ECOGRAPHY

Research

Multispectral canopy reflectance improves spatial distribution models of Amazonian understory species

Jasper Van doninck, Mirkka M. Jones, Gabriela Zuquim, Kalle Ruokolainen, Gabriel M. Moulatlet, Anders Sirén, Glenda Cárdenas, Samuli Lehtonen and Hanna Tuomisto

J. Van doninck (https://orcid.org/0000-0003-2177-7882), M. M. Jones (https://orcid.org/0000-0002-8157-8730), G. Zuquim (https://orcid.org/0000-0003-0932-2308), K. Ruokolainen (https://orcid.org/0000-0002-7494-9417), G. M. Moulatlet (https://orcid.org/0000-0003-2571-1207), A. Sirén (https://orcid.org/0000-0003-4159-4506), G. Cárdenas (https://orcid.org/0000-0002-3441-4602) and H. Tuomisto (https://orcid.org/0000-0003-1640-490X) ✉ *(hanna.tuomisto@utu.fi), Dept of Biology, Univ. of Turku, Turku, Finland. JVd also at: Dept of Geography and Geology, Univ. of Turku, Turku, Finland. MMJ also at: Dept of Applied Physics, Aalto Univ. School of Science, Aalto, Finland. GMM also at: Univ. Regional Amazónica Ikiam, Tena, Napo, Ecuador. – S. Lehtonen (https://orcid.org/0000-0001-6235-2026), Biodiversity Unit, Univ. of Turku, Turku, Finland.*

Ecography
43: 128–137, 2020
doi: 10.1111/ecog.04729

Subject Editor: Nathalie Butt
Editor-in-Chief: Miguel Araújo
Accepted 28 August 2019

Species distribution models are required for the research and management of biodiversity in the hyperdiverse tropical forests, but reliable and ecologically relevant digital environmental data layers are not always available. We here assess the usefulness of multispectral canopy reflectance (Landsat) relative to climate data in modelling understory plant species distributions in tropical rainforests. We used a large dataset of quantitative fern and lycophyte species inventories across lowland Amazonia as the basis for species distribution modelling (SDM). As predictors, we used CHELSA climatic variables and canopy reflectance values from a recent basin-wide composite of Landsat TM/ETM+ images both separately and in combination. We also investigated how species accumulate over sites when environmental distances were expressed in terms of climatic or surface reflectance variables. When species accumulation curves were constructed such that differences in Landsat reflectance among the selected plots were maximised, species accumulated faster than when climatic differences were maximised or plots were selected in a random order. Sixty-nine species were sufficiently frequent for species distribution modelling. For most of them, adequate SDMs were obtained whether the models were based on CHELSA data only, Landsat data only or both combined. Model performance was not influenced by species' prevalence or abundance. Adding Landsat-based environmental data layers overall improved the discriminatory capacity of SDMs compared to climate-only models, especially for soil specialist species. Our results show that canopy surface reflectance obtained by multispectral sensors can provide studies of tropical ecology, as exemplified by SDMs, much higher thematic (taxonomic) detail than is generally assumed. Furthermore, multispectral datasets complement the traditionally used climatic layers in analyses requiring information on environmental site conditions. We demonstrate the utility of freely available, global remote sensing data for biogeographical studies that can aid conservation planning and biodiversity management.

Keywords: Amazonia, CHELSA, ferns, Landsat, remote sensing, soils, species distribution modelling

Introduction

Species distribution models (SDMs) are widely used in ecology, biogeography and conservation biology to prioritise conservation actions, forecast climate change impacts, predict biological invasions and test biogeographic hypotheses (Rylands 1990, Guisan and Zimmermann 2000). Species distribution modelling is particularly relevant for large areas with a low density of field observation points, such as Amazonia. Unfortunately, the accuracy of SDMs is constrained not only by the scarcity of verified species presence–absence data points, but also by limited availability and reliability of digital environmental layers (Araújo and Guisan 2006, Carneiro et al. 2016). In other words, areas that would benefit most from SDMs are typically those where the data are least adequate for the purpose.

Commonly used explanatory variables in SDMs are related to climate and elevation (Elith and Leathwick 2009, He et al. 2015). Climate is often thought to be the primary factor that limits species distributions at broad spatial scales (Mackey and Lindenmayer 2001). Digital climatic and topographic data layers have been freely available already for some time (Hijmans et al. 2005, Karger et al. 2017), so they can easily be incorporated in SDMs. Soil properties have also been found important determinants of species distribution for Amazonian plants (Phillips et al. 2003, Tuomisto et al. 2003a, b, 2016, Duque et al. 2005, Baldeck et al. 2016, Cámara-Leret et al. 2017). Digital soil maps have recently become available (Hengl et al. 2017, Zuquim et al. 2019b) and some of these have been found to improve predictions of plant species ranges in Amazonia (Velazco et al. 2017, Figueiredo et al. 2018, Zuquim et al. 2019a), but current global soil maps lack the ecologically most important variables and suffer from low accuracy, especially in poorly sampled areas (Moulatlet et al. 2017).

When digital layers of terrain and environmental characteristics are needed over large and inaccessible areas, remote sensing is a powerful data source and has emerged as a crucial tool in the global modelling of species' distributions, diversity and traits (He et al. 2015, Lausch et al. 2016, Rocchini et al. 2016). In the context of species distribution modelling, data layers derived from low to medium spatial resolution multispectral satellite data can act as biotic predictor variables. For example, data layers derived from the MODIS or AVHRR satellite sensors have been used as a descriptor of habitat in the distribution modelling of various animal species (He et al. 2015). MODIS-derived metrics were also successfully used in SDMs of timber species across Amazonia (Saatchi et al. 2008) and in SDMs of plant species across South America (Buermann et al. 2008). Special attention should be paid to spatial scale when applying SDMs of canopy trees. In particular, the situation should be avoided that the spatial resolution of remote sensing data is in the same order or smaller than the size of individual tree canopies. Each pixel would then represent the reflectance properties of just one canopy species rather than of the forest canopy in general, which would result in mapping the actual presence of a species instead of modelling habitat suitability (Bradley et al. 2012).

