International Journal of Applied Earth Observation and Geoinformation 148 (2026) 105226

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

Mapping tree species diversity across the Amazon using remote sensing, diverse environmental data and machine learning

Shoyo Nakamura [a,*], Nandika Tsendbazar [b], Hans ter Steege [c,d], Lars Hein [a], Jannik Schultner [a]

[a] Earth Systems and Global Change Group, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands
[b] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands
[c] Naturalis Biodiversity Center, PO Box 9517, 2300 RA Leiden, the Netherlands
[d] Quantitative Biodiversity Dynamics, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Tree species diversity
Amazon rainforest
Sentinel-1
Sentinel-2
Machine learning
Environmental drivers

ABSTRACT

Tropical forests are biodiversity hotspots facing increasing threats from climate change and anthropogenic pressures. Tree species diversity is a key indicator of forest biodiversity, making accurate and spatially detailed information essential for effective monitoring and conservation. For large-scale tree species diversity mapping, satellite data and machine learning offer great potential, but challenges remain, including limited field data, coarse spatial resolution, and complex relationships between in-situ diversity and remotely sensed metrics. This study investigates the potential of integrating remote sensing, environmental data and machine learning for tree species diversity mapping across the Amazon. Specifically, we 1) compared performances of widely used algorithms (random forest, extreme gradient boosting, artificial neural networks, support vector regression and ordinary least squares), 2) identified key environmental predictors, and 3) produced 1-km resolution maps to assess their spatial patterns. Our results show that extreme gradient boosting outperformed other algorithms. Climate and soil emerged as the most influential drivers of broad-scale diversity patterns, while remote sensing metrics were also influential in adding fine-scale spatial patterns. High tree diversity was associated most with climate stability, high precipitation, soil characteristics, and spatial heterogeneity of vegetation indices. The resulting maps aligned well with field observations and prior studies, while showing the potential of fine-scale mapping with competitive accuracy and low uncertainty using remotely sensed data. These findings thus demonstrate the effectiveness of integrating remote sensing, environmental data and machine learning for tree species diversity mapping in tropical forests, in support of biodiversity assessments and conservation measures.

1. Introduction

The era of the sixth mass extinction urgently demands effective biodiversity monitoring and conservation (Barnosky et al., 2011). In response, global efforts have emerged to assess and reverse biodiversity loss, as reflected in the development of Essential Biodiversity Variables as a standardized set of biological measurements and the Kunming-Montreal Global Biodiversity Framework as policy targets and strategies across scales (Skidmore et al., 2021, CBD, 2022). Tropical forests represent one of the most critical regions for biodiversity conservation, being a major reservoir of the planet's biodiversity within less than 10% of the global land surface (Giam, 2017, Pinon et al., 2024, De Lima et al., 2020, Liang et al., 2022). Within these ecosystems, tree species diversity serves as a key biodiversity indicator and plays a crucial role in guiding conservation efforts and sustainable forest management (Fagua et al., 2021). Trees not only provide essential resources and habitats for the majority of known terrestrial species, but also support various ecosystem services (Ampoorter et al., 2020, De Lima et al., 2020). However, tree species diversity in tropical forests is severely threatened due to climate change and anthropogenic pressures (Gomes et al., 2019). The challenges underscore the urgent need for developing effective monitoring approaches to track the status of tree diversity, particularly through spatial quantification and mapping, and to understand its underlying drivers.

* Corresponding author.
E-mail addresses: shoyo.nakamura@wur.nl (S. Nakamura), nandin.tsendbazar@wur.nl (N. Tsendbazar), hans.tersteege@naturalis.nl (H. ter Steege), lars.hein@wur.nl (L. Hein), jannik.schultner@wur.nl (J. Schultner).

https://doi.org/10.1016/j.jag.2026.105226
Received 24 November 2025; Received in revised form 4 February 2026; Accepted 1 March 2026
Available online 6 March 2026
1569-8432/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

To assess tree species diversity, field surveys have been common, but they are costly, labor-intensive, and time-consuming (Kacic and Kuenzer, 2022). A promising alternative is the use of remote sensing due to its spatial coverage, cost-effectiveness, and vegetation-relevant spectral information (Kacic and Kuenzer, 2022). A common approach based on remote sensing involves modeling individual species distributions, which are then used to evaluate species diversity (Grabska et al., 2020, Liu et al., 2023). However, this method requires a comprehensive or representative species list, reliable and sufficient field data for each species, and robust modeling techniques, which are often difficult to meet (Andermann et al., 2022); as a result, many studies focus on common, dominant or well-modelled species (Andermann et al., 2022, Grabska et al., 2020). Another method is the direct prediction of diversity indices from remotely sensed data using machine learning, which can learn complex patterns (Ming et al., 2024) and allows for model interpretability through explainable AI tools, such as SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). Previous studies have employed various algorithms, including random forest (RF) (Breiman, 2001), extreme gradient boosting (XGB) (Chen and Guestrin, 2016), artificial neural network (ANN), support vector regression (SVR) (Cristianini and Ricci, 2008) and ordinary least squares (OLS), though their effectiveness varies across studies (Fagua et al., 2021, Liu et al., 2023, Bai et al., 2024). Moreover, direct predictions of large-scale tree diversity are still constrained by limited field data and the inherent complexity of linking remotely sensed signals to in-situ diversity (Ming et al., 2024, Hoffmann et al., 2022).

High-resolution satellites, such as Sentinel-2 and Sentinel-1, are especially suitable for large-scale tree diversity mapping, given the practical balance of resolution, spatial coverage, and cost (Liu et al., 2023). Capturing in-situ tree diversity using Sentinel-1/Sentinel-2 data involves the computation of relevant features (Liu et al., 2023). Vegetation indices (VIs) are commonly used to represent vegetation characteristics such as health, water and nutrient stress, and biomass, which have been linked to tree diversity, but studies vary on which VIs are most relevant (Hoffmann et al., 2022, Liu et al., 2023, Ming et al., 2024), and their effectiveness in tropical rainforest contexts remains unclear. Another method involves temporal variability from multi-temporal satellite data within a certain time period to capture different seasonal dynamics of different species. For instance, Liu et al. (2023) demonstrated the utility of such metrics from both Sentinel-1 and Sentinel-2, though they remain underexplored. Additionally, spatial variability associated with the spectral variability hypothesis has shown relevance in explaining in-situ tree diversity (Fassnacht et al., 2022). Spectral variability is often driven by functional and morphological differences among tree species, and the hypothesis thus assumes that greater spectral heterogeneity implies higher plant species diversity (Fassnacht et al., 2022). However, this relationship is inconsistent and the direct prediction of tree diversity remains limited (Hoffmann et al., 2022), especially in heterogeneous landscapes and rugged terrains (Xi et al., 2023). Moreover, few studies have jointly applied VIs, temporal metrics, and spectral heterogeneity in a unified framework.

Beyond remote sensing, integrating complementary environmental data has shown considerable promise since environmental factors play a critical role in shaping tree diversity (Liang et al., 2022). For example, climate affects tree mortality, growth and ecological interactions; topographic conditions shape environmental gradients, which facilitate niche partitioning; soil determines belowground conditions such as nutrient and water availability; and human activities have reshaped ecosystems through intensive natural resource use (Liang et al., 2022). Beyond boosting performance, combining remotely sensed vegetation data with biophysical environmental datasets can offer deeper insights into the ecological drivers of tree diversity (Xi et al., 2023). In the tropics, these drivers interact in particularly complex ways, and it remains unclear to what extent each factor explains in-situ tree diversity (Liang et al., 2022).

Among tropical forests, the Amazon rainforest, a global biodiversity hotspot, harbors approximately 13% of the world's trees and nearly half of all tropical forest species (Crowther et al., 2015). However, Amazonian tree diversity faces severe threats from climate change and deforestation, with alarming projections of substantial biodiversity losses in the near future (Gomes et al., 2019, Flores et al., 2024). In response, efforts have been made to map and understand tree diversity across the region; early predictive maps relied on statistical interpolation of field data at a coarse 1-degree resolution (Ter Steege et al., 2003, Stropp et al., 2009) or simple machine learning approaches (Saatchi et al., 2008, Liang et al., 2022), though limited field data and the quality of predictor variables have constrained the scope of output maps. Statistical interpolation with extensive field data has helped to improve accuracy and spatial resolution (Ter Steege et al., 2023), though their accuracy and spatial resolution could further be improved to identify fine-grained spatial patterns.

This study aimed to explore the potential of remote sensing, environmental data and machine learning for direct prediction and large-scale mapping of tree species diversity across the Amazon. Specifically, integrating 1,389 field inventory plots, Sentinel-1, Sentinel-2 and environmental data for mapping species richness per ha and Fisher's alpha diversity, we aimed to: 1) compare the performances of common machine learning algorithms to identify the best-performing algorithm, and using this algorithm, 2) identify key predictors for better modeling interpretation and ecological understanding and 3) produce 1-km resolution maps to capture spatial patterns and assess the effectiveness of the approach.

2. Materials and methods

We developed a comprehensive workflow (Fig. 1) to effectively predict and map tree species diversity across the Amazon. The workflow consisted of three main steps following an initial preparation of field, remote sensing, and environmental datasets: 1) performance comparison of commonly used machine learning algorithms; 2) identification of key predictor variables and variable groups; and 3) mapping and spatial assessment across forest types and regions. The subsequent sections describe each of these steps in detail, following an overview of the study area. All analyses were conducted in Python, including Google Earth Engine (GEE) module (Gorelick et al., 2017) for data collection and processing and Tensorflow module for model implementation.

