Spatial predictions of tree density and tree height across Mexico´s forests using ensemble learning and forest inventory data (2009-2014)

Aylin Barreras[1], José Alanís de la Rosa[2], Rafael Mayorga[2], Rubi Cuenca[2], César Moreno-G[2], Carlos Godinez[2], Carina Delgado[2], Maria de los Ángeles Soriano-Luna[2], Stephanie George[2], Metzli Aldrete Leal[2], Sandra Medina[2], Johny Romero[2], Sergio Villela[2], Andrew Lister[3], Rachel Sheridan[3], Rafael Flores[3], Tom Crowther[4], and Mario Guevara[1]

[1]Universidad Nacional Autónoma de México Centro de Geociencias
[2]Comisión Nacional Forestal
[3]US Forest Service International Programs
[4]ETH Zurich

October 7, 2022

Abstract

The National Forestry Commission of Mexico continuously monitors forest structure within the country's continental territory by the implementation of the National Forest and Soils Inventory (INFyS). Due to the challenges involved in collecting data exclusively from field surveys, there are spatial information gaps for important forest attributes. This can produce bias or increase uncertainty when generating estimates required to support forest management decisions. Our objective is to predict the spatial distribution of tree height and tree density in all Mexican forests. We performed wall-to-wall spatial predictions of both attributes in 1-km grids, using ensemble machine learning across each forest type in Mexico. Predictor variables include remote sensing imagery and other geospatial data (e.g., vegetation indexes, surface temperature). Training data is from the 2009-2014 cycle (n>26,000 sampling plots). Spatial cross validation suggested that the model had a better performance when predicting tree height r2=0.4 [0.15,0.55] (mean[min, max]) than for tree density r2=0.2[0.10,0.31]. Maximum values of tree height were for coniferous forests, coniferous-broadleaf forests and cloud mountain forest (˜36 m, 30 m and 21 m, respectively). Tropical forests had maximum values of tree density (˜1370 trees/ha), followed by tropical dry forest (1006 trees/ha) and coniferous forest (988 trees/ha). Although most forests had relatively low values of uncertainty, e.g., values <40%, arid and semiarid ecosystems had high uncertainty in both tree height and tree density predictions, e.g., values >60%. The applied open science approach we present is easily replicable and scalable, thus it is helpful to assist in the decision-making and future of the National Forest and Soils Inventory. This work highlights the need for technical capabilities aimed to use and resignify all the effort done by the Mexican Forestry Commission in implementing the INFyS.

Title: *Spatial predictions of tree density and tree height across Mexico´s forests using ensemble learning and forest inventory data (2009-2014)*

Author list: Aylin Barreras[1-3], José Armando Alanís de la Rosa[2], Rafael Mayorga Saucedo[2], Rubi Angélica Cuenca Lara[2], César Moreno García[2], Carlos Isaías Godínez Valdivia[2], Carina Edith Delgado Caballero[2], Maria de los Ángeles Soriano Luna[2-3], Stephanie Patricia George[2], Metzli Ileana Aldrete Leal[2], Sandra Liliana Medina Casillas[2], Johny Romero Correa[2], Sergio Armando Villela Gaytán[2], Andrew Lister[3], Rachel Sheridan[3], Rafael Flores[3], Thomas W Crowther[4], Mario Guevara[1,5,6*]

Institutional addresses:

[1]Centro de Geociencias, Universidad Nacional Autónoma de México, Campus Juriquilla, Qro. MX.

[2]Comisión Nacional Forestal (CONAFOR). Periférico Poniente 5360, Zapopan, Jalisco, México 45019

[3]US Forest Service, International Programs. 1Thomas Circle NW, Suite 400, Washington, D.C. USA 20005

[4]Institute of Integrative Biology, ETH Zurich, Universitätstrasse 16, 8006, Zürich, Switzerland

[5]University of California, Riverside, Department of Environmental Sciences, 900 University Ave., Riverside, CA 92521, USA

[6]U.S. Salinity Laboratory, Agricultural Research Service, United States Department of Agriculture, 450 West Big Springs Rd., Riverside, CA 92507, USA

Corresponding authors:mguevara@geociencias.unam.mx

Abstract

The National Forestry Commission of Mexico continuously monitors forest structure within the country's continental territory by the implementation of the National Forest and Soils Inventory (INFyS). Due to the challenges involved in collecting data exclusively from field surveys, there are spatial information gaps for important forest attributes. This can produce bias or increase uncertainty when generating estimates required to support forest management decisions. Our objective is to predict the spatial distribution of tree height and tree density in all Mexican forests. We performed wall-to-wall spatial predictions of both attributes in 1-km grids, using ensemble machine learning across each forest type in Mexico. Predictor variables include remote sensing imagery and other geospatial data (e.g., vegetation indexes, surface temperature). Training data is from the 2009-2014 cycle (n>26,000 sampling plots). Spatial cross validation suggested that the model had a better performance when predicting tree height r^2=0.4 [0.15,0.55] (mean [min, max]) than for tree density r^2=0.2[0.10,0.31]. Maximum values of tree height were for coniferous forests, coniferous-broadleaf forests and cloud mountain forest (˜36 m, 30 m and 21 m, respectively). Tropical forests had maximum values of tree density (˜1370 trees/ha), followed by tropical dry forest (1006 trees/ha) and coniferous forest (988 trees/ha). Although most forests had relatively low values of uncertainty, e.g., values <40%, arid and semiarid ecosystems had high uncertainty in both tree height and tree density predictions, e.g., values >60%. The applied open science approach we present is easily replicable and scalable, thus it is helpful to assist in the decision-making and future of the National Forest and Soils Inventory. This work highlights the need for technical capabilities aimed to use and resignify all the effort done by the Mexican Forestry Commission in implementing the INFyS.

Keywords

Tree height, tree density, spatial prediction, ensemble machine learning, forest inventory.

1. Introduction

Forest inventories continuously monitor the status of forested ecosystems through the implementation of field campaigns for data collection and subsequent analysis (Smith, 2002). As forests play a key role in maintaining ecological stability, national forest inventories are playing an increasingly-important role in driving academic and governmental decision making (Saarela et al., 2020). For example, Mexico's National Forest and Soils Inventory (INFyS) is a pillar of its measurement, reporting and verification system (MRV), and the foundation for the national inventory of greenhouse gasses (GHG) emissions in the Land Use, Land-Use Change and Forestry (LULUCF) sector and for the national forest reference emissions level (FREL). MRV and FREL are components of a carbon accounting system used by the United Nations to incentivize practices that lower carbon emissions (Mitchell et al., 2017). National forest inventories usually focus on collecting field data over large geographic areas. Developing analytical tools that enhance the accessibility and understanding of nation-wide forest inventory data is critical for democratizing information about forest structure at national and international scales.