Over densely forested areas, such as Amazonia, remotely sensed canopy reflectance is largely a function of canopy structure, tree species composition and physiological condition. Several studies at local and landscape scales have found canopy surface reflectance to be highly correlated with patterns in species composition and turnover in several understory plant groups (Tuomisto et al. 1995, 2003a, b, Salovaara et al. 2005, Thessler et al. 2005, Higgins et al. 2011, 2012, Sirén et al. 2013). Although this may sound surprising, the explanation is logical: canopy trees define many properties of the understory habitat, and the distributions of both canopy trees and understory plants are associated with variation in soil properties and drainage (Ruokolainen et al. 2007).

Multispectral images with an intermediate spatial resolution (10–100 m), such as Landsat and Sentinel-2, have two main advantages for species distribution modelling: their pixel size matches well the typical spatial resolution of field surveys of plant communities, and they provide complete and global spatial coverage. Low resolution imagery (250–4000 m), such as MODIS or AVHRR, has a scale mismatch with field data, and high-resolution imagery (< 10 m) is either very expensive, lacks wide coverage or both. Despite this, we have seen as yet no studies using medium resolution multispectral data in basin-wide SDMs for Amazonian plants, and only a few at the local scale (Figueiredo et al. 2015, Pérez Chaves et al. 2018).

Traditional disadvantages of medium resolution imagery for basin-wide studies include a rather high data volume and problems related to persistent cloud cover, variable aerosol concentrations and effects of illumination and viewing geometry (Toivonen et al. 2006). Increases in computing and data storage capacity are, however, making the data volume problem obsolete, and recent advances in data access policy and algorithms have made it possible to produce radiometrically consistent image composites at the extent of the entire Amazon basin (Van doninck and Tuomisto 2018, Tuomisto et al. 2019a).

In this study, our aim is to assess to what degree spectral data provided by the Landsat sensors can complement traditionally used climatic data layers in ecological modelling at high taxonomic resolution, using an extensive dataset of understory ferns and lycophytes as model group. We compare how models that include Landsat TM/ETM+ data perform when compared to models using only climatic data. In order to obtain results that are generalisable over different modeling techniques and ecological questions, we made the models with three different SDM algorithms (Random Forests, GLM and MaxEnt), and additionally compared species accumulation along gradients defined either by climatic or by surface reflectance data.

Material and methods

Field data

We combined data from field inventories carried out between 1991 and 2013 by researchers working in either the Brazilian

biodiversity research program (PPBio) or the Univ. of Turku Amazon research team (UTU). The same dataset was used in Tuomisto et al. (2019a), where more details on the field methodology can be found. The PPBio inventories comprise 309 plots, each 250 m by 2 m in size. The UTU inventories were originally done using transects either 5 m wide and 500 or 1300 m long, or 2 m wide and up to 43 km long. To match PPBio plot size, the UTU transects were subdivided into 1822 plots of 500 m^2 (5 m by 100 m) and 849 plots of 400 m^2 (2 m by 200 m). The resulting database contains a total of 2980 georeferenced field plots that collectively span a large part of the Amazonian lowland forests (Fig. 1).

In each PPBio and UTU plot, all terrestrial fern and lycophyte individuals with at least one leaf (leafy stem in the case of lycophytes) longer than 10 cm were counted and identified to species. Species that are mostly epiphytic or hemiepiphytic were excluded. Species identifications across the UTU and PPBio data were harmonized by HT and GZ on the basis of either voucher specimens or photographs. Closely related species that had not been reliably separated in the field were lumped to ensure a consistent taxonomy over the entire dataset, which resulted in presence/absence or abundance information for 214 species. Ferns and lycophytes are a practical example group, because some species are present in practically all Amazonian forests, and local species composition varies in response to environmental conditions, especially soil properties (Tuomisto et al. 2003a, b, 2016, 2019a, Zuquim et al. 2014, 2019b).

Composite surface soil samples (topmost layer of the mineral soil down to 5–10 cm depth) were collected in 1572 of the plots and the concentration of exchangeable bases (Ca+Mg+K in cmol(+)/kg) was measured in the soil samples. The optimum and tolerance of each species along the soil base cation concentration gradient were calculated using presence/absence data. The optimum was defined as the average soil cation concentration in those plots where the species occurred and tolerance as the root mean squared error between the species optimum and the observed soil cation concentration value for each occurrence (ter Braak and van Dam 1989). Details on the field and laboratory methodology and on calculation of soil base cation concentration optimum and tolerance are available in Tuomisto et al. (2003a, b), Zuquim et al. (2014).

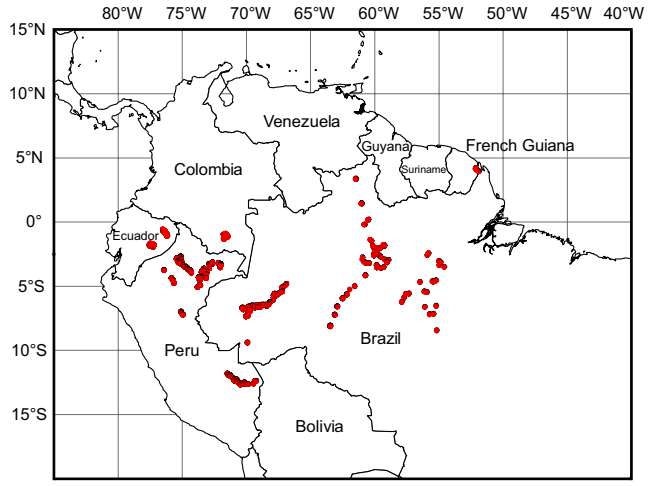

Figure 1. Geographical distribution of the fern and lycophyte inventory plots.