2.1. Study area

The study area covered approximately 7 million km^2, corresponding to the Amazonian biogeographic boundary (−79.50°W, −43.40°W, 10.06°N, −18.16°S) (Fig. 2) as defined by the Red Amazonica de Informacion Socioambiental Georreferenciada (RAISG, 2025). This region spans parts of Bolivia, Brazil, Colombia, Ecuador, Peru, French Guiana, Suriname, Guyana, and Venezuela. The forest can be categorized into four major types based on soil characteristics (Ter Steege et al., 2023): white sand forest, floodplain forest, swamp, and terra-firme. Additionally, the study area encompassed six major biogeographic sub-regions (Ter Steege et al., 2023): Guyana Shield, Central Amazonia, Eastern Amazonia, Southern Amazonia, Northwestern Amazonia, and Southwestern Amazonia.

2.2. Field data

We used plot-based tree diversity data from the Amazon Tree Diversity Network published by Ter Steege et al. (2023). Briefly, the dataset contained plot locations, sub-regions, forest types, inventory years (inventoried between 1934 and 2021), Fisher's alpha per plot and species richness per ha for 2,046 plots in undisturbed, old-growth forests. Fisher's alpha is a widely used local diversity metric, considering both species richness and evenness, calculated by iteratively solving the equation $\alpha = S / \ln(1 + N / \alpha)$, where N is the total number of individual

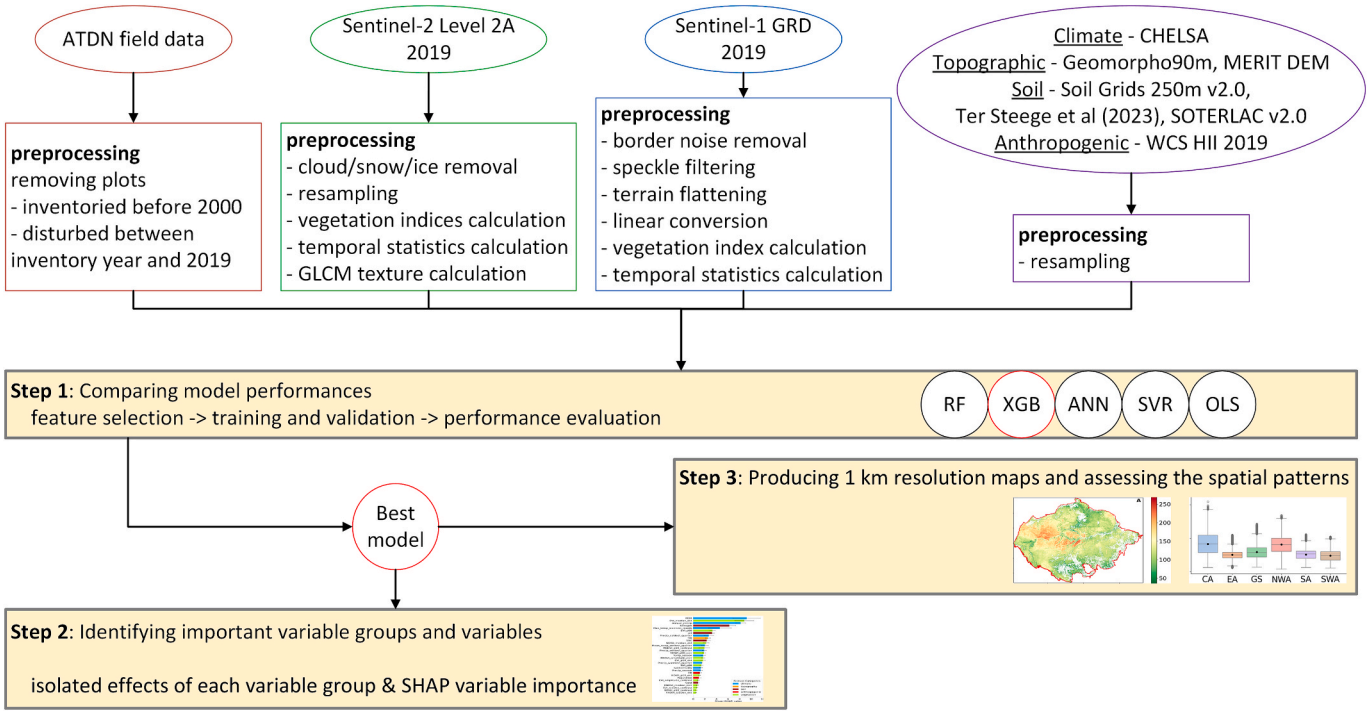

Fig. 1. Overall workflow for tree species diversity mapping approaches incorporating data preparation, model performance comparisons, identification of important variable groups and variables, and 1 km resolution mapping. RF: random forest; XGB: extreme gradient boosting; ANN: artificial neural networks; SVR: support vector regression; OLS: ordinary least squares.

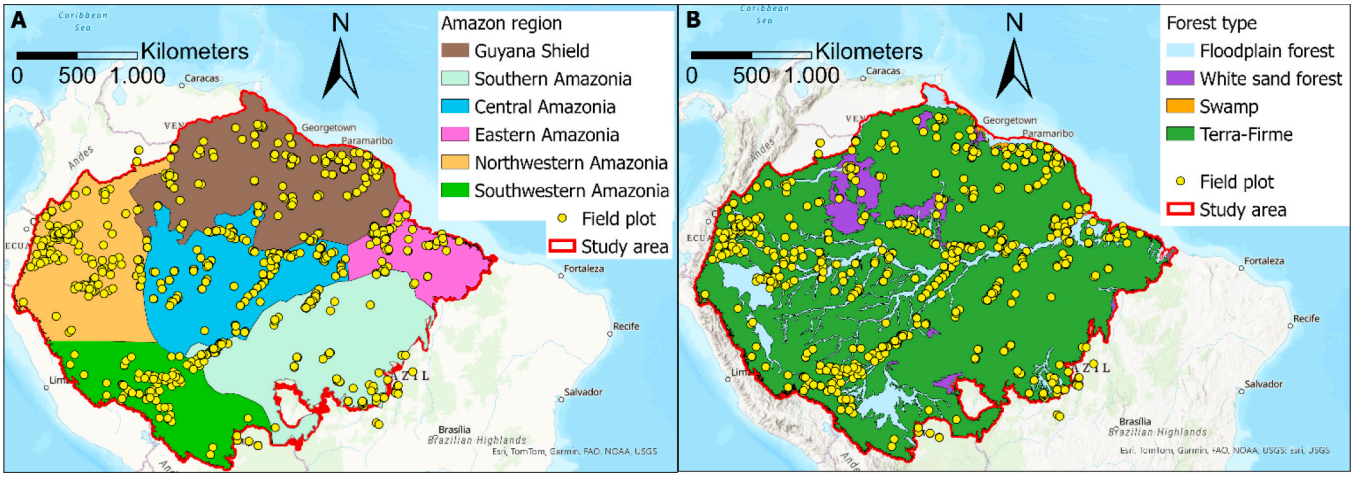

Fig. 2. Study area showing field plot locations and distributions of each sub-region (A) and forest type (B).

stems and S is the number of morpho-species recorded per plot. Species richness indicates the number of species, estimated using the expression $\alpha \times \ln(1 + N_{ha}/\alpha)$, where N_{ha} is the number of stems per ha (Fisher et al., 1943). Trees were defined as free-standing woody individuals with a diameter at breast height of $\geq$ 10 cm. The dataset included trees identified to at least the morpho-species level, from plots ranging in size from 0.5 to 2 ha. From the datasets, we omitted plots inventoried before 2000 for geolocation accuracy and plots disturbed between the inventory years and the year 2019 for temporal alignment with satellite data using the Hansen Global Forest Change v1.11 product (Hansen et al., 2013). After filtering, a total of 1,389 plots remained for analysis, distributed across the Amazon (Fig. 2).

2.3. Remote sensing and environmental data

2.3.1. Sentinel-2 data, preprocessing and feature calculation

We collected and preprocessed Sentinel-2 Level-2A surface reflectance imagery, atmospherically corrected and orthorectified products, on GEE for the year 2019, the earliest year with complete coverage of the study area. We excluded images with $\geq$ 70% cloud coverage and pixels classified as saturated/defective, dark area, cloud shadows, clouds with medium or high probability, cirrus and snow/ice using the Scene Classification outputs from the Sen2Cor algorithm (Louis et al., 2016). We collected a total of 52,954 images for both 10 m bands (Blue, Green, Red and Near-Infrared1 (NIR1)) and 20 m bands (Red-edge-2 and Shortwave infrared2 (SWIR2)), where 20 m bands were resampled to 10 m resolution. Using these bands, we calculated four VIs (Table 1): Enhanced Vegetation Index (EVI) (Huete et al., 2002), Red-edge Normalized

Table 1
List of metric calculations for remote sensing data.

