using Landsat ETM+ satellite imagery. Remote Sensing of Environment 97 (1), 39–51.

Sirén, A., Tuomisto, H., Navarrete, H., 2013. Mapping environmental variation in lowland Amazonian rainforests using remote sensing and floristic data. International Journal of Remote Sensing 34 (5), 1561–1575.

Soberón, J.M., 2010. Niche and area of distribution modeling: a population ecology perspective. Ecography 33 (1), 159–167.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148 (1), 1–13.

Syfert, M.M., Smith, M.J., Coomes, D.A., Roberts, D.L., 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. PLOS ONE 8 (2), e55158.

Tuomisto, H., Poulsen, A.D., Ruokolainen, K., Moran, R.C., Quintana, C., Celi, J., Cañas, G., 2003a. Linking floristic patterns with soil heterogeneity and satellite imagery in Ecuadorian Amazonia. Ecological Applications 13 (2), 352–371.

Tuomisto, H., Ruokolainen, K., Aguilar, M., Sarmiento, A., 2003b. Floristic patterns along a 43-km long transect in an Amazonian rain forest. Journal of Ecology 91, 743–756.

Tuomisto, H., Van doninck, J., Ruokolainen, K., Moulatlet, G.M., Figueiredo, F.O.G., Sirén, A., Cárdenas, G., Lehtonen, S., Zuquim, G., 2019. Discovering floristic and geoecological gradients across Amazonia. Journal of Biogeography 46 (8), 1734–1748.

Turner, W., 2014. Sensing biodiversity. Science 346 (6207), 301–302.

Van doninck, J., Jones, M.M., Zuquim, G., Ruokolainen, K., Moulatlet, G.M., Sirén, A., Cárdenas, G., Lehtonen, S., & Tuomisto, H. 2020. Multispectral canopy reflectance improves spatial distribution models of Amazonian understory species. *Ecography* 43: 128–137.

Valavi, et al., 2021. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs. https://doi.org/10.1002/ecm.1486.

Van doninck, J., Tuomisto, H., Nagendra, H., Rocchini, D., 2018. A Landsat composite covering all Amazonia for applications in ecology and conservation. Remote Sensing in Ecology and Conservation 4 (3), 197–210.

Veloz, S.D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography* 36: 2290–2299.

Vormisto, J., Tuomisto, H., Oksanen, J., 2004. Palm distribution patterns in Amazonian rainforests: What is the role of topographic variation? Journal of Vegetation Science 15, 485–494.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., & NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14: 763–773.

Zuleta, D., Russo, S.E., Barona, A., Barreto-Silva, J.S., Cardenas, D., Castaño, N., Davies, S.J., Detto, M., Sua, S., Turner, B.L., & Duque, A. 2018. Importance of topography for tree species habitat distributions in a terra firme forest in the Colombian Amazon. *Plant and Soil.* doi: 10.1007/s11104-018-3878-0.