Forest tree species distribution for Europe 2000–2020: mapping potential and realized distributions using spatiotemporal Machine Learning

Carmelo Bonannella (carmelo.bonannella@opengeohub.org)
OpenGeoHub https://orcid.org/0000-0002-5391-8427
Tomislav Hengl
OpenGeoHub https://orcid.org/0000-0002-9921-5129
Johannes Heisig
University of Münster https://orcid.org/0000-0003-3586-3001
Leandro Parente
OpenGeoHub https://orcid.org/0000-0003-1589-0467
Marvin N Wright
University of Bremen https://orcid.org/0000-0002-8542-6291
Martin Herold
Wageningen University & Research https://orcid.org/0000-0003-0246-6886
Sytze de Bruin
Wageningen University & Research https://orcid.org/0000-0002-6884-2832

Research Article

Keywords: Species distribution model, Ecological niche, Ensemble modeling, Machine learning, Imbalanced data, Presence-absence, Spatiotemporal modeling, Stacked generalization, High resolution, Tree species

Posted Date: June 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1252972/v3

Version of Record: A version of this preprint was published at PeerJ on July 25th, 2022. See the published version at https://doi.org/10.7717/peerj.13728.

Forest tree species distribution for Europe 2000–2020: mapping potential and realized distributions using spatiotemporal Machine Learning

Carmelo Bonannella[1, 2], Tomislav Hengl[1], Johannes Heisig[3], Leandro Parente[1], Marvin N Wright[4, 5], Martin Herold[2, 6], and Sytze de Bruin[2]

[1]**OpenGeoHub, Wageningen, The Netherlands**

[2]**Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen, The Netherlands**

[3]**Institute for Geoinformatics, University of Münster, Münster, Germany**

[4]**Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany**

[5]**University of Bremen, Bremen, Germany**

[6]**GFZ German Research Centre for Geosciences, Section 1.4 Remote Sensing and Geoinformatics, Telegrafenberg, Potsdam, 14473, Germany**

Corresponding author:

Carmelo Bonannella[1]

Email address: carmelo.bonannella@opengeohub.org

ABSTRACT

This paper describes a data-driven framework based on spatiotemporal machine learning to produce distribution maps for 16 tree species (*Abies alba* Mill., *Castanea sativa* Mill., *Corylus avellana* L., *Fagus sylvatica* L., *Olea europaea* L., *Picea abies* L. H. Karst., *Pinus halepensis Mill.*, *Pinus nigra J. F. Arnold*, *Pinus pinea* L., *Pinus sylvestris* L., *Prunus avium* L., *Quercus cerris* L., *Quercus ilex* L., *Quercus robur* L., *Quercus suber* L. and *Salix caprea* L.) at high spatial resolution (30 m). Tree occurrence data for a total of 3 million of points was used to train different algorithms: random forest, gradient-boosted trees, generalized linear models, k-nearest neighbors, CART and an artificial neural network. A stack of 305 coarse and high resolution covariates representing spectral reflectance, different biophysical conditions and biotic competition was used as predictors for realized distributions, while potential distribution was modelled with environmental predictors only. Logloss and computing time were used to select the three best algorithms to tune and train an ensemble model based on stacking with a logistic regressor as a meta-learner. An ensemble model was trained for each species: probability and model uncertainty maps of realized distribution were produced for each species using a time window of 4 years for a total of 6 distribution maps per species, while for potential distributions only one map per species was produced. Results of spatial cross validation show that the ensemble model consistently outperformed or performed as good as the best individual model in both potential and realized distribution tasks, with potential distribution models achieving higher predictive performances (TSS = 0.898, $R^2_{logloss}$ = 0.857) than realized distribution ones on average (TSS = 0.874, $R^2_{logloss}$ = 0.839). Ensemble models for *Q. suber* achieved the best performances in both potential (TSS = 0.968, $R^2_{logloss}$ = 0.952) and realized (TSS = 0.959, $R^2_{logloss}$ = 0.949) distribution, while *P. sylvestris* (TSS = 0.731, 0.785, $R^2_{logloss}$ = 0.585, 0.670, respectively, for potential and realized distribution) and *P. nigra* (TSS = 0.658, 0.686, $R^2_{logloss}$ = 0.623, 0.664) achieved the worst. Importance of predictor variables differed across species and models, with the green band for summer and the Normalized Difference Vegetation Index (NDVI) for fall for realized distribution and the diffuse irradiation and precipitation of the driest quarter (BIO17) being the most frequent and important for potential distribution. On average, fine-resolution models outperformed coarse resolution models (250 m) for realized distribution (TSS = +6.5%, $R^2_{logloss}$ = +7.5%). The framework shows how combining continuous and consistent Earth Observation time series data with state of the art machine learning can be used to derive dynamic distribution maps. The produced predictions can be used to quantify temporal trends of potential forest degradation and species composition change.