Category	Metrics	Description
VI	EVI	$2.5 \times (NIR1 - Red)/((NIR1 + 6 \times Red - 7.5 \times Blue + 1 + 6 \times Red - 7.5 \times Blue) + 1)$
	RNDVI	$(NIR1 - RedEdge2)/(NIR1 + RedEdge2)$
	NDWI	$(NIR1 - SWIR1)/(NIR1 + SWIR1)$
	RVI	$4 \times VH/(VV + VH)$
Temporal statistics	median	Median value across 2019 images
	stdDev	Standard deviation across 2019 images
	p90	90 percentile value as a proxy of the maximum value across 2019 images
	p10	10 percentile value as a proxy of the minimum value across 2019 images
	amplitude	Difference between p90 and p10 across 2019 images
GLCM texture[a]	Entropy	Randomness in the texture pattern distribution, given by $\sum_{ij=0}^{N=1} p(i,j) log(p(i,j))$
	Contrast	Intensity difference between neighboring pixels, given by $\sum_{ij=0}^{N=1} p(i,j)(i-j)^2$
	Correlation	Similarity of neighboring pixel values, given by $\sum_i \sum_j (i-\mu_i)\left(j-\mu_j\right) p(i,j)/\sigma_i \sigma_j$

[a] N is the number of grey levels; i and j are the row and column indices of the GLCM matrix; p(I,j) is the probability of grey levels i and j occurring as adjacent pixel pairs; σ_i and σ_j are the standard deviations of the marginal probabilities of rows and columns, respectively; μ_i and μ_j are the average intensity for pixels in rows and columns, respectively.

Vegetation Index (RNDVI) (Gitelson and Merzlyak, 1994), and Normalized Difference Water Index (NDWI) (Gao, 1996). EVI is robust against background noise and atmospheric influences, and has been proven to be useful for analysis in dense vegetation (Huete et al., 2002). RNDVI and NDWI are sensitive to key canopy traits, including chlorophyll content and water availability (Xiao et al., 2020, Gao, 1996). These indices are based on red-edge and SWIR bands, which, along with their associated VIs, have demonstrated variable yet significant utility in tree diversity prediction (Liu et al., 2023, Ming et al., 2024). For each VI, we calculated five temporal statistics for the year 2019 (median, standard deviation, 10th percentile, 90th percentile, and amplitude; Table 1) to characterize phenological dynamics (Grabska et al., 2020). Then, from each statistical combination, we computed 9 × 9 Gray-Level-Co-Occurrence-Matrix (GLCM) textures (Haralick et al., 2007) centered on each plot location to capture spatial heterogeneity (Table 1). For that, we used GEE's ee.Image.glcmTexture function to compute the texture metrics across the four standard directions (0, 45, 90 and 135 degrees) and subsequently averaged them. The GLCM textures have three main categories (contrast, orderliness and descriptive statistics) and are known to be highly correlated with each other, especially within the same category (Ozdemir and Karnieli, 2011). To avoid variable cross-correlation, following Ozdemir and Karnieli (2011), we selected three complementary texture features, one from each texture category: contrast from the contrast category, entropy from the orderliness category, and correlation from the descriptive statistics category (Ozdemir and Karnieli, 2011). This resulted in 45 optical remote sensing variables (3 VIs × 5 temporal statistics × 3 GLCM textures; Table S1), all resampled to 1 km resolution.

2.3.2. Sentinel-1 data, preprocessing and feature calculation

We collected Sentinel-1 Ground Range Detected (GRD), dual-polarization (VV and VH) imagery acquired in 2019 in Interferometric Wide swath mode. A total of 10,281 images from both ascending and descending orbits were used. The images have approximately 20 × 22 m spatial resolution and 10 m pixel spacing. The images provided on GEE were preprocessed with orbit file application, thermal noise removal, GRD border noise removal, radiometric calibration to sigma naught, and range-doppler terrain correction. We followed the workflow of Mullissa et al. (2021) to further preprocess the images by applying border noise removal, 3 × 3 refined Lee speckle filtering, and radiometric terrain correction. Using the preprocessed data in linear format, we computed the radar vegetation index (RVI), which has been used for vegetation studies, such as tree species mapping and water content estimation (Schulz et al., 2024, Kim et al., 2011). For RVI, we computed the same five temporal statistics as above, yielding 5 radar-based variables (Table S1), all resampled to 1 km resolution.

2.3.3. Environmental data

We compiled 34 environmental predictor variables encompassing climate, topographic, soil and anthropogenic factors that influence tree diversity by shaping environmental gradients and conditions. A brief description of each variable is provided in Table 2 (see Table S1 for more details). Specifically, we used 1) seven bioclimatic variables from Climatologies at high resolution for the Earth's land surface areas (CHELSA) V2.1 that capture annual means and seasonality (less sensitive to artefacts) at 30 arc-second resolution (Karger et al., 2017, Karger et al., 2021); 2) 13 topographic variables at approximately 90 m resolution directly from Multi-Error-Removed Improved-Terrain (MERIT) digital elevation model (DEM) (Yamazaki et al., 2017) and Geomorpho90m (Amatulli et al., 2020); 3) two soil variables (nitrogen and soil organic carbon content) at 250 m resolution from Soil Grids v2.0 (Poggio et al., 2021), two soil variables (pH and sum of bases as a soil fertility indicator) at 0.1 degree resolution from Ter Steege et al (2023) and three rasterized data of underrepresented forest types based on the soil types (floodplain forest, white sand forest and swamp) at 1 km resolution (Dijkshoorn et al., 2005); and 4) seven anthropogenic variables at 1 km resolution for the 2019 Human Impact Index dataset (Venter et al., 2016). All variables were resampled to 1 km resolution.

2.4. Regression analysis and model comparison

2.4.1. Models

Our study compared the performance of five widely used machine learning algorithms: RF, XGB, ANN, SVR and OLS. *RF*, an ensemble of decision trees, is known for its robustness, scalability and ability to model non-linear relationships (Liang et al., 2022, Ming et al., 2024). *XGB* is a widely used ensemble of decision trees that sequentially correct residual errors from previous trees through gradient boosting (Liu et al., 2023). *ANN*, composed of an input layer, hidden layers and an output layer, can effectively model complex relationships with high-quality data (Hoffmann et al., 2022, Andermann et al., 2022). We used rectified linear unit activation functions in the hidden layers and a linear activation function in the output layer. *SVR* applies kernel functions to model non-linearities in high-dimensional data (Bai et al., 2024). *OLS* is a traditional linear regression, which served as a benchmark to evaluate the relative performance of the more complex models.

2.4.2. Feature selection

Removing redundant and irrelevant variables helps simplify models, reduce computational load and avoid overfitting. We performed a two-step feature selection process. First, we calculated Pearson's cross-

Table 2
List of environmental data (see Table S1 for more details).

Category	Metric examples	Ecological importance (Liang et al., 2022)
Climate	Annual mean temperature	Controls growth, mortality and physiological limits
Topography	Elevation, slope	Shapes environmental gradients and niche partitioning
Soil	pH, sum of bases	Reflects nutrient availability and soil fertility
Anthropogenic	Road, population density	Reflects pressures from human activities

correlation to identify variable pairs with high correlation ($r \geq 0.8$) and removed those of lower Spearman's correlation with in-situ diversity data. Second, we employed recursive feature elimination with three repetitions of 10-fold cross-validation, using each algorithm as the base model. Since ANN cannot be directly used with recursive feature elimination, we used RF as the base model instead, as RF demonstrated higher global accuracy in recursive feature elimination than other algorithms (Niquini et al., 2023). Depending on the algorithm, the number of selected variables ranged from 5 to 53.

2.4.3. Regression analysis and performance evaluation

Using the selected features, we applied 10-fold nested cross-validation for each of the five models to evaluate model performance while mitigating overfitting. In the outer loop, the dataset was split into training and test sets in an 8:2 ratio, stratified by the forest types, to evaluate model performance on unseen data. In the inner loop, 20% of the training set was used as a validation set to tune hyperparameters. Since ANN, SVR and OLS are sensitive to the scale of input variables, all variables were rescaled to the [0, 1] range prior to training these models.

For hyperparameter tuning, we used the Optuna framework (Akiba et al., 2019). Optuna employs Bayesian optimization to efficiently search for the optimal hyperparameter combination by maximizing a predefined objective function, which, in this study, was negative root mean squared error. This approach is efficient and effective compared to traditional methods such as random search (Akiba et al., 2019). The hyperparameter search spaces are summarized in Table 3.

Model performances were evaluated using R2 and root mean squared error (RMSE). Based on these scores, we determined the best-performing algorithm for further analysis. Additionally, Moran's I was calculated on the residuals of model predictions to assess spatial autocorrelation.

2.5. Important variables and variable groups

To interpret the predictions from the best-performing algorithm, we computed SHAP (Lundberg and Lee, 2017) values for each variable and assessed their contributions to model predictions. SHAP provides a unified framework for feature attribution by assigning each feature a unique contribution value for individual predictions.

Additionally, we assessed the isolated predictive power of each variable group by running the best-performing algorithm using only one variable group at a time. This allowed us to assess the relative importance of each group for tree species diversity prediction.

Table 3
Model hyperparameters and their search space. RF: random forest; XGB: extreme gradient boosting; ANN: artificial neural networks; SVR: support vector regression; OLS: ordinary least squares.