Environmental datasets

Climatic data

We used the nineteen bioclimatic layers of the CHELSA (climatologies at high resolution for the earth's land surface areas) dataset (Karger et al. 2017) as climatic environmental variables, which are derived from monthly minimum, maximum and mean temperature and mean precipitation values. For each field plot, we extracted the bioclimatic variable values of the 30 arcsec (approximately 1 km) grid cell overlapping the plot midpoint.

Canopy reflectance

We obtained surface reflectance values from a new Amazon-wide Landsat TM/ETM+ image composite (Van doninck and Tuomisto 2018). This product combines all Landsat TM/ETM+ acquisitions with less than 60% cloud cover from the dry-season months July, August and September of the years 2000–2009. The compositing process applied a correction for reflectance anisotropy that was calibrated for tropical forests (Van doninck and Tuomisto 2017a) and a multi-dimensional median compositing that took advantage of the large number of multitemporal observations per pixel (Van doninck and Tuomisto 2017b). Consequently, this image composite obtained an exceptionally high radiometric consistency compared to other products over Amazonia (Van doninck and Tuomisto 2018).

We extracted surface reflectance values of Landsat bands 3 (red), 4 (NIR), 5 (SWIR1) and 7 (SWIR2) for each field plot using a 15 by 15 pixel window centered on the plot midpoint. Bands 1 (blue) and 2 (green) were not used here, because these short wavelengths visibly retained significant residual atmospheric contamination. Non-forested pixels were identified using an unsupervised k-means nearest neighbour clustering with post-classification interpretation and excluded from the analyses. For each field plot, we calculated per-window median and standard deviation surface reflectance of the forested pixels corresponding to that plot.

Environmental variable reduction

Combining the nineteen bioclimatic layers and the eight layers derived from Landsat gave a total of 27 environmental variables, many of which were strongly correlated. We performed variable selection to obtain a smaller set of environmental variables to be used throughout this study. To ensure that these were relevant predictors of fern and lycophyte species occurrence, we first ran a principal coordinates analysis (PCoA) of all the field sites using extended Sørensen dissimilarities (De'ath 1999) and retained the first three ordination axes, accounting for 38, 13 and 8% of the total variation. We then correlated each axes with each environmental variable in turn and ordered the environmental variables according to

Table 1. Selected environmental variables, maximum correlation with any of the floristic PCoA axes (r_{max}), and variance inflation factor (VIF).

Environmental variable	r_{max}	VIF
CHELSA – Bio5 max. temperature of warmest month	0.65	10.62
Landsat – Band 7 median reflectance	0.55	1.94
CHELSA – Bio14 precipitation of driest quarter	0.54	2.43
CHELSA – Bio11 mean temperature of coldest quarter	0.55	7.44
Landsat – Band 5 standard deviation reflectance	0.39	1.32
CHELSA – Bio2 mean diurnal range	0.37	6.26
Landsat – Band 4 median reflectance	0.37	1.79

decreasing maximum correlation with any of the three PCoA axes (r_{max}). The variable with the largest r_{max} was retained, and all environmental variables with a correlation coefficient larger than 0.78 with this variable were removed to avoid covariation of explanatory variables. This process was repeated for the remaining variables, until a minimum threshold of 0.3 was reached. This gave a set of seven environmental variables, including four climatic and three remote sensing variables (Table 1). Taken together, these are strongly correlated with the understory floristic variability across the study sites. Overall, the Landsat variables correlated most strongly with the first PCoA axis and the CHELSA variables with the second PCoA axis. None of the 27 original environmental variables correlated significantly with the third PCoA axis.

Species accumulation curves

To assess how species accumulate in relation to gradients defined by the selected environmental variables, we first selected one plot at random, and then added the plot that had the largest environmental dissimilarity to it (Euclidean distance, with all the variables standardized to zero mean and unit standard deviation). The remaining plots were then added iteratively such that the plot with the largest environmental dissimilarity to any of the already selected plots was added next. Separate species accumulation curves were derived using dissimilarities based on the climatic data (four CHELSA variables) and the reflectance data (three Landsat variables), as well as using entirely random plot selection. This process was repeated 50 times for each of the three species accumulation curves.

Species distribution modelling

In order to evaluate the relative importance of climatic and remote sensing variables for modelling terrestrial ferns and lycophytes, we validated species distribution models (SDMs) based on three sets of environmental data layers: the four selected CHELSA variables only, the three selected Landsat variables only and all seven environmental variables together (Table 1). We used three established techniques: Random Forests (RF) (Breiman 2001), Generalized Linear Models (GLM) and Maximum Entropy (MaxEnt) (Phillips et al. 2006). While RF and GLM are based on presence/absence data, which can be extracted from the field plot data, MaxEnt is based on presence/background data. We extracted background data from across the entire Amazon biome using stratified random sampling in tiles of 2.5 degrees by 2.5 degrees. In each of these tiles, 100 coordinates were randomly selected, and climatic and remote sensing variables were extracted using the same methodology as for the field plots. After removal of non-forested samples, 7571 background points were retained for analysis.

Species distribution modelling was performed in the R programming language (ver. 3.5.0), using the randomForest, stats and maxnet packages for RF, GLM and MaxEnt modelling, respectively. For RF modelling, the number of trees was set to 1000, and the number of variables sampled as candidates at each split was set to the square root of the total number of environmental variables used, rounded up. For GLM, a binomial link function was used. Both GLM and MaxEnt used the linear and quadratic terms of the (normalized) environmental variables.

We evaluated the discriminatory capacity of the different modelling techniques and environmental datasets using 10-fold cross-validation. We split the plot data by longitude rather than at random to obtain geographically separated folds which reduces the effects of spatial autocorrelation. We restricted the analysis to those 69 species that were sufficiently prevalent (present in more than 40 plots in the whole dataset) and sufficiently well distributed spatially (present in more than 20 plots in each of the ten calibration sets).