Forest inventories based on a statistical sample are used to estimate mean or total amounts of forest inventory attributes within the population of interest (Tomppo, Haakana, et al., 2008). However, field surveys can be costly, time consuming and logistically-challenging. Furthermore, collecting data exclusively from field surveys can result in designs that do not satisfy the statistical assumptions and can have limited sample sizes due to the phenomenon of non-response, which occurs when field plots that were part of the design cannot be accessed. Improper management of nonresponse can produce bias or increase uncertainty when generating estimates (McRoberts et al., 2005). Emerging satellite and machine learning (ML) technologies give us the opportunity to build standardized analytical tools that can mitigate problems associated with non-response and produce maps that serve for multiple purposes (Tomppo, Olsson, et al., 2008).

Technologies for mapping forest attributes have evolved through the modeling of attributes contained in field data with remotely-sensed satellite data, and then the use of these models to predict the spatial distribution of forest attributes (Schumacher et al., 2020; Wang et al., 2009). The integration of both data sources has been widely applied to better visualize national-scale estimates, reduce uncertainty, and improve dataset robustness (Haakana et al., 2019; Ohmann et al., 2014; Saarela et al., 2020; Tomppo et al., 2010). This approach has played a key role in modeling national estimates of forest structure such as aboveground biomass (AGB) as well as attributes such as forest age (Saarela et al., 2020; Schumacher et al., 2020). Both tree height and tree density are drivers of AGB and bioenergy potential in forest ecosystems. To obtain accurate spatial predictions of forest attributes, many studies employ ML models using a multivariate approach (Khaledian & Miller, 2020; Li et al., 2020; Soriano-Luna et al., 2018; Wadoux et al., 2020). ML is a field of artificial intelligence (AI), and one of its main objectives is to identify and model relationships between dependent data (such as forest inventory attributes) and independent data (such as remote sensing), and apply these models to generate predictions in a semi-autonomous approach (James et al., 2013a). The performance of different types of ML models often varies when modeling forest attributes. For example, spatially explicit estimates of AGB varied by as much as 19% when performing linear (LM), generalized additive (GAM) and random forest (RF) empirical models in a temperate forest in central Mexico (Soriano-Luna et al., 2018). The three fitted AGB models performed well when predicting AGB spatial distribution, but GAM was better for representing AGB variations across the landscape. Thus, different ML models yield different results and studies use multiple models or algorithms to identify the best solutions for predicting forest attributes or specific response variables, as no silver bullets exist in ecological modeling (Qiao et al., n.d.).

One commonly-used set of ML approaches used to perform spatial prediction are ensemble learners, which integrate multiple ML models and algorithms (Holloway & Mengersen, 2018). Ensemble ML models are used in mapping forest attributes because they offer improvements in accuracy to independent algorithms (Healey et al., 2018). Examples of popular ensemble ML algorithms include RF (Breiman, 2001), which applies a bagging method, and Super Learner, which applies a stacked method and uses cross-validation to estimate the performance of multiple ML models (Polley & Laan, 2010). The latter has been shown to outperform the individual algorithms used to build the model (Davies & van der Laan, n.d.; Taghizadeh-Mehrjardi et al., 2021).

Forests in Mexico are a critical natural resource, containing vast amounts of biodiversity and providing ecosystem goods and services (e.g., timber production, water security, soil conservation) as well as economic benefits. The National Forestry Commission of Mexico (CONAFOR) has been in charge of implementing the INFyS from 2004 to the present. The INFyS is a national program in which a stratified, systematic sample of permanent ground plots is used to measure trees (e.g., height, diameter at breast height, count) and site (e.g., forest type, site class, topographic data) variables across all forest lands every 5 years (CONAFOR, 2017).

The main goal of this study is to develop a methodological framework with which CONAFOR can generate country-level maps of INFyS forest attributes. Specifically, this involves operationalizing methods based on integrating field data with remote sensing data in an ensemble ML framework to map forest attributes. We are starting with tree height and tree density, as these are key components of forest structure and can be useful to provide information that helps mitigate impacts of nonresponse, and in the estimation of

AGB, carbon storage and forest productivity over time (Humagain et al., 2017; Pirotti, 2010; Selkowitz et al., 2012). Accurate spatial predictions of such structural variables are fundamental for the management and conservation of forest ecosystems, as they are important constituents in the study of land-atmosphere interactions, carbon cycling, assessment of fire hazards and timber volume estimation (Chopping et al., 2008; Selkowitz et al., 2012). By developing workflows and products based on INFyS data, this study aims to support CONAFOR in generating information that will be used by decision makers to manage forests more effectively, preserve the country's forest patrimony, and improve national and international reporting associated with MRV and FREL. We envision this methodology could be further applied for several other forest attributes such as AGB, carbon storage, and timber volume, among others, and improve Mexico's national estimates of other relevant forest attributes.

2. Material and Methods

2.1 Study area

The study was conducted at a national scale and included all forest types in Mexico (Fig 1). The country is located between latitudes 32° and 14° N, where the Nearctic and Neotropical biogeographic zones converge. Due to its geographical location, the territory has complex topographic and climatic characteristics (CONABIO, 1998). From the arid zones in the northwest to the humid rainforest in the southeast, forest ecosystems in Mexico are very diverse. They comprise a vast variety of vegetation, having tree heights ranging between 60m in coniferous forests to 1.3 m in xerophilous scrubs (CONAFOR, 2017). Tree species of economic interest include mahogany (*Swietenia macrophylla*) and cedar (*Cedrela odorata*), which are typical of tropical forests.

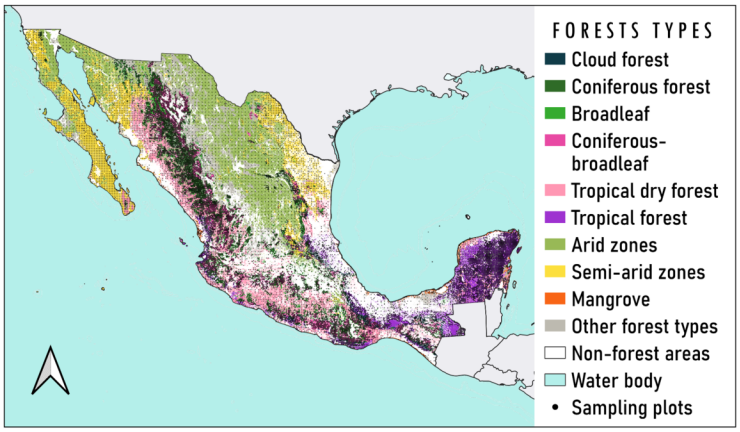

Fig 1. Map of Mexico forest types and INFyS sampling plots (black dots). Prepared from the Land Use and Vegetation map, scale 1:250,000, Series VI, Instituto Nacional de Geoestadística y Geografía (CONAFOR, 2017; INEGI, 2017).