Model	Hyperparameter	Search space
RF	Number of trees	50, 100, 150, 200
	Maximum depth of each tree	10, 20, 30, no limit
	Minimum number of samples required to split an internal node	2, 5, 10
XGB	Number of trees	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000
	Maximum depth of each tree	3, 4, 5, 6, 7, 8, 9, 10
	eta: Learning rate	Between 0.01 and 0.3 on a logarithmic scale
	subsample: Fraction of training data used per tree	Between 0.5 and 1.0
	col_sample_bytree: Fraction of features used per tree	Between 0.3 and 1.0
	Minimum sum of instance weights in a child node	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	gamma: Minimum loss reduction required for a further split	Between 0 and 5
ANN	Number of hidden layers	1, 2, 3, 4, 5
	Number of neurons in each hidden layer	32, 64, 96, 128
	Learning rate	Between 0.0001 and 0.01 on a logarithmic scale
	Batch size	32, 64, 128, 256
SVR	C: Regularization parameter controlling flexibility vs. overfitting	Between 0.1 and 100 on a logarithmic scale
	epsilon: Margin where small errors are not penalized	Between 0.01 and 1.0 on a logarithmic scale
	Kernel type	rbf, linear

2.6. Tree species diversity mapping

Using the 10 models from 10-fold cross-validation, we predicted species richness and Fisher's alpha across the Amazon at a 1 km spatial resolution. To restrict predictions to tree-covered areas, we masked out areas except those classified as tree cover and mangroves using the World Cover 2020 (Zanaga et al., 2021). From the resulting ten predictions, we computed the mean and the coefficient of variation. Additionally, to characterize the spatial heterogeneity of tree diversity and assess the effectiveness of the modeling approach, we analyzed the predicted diversity ranges within each forest type and each sub-region. Furthermore, to investigate spatial patterns associated with key variable groups, we produced tree diversity maps using different combinations of only the variable groups identified as important in the isolated predictive power analysis.

3. Results

3.1. Comparison of the model performances

We ran five different models to compare their performances. XGB and RF exhibited higher accuracy (species richness: RMSE = 33 & 35, R^2 = 0.77 & 0.75; Fisher's alpha: RMSE = 23 & 25 and R^2 = 0.75 & 0.71) than ANN, SVR and OLS (Fig. 3). Compared to RF, XGB achieved slightly better mean and median accuracy with smaller variance (Fig. 3). Therefore, we chose XGB as the best-performing algorithm for further analysis.

Both diversity metrics tended to be overestimated in lower-diversity forests and underestimated in higher-diversity areas in terra-firme forests (Fig. 4, Fig. S2). Yet, residuals exhibited minimal spatial autocorrelation (Moran's $I < 0.05$ with 4–5 significant folds), suggesting limited spatial structure in prediction errors. At the forest-type level (Fig. S3), predictions for terra-firme showed high accuracy, whereas the other less dominant forest types exhibited lower predictive performance. Regionally (Fig. S3), Central Amazonia displayed the highest model accuracy, while other sub-regions showed moderate performance.

3.2. Importance of different predictors and predictor groups

Using XGB, we assessed the importance of each variable group by analyzing its isolated effects on model performance (Fig. 5). The model incorporating all variable groups achieved the highest accuracy for both diversity metrics. Among the individual groups, climate, soil and vegetation factors yielded the highest accuracy. In contrast, topographic and anthropogenic variables demonstrated the weakest predictive power.

Regarding individual variable importance by SHAP analysis, both tree diversity metrics were influenced by similar environmental factors (Fig. 6). Key predictors across both biodiversity metrics included temperature, precipitation, spatial and temporal metrics of EVI and other VIs and soil characteristics. Among remote sensing-derived features, contributions came from different VIs (EVI, NDWI, RNDVI, RVI), temporal features (median, p90, p10, amplitude), and textures (entropy, contrast, correlation). Higher species richness was mainly explained by lower variability in temperature and precipitation, higher precipitation, non-underrepresented forest types (floodplain, white sand and swamp), lower sum of bases, and lower texture values of VIs, although a partial positive contribution was observed for higher entropy of EVI_p10. Fisher's alpha showed a similar pattern, being positively linked to higher precipitation, lower variability of temperature and precipitation, lower

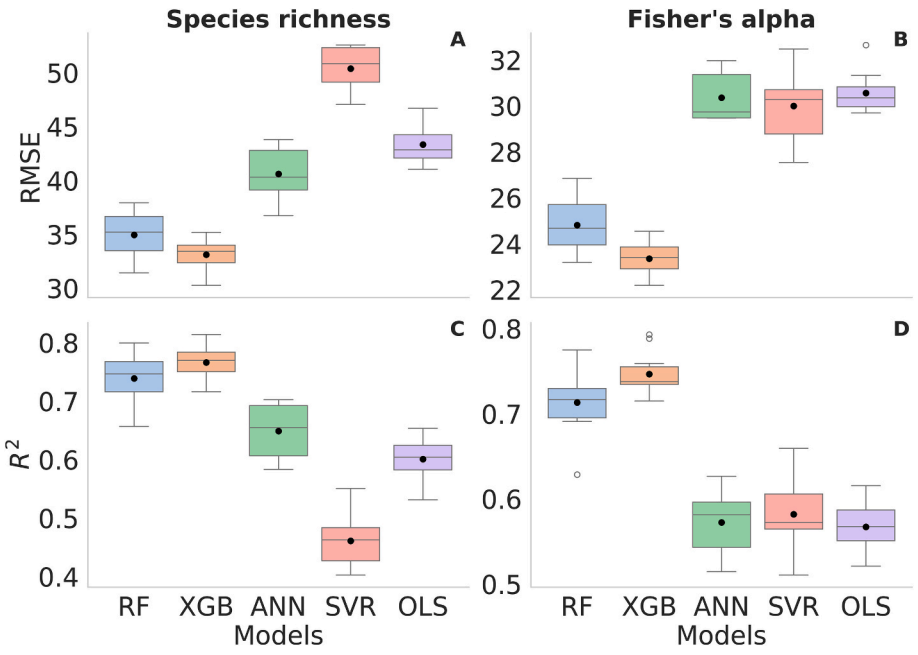

Fig. 3. Performance metric ranges of the five models (RF: random forest; XGB: extreme gradient boosting; ANN: artificial neural networks; SVR: support vector regression; OLS: ordinary least squares) from 10-fold cross-validation in terms of RMSE (A, B) and R^2 (C, D). Black dots are means.

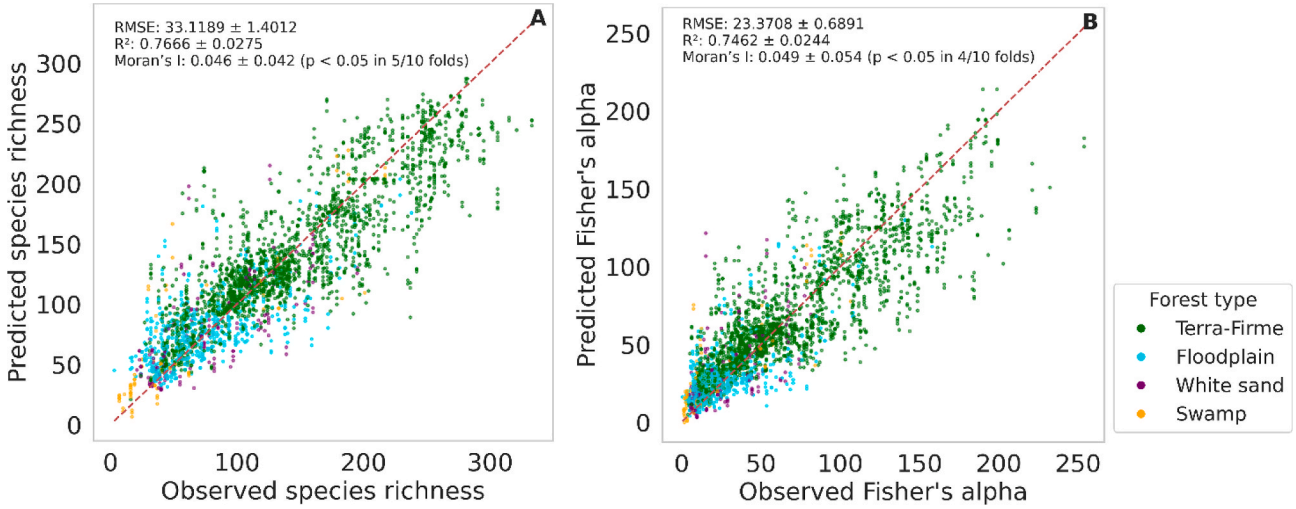

Fig. 4. Scatter plots between observed and predicted values of species richness (A) and Fisher's alpha (B) based on RF. Red, broken lines are 1:1 lines.

pH, non-underrepresented forest types, and lower texture values of VIs, but with a partial positive influence from higher entropy of NDWI_amplitude.

3.3. Spatial patterns of tree species diversity

Using XGB, we generated predictive maps and the coefficient of variation at 1 km spatial resolution (Fig. 7). Predicted species richness and Fisher's alpha showed similar spatial patterns, with the highest tree diversity found in the terra-firme forests in the central to northwestern regions and French Guiana (Figs. 7 & 8). In contrast, lower tree diversity was observed in the sandy forests of the Guiana Shield and Upper Rio Negro, floodplains along rivers, swamps, and southern Amazonia (Figs. 7 & 8). The coefficient of variation of the 10-fold cross-validation and was low across the study area (Fig. 7), indicating low prediction uncertainty by model settings. The residuals were also low across the study area, with little spatial structure (Fig. S4 & S5).

The predictive maps reflected variations in the spatial patterns of species richness and Fisher's alpha across forest types and sub-regions (Fig. S3). Among the forest types, terra-firme exhibited the highest predicted species richness and Fisher's alpha, whereas swamp and floodplain forests showed relatively lower diversity. Regionally, Central Amazonia and Northwestern Amazonia displayed higher species richness and Fisher's alpha than the other sub-regions.