For each species, we validated the discriminating capacity of the nine combinations of modelling technique and predictor variables using the threshold-independent area under the receiver operating characteristic (ROC) curve (AUC), and the threshold-dependent True Skill Statistic (TSS) (Allouche et al. 2006). The threshold required for calculation of the TSS was set to maximize the sum of specificity and sensitivity, effectively maximizing TSS.

Data deposition

Plot data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.v7fp8ms> (Tuomisto et al. 2019b). Landsat TM/ETM+ composite image available from Fairdata.fi: <http://urn.fi/urn:nbn:fi:att:71ba2590-7112-4669-a4b3-a427c85c7a86>.

Results

Species accumulation curves

At relatively small sample sizes, the most efficient method of finding previously undetected species was to accumulate plots such that the within-sample climatic heterogeneity was maximised (Fig. 2). However, the CHELSA accumulation

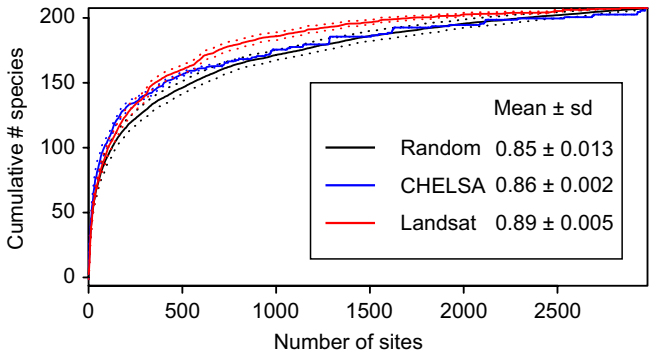

Figure 2. Species accumulation curves for ferns and lycophytes in Amazonia obtained by adding field plots either in a random order or such that the environmental heterogeneity of the selected set of plots is maximised (based on either four CHELSA climate variables or three Landsat reflectance variables). Solid lines indicate mean values obtained with 50 random initial sites, dashed lines indicate standard deviations. Mean and standard deviation of the species accumulation curve integrals (expressed as proportion of total graph area) over the 50 runs are provided in the inset.

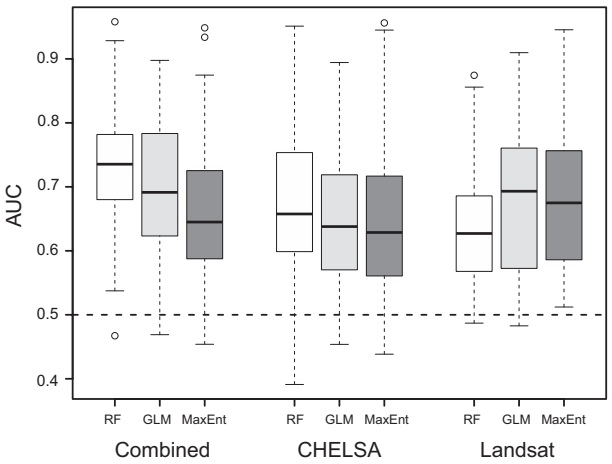

Figure 3. AUC of 69 species distribution models for ferns and lycophytes in Amazonia based on three modelling techniques and using either the combined set of seven environmental predictor variables (listed in Table 1), the four CHELSA climatic variables only, or the three Landsat surface reflectance variables only.

curve only had the highest species richness until approximately 300 field plots had been sampled, at which point the Landsat curve crossed the CHELSA curve. After that, maximising heterogeneity in Landsat reflectance was a superior method for detecting new species. This was especially the case in samples larger than 800 plots, where the number of species added by maximising remaining climatic differences was almost identical to that obtained by random sampling.

The largest gain in cumulative species number when Landsat data were used to inform field plot selection was observed between 750 and 1000 plots, when approximately 13 more species were detected compared to plot selection based on climatic differences, and up to 16 more species compared to plot selection at random. Overall, the integral of the Landsat-based curve was significantly larger ($p < 0.001$ in a paired t-test) than the integrals of the other two curves.

Discriminatory capacity of modelling techniques and environmental datasets

The discriminatory capacity of the SDMs was found to vary considerably among the 69 modelled species, with threshold-independent AUC values ranging from below 0.5 (equivalent to random prediction) to above 0.9. Variation was large for each combination of modelling technique and predictive variables, although some combinations produced generally better models than others (Fig. 3). Validation using the threshold-dependent TSS yielded comparable results (Supplementary material Appendix 1 Fig. A1).

The highest average AUC was obtained in a Random Forests model based on the combined set of four climatic and three remote sensing variables (Fig. 3). This combination significantly outperformed all other possible combinations ($p < 0.001$ for each paired t-test). When using Random Forests, combining climatic and remote sensing variables clearly improved the discriminatory capacity of the SDMs compared to using CHELSA and Landsat variables separately. This was not the case for SDMs based on GLM or MaxEnt, in which the median AUC over all species was lower for the combined models than for the corresponding Landsat-only models.

Interestingly, there was a significant positive correlation between the AUC of a CHELSA-only RF SDM and the soil tolerance of the corresponding species ($r = 0.27$; Fig. 4).

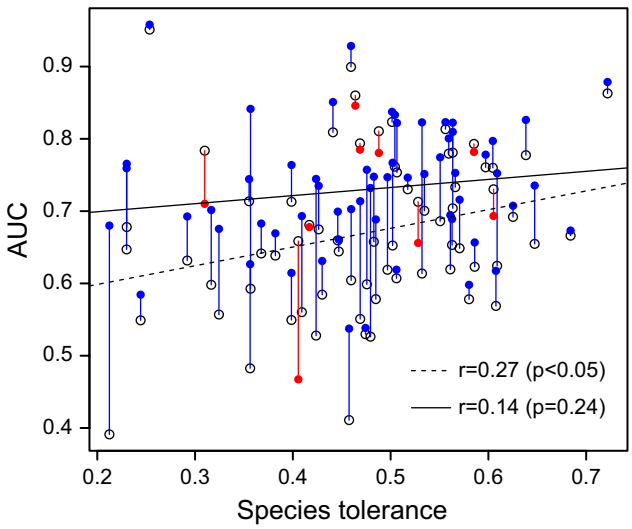

Figure 4. AUC of 69 species distribution models for ferns and lycophytes in Amazonia plotted against the corresponding species tolerance along a soil base cation concentration gradient. All SDMs used Random Forests modelling and either only the four CHELSA variables (open circles) or both these and three Landsat variables (filled circles). Blue circles indicate an increase in discriminatory capacity of the model when Landsat variables were added, red circles a decrease. Dashed and solid lines correspond to open and filled circles, respectively.