2.2 Mexico National Forest and Soils Inventory data

Tree height and tree density models were developed using plot level data collected between 2009 and 2014 and obtained from the INFyS database. The sampling design considered a total of 26,220 plots distributed across the Mexican territory during 2009-2014 (Fig 1), however 9.5% of the total plots were categorized

as inaccessible sites and another 5.1% were obtained from remote sensing data (CONAFOR, 2017). The number of sampled plots for each forest ecosystem were 2,606 for coniferous forests; 4,111 for coniferous-broadleaf; 3,249 for broadleaf forest; 483 for cloud mountain forest; 3,724 for tropical forest; 1,466 for tropical dry forests; 240 for arid zones; 1,334 for semiarid zones and 157 for mangrove forests. Data available from the Environmental Data Initiative (EDI): https://doi.org/10.6073/pasta/4620375aea631ab6a09cb573c7bf8aff (Barreras et al., 2022) and at the official web page https://snmf.cnf.gob.mx/datos-del-inventario/.

Sampled plots are distributed across all land cover types, ecological stages, and land tenure classes (e.g., private, social, government). Plot distribution is accomplished through systematic, pre-stratified sampling with 5x5, 10x10 and 20x20 km spacing in temperate and tropical forests, dry and semiarid vegetation communities, and arid vegetation strata, respectively. These strata are derived from a forest type map created by the Mexican government (Fig 1, INEGI, 2017), hereafter referred to as the INEGI map. The plot is considered a cluster design with 4 circular subplots, 3 of which are configured in a triangular array around a central subplot. Primary subplots, where trees with a diameter at breast height (dbh, 1.3 m above ground) [?] 7.5 cm are measured, have a radius of 12.56 meters and are 400 m^2in area; spacing between adjacent primary subplot centers is 45.14 m. (CONAFOR 2017). For the purpose of this study, a tree is defined as those greater than or equal to 7.5 cm dbh.

2.3 Remotely-sensed data as model predictors

As a cloud-based platform, Google Earth Engine (GEE) provides easy access to an extensive catalog of satellite imagery and other geospatial data for scientific, business and government users (Gorelick et al., 2017). We obtained a combination of topographic, climatic, and vegetation derived variables with pixel sizes of 1000 m (Supplementary Table S1) for the period of 2009 to 2014 from GEE to assemble a nation-wide geospatial dataset to use as predictors in tree height and tree density models.

Datasets included WorldClim V1; a set of bioclimatic variables derived from the monthly temperature and rainfall (Hijmans, 2005); time-series analysis of Landsat images from the Hansen Global Forest Change v1.8 (2000-2020) dataset (Hansen et al., 2013); 4-day composite dataset from Moderate Resolution Imaging Spectro-radiometer (MODIS) sensors with fraction of photosynthetic active radiation and leaf area index at 500-m resolution (Myneni, Ranga et al., 2015) and the Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Emissivity Database (2000-2008) (Hulley et al., 2009, 2012, 2015; Hulley & Hook, 2008, 2009, 2011; NASA JPL, 2014) (Supporting Information, Table S1). All covariates were resampled to 1000 m. The resampling was done with conventional bilinear interpolation as implemented in GEE. Data available from Zenodo under the name "Nationwide geospatial dataset of environmental covariates at 1km resolution in Mexico" (https://doi.org/10.5281/zenodo.7130164) (Barreras & Guevara, 2022).

We reduced the number of potential predictor layers to 6, through a culling process guided by an analysis of the correlation between each potential predictor and the target variables (tree height and density). We ranked the variables based on the magnitude of their correlation coefficients. The intent of this data reduction step was to improve the efficiency of our modeling framework. These univariate correlation results are only included to give a sense of the directionality of the relationships with target variables, but do not suggest causality.

2.4 Spatial prediction using LANDMAP

To determine the best model, we applied the Super Learner ensemble algorithm as implemented in the LANDMAP package v0.0.14 for R v4.1.0, which provides a strategy for automated mapping by performing spatial prediction using raster data as predictors (Hengl et al., 2018, 2021; Polley & Laan, 2010; RStudio Team, 2021) (https://github.com/Envirometrix/landmap). The Super Learner ensemble ML algorithm developed by Polley & Laan (2010), estimates the performance of multiple ML models by using cross-validation. It develops an ensemble of the optimal weighted averages from the models using the test data performance (van der Laan et al., 2007). The LANDMAP package has 41 different predictive algorithms available. The methods implemented in the model ensemble were decision trees-based methods (random forest), kernel-based methods (support vector machines), methods based on neural networks, and generalized linear models. We

assumed that different methods describe relationships in our data in a different manner.

We took advantage of the geographical distances in our training data and used the oblique geographic coordinates technique to assume there is no collinearity between covariates, as used by previous studies (Moller et al., 2020). We expressed the uncertainty of our estimates in percentage form as the range of the 68% prediction intervals divided by their mean prediction for each pixel, as performed by Viscarra Rossel et al. (2014). We used a 5-fold spatial cross validation (spCV) approach to assess the predictive accuracy of our modeling framework (Brenning, 2012; James et al., 2013b; Wadoux et al., 2021). The spCV yields model independent residuals required to compute map quality indicators such as: the coefficient of determination (r^2) and root mean square error (RMSE). To compare model accuracy among different forest types we used Taylor diagrams (Wadoux et al., 2022).

3. Results

3.1 Descriptive statistics of sampled inventory data

Maximum field measurements of tree height were 36 m and found in coniferous forests and coniferous-broadleaf forests, and minimum tree heights were 1 m and found in all forest types. Mean tree heights measured in the field ranged from 5-10 m, with the exception of arid and semi-arid zones, where trees had an average height of ~4 m (Fig 2).