Maps generated using different combinations of important variable groups (i.e. climate, soil and vegetation) further identified the spatial contributions of each group (Fig. S6). Predictions based on climate and/or soil variables displayed smooth spatial gradients, while those based on vegetation variables captured fine-scale patterns. However, predictions using soil or vegetation variables showed noise, such as abrupt spatial gradients from spatial interpolation or swath-edge artifacts. Incorporating additional variable groups reduced such noise, with the combined use of all variable groups producing the best maps.

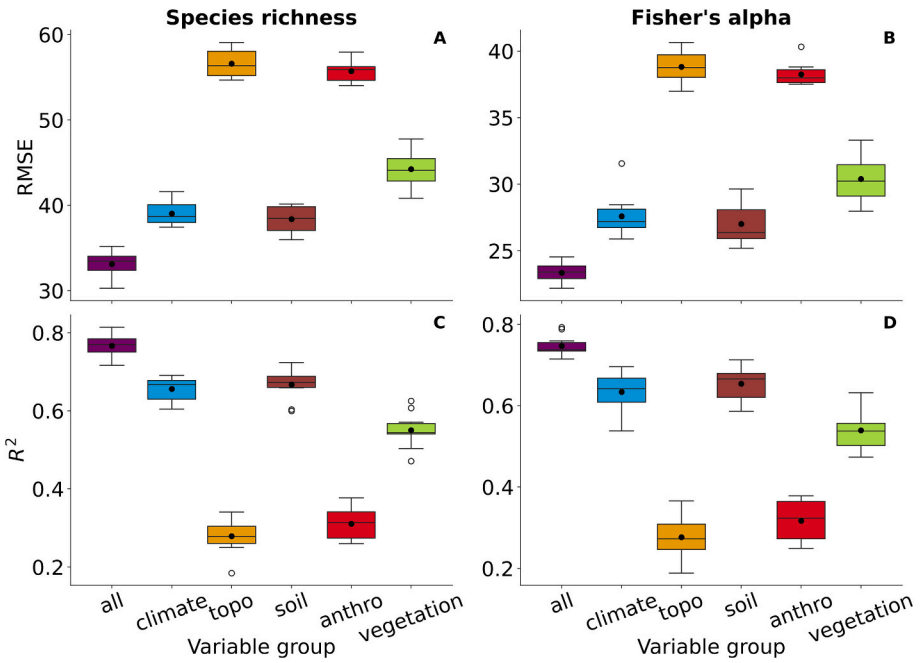

Fig. 5. RMSE (A, B) and R^2 (C, D) from 10-fold cross-validation in six scenarios: using all variables, climate variables only, topographic (topo) variables only, soil variables only, anthropogenic (anthro) variables only, and remote sensing-based vegetation variables only. Black dots are means.

4. Discussion

To explore the potential of remote sensing, environmental data and machine learning for tree species diversity prediction and mapping, this study 1) compared the performances of common machine learning algorithms, and the best-performing algorithm, 2) assessed the importance of different environmental factors and variables, and 3) produced 1 km resolution maps to analyze spatial patterns. Below, we discuss our results and their implications.

4.1. Machine learning prediction of Amazon tree species diversity

Among the tested algorithms, XGB demonstrated the highest accuracy and comparatively small variation (Fig. 3), aligning with several studies highlighting the superior performance of XGB for tree and plant species diversity prediction (Liu et al., 2023, Zhao et al., 2022). XGB benefits from its gradient boosting framework with an ensemble of decision trees and regularization techniques to capture complex relationships, improve generalization and avoid overfitting (Zhao et al., 2022, Chen and Guestrin, 2016). RF attained comparable accuracy to XGB, but slightly lower accuracy scores and higher variability. RF can model complex, non-linear relationships and is robust against overfitting, for which previous studies reported the excellent performance of RF for tree species diversity prediction (Fagua et al., 2021, Ming et al., 2024). In this study, however, the gradient boosting, regularization and flexibility of XGB may have made a difference (Li et al., 2021, Narin, 2025). ANN showed moderate predictive performance, although ANN could outperform other machine learning algorithms by learning complex relationships solely from training data (Xi et al., 2023). Its dependence on large training data with good quality may still limit its utility in data-scarce contexts like this study. SVR performed poorly, despite its ability to model non-linear relationships (Bai et al., 2024), supporting the advantages of tree-based models, such as RF and XGB (Liu et al., 2023, Fagua et al., 2021) over SVR. Finally, the low accuracy of OLS underscores the necessity of capturing complex relationships, justifying the use of non-linear models such as RF.

All models exhibited a consistent bias, also observed in other studies on machine learning-based tree species diversity mapping (Liu et al., 2023, Hoffmann et al., 2022): overestimation in low-diversity areas and underestimation in high-diversity areas (Fig. 4; Fig. S2). The bias may have resulted from the averaging effects of machine learning models, especially when the highest and lowest tree diversity values are underrepresented in the training data or when tree-based ensemble models are used (Belitz and Stackelberg, 2021). Alternatively, the models and predictors may not be able to fully capture the tree diversity variabilities in the highest and lowest tree diversity areas (Hoffmann et al., 2022), especially when they are spatially close. Future improvements for the highest and lowest tree diversity zones could include targeted sampling in underrepresented forest types, potentially with separate analysis by forest types, exploration of more informative predictors that better distinguish these extreme values, or integration of spatial modeling approaches to leverage spatial patterns and structure (Zhao et al., 2022, Stropp et al., 2009, Ter Steege et al., 2023).

4.2. Key environmental predictors

Using XGB, our factorial analysis with different variable groups highlighted the isolated effects of various environmental factors (Fig. 5). Climate, soil and vegetation emerged as the most influential factors, whereas anthropogenic and topographic factors were less important. These findings support earlier research emphasizing climate as a dominant driver of tree diversity in the tropics (Liang et al., 2022) and the Amazon (Ter Steege et al., 2023), given that temperature and precipitation are impactful factors of tree species distribution (Gomes et al., 2019, Toledo et al., 2012). Additionally, the importance of vegetation features supports the utility of satellite data to enhance model performance as observed in Saatchi et al. (2008) and Fagua et al. (2021), although remote sensing-based features alone have limited modeling capacity (Hoffmann et al., 2022). Soil characteristics, related to soil type, nutrient availability, hydrological conditions and acidity, determine fundamental conditions that shape forest types and influence the survival of many tree species (Toledo et al., 2012, Liang et al., 2022, Ter Steege et al., 2023). While anthropogenic factors are generally critical due to widespread human-induced biodiversity loss (Barnosky et al., 2011), their influence was minimal in this study, likely due to the focus on undisturbed forest plots (Ter Steege et al., 2023). Focusing solely on