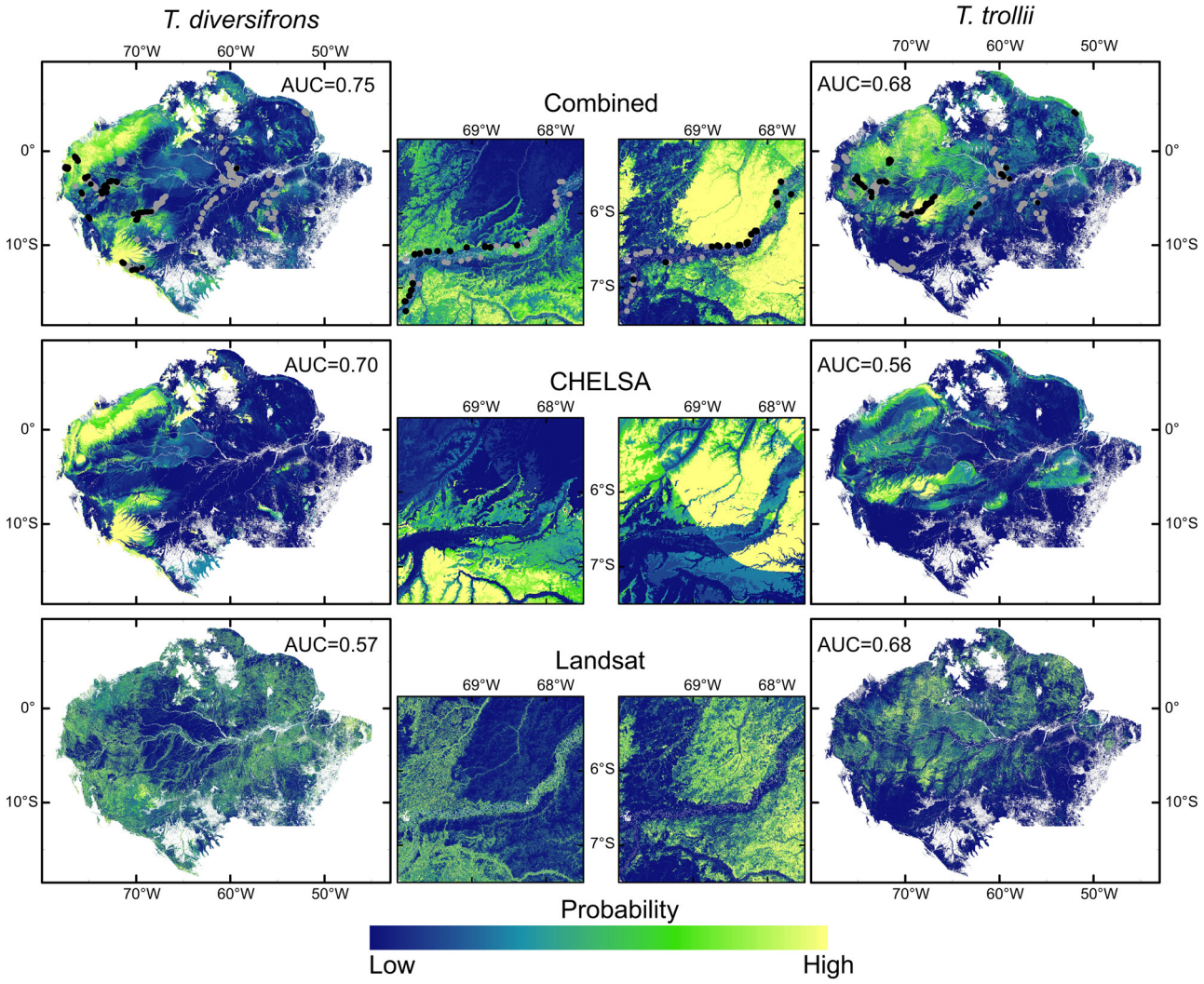

Figure 5. Predicted probability of occurrence of *Trichomanes diversifrons* (left columns) and *T. trollii* (right columns) in Amazonia (outer columns) and in an area along the middle Juruá (inner columns) when using the Random Forests method with the predictor variables consisting of either all seven environmental predictor variables (top row), four CHELSA climate variables only (middle row), or three Landsat surface reflectance variables only (bottom row). Circles indicate quantitative field inventory plots in which the species was observed (black) and was not observed (grey).

When Landsat variables were added to these SDMs, AUC values increased for 60 out of 69 species, and the increase was stronger for species with a narrow tolerance (soil specialists) than for species with broad tolerance (soil generalists). As a result, the AUCs of SDMs based on all seven environmental variables were independent of the degree of species tolerance along the soil base cation concentration gradient.

Comparison of SDM maps produced with the different explanatory variable combinations illustrates the added value of remotely sensed surface reflectance data when predicting understory plant species distributions. As an example, we show RF-predicted probabilities of species occurrence for two closely related terrestrial fern species with contrasting edaphic preferences: *Trichomanes diversifrons* and *T. trollii*. Both species occurred in similar climates and often co-occurred within a region, but *T. diversifrons* was found on soils with a relatively high base cation concentration (optimum = 1.06 cmol(+)/kg) whereas *T. trollii* grew on soils with relatively low soil base cation concentration (optimum = 0.07 cmol(+)/kg). At the biome extent (Fig. 5), the general pattern of predictions of the climate-only model for both species roughly corresponded to observed presence and absence records, obtaining an AUC of 0.70 for *T. diversifrons*. When zooming in, it becomes clear that at the regional extent (about 300 km) the climate-only models created artefacts related to the low thematic resolution of the climatic data in this climatically relatively uniform area. Field-observed species occurrences were consistently structured in relation to a geological and edaphic boundary that runs roughly perpendicularly across the Juruá river (Higgins et al. 2011, Tuomisto et al. 2016), and SDMs that included Landsat layers were able to more accurately predict species distributional patterns in this region (Fig. 5).