1 INTRODUCTION

Reforestation and forest restoration are considered key strategies for tackling climate change by enhancing CO_2 sequestration (Lefebvre et al., 2021; Domke et al., 2020; Nave et al., 2019). Under the European Green Deal and the European biodiversity strategy for 2030, the European Union has committed to plant at least 3 billion additional trees by 2030 (European Commission, 2021). At the same time, tree deaths due to bark beetle infestations and increased drought fueled by a warming climate have reduced the total forest area of Germany by 2.5% since 2018 (Popkin, 2021). Obtaining reliable information on forest tree species distribution in both space and time is now urgently required for stakeholders and decision-makers in order to develop effective forest management and adaptation strategies (Keenan, 2015).

Understanding the range, constraints and drivers of species distribution has always been a primary goal of ecology (Andrewartha and Birch, 1954). However, only with the advent of Geographical Information

Systems (GIS) and the usage of extensive digital maps of environmental variables were ecologists able to access powerful enough tools to study species distributions at landscape scales (Franklin, 1995). Progress in this direction has given rise to a new field called Species Distribution Modelling (SDM) (Franklin, 2010): maps of species ecological niches are made by associating values of different predictors to known locations of the target species and then used to predict distribution in geographic space where no field data for the target species is available. Commonly, SDMs rely on climatic or bioclimatic factors at a coarse spatial resolution ($\geq$1 km) while in the temporal dimension long time averages (30-–50 years) are often used (Iturbide et al., 2018a). Whilst forest distribution maps are often used to guide management decisions happening at local scales, the potential impact of differences in resolution of the predictor variables on the results is often overlooked (Porfirio et al., 2014). For conservation purposes, previous studies have shown how distribution maps with high spatial resolution ($<$ 100 m) and slightly lower prediction accuracy are actually more useful than coarser ($>$ 250 m) but more accurate maps (Manzoor et al., 2018; Guisan et al., 2013; Gottschalk et al., 2011; Prates-Clark et al., 2008). Therefore, even at the cost of overall map accuracy, finer spatial resolution maps are more valuable for practical use. For these reasons, Earth Observation (EO) data, and specifically the use of high spatial resolution data, have grown in use for SDM applications (Gelfand and Shirota, 2021; Pérez Chaves et al., 2018; Hefley and Hooten, 2016).

In addition to the clear need for finer spatial resolution mapping, there are similar needs to drive research towards producing finer temporal resolution mapping. This is due to the recent and relatively swift change in disturbance regimes and weather patterns, which are significantly altering the ecological niches of tree species on a temporal scale of less than a few decades instead of centuries. The impact of these changes on tree species has become more noticeable from year to year, with growth decline (Martinez del Castillo et al., 2022) and increased mortality rates (Senf et al., 2021, 2018) demonstrated in literature as already occurring across large forested areas. Including the temporal domain in tree species distribution studies is therefore fundamental to capture the temporal evolution of these change processes. However no general consensus has yet been reached on the influence of these new high spatial and temporal resolution data sources on SDM performances. The inclusion of spatiotemporal data sources in SDM studies requires taking an additional effort when choosing the appropriate modeling technique, a task that has proved to be difficult even with traditional spatial-only data sources (Elith and Graham, 2009), let alone when also attempting to include the temporal dimension.

Aside from spatial and temporal considerations of predictor variables and species observations, in the last decade ecologists have conducted hundreds of studies purely to determine which modeling methods best suit the needs of SDM. Model choices have thus far proven to be highly impactful, with distribution maps derived with different models from the same dataset leading to quite opposite conclusions (Araújo and New, 2007; Pearson et al., 2006). Inter-model variability in projections has been tackled using ensemble modeling, where numerous independent models are fit using a range of methods applied to the same input data while the outputs of the individual models are aggregated into the final prediction. Ensemble modeling is a solution to high model variance and it has been demonstrated that reducing variance also reduces the effect of model overfitting and extrapolation (Zhou, 2019). This is achieved at the cost of increased model complexity, reduced model interpretability, and increased computational time

(Zhou, 2019). As such, the few examples of ensemble modeling approaches that have been investigated for SDM applications are limited to *mean*, *median* and *weighted average* approaches (Hao et al., 2019). These approaches are intuitively simple to implement and interpret, and involve, in the first two cases, just taking the mean or median of the predictions of the individual models as the final prediction. The weighted average approach is similar but scales the predictions by weights assigned based on predictive performances of the models obtained from cross validation. A robust ensemble technique that, to our knowledge, has not been tested yet for SDM is *stacking* or *stacked generalization*. In this approach outputs made by the individual models are the inputs of a *meta-learner* (i.e. a model that learns from other models) which then produces the final prediction (Wolpert, 1992).