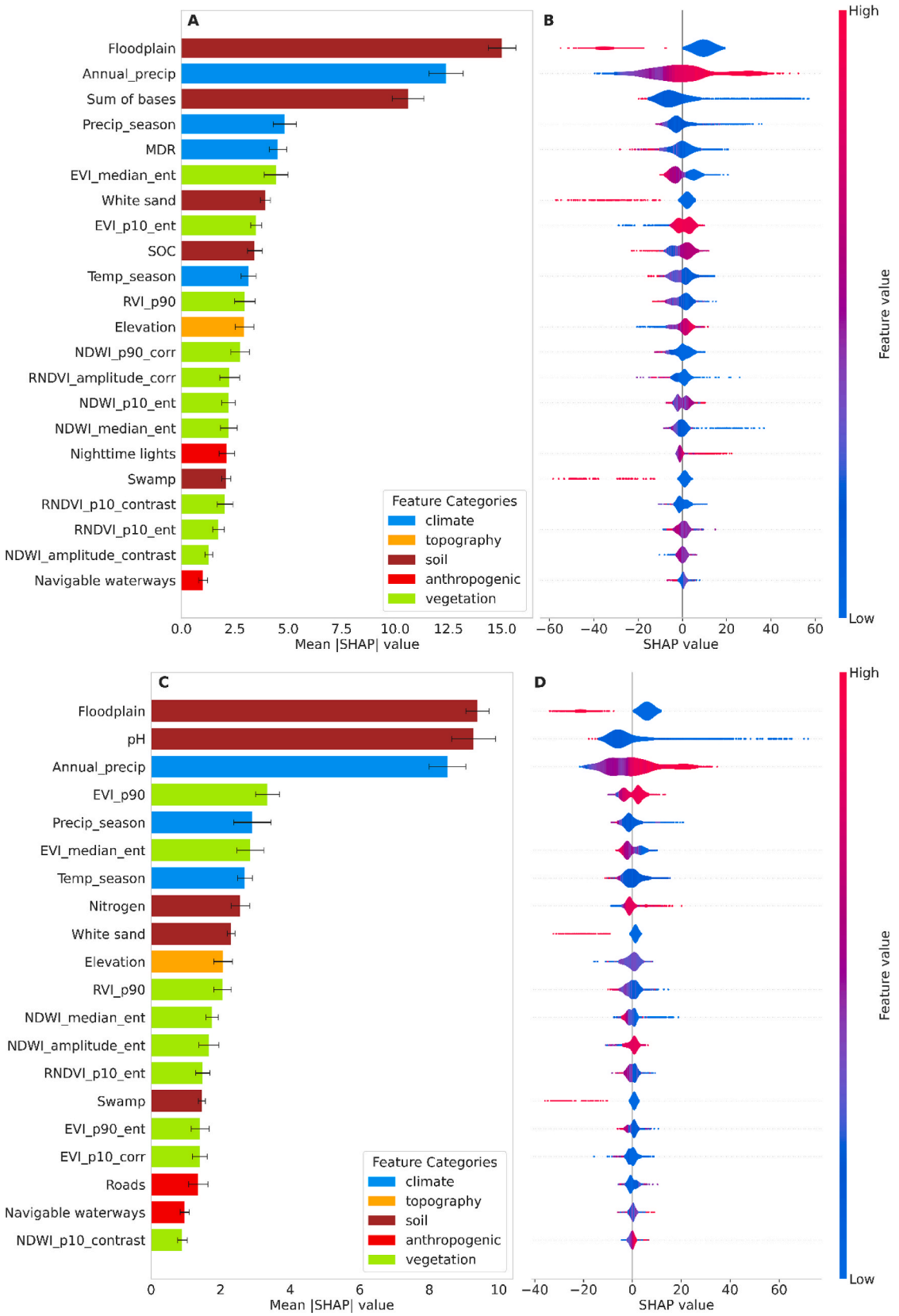

Fig. 6. Mean absolute SHAP and SHAP summary plot for species richness (A, B) and Fisher's alpha (C, D). MDR: Mean Diurnal Range; SOC: Soil Organic Carbon; EVI: Enhanced Vegetation Index; RNDVI: Red-edge Normalized Vegetation Index; NDWI: Normalized Difference Water Index; ent: entropy; corr: correlation; stdDev: standard deviation; p10: 10 percentile; p90: 90 percentile. See Table S1 for more information about each variable.

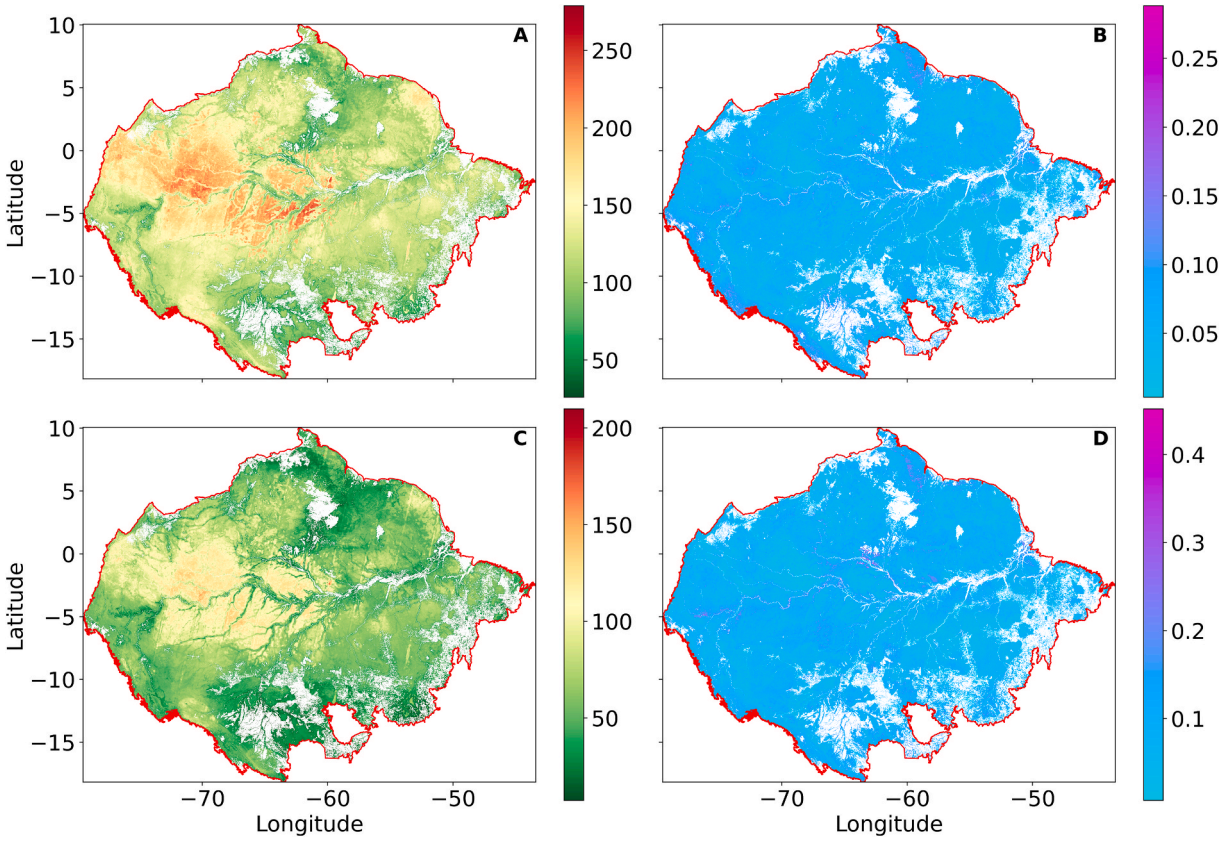

Fig. 7. Prediction maps of Amazon tree species diversity: Mean and coefficient of variation for species richness (A, B) and for Fisher's alpha (C, D).

degraded regions, such as Southern Amazonia, may result in stronger anthropogenic signals (De Lima et al., 2020). The limited importance of topographic variables indicates the significance of broader-scale factors directly influencing species survival, such as climate and soil (Ter Steege et al., 2023, Toledo et al., 2012). However, topography can shape local tree diversity by influencing microclimates and microhabitats (Liu et al., 2023, Baldeck et al., 2013), with potentially stronger impacts at local scales (Marca-Zevallos et al., 2022).

The SHAP analysis revealed the relative contributions of remote sensing features across sensors, bands and processing methods (Fig. 6). Sentinel-2 variables were particularly influential, consistent with previous studies emphasizing their utility in capturing key biochemical and biophysical attributes of vegetation (Liu et al., 2023, Xi et al., 2023). Among these, EVI was the most impactful, reflecting its sensitivity to vegetation greenness and its robustness in dense tropical canopies (Hoffmann et al., 2022). Multiple NDWI and RNDVI variables also contributed, indicating the complementarity and relevance of canopy water and pigment concentration (Gao, 1996, Gitelson and Merzlyak, 1994). In contrast, Sentinel-1 radar features contributed little, despite its utility shown in earlier work (Bae et al., 2019, Fagua et al., 2021). This discrepancy may be due to redundancy in this study between radar and other features, such as multiple VIs, which can also reflect canopy moisture and structure. Temporal metrics alone were less informative except maximum (p90) of EVI for Fisher's alpha, which indicates that high tree diversity is associated with high peak greenness. However, multiple GLCM textures derived from different temporal features contributed, capturing spatial heterogeneity during different phenological phases. Interestingly, lower texture values were mostly associated with higher tree diversity, contrary to the spectral variability hypothesis (Pinon et al., 2024). Our results may reflect the spectral homogeneity of undisturbed, biodiverse dense forests at 10 m resolution, while forests that are degraded, fragmented, or located near ecological edges or adjacent to non-forest areas exhibit greater spectral variability but lower diversity. Additionally, the hypothesis may further be constrained by relatively coarse resolution, limited plot size of field data and high variation in in-situ tree diversity (Pinon et al., 2024).

Beyond remote sensing features, SHAP also highlighted key environmental variables (Fig. 6). Among such climate variables, low variability in temperature and precipitation, together with high annual precipitation, were most associated with higher tree diversity. These results reinforce the role of climate stability and water availability as primary ecological constraints of tree diversity, and alteration of these conditions by climate change can thus be a major threat (Toledo et al., 2012, Gomes et al., 2019). Soil characteristics also played a critical role. Underrepresented forest types (i.e. floodplain, white sand and swamp) were linked to lower tree diversity, reflecting the strong effect of soil characteristics on tree diversity (Ter Steege et al., 2023). Additionally, low soil fertility, indicated by a low sum of bases, pH and soil organic carbon contributed to high tree diversity, consistent with nutrient-poor but species-rich forests in the central Amazonia (Quesada et al., 2010). However, soil fertility is not a consistent predictor across the Amazon as Ter Steege et al. (2023) concluded; other regions (e.g. northwestern Amazonia and Guiana Shield) also have high diversity on relatively fertile soils. These results suggest the importance of considering multiple variables and their interactions, rather than soil fertility alone. On the other hand, the finding that low pH soils support greater diversity is consistent with the favorability of many tropical species adapted to acidic conditions (Zhou et al., 2024). These insights lay the foundation to enhance model interpretability, understand the underlying drivers of tree diversity, and improve conservation strategies under increasing pressures.