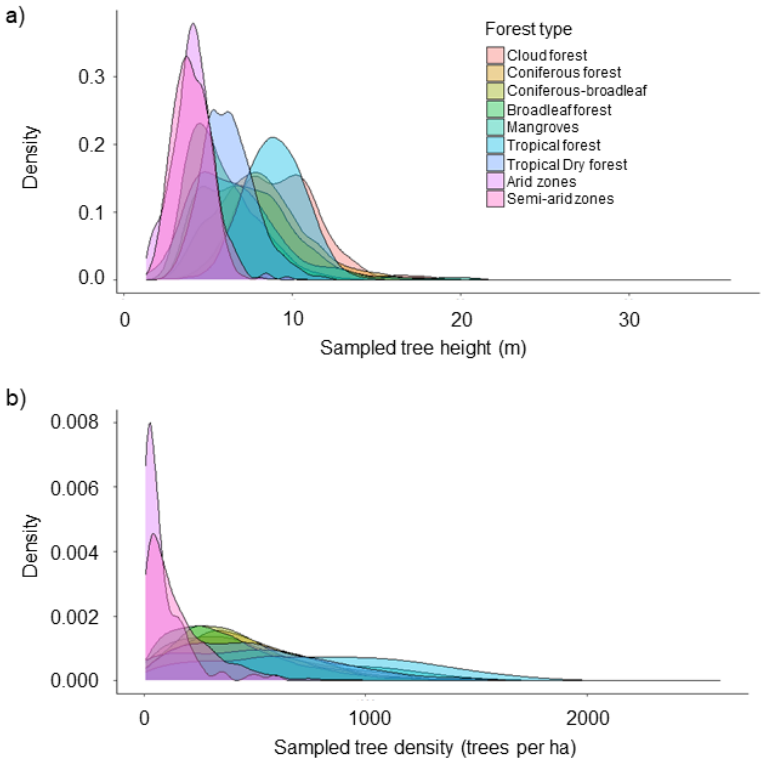

Fig 2. Density plots of a) sampled tree height and b) sampled tree density for each forest type. Data obtained from the National Forest and Soils Inventory (INFyS, CONAFOR 2017).

Mean field estimates of tree density were higher in tropical forests, with an average of ˜790 trees per ha. Generally, the other forest types had an average of ˜400 to 500 trees per ha, even mangroves which had a relatively small number of sampled plots (157). Arid and semi-arid zones had an average of ˜88 and ˜162 trees per ha, respectively (Fig 2).

3.2 Model predictors

To reduce the number of model predictors we performed a correlation analysis between the remote-sensed variables and each sampled variable (tree height and tree density) (Table S2, supporting information). Best correlated predictors were the same for both tree height and density: 1) tree canopy cover (r= 0.62 and r= 0.50, respectively) (Hansen et al., 2013), 2) fraction of photosynthetic active radiation (FPAR) (r= 0.61 and r= 0.49, respectively), 3) leaf area index (LAI) (r= 0.56 and r= 0.50, respectively) (Myneni, Ranga et al.,

2015), 4) mean land surface temperature (r= -0.42 and r= -0.25, respectively) (Hulley et al., 2009, 2012, 2015; Hulley & Hook, 2008, 2009, 2011; NASA JPL, 2014), 5) temperature annual range (r= -0.50 and r= -0.26, respectively), and 6) temperature seasonality (r= -0.47 and r= -0.23, respectively) both from Hijmans, 2005.

3.3 Model summaries and evaluation

Descriptive statistics for each forest-type predicted tree height and density map are presented in Table 1. The mean, maximum and minimum values of predictions generally aligned with those of the plots. Tree height models with the highest r^2 values were coniferous-broadleaf forest, broadleaf forest, and mangroves (Table 1) but appeared to have the best performance for tropical dry forests as it had relatively small RMSE and a correlation of ~0.65 (Fig 3a).

Table 1. Summary statistics of tree height and tree density predictions for all forest ecosystems in Mexico. *r^2*: *coefficient of determination,* RMSE*: root mean square error.*

Forest type	Tree height mean [m]	Tree height r^2	Tree height RMSE	Tree density mean [trees/ ha]	Tree density r^2	Tree density RMSE
Coniferous forest	7.98	0.45	2.34	402.38	0.25	266.24
Coniferous and broadleaf forests	7.23	0.55	2.00	360.45	0.14	261.09
Broadleaf forest	5.71	0.54	1.42	309.21	0.21	256.93
Cloud forest	9.17	0.15	2.37	366.11	0.10	328.97
Tropical forest	8.90	0.26	1.57	700.54	0.31	353.94
Tropical dry forest	6.13	0.44	1.19	438.05	0.23	294.65
Arid zones	3.92	0.22	1.11	52.66	0.17	107.82
Semi-arid zones	3.99	0.36	0.94	106.30	0.18	161.55
Mangroves	6.68	0.52	2.21	395.70	0.12	433.16

Moreover, the model performed tree density predictions with higher r^2 in tropical forests, coniferous forests, and tropical dry forests. Notwithstanding having the highest r^2 in tropical forests, it also was the forest ecosystem with the highest prediction error (Fig 3b).

On average, predicted tree height ranged between 4 to 9 m in all forest ecosystems (averaged from all pixel values). Cloud forest, arid and semi-arid zones had smaller r^2 for both target variables, which could be related to the smaller amounts of sampled data in these forest types. However, arid and semi-arid zones seemed to have the smallest error in both tree height and tree density predictions (Fig 3). A Taylor diagram is a graphical approach that quantifies how closely the predicted values match the observed values and uses correlation (r), standard deviation of the error (SDE) and standard deviation of observed (σ_z) and predicted (σ) values as evaluators (Wadoux et al., 2022). According to Taylor diagrams, the model had a better predictive performance for tropical dry forest and broadleaf forest when predicting tree height as stated by its correlation and RMSE together (Fig 3a). The model seemed to have the best predictive performance for tropical forest when predicting tree density, nonetheless, all forest types had a similar performance (Fig 3b).

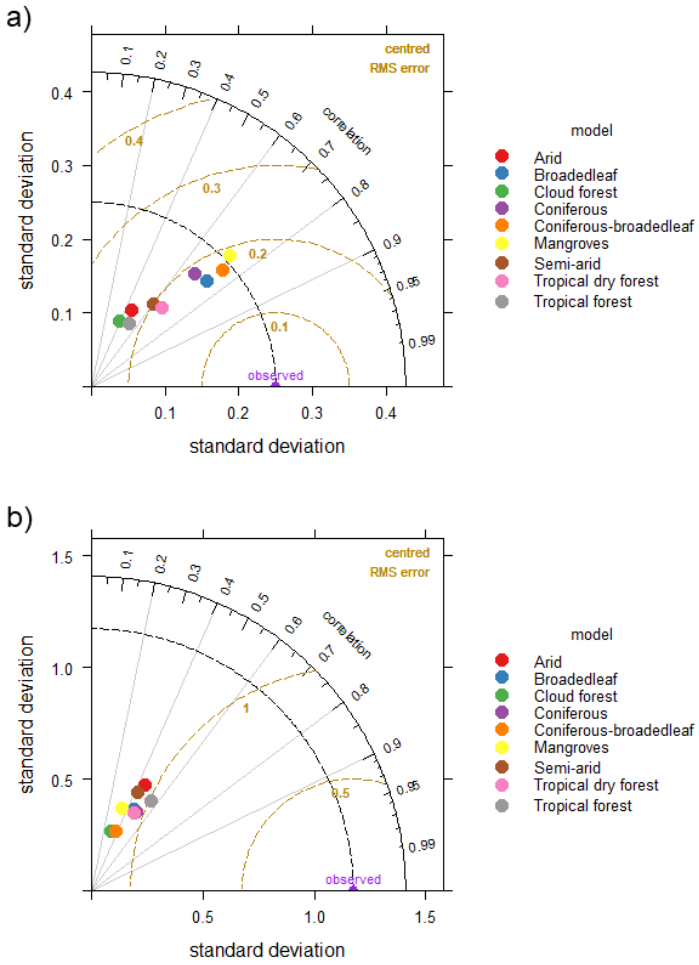

Fig 3. Taylor diagram of Mexico forest types for predicted a) tree height and b) tree density. They are shown in a polar coordinate system, the radial distance from the origin is assigned to the standardized standard deviation ($\sigma^* = \sigma / \sigma_z$) and the angular position to the correlation coefficient between the map with predicted values and the map with the true values. The brown dashed lines represent the distance in standardized SDE from the reference point (observed, purple point) (Wadoux et al., 2022).