We tested this ensemble technique on European forest tree species distribution. There is no shortage of information on European tree species distribution: the European Atlas of Forest Tree species is among one of the largest data sources with information on forest tree species for Europe (San-Miguel-Ayanz et al., 2016). It describes in detail the autoecology of 76 different forest tree species and provides geographical information on each species in the form of chorological maps, probability of presence maps and maximum habitat suitability maps. Recently, the Atlas has been further expanded with future projections in different climatic scenarios (Mauri et al., 2022). While these predictions are certainly useful to determine potential species composition of European forests, new methods are now needed to deal with the more and more attention to reproducibility of studies (Fidler et al., 2017), increasing spatial and temporal resolution of predictor variables (Zhu et al., 2019) and availability of ecological *"big data"* (i.e. gathered by multiple sources such as sensors, cameras etc.) (Hampton et al., 2013). Furthermore, SDM studies use high-dimensional data which is often non-linear and does not meet assumptions of conventional statistical procedures (Zhang and Li, 2017). For this reason, and thanks to the exponential increase in computing power of the last decade (Gorelick et al., 2017), solutions such as Machine Learning (ML) algorithms have recently become very popular for SDM studies. ML tries to learn the relationship between the response and the predictors through the observation of dominant patterns (Breiman, 2001b). Contrary to traditional statistical models, no kind of ecological assumptions are explicitly embedded in ML algorithms: ML can be especially useful when dealing with data gathered without a specific and rigorous sampling design (Ij, 2018). ML algorithms have great potential to analyze the large amount of data available nowadays, enabling the mapping and monitoring of changes on multiple geographical scales in a timely manner through reproducible research (Gobeyn et al., 2019).

In this sense, the objectives of this study were (**a**) to test different ML algorithms to develop a framework for modeling species distribution in space-time, (**b**) to assess the importance of various sources of EO data on model performances for mapping tree species distributions and (**c**) to explore and quantify the specific importance of high resolution data on model performances.

2 MATERIALS AND METHODS

2.1 General workflow

We modeled potential and realized distribution for 16 forest tree species for continental Europe for the time period January 2000 – December 2020 using a spatio-temporal ML approach. The general workflow

used to derive the distribution maps is shown in Fig. 1. We modeled the potential distribution as a baseline to assess the importance of EO data sources: we used only environmental predictors (i.e. temperature, precipitation, wind speed, water vapor and topographical variables) and environmental absences (i.e. location with known environmental conditions not suitable for the target species, following the definition used by Lobo et al. (2010)) to produce a neutral model for baseline species potential.