4.3. Spatial patterns of Amazon tree species diversity

Our 1 km resolution prediction maps showed spatial patterns of tree diversity at comparably high resolution with competitive accuracy and

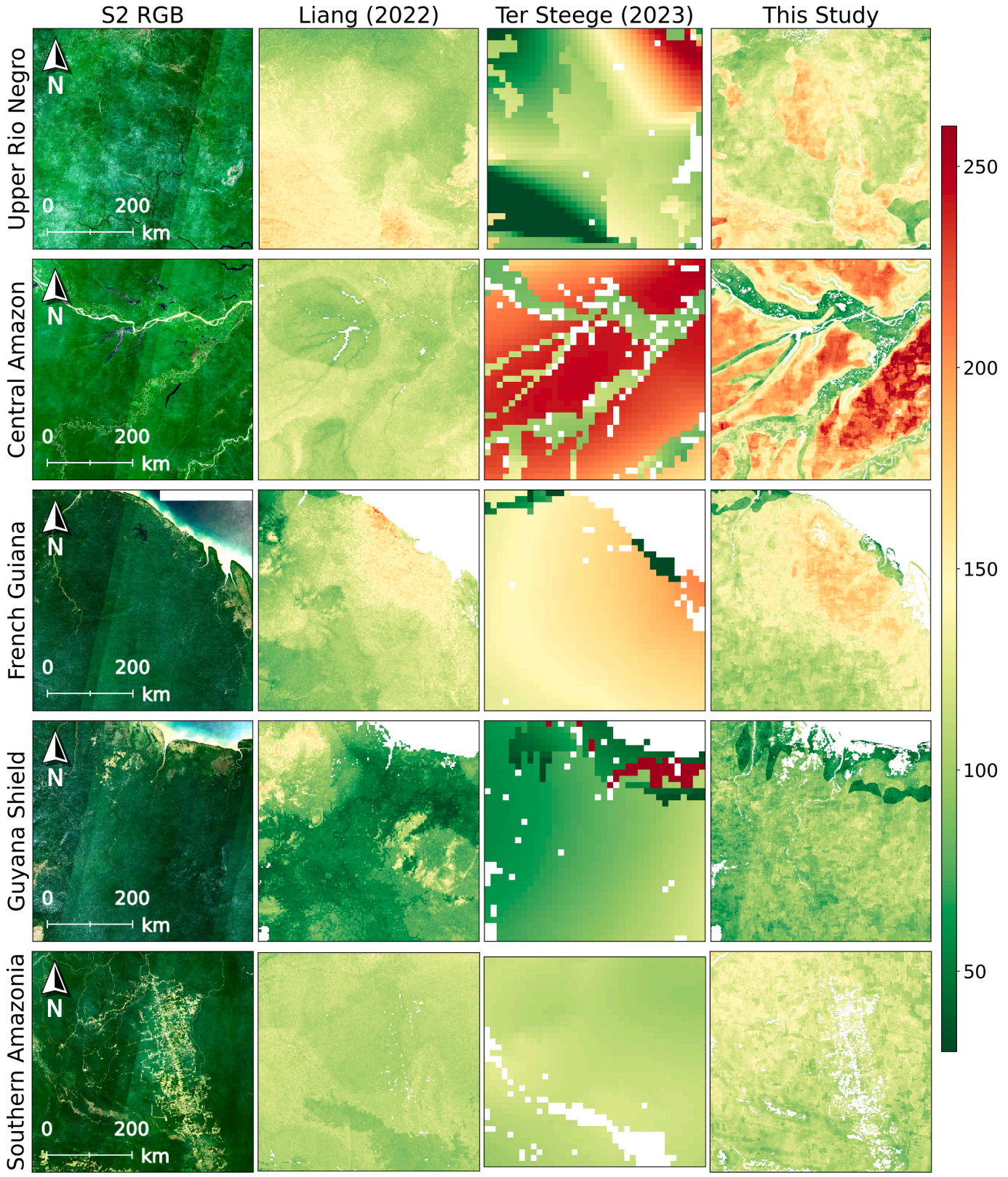

Fig. 8. Zoomed-in examples of comparative maps in five key areas across different forest types and regions of the Amazon. From left to right, each panel shows Sentinel-2 images (RGB = B4, B3, B2) in 2019 at 100 m resolution as references and the corresponding maps of tree species richness per ha derived from Liang et al. (2022), Ter Steege et al. (2023) and this study.

overall low uncertainty (Figs. 7 & 8). High tree diversity was observed in terra-firme forests in the central to northwestern Amazonia and French Guinea, while low tree diversity appeared in Guyana sandy forests, floodplains along the watersheds, and Southern Amazonia, broadly aligning with those reported by Ter Steege et al. (2023) and field plots.

Mapping with different combinations of variable groups identified the spatial patterns captured by each group (Fig. S6). Among the most influential predictors (i.e. climate, soil, and vegetation), climate, soil and their combination produced smooth spatial gradients, indicating that they capture broad-scale patterns of tree diversity as described in some previous studies (Stropp et al., 2009, Ter Steege et al., 2023). In contrast, remote sensing-based vegetation features added fine-scale spatial variation while remaining consistent with the broader diversity patterns, highlighting their potential for high-resolution tree diversity mapping (Liu et al., 2023, Hoffmann et al., 2022). Predictions generated using only soil or only vegetation variables displayed noise, but combining multiple variable groups reduced these artefacts, demonstrating the effectiveness of integrating complementary groups of predictors.

Among the forest types, terra-firme showed the highest tree species diversity and prediction accuracy (Fig. S3), likely reflecting its dominance in the training data (Fig. S1). In contrast, underrepresented forest types, such as non-flooded white sand forest and seasonally flooded floodplain forest, tended to be overestimated (Fig. 4; Fig. S3), diverging from observed patterns (Ter Steege et al., 2023). Although ecologically distinct, underrepresented forest types may still share similar environmental predictors with terra-firme forests, making them harder to differentiate in models trained on imbalanced data. The scarcity of field data from these forest types further constrains the models to learn their unique signatures and hampers robust accuracy assessment. At the sub-regional scale, Central Amazonia and Northwestern Amazonia displayed elevated tree diversity compared to the other sub-regions (Fig. S3), aligning with spatial patterns of biodiversity hotspots in terra-firme forests in these regions.

An important source of uncertainty can also arise from the scale mismatch between field plots (0.5–2 ha) and prediction maps (1 km). This approach assumes that diversity within plots is representative of diversity within larger 1 km grids, an assumption commonly applied in large-scale mapping studies (Ter Steege et al., 2023; Liang et al., 2022), but it may not hold in heterogeneous or transitional areas. Although our maps are at a higher resolution than those of previous studies, such as Liang et al. (2022) and Ter Steege et al. (2023), local variability may still be smoothed in the prediction maps, which requires caution for map interpretation. Nonetheless, the strong agreement with the large-scale maps from Ter Steege et al. (2023) and stable cross-validation performance indicate the reliability of our maps at the 1 km scale.

Overall, our study demonstrated the effectiveness of integrating remote sensing, environmental data and machine learning for large-scale, high-resolution tree species diversity mapping. The resulting maps show the spatial patterns of tree diversity, offering fine-scale insights for biodiversity assessments, forest management and conservation measures. They may also provide a foundation for future research exploring the links between tree species diversity and other forest attributes, such as other species diversity, forest disturbances or carbon stocks. While the field dataset covers the entire Amazon, field data scarcity in underrepresented forest types and the highest tree diversity zones warrants caution for interpretation in those areas. Nevertheless, the predicted maps offer valuable insights across the Amazon with competitive overall accuracy. Importantly, this modeling approach links in-situ diversity and high-resolution, spatially and temporally explicit remote sensing, alongside environmental data; therefore, it holds potential for long-term, high-resolution and large-scale tree diversity analysis in the future, for tracking biodiversity changes under increasing threats.

5. Conclusion

This study demonstrates the potential of remote sensing and diverse environmental data with machine learning for Amazon tree species diversity mapping with competitive accuracy and resolution. Based on the model performance comparison, predictor importance analysis and mapping, key findings include: 1) XGB offers accurate and stable implementation for large-scale tree diversity mapping; 2) Climate (especially precipitation and climate stability), vegetation and soil characteristics are the most important predictors of tree diversity; 3) Sentinel-2-based vegetation indices, along with their spatial textures and different phenological features, may introduce the fine-scale tree diversity information; 4) The roles of Sentinel-1 and the spectral variability hypothesis were limited, suggesting the need for refinement or reevaluation in this context; 5) The resulting 1 km maps show ecological gradients and hotspots. The findings also underscore the limitations in accuracy and ecological interpretability in areas with sparse field data, requiring future studies with more field data from underrepresented forest types and megadiverse terra-firme forests. Overall, this study presents a robust, accurate and transferable method for large-scale tree species diversity mapping, with a potential future integration into other data-scarce tropical forests, repeated monitoring, and conservation efforts.

CRediT authorship contribution statement

Shoyo Nakamura: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Nandika Tsendbazar:** Writing – review & editing, Supervision, Conceptualization. **Hans ter Steege:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Lars Hein:** Funding acquisition, Conceptualization. **Jannik Schultner:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the European Commission (Horizon Europe ID: 101060415). We thank the Google Earth Engine team for providing computational resources to complete the analysis. We also thank Sytze de Bruin, Dainius Masiliunas, Yang Li and Sietse van der Woude for insightful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2026.105226.

Data availability

The field plot data and raster maps of species richness and Fisher's alpha are publicly available under license CC BY 4.0 at https://doi.org/10.6084/m9.figshare.29352821. Additional data is available upon reasonable request.

References

Akiba, T., Sano, S., Yanase, T., et al., 2019. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631.