3.4 Mapping Mexico forests tree height and tree density

The models were used to generate a spatially continuous national map of total tree height (Fig 4) and the total number of trees (Fig 5), both at a 1000-m resolution, along with their associated uncertainties. At the forest type level, maximum pixel values of tree height exist in coniferous, coniferous-broadleaf forests and cloud mountain forests (˜36 m, 30 m and 21 m, respectively). These types of forest ecosystems constitute Mexico's mountain chains Sierra Madre Oriental and Sierra Madre Occidental. Moreover, the smallest tree heights were predicted in arid and semi-arid zones, having a mean of ˜4 m. The model had the highest

uncertainty when predicting tree height in arid zones (60-80%), the latter could be related to the limited sample size we had for that specific forest type (Fig 4b). Lower uncertainty was observed for tropical forest, tropical dry forest, coniferous and broadleaf forest.

Tropical forests had the maximum pixel values of tree density (˜1370 trees/ha), followed by tropical dry forest (1006 trees/ha), coniferous forest (988 trees/ha) and coniferous-broadleaf forest (931 trees/ha) (Fig 5a). Uncertainty in predictions of tree density was relatively low across all forest types (Fig 5b). Similar to what we observed for tree height, the highest uncertainty estimates were found for arid and semi-arid ecosystems but were of a higher magnitude (>80%).

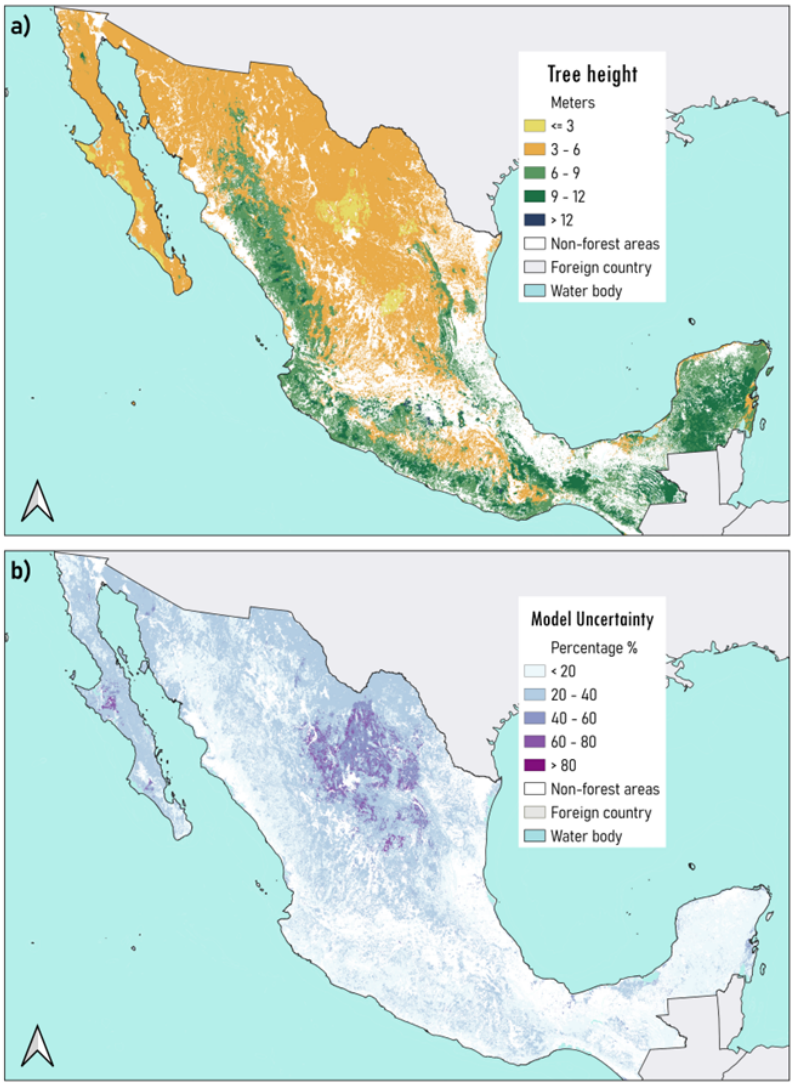

Fig 4. National maps of a) predicted mean tree height and b) its associated uncertainty across all Mexico's forest ecosystems.