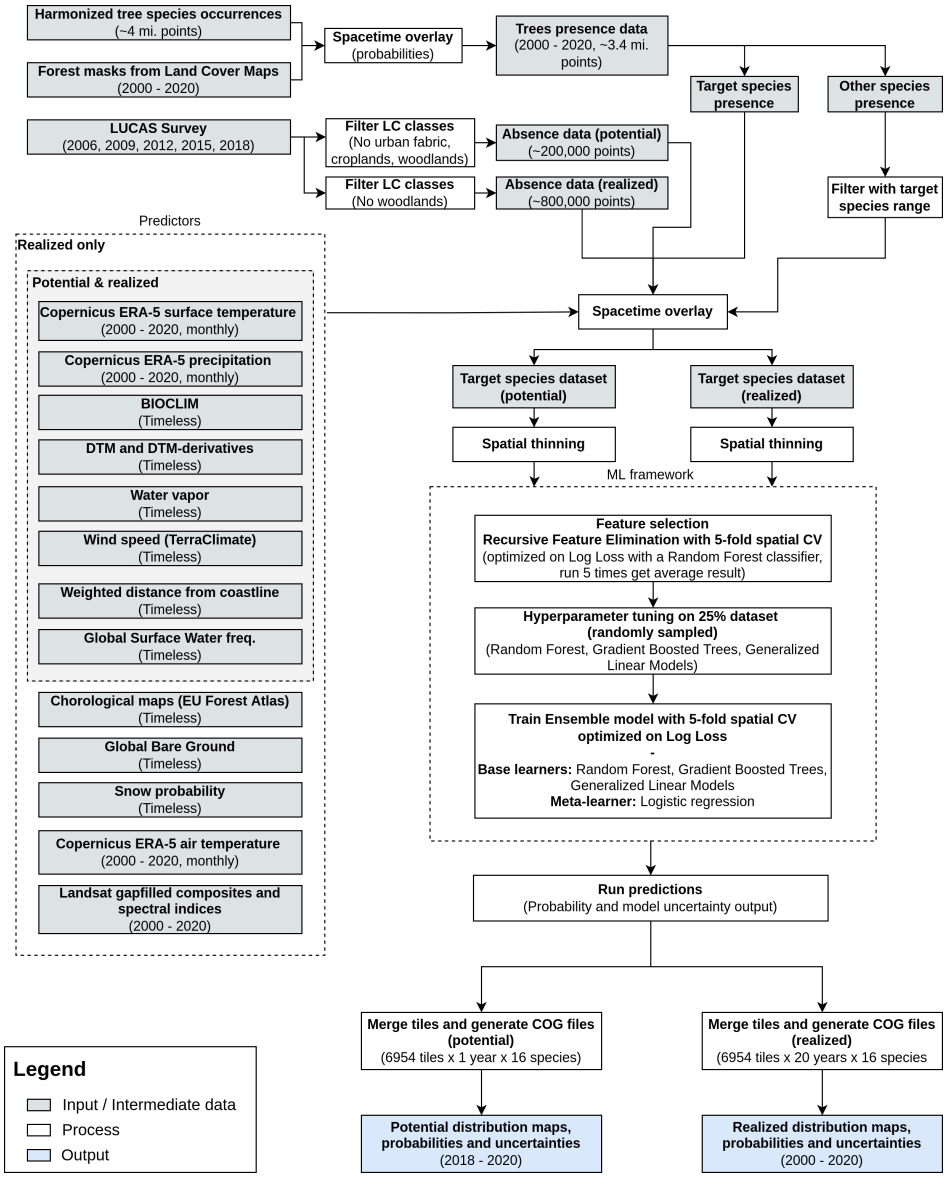

Figure 1. General workflow illustrating the preparation of the point data, the predictor variables used, model building (feature selection — hyperparameter optimization — training) and preparation of distribution maps for one species. The process was identically replicated for all the species.

As an additional source of homogeneously distributed true absence data we used the Land Use/Cover Area Survey (LUCAS) (EUROSTAT, 2017) dataset: in-situ observations of land use and land cover distributed on a 2×2 km grid covering the whole European Union (see d'Andrimont et al. (2021) for more information and https://ec.europa.eu/eurostat/web/lucas/data/lucas-grid for the official

grid). Final prediction maps show the probability of presence (0–100%) of at least one individual of the target species in the area covered by a 30 m pixel. Probability of presence is relative to the mapped target species, irrespective of the potential co-occurrence of other species in the same 30 m pixel and should not be confused with the absolute abundance or proportion of each species in the pixel area. The sum of the presence probabilities of different species in the same pixel can thus exceed 100%. We produced one potential distribution map and six realized distribution maps for each species: the assumption is that the conditions in the study area that determine the potential distribution of the species did not change over the time period analyzed; this does not hold for the realized distribution. We split the time period analyzed in six time windows according to the following scheme: (1) 2000–2002, (2) 2002–2006, (3) 2006–2010, (4) 2010–2014, (5) 2014–2018 and (6) 2018–2020.

To ensure transparent reporting and reproducibility, we described the dataset according to the ODMAP protocol suggested by Zurell et al. (2020). We implemented the workflow in the Python (Van Rossum and Drake, 2009) and R (R Core Team, 2021) programming languages. More technical details on preprocessing steps and packages used according to ODMAP (Zurell et al., 2020) are presented in Table S1 (found in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.1) (Bonannella et al., 2022).

2.2 Study area

The study area covers the European continent, that is all countries included in the Corine Land Cover (CLC) database (Büttner et al., 1998) except Turkey (Fig. 2). European forests cover 33% of the continent's land area. Owing to the variety of climatic conditions across both latitudinal and longitudinal gradients, twelve out of the 20 FAO Forest Ecological Zones are represented in European forests (de Rigo et al., 2016). The European Atlas of Forest Tree Species (San-Miguel-Ayanz et al., 2016) reports detailed information for a total of 76 forest tree species. From those, a selection of 16 were chosen and modelled in this study. The complete list of species is presented in Table S1.

2.3 Training points

2.3.1 Preparing and combining legacy occurrence points

A total of 2,454,997 tree species occurrence points from three different sources were gathered. The majority of points (71%) comes from the Global Biodiversity Information Facility (GBIF). National Forest Inventory (NFI) data from multiple EU member states published by Mauri et al. (2017) forms another 23% of the dataset. The remaining 6% comes from the LUCAS dataset.