Among-species differences in the predictive capabilities of SDMs were very large, even when the same modelling method and explanatory variables were used. However, the reasons for these differences are not clear, as model performance was independent of both species prevalence (the number of plots in which the species was observed) and of the abundance of the species at plots where it was present (Supplementary material Appendix 1 Fig. A2, A3).

Discussion

Predictive power of multispectral and climatic layers

Our results show that canopy surface reflectance obtained by multispectral sensors can provide studies of tropical ecology, as exemplified by SDMs, much higher thematic (taxonomic) detail than is generally assumed. Furthermore, multispectral datasets complement the traditionally used climatic layers in analyses requiring information on environmental site conditions. Here we demonstrate the utility of remote sensing data for biogeographical studies that can aid conservation planning and biodiversity management.

Medium spatial resolution multispectral sensors are already routinely used for the monitoring of dynamic processes in Amazonia related to deforestation and forest disturbances (Hansen et al. 2013). Several studies have documented that these datasets are also useful for identifying and mapping patterns in species composition and species turnover in Amazonia (Tuomisto et al. 2003a, b, Salovaara et al. 2005, Higgins et al. 2011, 2012) and tropical forests in general (Helmer et al. 2015) at local to regional extents. However, until recently (Tuomisto et al. 2019a), no studies have reported the use of medium resolution multispectral data to model Amazonian species distributions or diversity at the basin-wide extent. This can be mainly attributed to challenges that emerge when the spectral similarity among tropical forest types is combined with artefacts introduced by illumination and viewing geometry, clouds and other atmospheric contamination, and image compositing (Toivonen et al. 2006, Helmer et al. 2015, Muro et al. 2016). As a result, multispectral data have sometimes been deemed inadequate for these types of applications (Nagendra and Rocchini 2008, Lausch et al. 2016).

Using a newly developed Landsat TM/ETM+ image composite for Amazonia (Van doninck and Tuomisto 2018), we here provide evidence that multispectral data are ecologically relevant for thematically detailed studies at the basin-wide extent and effectively complement climatic data. The ordination analyses showed that the main floristic gradient (PCoA axis 1) in our study was most strongly related to Landsat variables and the secondary gradient (PCoA axis 2) to climatic variables (Table 1, Tuomisto et al. 2019a). The information value of Landsat data was also clearly evident from the species accumulation curves (Fig. 2). Adding sites based on maximum climatic distance resulted in the highest initial rate of species accumulation, but this was soon overtaken by accumulation based on Landsat dissimilarities as the number of plots increased. The species accumulation curve based on Landsat data stayed above the random curve throughout, indicating that Landsat data continued to consistently identify compositional dissimilarities as survey sites were added (Tuomisto et al. 2003a, Rocchini et al. 2005). This is useful when planning new field surveys: spectral dissimilarity to previously sampled sites can be used to estimate the likelihood of finding new species or forest types. Finally, species distribution models that included Landsat-derived data layers as predictor variables had a better discriminatory ability than climate-only models (Fig. 3), again supporting the view that Landsat reflectance indicates environmental variation of ecological relevance for understory ferns and lycophytes.

Environmental data layers derived from multispectral sensors can be expected to have the greatest added value to climatic data in SDMs when they provide distinct information, in areas of broadly similar climate where features other than climate alter vegetation characteristics (Bradley et al. 2012). This was clearly the case in our study over lowland Amazonia, where the climatic gradient is relatively short and meteorological stations are sparse, which limits the quality of CHELSA and other climatic data layers. The limitations of using climate-only data in tropical ecology studies were apparent in the species accumulation curves, which were indistinguishable from random curves when large numbers of plots were sampled (Fig. 2), and from the overall lower discriminatory capacities of SDMs based on climatic data alone (Fig. 3, Supplementary material Appendix 1 Fig. A1).

Translating environmental predictor variables derived from multispectral imagery into meaningful ecological entities is challenging (He et al. 2015), especially compared to the more straightforward interpretation of bioclimatic variables. The multispectral variables used in this study (Table 1) are influenced by canopy properties such as tree species composition, leaf chemistry, physiology and branching structure. While it is possible that some of these features directly impact understory habitat suitability, e.g. through light availability, it is more likely that these canopy properties are themselves affected by the same environmental variables that the understory plants react to, such as soil base cation concentration, drainage and climate (Tuomisto et al. 2003a, Ruokolainen et al. 2007). In Amazonia, where soil properties vary more than climate does, the Landsat TM/ETM+ surface reflectance mostly serves as a proxy for soil properties even at the basin-wide extent (Van doninck and Tuomisto 2018, Tuomisto et al. 2019a). This was most noticeable for the modelling technique with the largest overall discriminatory capacity (RF), where adding surface reflectance data to climate-only models improved model performance especially for soil specialist species (Fig. 4). Additionally, examples for two species with similar climatic but distinct edaphic niches showed more realistic predictions of species distribution when adding remote sensing layers to climate-only models, especially in an area encompassing a known geological boundary (Fig. 5).