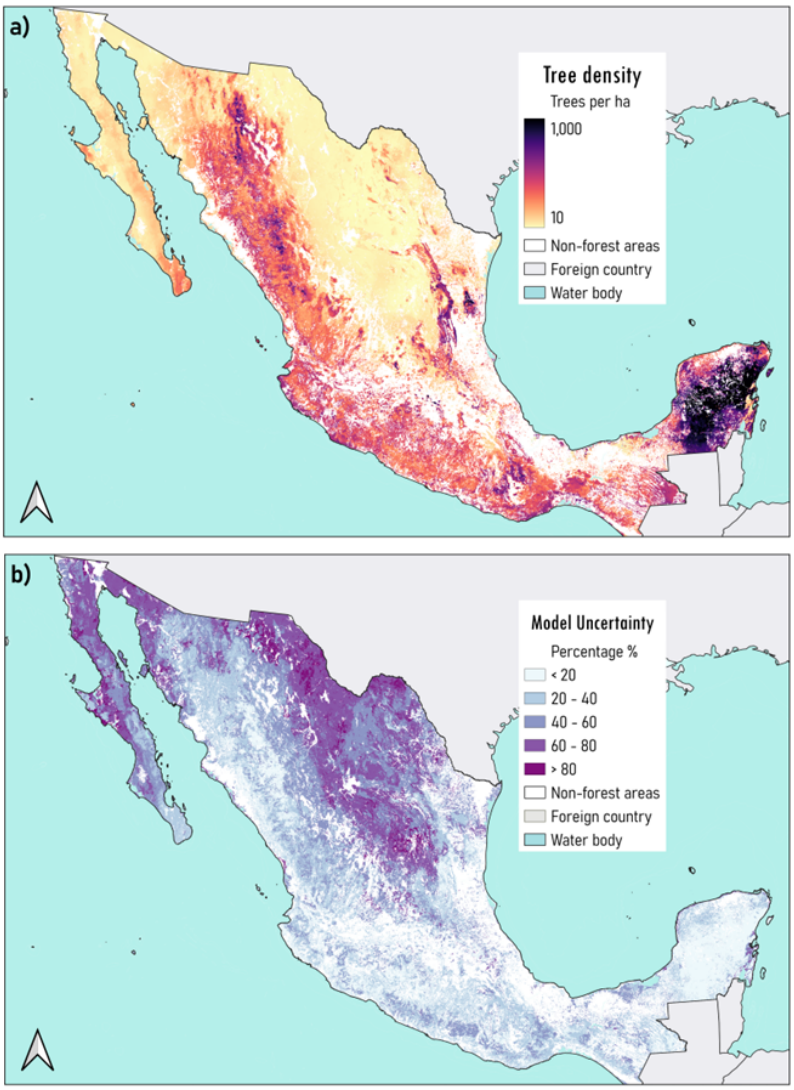

Fig 5. National maps of a) predicted mean tree density and b) its associated uncertainty across all Mexico's forest ecosystems.

4. Discussion

Over the last 20 years, CONAFOR has invested significant time and resources to produce forest inventory data that accurately represents all forest ecosystems in Mexico. To further expand the utility of this data, we developed an analytical framework to model, predict, and map forest structural attributes (tree density and height) across the country. By exploiting the available open access of remotely sensed data (e.g., mean land surface temperature, LAI, NPP, FPAR) (Gorelick et al., 2017), the ensemble machine learning method in the LANDMAP package v0.0.14 for R v4.1.0 (Hengl et al., 2021; RStudio Team, 2021), and the INFyS data (CONAFOR, 2017), we have modeled and performed predictions of tree height and tree density across Mexico. Results suggest that the ensemble ML algorithm had a better performance when predicting tree height than tree density (Table 1). In addition to providing numerical estimates, these maps are user-friendly

devices that help users visualize forest structures across Mexico.

Mapping forest attributes along with associated uncertainties at a national scale requires substantial computational resources. We simplified our approach by modeling at a 1000-m resolution and reducing the number of model predictors, thus reducing computing costs and still displaying valuable nation-wide maps for biodiversity studies and ecologic matters. Nevertheless, previous studies have shown that high resolution satellite data (e.g., 30 m) has helped achieve an increase in predictive ability (Hengl et al., 2021). It is important to acquire sufficient computational resources for the project's next stage and perform predictions with high-resolution covariates. Both tree height and density had strong univariate correlations with remotely-sensed predictors like canopy cover, FPAR and LAI. Previous studies have shown that, using more than one vegetation trait as model predictors can reduce prediction uncertainty when mapping forest attributes (Saarela et al., 2020). These results give a sense of the directionality of the relationships between the modeled attributes and its environment and strengthen the conviction of monitoring forest change through time.

The range of mean predicted values for tree height were consistent with forest inventory data (˜5-10 m). These results suggest that predictions using the Super Learner model reflected the input data adequately. On average, cloud mountain forest is the forest type with the tallest trees in Mexico (Table 1). This particular forest belongs to humid and temperate areas; it has the largest aerial biomass density and the greatest timber volume of all Mexico forest types but it accounts for only ˜1% of the national forest area (Villaseñor & Gual, 2014). According to CONAFOR (2017), more than half of its vegetation is in early stages of succession, with high densities of young and smaller trees. Maps of tree height, therefore, can indicate areas that deserve more attention, such as the wide exploitation of cloud mountain forest goods. Estimates of tree height are also critical for the evaluation of forest structure (e.g., successional stages) and projecting Mexican forests growth trajectories under different management scenarios.

Mean predicted tree density values were smaller than the field-sampled inventory data (Table 1). Globally, 42.8% of the planet's trees exist in tropical and subtropical regions (Crowther et al., 2015). Generally, optimal conditions for tree growth are warm temperatures and moisture availability (Leathwick & Austin, 2001). In accordance with this assumption, tropical forests, which develop in a warm and moist environment, have the highest tree density of all Mexico forest types (maximum values of ˜1370 trees/ha). The highest forest densities can be observed in the Calakmul rainforest area located within the Yucatán Península, in the southeast of Mexico (Fig 5a). The Calakmul rainforest is part of an important ecological gradient, the Mesoamerican Biological Corridor. The conservation of this ecologically important region has been a challenge due to continuous forest disturbances. Tree density has been used as an indicator for forest degradation on tropical ecosystems (Román-Dañobeytia et al., 2014), therefore we encourage the long-term monitoring of tropical forest structure.

For both target variables, uncertainty in our predictions was below 50% in most forests. Our uncertainty maps also show areas where the model performs poorly, especially in northern areas which consist of arid and semi-arid ecosystems (>80% uncertainty). These ecosystems have fewer sampling plots, which leaves less training data for modeling over a considerably large area of Mexico. The diversity of Mexican forests and the limited land access imply a logistics challenge for the forest inventory and this causes an under-representation of specific forest areas. One potential use for our uncertainty maps is for the INFyS to identify certain areas that require more sampling plots (e.g., arid and semi-arid ecosystems) and to identify new sampling locations on the areas with poor modeling accuracy (e.g., areas with high uncertainty).

Data from this study was managed under the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) for scientific data management by setting up an open-access online data repository available at the Environmental Data Initiative (EDI): https://doi.org/10.6073/pasta/4620375aea631ab6a09cb573c7bf8aff. Having well-documented methods, FAIR research protocols, and a good documentation of forest inventory data for all users can help advance the science and policy relevant to forestry research and management.

Continuous improvement in the study design we present here is encouraged in order to improve the accuracy

of predictions. For instance, we suggest acquiring remote sensing data at a higher resolution, increasing computational capacity, assessing new spatial prediction models, and locating new sampling sites in ecosystems with poor map quality indicators (e.g. r^2, RMSE) or uncertainties >80%. Finally, the results of this study can facilitate the understanding of Mexican forest ecosystems by further applying this methodological framework for the mapping of other forest attributes such as AGB, soil and vegetation carbon storage and their associated functional traits. To achieve this, it is important to continue with active forest inventory campaigns that facilitate the estimation of forest structure patterns through time.