Entries were filtered for species included in the European Atlas of Forest Tree Species (San-Miguel-Ayanz et al., 2016). Occurrences with a taxonomy rank other than species or genus were omitted. Same applies to points which had flags indicating serious location issues (i.e. missing coordinates). Geometries were re-projected to coordinate reference system ETRS89 / LAEA Europe (EPSG: 3035). A high resolution land mask for Europe (Hengl et al., 2020) was applied to further exclude misplaced occurrence points. GBIF taxon and genus keys were derived for the other two data sources. Quality flag variables for location accuracy and date were established from existing metadata to indicate potentially problematic entries. The harmonized point dataset has information on species and genus (including respective GBIF

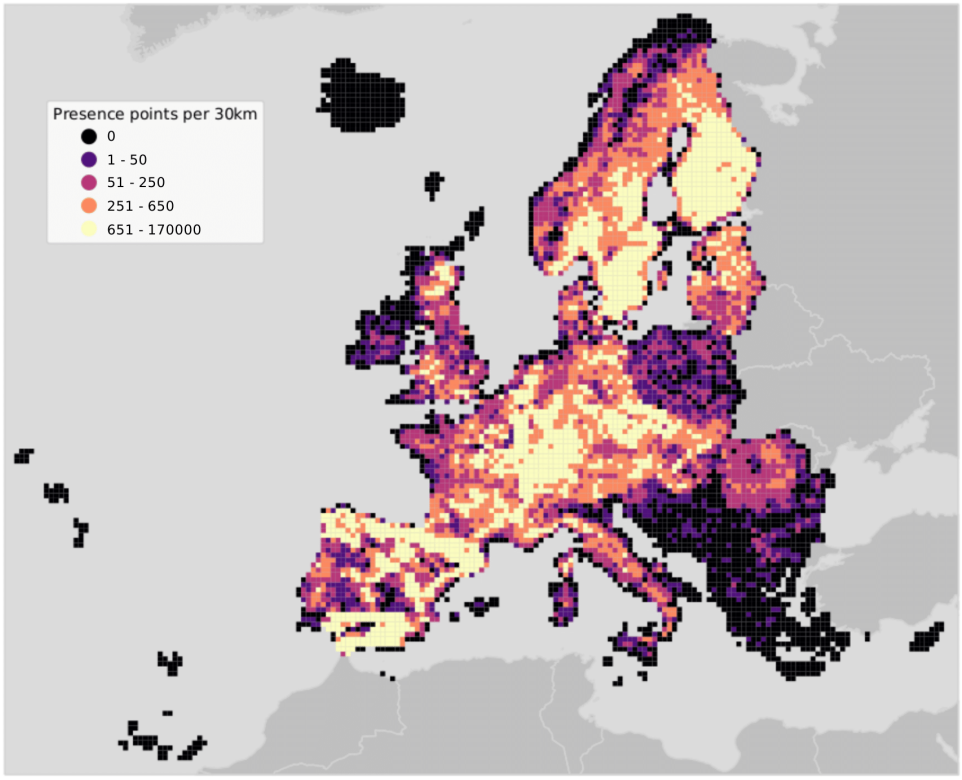

Figure 2. Map of the study area showing presence points only. Points are aggregated at a coarse resolution (30 km) scale and absence points are omitted for visualization purposes.

keys), year of observation, country, original data source, citation, and license among other auxiliary variables.The dataset was published separately (Heisig and Hengl, 2020).

We used yearly forest masks derived from Witjes et al. (2022) to decide upon including point data lacking the year of observation. Witjes et al. (2022) provides yearly probability maps at 30 m for the 2000–2020 for 43 land cover classes according to the CLC level 3 legend. We overlaid the points with the probability maps with prevalent forest (classes: 311, 312, 313 and 323) or woodland-shrub (324, 333) cover. Points were used only if the probability value extracted for at least one of the classes was $\geq$50% for all the years considered. Each unique combination of longitude, latitude and year was then considered as an independent sample. An additional quality flag was added to distinguish points coming from this operation and the points with original year of observation coming from source datasets.

2.3.2 Preparing non-occurrence points

A total of 883,630 land cover points were gathered from the LUCAS database as provided by Eurostat and used as absence data. All LUCAS survey data (2006, 2009, 2012, 2015 and 2019) was used: each survey was first downloaded individually and then aggregated. As for the occurrence points, spatial and temporal information were used to uniquely identify one observation. All main land cover classes were used for selecting observations for the absence dataset with the exception of class C (Woodland class),

as points belonging to that class already served the selection of presence data. For potential distribution only, points coming from human influenced land cover classes (class A and B) were also excluded. This choice was taken assuming that cities (class A) and croplands (class B) could be suitable areas for the target species if only environmental criteria are met. Two presence-absence datasets were produced for each species, one to be used for potential distribution and one for realized distribution. Locations in space and time of the target species were considered to be true presences, while presence locations of other species and observations from the LUCAS dataset were assumed to be the true absence locations. Presence locations of other species used as true absence were additionally filtered by overlaying them with a rasterized chorological map downloaded from the European Atlas of Forest Tree Species portal for each of the target species. Only points falling outside the geographical extent of the target species chorological map were used as absence locations for modeling.