Variability in model performance among species

We observed high variability in the discriminatory capacity of SDMs both among the three modelling techniques and among the 69 fern and lycophyte species (Fig. 3). Among the modelling techniques, MaxEnt resulted in the lowest overall AUCs. A possible reason for this is that background samples required for calibrating the MaxEnt models were extracted over the entire Amazon biome, and potentially included reference samples that were environmentally more distant from the validation samples than those used in the models based on observed species presences and absences. Our study did not take advantage of one of the main strengths of MaxEnt: the possibility to use presence-only data. Calibration datasets for MaxEnt could be easily extended using observations registered in herbaria or online repositories. However, this would have required significant additional taxonomic standardization efforts.

Even for the combination of modelling technique and environmental datasets that was best overall, AUC values varied strongly for the different species (Fig. 3). Using AUC to validate SDMs has been criticised on the grounds that AUC should not be used to compare modelling performance among species that differ in their relative area of occurrence (Lobo et al. 2008). By design of the metric, rare species are expected to have higher AUC values. Such a pattern was not, however, observed in our analyses (Supplementary material Appendix 1 Fig. A2), and results using TSS instead of AUC as a metric of model performance were similar (Supplementary material Appendix 1 Fig. A1, A3).

An important issue to consider in species distribution modelling is the uncertain nature of recorded absences, which may be due to species' rarity in spite of favourable conditions (Lobo et al. 2010). Uncertainty of absences can be expected to be an issue in our dataset, given the high floristic diversity of Amazonian forests and the use of relatively small plots of 400–500 m^2. Earlier studies on ferns and lycophytes have used a sampling unit size of 2500 m^2, as the species accumulation by that time has generally slowed down (Tuomisto and Poulsen 1996). Species that are abundant when environmental conditions are suitable can be expected to be less prone to false absences, leading back to the expectation that AUC should be influenced by species' abundance (Lobo et al. 2008). Again, however, our results did not show such an effect (Supplementary material Appendix 1 Fig. A2, A3).

In order to highlight the added value of multispectral remote sensing in tropical biodiversity modelling, all modelling in this study was based on a limited and fixed set of four climatic variables from the CHELSA dataset and three spectral layers from Landsat. While this set of variables described well the general compositional trends of fern and lycophyte species across lowland Amazonia (Table 1), they are not necessarily the most relevant predictive variables for each individual species. Including species-specific climatic and spectral variables, and additional relevant environmental layers might further improve predictive model accuracies. Moreover, the improved radiometric and spectral resolution of newer generation multispectral sensors such as Landsat OLI and Sentinel 2 can be expected to make these even more suitable for these types of applications, than Landsat TM/ETM+.

Conclusions

We found multispectral satellite data improved basin-wide distribution models of terrestrial fern and lycophyte species, and predict that they will be found informative for many other groups of tropical forest taxa as well. Obviously, this goes for the canopy species themselves, but also for any plant or animal group whose distribution is directly or indirectly related to canopy structure and species composition.

Acknowledgements – Numerous people have contributed to the data by participating in field work, helping with practical arrangements, or sharing their expertise for species identification or soil analysis. This work also contributes to the technical series of the biological dynamics of forest fragments project.
Funding – We made use of computing resources provided by the open geospatial information infrastructure for research (oGIIR, urn:nbn:fi:research-infras-2016072513) funded by the Academy of Finland, and by CSC – IT Center for Science, Finland. Funding has been provided by numerous agencies, most recently by the Academy of Finland (grants 139959 and 273737 to HT, grant 296406 to Risto Kalliola), Finnish Cultural Foundation (grant to GZ) and several grants associated with the Brazilian Program in Biodiversity Research (PPBio) from Brazilian agencies like FAPEAM and CNPq.
Author contributions – HT, JVD and MJ conceived the ideas and designed methodology; HT, GZ, KR, GMM, AS, GC and SL collected the data; JVD, HT, MJ and GZ analysed the data; JVD led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.
Permits – We thank the several national authorities for granting the permits to carry out field work and collect voucher specimens.

References

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Baldeck, C. A. et al. 2016. Environmental drivers of tree community turnover in western Amazonian forests. – Ecography 39: 1089–1099.

Bradley, B. A. et al. 2012. Species detection vs. habitat suitability: are we biasing habitat suitability models with remotely sensed data? – Ecol. Model. 244: 57–64.

Breiman, L. 2001. Random Forests. – Mach. Learn. 45: 5–32.

Buermann, W. et al. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. – J. Biogeogr. 35: 1160–1176.

Cámara-Leret, R. et al. 2017. Modelling responses of western Amazonian palms to soil nutrients. – J. Ecol. 105: 367–381.

Carneiro, L. R. d. A. et al. 2016. Limitations to the use of species-distribution models for environmental-impact assessments in the Amazon. – PLoS One 11: e0146543.