Amatulli, G., Mcinerney, D., Sethi, T., et al., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7.
Ampoorter, E., Barbaro, L., Jactel, H., et al., 2020. Tree diversity is key for promoting the diversity and abundance of forest-associated taxa in Europe. Oikos 129, 133–146.
Andermann, T., Antonelli, A., Barrett, R. L., et al. 2022. Estimating Alpha, Beta, and Gamma Diversity Through Deep Learning. *Frontiers in Plant Science,* 13.
Bae, S., Levick, S.R., Heidrich, L., et al., 2019. Radar vision in the mapping of forest biodiversity from space. Nat. Commun. 10.
Bai, J., Ren, C., Shi, X., et al., 2024. Tree species diversity impacts on ecosystem services of temperate forests. Ecol. Ind. 167, 112639.
Baldeck, C.A., Harms, K.E., Yavitt, J.B., et al., 2013. Soil resources and topography shape local tree community structure in tropical forests. Proc. R. Soc. B Biol. Sci. 280, 20122532.
Barnosky, A.D., Matzke, N., Tomiya, S., et al., 2011. Has the Earth's sixth mass extinction already arrived? Nature 471, 51–57.
Belitz, K., Stackelberg, P., 2021. Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. Environ. Model. Software 139, 105006.
Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
CBD 2022. Kunming-Montreal Global Biodiversity Framework. *Convention of Biological Diversity (CBD), Montreal, Canada*.
Chen, T. & Guestrin, C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016-08-13 2016. ACM.
Cristianini, N. & Ricci, E. 2008. Support Vector Machines. *Encyclopedia of Algorithms.* Springer US.
Crowther, T.W., Glick, H.B., Covey, K.R., et al., 2015. Mapping tree density at a global scale. Nature 525, 201.
De Lima, R.A.F., Oliveira, A.A., Pitta, G.R., et al., 2020. The erosion of biodiversity and biomass in the Atlantic Forest biodiversity hotspot. Nat. Commun. 11.
Dijkshoorn, K., Huting, J., Tempel, P., 2005. Update of the 1: 5 million soil and terrain database for Latin America and the Caribbean (SOTERLAC). ISRIC Rep 1, 25.
Fagua, J.C., Jantz, P., Burns, P., et al., 2021. Mapping tree diversity in the tropical forest region of Chocó-Colombia. Environ. Res. Lett. 16, 054024.
Fassnacht, F.E., Müllerová, J., Conti, L., et al., 2022. About the link between biodiversity and spectral variation. Appl. Veg. Sci. 25.
Fisher, R.A., Corbet, A.S., Williams, C.B., 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. 42–58.
Flores, B.M., Montoya, E., Sakschewski, B., et al., 2024. Critical transitions in the Amazon forest system. Nature 626, 555–564.
Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266.
Giam, X., 2017. Global biodiversity loss from tropical deforestation. Proc. Natl. Acad. Sci. 114, 5775–5777.
Gitelson, A., Merzlyak, M.N., 1994. Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and Acer platanoides L. leaves. Spectral features and relation to chlorophyll estimation. J. Plant Physiol. 143, 286–292.
Gomes, V.H.F., Vieira, I.C.G., Salomão, R.P., et al., 2019. Amazonian tree species threatened by deforestation and climate change. Nat. Clim. Chang. 9, 547–553.
Gorelick, N., Hancher, M., Dixon, M., et al., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.
Grabska, E., Frantz, D., Ostapowicz, K., 2020. Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the Polish Carpathians. Remote Sens. Environ. 251, 112103.
Hansen, M.C., Potapov, P.V., Moore, R., et al., 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 342, 850–853.
Haralick, R.M., Shanmugam, K., Dinstein, I.H., 2007. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 610–621.
Hoffmann, J., Muro, J., Dubovyk, O., 2022. Predicting Species and Structural Diversity of Temperate Forests with Satellite Remote Sensing and Deep Learning. Remote Sens. (Basel) 14, 1631.
Huete, A., Didan, K., Miura, T., et al., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.
Kacic, P., Kuenzer, C., 2022. Forest Biodiversity monitoring based on Remotely Sensed Spectral Diversity—A Review. Remote Sens. (Basel) 14, 5363.
Karger, D.N., Conrad, O., Böhner, J., et al., 2017. Climatologies at high resolution for the earth's land surface areas. Sci. Data 4, 170122.
Karger, D.N., Conrad, O., Böhner, J., et al., 2021. Climatologies at high resolution for the earth's land surface areas. EnviDat.
Kim, Y., Jackson, T., Bindlish, R., et al., 2011. Radar vegetation index for estimating the vegetation water content of rice and soybean. IEEE Geosci. Remote Sens. Lett. 9, 564–568.
Li, C., Zhou, L., Xu, W., 2021. Estimating Aboveground Biomass using Sentinel-2 MSI Data and Ensemble Algorithms for Grassland in the Shengjin Lake Wetland. China. Remote Sensing 13, 1595.
Liang, J., Gamarra, J.G.P., Picard, N., et al., 2022. Co-limitation towards lower latitudes shapes global forest diversity gradients. Nat. Ecol. Evol. 6, 1423–1437.
Liu, X., Frey, J., Munteanu, C., et al., 2023. Mapping tree species diversity in temperate montane forests using Sentinel-1 and Sentinel-2 imagery and topography data. Remote Sens. Environ. 292, 113576.
Louis, J., Debaecker, V., Pflug, B., et al. Sentinel-2 Sen2Cor: L2A processor for users. Proceedings living planet symposium 2016, 2016. Spacebooks Online, 1-8.
Lundberg, S. M. & Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems,* 30.
Marca-Zevallos, M. J., Moulatlet, G. M., Sousa, T. R., et al. 2022. Local hydrological conditions influence tree diversity and composition across the Amazon basin. *Ecography,* 2022.
Ming, L., Liu, J., Quan, Y., et al., 2024. Mapping tree species diversity in a typical natural secondary forest by combining multispectral and LiDAR data. Ecol. Ind. 159, 111711.
Mullissa, A., Vollrath, A., Odongo-Braun, C., et al., 2021. Sentinel-1 SAR Backscatter Analysis Ready Data Preparation in Google Earth Engine. Remote Sens. (Basel) 13, 1954.
Narin, O.G., 2025. Gap filling of water level time series with water area using remote sensing data: a comparative performance analysis of polynomial functions, XGBoost, Random Forest and support Vector Machine. Hydrol. Sci. J. 70, 750–760.
Niquini, F.G.F., Branches, A.M.B., Costa, J.F.C.L., et al., 2023. Recursive Feature Elimination and Neural Networks Applied to the Forecast of Mass and Metallurgical Recoveries in a Brazilian Phosphate Mine. Minerals 13, 748.
Ozdemir, I., Karnieli, A., 2011. Predicting forest structural parameters using the image texture derived from WorldView-2 multispectral imagery in a dryland forest, Israel. Int. J. Appl. Earth Obs. Geoinf. 13, 701–710.
Pinon, T.B.M., Mendonça, A.R.D., Silva, G.F.D., et al., 2024. Biodiversity from the Sky: Testing the Spectral Variation Hypothesis in the Brazilian Atlantic Forest. Remote Sens. (Basel) 16, 4363.
Poggio, L., De Sousa, L.M., Batjes, N.H., et al., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240.
Quesada, C.A., Lloyd, J., Schwarz, M., et al., 2010. Variations in chemical and physical properties of Amazon forest soils in relation to their genesis. Biogeosciences 7, 1515–1541.
RAISG. 2025. Available: https://www.amazoniasocioambiental.org/en/ [Accessed 2025].
Saatchi, S., Buermann, W., Ter Steege, H., et al., 2008. Modeling distribution of amazonian tree species and diversity using remote sensing measurements. Remote Sens. Environ. 112, 2000–2017.
Schulz, C., Förster, M., Vulova, S.V., et al., 2024. Spectral-temporal traits in Sentinel-1 C-band SAR and Sentinel-2 multispectral remote sensing time series for 61 tree species in Central Europe. Remote Sens. Environ. 307, 114162.
Skidmore, A.K., Coops, N.C., Neinavaz, E., et al., 2021. Priority list of biodiversity metrics to observe from space. Nat. Ecol. Evol. 5, 896–906.
Stropp, J., Ter Steege, H., Malhi, Y., 2009. Disentangling regional and local tree diversity in the Amazon. Ecography 32, 46–54.
Ter Steege, H., Pitman, N., Sabatier, D., et al., 2003. A spatial model of tree a-diversity and tree density for the Amazon. Biodivers. Conserv. 12, 2255–2277.
Ter Steege, H., Pitman, N.C.A., Do Amaral, I.L., et al., 2023. Mapping density, diversity and species-richness of the Amazon tree flora. Commun. Biol. 6.
Toledo, M., Peña-Claros, M., Bongers, F., et al., 2012. Distribution patterns of tropical woody species in response to climatic and edaphic gradients. J. Ecol. 100, 253–263.
Venter, O., Sanderson, E.W., Magrach, A., et al., 2016. Global terrestrial Human Footprint maps for 1993 and 2009. Sci. Data 3, 160067.
Xi, Y., Zhang, W., Brandt, M., et al., 2023. Mapping tree species diversity of temperate forests using multi-temporal Sentinel-1 and-2 imagery. Sci. Remote Sens. 8, 100094.
Xiao, C., Li, P., Feng, Z., et al., 2020. Sentinel-2 red-edge spectral indices (RESI) suitability for mapping rubber boom in Luang Namtha Province, northern Lao PDR. Int. J. Appl. Earth Obs. Geoinf. 93, 102176.
Yamazaki, D., Ikeshima, D., Tawatari, R., et al., 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44, 5844–5853.
Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., et al. 2021. ESA WorldCover 10 m 2020 v100.
Zhao, Y., Yin, X., Fu, Y., et al., 2022. A comparative mapping of plant species diversity using ensemble learning algorithms combined with high accuracy surface modeling. Environ. Sci. Pollut. Res. 29, 17878–17891.
Zhou, X., Tahvanainen, T., Malard, L., et al., 2024. Global analysis of soil bacterial genera and diversity in response to pH. Soil Biol. Biochem. 198, 109552.