2.3.3 Spatial thinning

Combining different data sources to generate the tree occurrence points produced a dataset with unknown sampling design, while LUCAS points are regularly distributed across the whole study area. To overcome the problem of uneven sampling intensity and spatial clustering, we applied a spatial thinning procedure using the *spThin* R package (Aiello-Lammens et al., 2015). A distance of 2 km was considered as minimum distance between the points, to harmonize the sampling intensity between presence and absence data. The procedure was repeated 10 times: at each iteration, the algorithm randomly removes one observation from the dataset until no observation is left with a nearest neighbor closer than the thinning distance. Among the 10 datasets obtained, the one with the largest number of records was retained and used for modeling. However, the package was not developed for large datasets: the implementation of the thinning algorithm cannot be processed in parallel and computation time can take even days with a number of observations ≥ 3000. Due to these computational constraints, we first overlaid the points with a 10×10 km grid and ran the thinning procedure per tile. Results of this operation are shown in Table S2 and Fig. S1 (found in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.2) (Bonannella et al., 2022).

2.4 Predictor variables

A total of 305 harmonized variables covering continental Europe at different spatial resolution were used as predictors to model the realized distribution of the species. In this study we included both dynamic (i.e. time-series of data of different temporal resolution) variables covering the time period January 2000 – December 2020 and static (i.e. variables not expected to change during the modelled time period) variables. A subset of only 103 variables were used instead to model the potential distribution (see Fig. 1). All data was reprojected in the coordinate reference system ETRS89 / LAEA Europe (EPSG: 3035) before the analysis.

2.4.1 Dynamic data

We used a reprocessed version of Landsat ARD data provided by Global Land Analysis and Discovery (GLAD) (Potapov et al., 2020): time series used in this study covers the period 1999–2020. Cloud and cloud shadow pixels were removed from the images, maintaining only the quality assessment-QA values

labeled as clear-sky. Afterwards, individual images were averaged by season according to three different quantiles (25th, 50th and 75th) and the following calendar dates for all periods:

- Winter: December 2 of previous year until March 20 of current year,
- Spring: March 21 until June 24 of current year,
- Summer: June 25 until September 12 of current year,
- Fall: September 13 until December 1 of current year.

84 images (3 quantiles × 4 seasons × 7 Landsat bands) were produced for each year. Missing values were imputed using the *Temporal Moving Window Median* algorithm. For more details on the preprocessing of Landsat data for this study see Witjes et al. (2022). 7 different spectral indices (see Table 1) were computed for each year and season using the 50th quantile only, for a total of 7 x 4 = 28 spectral indices variables per year.

Table 1. Table with Landsat-derived spectral indices used in this study.

Spectral Index	Abbreviation	Formula	Reference
Enhanced Vegetation Index	EVI	$2.5 \times \frac{NIR - RED}{NIR + 6 \times RED - 7.5 \times BLUE + 1}$	(Huete et al., 2002)
Enhanced Vegetation Index 2	EVI2	$2.5 \times \frac{NIR - RED}{NIR + 2.4 \times RED + 1}$	(Jiang et al., 2008)
Modified Soild Adjusted Vegetation Index	MSAVI	$\frac{(2 \times NIR + 1) - \sqrt{(2 \times NIR + 1)^2 - 8 \times (NIR - RED)}}{2}$	(Qi et al., 1994)
Normalized Burned Ratio	NBR	$\frac{NIR - SWIR2}{NIR + SWIR2}$	(Key and Benson, 1999)
Normalized Difference Vegetation Index	NDVI	$\frac{NIR - RED}{NIR + RED}$	(Tucker, 1979)
Normalized Difference Wetness Index	NDWI	$\frac{NIR - SWIR1}{NIR + SWIR1}$	(Gao, 1996)
Soil Adjusted Vegetation Index	SAVI	$(1 + 0.5) \times \frac{NIR - RED}{(NIR + RED + 0.5)}$	(Huete, 1988)

A reprocessing of the ERA5 Land hourly dataset has been used to have monthly aggregates of air temperature (2 meters above ground), surface temperature and precipitation. Original ERA5 data was aggregated to daily data, and subsequently to monthly data, with increased resolution (1 km) using CHELSA data (Karger et al., 2020): in this way the general spatial and temporal pattern of ERA5 Land dataset was kept while using the fine spatial detail coming from the CHELSA dataset. For air and surface temperature we obtained the monthly minimum, mean and maximum, while for precipitation the monthly sum for a total of 84 climatic time series layers.