Here we develop a methodological framework for the spatial prediction of forest attributes, which assists the understanding of forest structure and expands institutional and technical capabilities for data analysis within the National Forestry Commission of Mexico. Out of ten forest ecosystems, our analyses show that the best predictive performance when mapping tree height was in tropical dry forest and broadleaf forest (model explained ˜50% of variance). The best predictive performance when mapping tree density was in tropical forest (model explained ˜30% of variance). For both target variables, uncertainties in our predictions were below 50% in most forests.

Our results suggest that an ensemble learning framework can be successfully used for the spatial prediction of forest attributes and can likely be improved by having a larger number of field observations and sufficient model predictors that reflect the environment of each forest ecosystem. In order to ensure best practices for forest management in Mexico, it is important that governmental and academic institutions work together to develop approaches. This strategy helps improve the quality and transparency of forestry datasets.

References

Barreras, A., Alanís De La Rosa, J. A., Cuenca Lara, R. A., Moreno García, C., Godínez Valdivia, C. I., Delgado Caballero, C. E., Soriano Luna, M. D. L. Á., George, S. P., Aldrete Leal, M. I., Medina Casillas, S. L., Romero Correa, J., Villela Gaytán, S. A., & Guevara, M. (2022).*National Forest and Soils Inventory of Mexico 2009-2014* [Data set]. Environmental Data Initiative. https://doi.org/10.6073/PASTA/4620375AEA631AB6A09CB573C7BF8AFF

Barreras, A., & Guevara, M. (2022). *Nationwide geospatial dataset of environmental covariates at 1km resolution in Mexico* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7130164

Breiman, L. (2001). Random Forests. *Machine Learning* ,*45* (1), 5–32. https://doi.org/10.1023/A:1010933404324

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *2012 IEEE International Geoscience and Remote Sensing Symposium* , 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393

Chopping, M., Moisen, G. G., Su, L., Laliberte, A., Rango, A., Martonchik, J. V., & Peters, D. P. C. (2008). Large area mapping of southwestern forest crown cover, canopy height, and biomass using the NASA Multiangle Imaging Spectro-Radiometer. *Remote Sensing of Environment* , *112* (5), 2051–2063. https://doi.org/10.1016/j.rse.2007.07.024

CONABIO, C. N. P. E. C. Y. U. D. L. B. (1998). *La diversidad biológica de México: Estudio de País 1998* (pp. 238–283).

CONAFOR, C. N. F. (2017). *Inventario Nacional Forestal y de Suelos, Informe de Resultados 2009-2014* .

Crowther, T. W., Glick, H. B., Covey, K. R., Bettigole, C., Maynard, D. S., Thomas, S. M., Smith, J. R., Hintler, G., Duguid, M. C., Amatulli, G., Tuanmu, M.-N., Jetz, W., Salas, C., Stam, C., Piotto, D., Tavani, R., Green, S., Bruce, G., Williams, S. J., ... Bradford, M. A. (2015). Mapping tree density at a global scale. *Nature* ,*525* (7568), 201–205. https://doi.org/10.1038/nature14967

Davies, M. M., & van der Laan, M. J. (n.d.). *Optimal Spatial Prediction Using Ensemble Machine Learning* . Retrieved March 25, 2022, from https://www.degruyter.com/document/doi/10.1515/ijb-2014-0060/html

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* , *202* , 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Haakana, H., Heikkinen, J., Katila, M., & Kangas, A. (2019). Efficiency of post-stratification for a large-scale forest inventory—Case Finnish NFI. *Annals of Forest Science* , *76* (1), 1–15. https://doi.org/10.1007/s13595-018-0795-6

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. *Science (New York, N.Y.)* , *342* (6160), 850–853. https://doi.org/10.1126/science.1244693

Healey, S. P., Cohen, W. B., Yang, Z., Kenneth Brewer, C., Brooks, E. B., Gorelick, N., Hernandez, A. J., Huang, C., Joseph Hughes, M., Kennedy, R. E., Loveland, T. R., Moisen, G. G., Schroeder, T. A., Stehman, S. V., Vogelmann, J. E., Woodcock, C. E., Yang, L., & Zhu, Z. (2018). Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment* , *204* , 717–728. https://doi.org/10.1016/j.rse.2017.09.029

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., ... Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports* , *11* (1), 6130. https://doi.org/10.1038/s41598-021-85639-y

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* , *6* , e5518. https://doi.org/10.7717/peerj.5518

Hijmans. (2005). *Very high resolution interpolated climate surfaces for global land areas* . International Journal of Climatology - Wiley Online Library. https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.1276

Holloway, J., & Mengersen, K. (2018). Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review.*Remote Sensing* , *10* (9), 1365. https://doi.org/10.3390/rs10091365

Hulley, G. C., & Hook, S. J. (2008). A new methodology for cloud detection and classification with ASTER data. *Geophysical Research Letters* , *35* (16). https://doi.org/10.1029/2008GL034644

Hulley, G. C., & Hook, S. J. (2009). The North American ASTER Land Surface Emissivity Database (NAALSED) Version 2.0. *Remote Sensing of Environment* , *113* (9), 1967–1975. https://doi.org/10.1016/j.rse.2009.05.005

Hulley, G. C., & Hook, S. J. (2011). Generating Consistent Land Surface Temperature and Emissivity Products Between ASTER and MODIS Data for Earth Science Research. *IEEE Transactions on Geoscience and Remote Sensing* , *49* (4), 1304–1315. https://doi.org/10.1109/TGRS.2010.2063034

Hulley, G. C., Hook, S. J., Abbott, E., Malakar, N., Islam, T., & Abrams, M. (2015). The ASTER Global Emissivity Dataset (ASTER GED): Mapping Earth's emissivity at 100 meter spatial scale. *Geophysical Research Letters* , *42* (19), 7966–7976. https://doi.org/10.1002/2015GL065564

Hulley, G. C., Hook, S. J., & Baldridge, A. M. (2009). Validation of the North American ASTER Land Surface Emissivity Database (NAALSED) version 2.0 using pseudo-invariant sand dune sites. *Remote Sensing of Environment* , *113* (10), 2224–2233. https://doi.org/10.1016/j.rse.2009.06.005

Hulley, G. C., Hughes, C. G., & Hook, S. J. (2012). Quantifying uncertainties in land surface temperature and emissivity retrievals from ASTER and MODIS thermal infrared data. *Journal of Geophysical Research: Atmospheres* , *117* (D23). https://doi.org/10.1029/2012JD018506

Humagain, K., Portillo-Quintero, C., Cox, R. D., & Cain, J. W. (2017). Mapping Tree Density in Forests of the Southwestern USA Using Landsat 8 Data. *Forests* , *8* (8), 287. https://doi.org/10.3390/f8080287

INEGI, I. N. de E. y. (2017). *Mapas de Uso de suelo y vegetación* . Instituto Nacional de Estadística y Geografía. INEGI. https://www.inegi.org.mx/temas/usosuelo/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). Introduction. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* (pp. 1–14). Springer. https://doi.org/10.1007/978-1-4614-7138-7_1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). Statistical Learning. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* (pp. 15–57). Springer. https://doi.org/10.1007/978-1-4614-7138-7_2