De'ath, G. 1999. Extended dissimilarity: a method of robust estimation of ecological distances from high beta diversity data. – Plant Ecol. 144: 191–199.
Duque, A. J. et al. 2005. Ferns and Melastomataceae as indicators of vascular plant composition in rain forests of Colombian Amazonia. – Plant Ecol. 178: 1–13.
Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.
Figueiredo, F. O. G. et al. 2018. Beyond climate control on species range: the importance of soil data to predict distribution of Amazonian plant species. – J. Biogeogr. 45: 190–200.
Figueiredo, S. M. d. M. et al. 2015. Predicting the distribution of forest tree species using topographic variables and vegetation index in eastern Acre, Brazil. – Acta Amazon. 45: 167–174.
Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.
Hansen, M. C. et al. 2013. High-resolution global maps of 21st-century forest cover change. – Science 342: 850–853.
He, K. S. et al. 2015. Will remote sensing shape the next generation of species distribution models? – Remote Sens. Ecol. Conserv. 1: 4–18.
Helmer, E. et al. 2015. Characterizing tropical forests with multispectral imagery. – In: Thenkabail, P. S. (ed.), Land resources: monitoring, modeling and mapping. Remote sensing handbook. Vol. 2. CRC press, pp. 367–396.
Hengl, T. et al. 2017. Soilgrids250m: global gridded soil information based on machine learning. – PLoS One 12: e0169748.
Higgins, M. A. et al. 2011. Geological control of floristic composition in Amazonian forests. – J. Biogeogr. 38: 2136–2149.
Higgins, M. A. et al. 2012. Use of Landsat and SRTM data to detect broad-scale biodiversity patterns in northwestern Amazonia. – Remote Sens. 4: 2401–2418.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Karger, D. N. et al. 2017. Climatologies at high resolution for the earth's land surface areas. – Sci. Data 4: 170122.
Lausch, A. et al. 2016. Linking Earth observation and taxonomic, structural and functional biodiversity: local to ecosystem perspectives. – Ecol. Indic. 70: 317–339.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.
Mackey, B. G. and Lindenmayer, D. B. 2001. Towards a hierarchical framework for modelling the spatial distribution of animals. – J. Biogeogr. 28: 1147–1166.
Moulatlet, G. M. et al. 2017. Using digital soil maps to infer edaphic affinities of plant species in Amazonia: problems and prospects. – Ecol. Evol. 7: 8463–8477.
Muro, J. et al. 2016. Floristic composition and across-track reflectance gradient in Landsat images over Amazonian forests. – ISPRS J. Photogram. Remote Sens. 119: 361–372.
Nagendra, H. and Rocchini, D. 2008. High resolution satellite imagery for tropical biodiversity studies: the devil is in the detail. – Biodivers. Conserv. 17: 3431–3442.
Pérez Chaves, P. et al. 2018. Using remote sensing to model tree species distribution in Peruvian lowland Amazonia. – Biotropica 50: 758–767.
Phillips, O. L. et al. 2003. Habitat association among amazonian tree species: a landscape-scale approach. – J. Ecol. 91: 757–775.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Rocchini, D. et al. 2005. Maximizing plant species inventory efficiency by means of remotely sensed spectral distances. – Global Ecol. Biogeogr. 14: 431–437.
Rocchini, D. et al. 2016. Satellite remote sensing to monitor species diversity: potential and pitfalls. – Remote Sens. Ecol. Conserv. 2: 25–36.
Ruokolainen, K. et al. 2007. Are floristic and edaphic patterns in Amazonian rain forests congruent for trees, pteridophytes and Melastomataceae? – J. Trop Ecol. 23: 13–25.
Rylands, A. B. 1990. Priority areas for conservation in the Amazon. – Trends Ecol. Evol. 5: 240–241.
Saatchi, S. et al. 2008. Modeling distribution of Amazonian tree species and diversity using remote sensing measurements. – Remote Sens. Environ. 112: 2000–2017.
Salovaara, K. J. et al. 2005. Classification of Amazonian primary rain forest vegetation using Landsat ETM+ satellite imagery. – Remote Sens. Environ. 97: 39–51.
Sirén, A. et al. 2013. Mapping environmental variation in lowland amazonian rainforests using remote sensing and floristic data. – Int. J. Remote Sens. 34: 1561–1575.
ter Braak, C. F. and van Dam, H. 1989. Inferring ph from diatoms: a comparison of old and new calibration methods. – Hydrobiologia 178: 209–223.
Thessler, S. et al. 2005. Mapping gradual landscape-scale floristic changes in Amazonian primary rain forests by combining ordination and remote sensing. – Global Ecol. Biogeogr. 14: 315–325.
Toivonen, T. et al. 2006. Across-path DN gradient in Landsat TM imagery of Amazonian forests: a challenge for image interpretation and mosaicing. – Remote Sens. Environ. 100: 550–562.
Tuomisto, H. and Poulsen, A. D. 1996. Influence of edaphic specialization on pteridophyte distribution in neotropical rain forests. – J. Biogeogr. 23: 283–293.
Tuomisto, H. et al. 1995. Dissecting Amazonian biodiversity. – Science 269: 63–66.
Tuomisto, H. et al. 2003a. Linking floristic patterns with soil heterogeneity and satellite imagery in Ecuadorian Amazonia. – Ecol. Appl. 13: 352–371.
Tuomisto, H. et al. 2003b. Floristic patterns along a 43-km long transect in an Amazonian rain forest. – J. Ecol. 91: 743–756.
Tuomisto, H. et al. 2016. A compositional turnover zone of biogeographical magnitude within lowland Amazonia. – J. Biogeogr. 43: 2400–2411.
Tuomisto, H. et al. 2019a. Discovering floristic gradients and biogeographic subdivisions across amazonia. – J. Biogeogr. 46: 1734–1748.
Tuomisto, H. et al. 2019b. Data from: Discovering floristic and geoecological gradients across Amazonia. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.v7fp8ms>.
Van doninck, J. and Tuomisto, H. 2017a. Evaluation of directional normalization methods for Landsat TM/ETM+ over primary Amazonian lowland forests. – Int. J. Appl. Earth Observ. Geoinform. 58: 249–263.
Van doninck, J. and Tuomisto, H. 2017b. Influence of compositing criterion and data availability on pixel-based Landsat

TM/ETM+ image compositing over Amazonian Forests. – IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 10: 857–867.
Van doninck, J. and Tuomisto, H. 2018. A Landsat composite covering all Amazonia for applications in ecology and conservation. – Remote Sens. Ecol. Conserv. 4: 197–210.
Velazco, S. J. E. et al. 2017. Using worldwide edaphic data to model plant species niches: an assessment at a continental extent. – PLoS One 12: e0186025.

Supplementary material (available online as Appendix ecog-04729 at <www.ecography.org/appendix/ecog-04729>). Appendix 1.

Zuquim, G. et al. 2014. Predicting environmental gradients with fern species composition in Brazilian Amazonia. – J. Veg. Sci. 25: 1195–1207.
Zuquim, G. et al. 2019a. The importance of soils in predicting the future of plant habitat suitability in a tropical forest. – Plant Soil, doi: 10.1007/s11104-018-03915-9.
Zuquim, G. et al. 2019b. Making the most of scarce data: mapping soil gradients in data-poor areas using species occurrence records. – Methods Ecol. Evol. 10: 788–801.