2.4.2 Static covariate datasets

As additional static covariates, we used the following datasets:

- 19 bioclimatic variables (Hijmans et al., 2005) for the period 1979 - 2013 to provide a baseline of the actual state of the climate; we used bioclimatic variables from the CHELSA dataset since it has been claimed to better match data from meteorological stations than WorldClim (Karger et al., 2017),

- 50 chorological maps downloaded from the European Atlas of Forest Tree Species web portal. Chorological maps provide a qualitative overview of the spatial distribution of a species over an area, differentiating between native and introduced. We considered both the native and introduced areas as the potential distribution of a species for the time period covered by the study,
- Global bare ground cover from Hansen et al. (2013). The layer provides information on bare ground cover on a percent (1–100) scale,
- Solar direct and diffuse irradiation,
- 13 cloud fraction layers (monthly averages and annual average) derived from MODIS (Wilson and Jetz, 2016),
- Digital terrain model (DTM) for Europe (Hengl et al., 2020) and DTM-derived (slope, hillshade) variables,
- Easterness, northness (Olaya, 2009), and positive and negative openness,
- Probability of surface water occurrence at 30 m resolution derived from Landsat time series (Pekel et al., 2016),
- Height above nearest drainage (HAND) and flow accumulation area at 90 m resolution from the MERIT Hydro global hydrography datasets,
- Long-term flood hazard map calculated on a 500 years time period (Dottori et al., 2016),
- Water vapor pressure (kPa) based on the WorldClim2.1 dataset (Fick and Hijmans, 2017),
- Long-term snow probability based on MODIS (MOD10A2) and available at `https://doi.org/10.5281/zenodo.5774953`,
- Monthly wind speed (1998–2018) from TerraClimate.

For more details on spatial and temporal resolution, preprocessing and data sources see the supplementary material.

2.5 Feature selection

Features for potential and realized distribution for each species were selected using the Recursive Feature Eliminitation (RFE) strategy, implemented in the scikit-learn library (Pedregosa et al., 2011). For each combination of species and modelled distribution we trained a random forest classifier (num.trees = 50, default values were used for the other parameters): RFE fits the model and removes the weakest feature until a specified number of features is reached, then ranks the importance of the features based on the model's coefficients (for regression-based models) or feature importance (for random forest).

The minimum number of features was not known before hand: to select this number, we ran the Recursive Feature Elimination with a spatial 5–fold Cross Validation (RFECV), using the logarithmic loss, or logloss, as a scoring estimator. Logloss is one of the most robust performance metrics when it comes to imbalanced datasets (Ferri et al., 2009). Logloss is indicative of how close the predicted probability for

an observation i is to the corresponding label y. For binary classification with label $y \in 0, 1$ the overall logloss was calculated as:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \ln[p(y_i)] + (1 - y_i) \cdot \ln[1 - p(y_i)] \quad (1)$$

where N is the total number of observations and $p(y_i)$is the predicted probability for an observation with $y_i = 1$. It follows that values close to 0 indicate high prediction performances, with logloss = 0 being a perfect match, and values that are positive to infinite are progressively worse scores. For comparison, the value of logloss for random assignment depends on the number of classes (a) and the prevalence of the classes (b): for binary classification and a balanced (50:50) dataset with $N = 10$ observations, the equation (1) gives a value of 0.69.

We ran the RFECV on a 25% random subsample for each species and modelled distribution; this operation was replicated 5 times. For each iteration we selected the minimum of the logloss function (see Fig. S2 in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.3) and the averaged result was then used as the minimum number of features for the RFE.

2.6 Model building and evaluation

2.6.1 Modeling methods

To build an ensemble model, we decided to compare predictive performances and computing time (hyperparameter tuning — cross validation — prediction time) of different machine learning algorithms on a random 25% subset of observations for both potential and realized distribution datasets. A detailed workflow of this process is shown in Fig. S3 (in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.4). We decided to conduct this test on seven different species: choice of the species was based on the spatial distribution of the training points and the ratio between presence and absence points. In this way, algorithms performances could be tested on different ecological conditions (latitudinal and longitudinal gradient) and imbalance of classes. The species selected were: *A. alba*, *C. sativa*, *F. sylvatica*, *P. abies*, *P. halepensis* and *P. sylvestris*.

We compared seven different algorithms: Random Forests (RF) (Breiman, 2001a), Gradient-boosted trees (GBT) (Friedman, 2002), Classification trees (CART) (Therneau and Atkinson, 2011), Generalized Linear Models (Nelder and Wedderburn, 1972) with Lasso regularization (Tibshirani, 1996) (just GLM from now on), C5.0 (Quinlan, 1986), K-nearest neighbor (KNN) (Fix and Hodges, 1989) and Artificial Neural Network (ANN) (Ripley and Venables, 2017). Analyses were conducted using the *mlr* package (Bischl et al., 2016). For each algorithm, a hyperparameter space was defined: combinations of hyperparameters were generated per model based on a grid search of 5 steps per hyperparameter. More details on the hyperparameter space are available in Table S3 (found in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.5).