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling* , *81* , 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Leathwick, J. R., & Austin, M. P. (2001). Competitive Interactions Between Tree Species in New Zealand's Old-Growth Indigenous Forests.*Ecology* , *82* (9), 2560–2573. https://doi.org/10.1890/0012-9658(2001)082[2560:CIBTSI]2.0.CO;2

Li, W., Niu, Z., Shang, R., Qin, Y., Wang, L., & Chen, H. (2020). High-resolution mapping of forest canopy height using machine learning by coupling ICESat-2 LiDAR with Sentinel-1, Sentinel-2 and Landsat-8 data. *International Journal of Applied Earth Observation and Geoinformation* , *92* , 102163. https://doi.org/10.1016/j.jag.2020.102163

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C., & Gormanson, D. D. (2005). Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Canadian Journal of Forest Research* , *35* (12), 2968–2980. https://doi.org/10.1139/x05-222

Mitchell, A. L., Rosenqvist, A., & Mora, B. (2017). Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for REDD+. *Carbon Balance and Management* , *12* (1), 9. https://doi.org/10.1186/s13021-017-0078-9

Møller, A. B., Beucher, A. M., Pouladi, N., & Greve, M. H. (2020). Oblique geographic coordinates as covariates for digital soil mapping.*SOIL* , *6* (2), 269–289. https://doi.org/10.5194/soil-6-269-2020

Myneni, Ranga, Knyazikhin, Yuri, & Park, Taejin. (2015).*MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006* [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD15A2H.006

NASA JPL. (2014). *ASTER Global Emissivity Dataset, 100-meter, HDF5* [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/COMMUNITY/ASTER_GED/AG100.003

Ohmann, J. L., Gregory, M. J., & Roberts, H. M. (2014). Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. *Remote Sensing of Environment* , *151* , 3–15. https://doi.org/10.1016/j.rse.2013.08.048

Pirotti, F. (2010). Assessing a Template Matching Approach for Tree Height and Position Extraction from Lidar-Derived Canopy Height Models of Pinus Pinaster Stands. *Forests* , *1* (4), 194–208. https://doi.org/10.3390/f1040194

Polley, E., & Laan, M. van der. (2010). Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series* . https://biostats.bepress.com/ucbbiostat/paper266

Qiao, H., Soberón, J., & Townsend Peterson, A. (n.d.). *No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation—Qiao—*

2015—Methods in Ecology and Evolution—Wiley Online Library . Retrieved September 30, 2022, from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12397

Román-Dañobeytia, F. J., Levy-Tacher, S. I., Macario-Mendoza, P., & Zúñiga-Morales, J. (2014). Redefining Secondary Forests in the Mexican Forest Code: Implications for Management, Restoration, and Conservation.*Forests* , *5* (5), 978–991. https://doi.org/10.3390/f5050978

RStudio Team. (2021). *RStudio: Integrated Development Environment for R* (1.4.1717). RStudio, PBC. http://www.rstudio.com/

Saarela, S., Wästlund, A., Holmström, E., Mensah, A. A., Holm, S., Nilsson, M., Fridman, J., & Ståhl, G. (2020). Mapping aboveground biomass and its prediction uncertainty using LiDAR and field data, accounting for tree-level allometric and LiDAR model errors.*Forest Ecosystems* , *7* (1), 43. https://doi.org/10.1186/s40663-020-00245-0

Schumacher, J., Hauglin, M., Astrup, R., & Breidenbach, J. (2020). Mapping forest age using National Forest Inventory, airborne laser scanning, and Sentinel-2 data. *Forest Ecosystems* , *7* (1), 60. https://doi.org/10.1186/s40663-020-00274-9

Selkowitz, D. J., Green, G., Peterson, B., & Wylie, B. (2012). A multi-sensor lidar, multi-spectral and multi-angular approach for mapping canopy height in boreal forest regions. *Remote Sensing of Environment* , *121* , 458–471. https://doi.org/10.1016/j.rse.2012.02.020

Smith, W. B. (2002). Forest inventory and analysis: A national inventory and monitoring program. *Environmental Pollution* , *116* , S233–S242. https://doi.org/10.1016/S0269-7491(01)00255-X

Soriano-Luna, M. D. los Á., Ángeles-Pérez, G., Guevara, M., Birdsey, R., Pan, Y., Vaquera-Huerta, H., Valdez-Lazalde, J. R., Johnson, K. D., & Vargas, R. (2018). Determinants of Above-Ground Biomass and Its Spatial Variability in a Temperate Forest Managed for Timber Production.*Forests* , *9* (8), 490. https://doi.org/10.3390/f9080490

Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Ghebleh Goydaragh, M., Schmidt, K., & Scholten, T. (2021). Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. *Geoderma* , *399* , 115108. https://doi.org/10.1016/j.geoderma.2021.115108

Tomppo, E., Haakana, M., Katila, M., & Peräsaari, J. (2008).*Multi-Source National Forest Inventory: Methods and Applications* . Springer Science & Business Media.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., & Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment* ,*112* (5), 1982–1999. https://doi.org/10.1016/j.rse.2007.03.032

Tomppo, E., Schadauer, K., McRoberts, R. E., Gschwantner, T., Gabler, K., & Ståhl, G. (2010). Introduction. In E. Tomppo, T. Gschwantner, M. Lawrence, & R. E. McRoberts (Eds.), *National Forest Inventories: Pathways for Common Reporting* (pp. 1–18). Springer Netherlands. https://doi.org/10.1007/978-90-481-3233-1_1

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology* , *6* (1). https://doi.org/10.2202/1544-6115.1309

Villaseñor, J., & Gual, M. (2014). *El bosque mesófilo de montaña y sus plantas con flores* (pp. 221–236).

Viscarra Rossel, R. A., Webster, R., Bui, E. N., & Baldock, J. A. (2014). Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change.*Global Change Biology* , *20* (9), 2953–2970. https://doi.org/10.1111/gcb.12569

Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling* , *457* , 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews* , *210* , 103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wadoux, A. M. J.-C., Walvoort, D. J. J., & Brus, D. J. (2022). An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. *Geoderma* , *405* , 115332. https://doi.org/10.1016/j.geoderma.2021.115332

Wang, G., Oyana, T., Zhang, M., Adu-Prah, S., Zeng, S., Lin, H., & Se, J. (2009). Mapping and spatial uncertainty analysis of forest vegetation carbon by combining national forest inventory data and satellite images.*Forest Ecology and Management* , *258* (7), 1275–1283. https://doi.org/10.1016/j.foreco.2009.06.056