2.6.2 Selecting component models

We evaluated each combination of hyperparameters by comparing logarithmic loss values during a 5–fold spatial cross validation replicated 5 times: we used spatial instead of normal cross validation for hyperparameter tuning because it reduces overoptimistic performance results in the presence of strong data clustering (Schratz et al., 2019). We used the tile ID produced in the tiling system for Europe as the blocking parameter in the training function in *mlr*. All the compared algorithms were used in *"probability"* mode, that is, predicting for each observation in the dataset a probability value for presence (class 1) and absence (class 0). Besides the performance achieved in the logloss metric, computing time for the hyperparameter tuning, a 5–fold spatial cross validation and prediction time for a 30 km tile were also considered as additional criteria: we calculated the computing time only for the species that had the highest computational costs (*P. sylvestris*). This gave us an estimate of how long the process of training each component model could take during the building of the ensemble model. We used logloss performance as the first criteria to choose the component models: only in the case of two or more methods performing within one standard deviation from the average performance, we chose the computationally fastest.

2.6.3 Training ensemble model using stacking

Stacked generalization involves combining predictions made by level 0 models and using them as training data for a level 1 model (or meta-learner from now on) (Wolpert, 1992). To limit overfitting in the training data, we used a 5–fold spatial cross validation: the out-of-fold predictions were used to build a level 1 training dataset for the meta-learner. We used logistic regression as a meta-learner, which is usually the most used model for classification problems (Gomes et al., 2012). Final predictions are delivered as probability maps (0–100%) for presence together with model uncertainty maps: we consider as model uncertainty the standard deviation of the predicted values of the base learners. The principle is that the higher the standard deviation the more uncertain the model is regarding the right value to assign to the pixel (Brown et al., 2020).

2.6.4 Variable importance assessment

To assess to what extent the three level 0 models used different parts of the available feature space and the agreement between these models, we compared the variable importance when possible. For RF and CART we used Gini importance, for C5.0 the *"percentage of training set samples that fall into all the terminal nodes after the split"*(Quinlan, 1986), for GBT the gain metric (Shi et al., 2019) and for GLM the coefficients for the minimum fitted value of λ (Hastie et al., 2016). To better analyze the results, we aggregated the whole set of variables in 7 macro-classes:

- Climate (i.e. precipitation, wind speed, water vapor, snow probability etc.),
- Temperature (i.e. time series of recorded temperatures for the observed time period),
- Bioclim (i.e. bioclimatic variables from CHELSA),
- Topography (i.e. DTM and DTM-derivative variables),
- Landsat band (i.e. all percentiles, all seasons),

- Distribution (i.e. species distribution maps from European Atlas of Forest Tree Species),
- Spectral index (i.e. spectral indices derived from Landsat bands).

2.6.5 Model evaluation

Predictive performance of the ensemble model was assessed through spatial 5–fold cross-validation repeated 5 times (Roberts et al., 2017) with logloss as performance metric. To investigate if the ensemble model outperformed the component models, we compared results of the spatial cross validation of the ensemble with the results of the component models. To be able to compare performances between different species, we converted logloss performances used the following formula:

$$R^2_{\text{logloss}} = 1 - \frac{Logloss_{\text{m}}}{Logloss_{\text{r}}} \tag{2}$$

where $Logloss_m$ is the performance achieved by the model and $Logloss_r$ is the value for random logloss, used as a baseline for predictive performances. Values close to 1 indicate high predictive performances, while values close to 0 indicate lower performances, with 0 meaning that the model is no better than a random guess. We also reported a threshold-dependent metric, the True Skill Statistic (TSS) and a threshold independent metric, the area under ROC curve (AUC), as they are commonly used metric to evaluate SDMs predictive performances (Chakraborty et al., 2021; Shabani et al., 2018). TSS was computed using the default threshold value (0.5) when assigning predicted probabilities values to the presence or absence class. Logloss is one of the least sensitive metric to prevalence (Ferri et al., 2009), hence our choice of logloss as a primary performance metric to compare different models coming from different training datasets.

To assess the effect of high resolution products on predictive performances, we excluded Landsat bands and Landsat-derived spectral indices from the list of predictors used for realized distribution. We then applied our spatio-temporal machine learning framework (feature selection — hyperparameter tuning — ensemble model training) on each species and ran a 5–fold spatial cross validation repeated 5 times to evaluate model performances. For the ensemble model we used the same component models (RF, GBT and penalized GLM) and meta-learner (logistic regression). Results of this analysis were then compared with the performances achieved by the ensemble models using Landsat data.

3 RESULTS

3.1 Spatio-temporal machine learning framework

Table 2 shows that RF on average had the highest predictive performances for all species, with GLM coming closer. RF scored the lowest logloss among all the other algorithms in 9 cases out of 14 and scored the same as GLM in 1 case out of 14. In the remaining cases, GLM scored the lowest logloss value, with RF scoring the second lowest. On average, GLM performed better in in potential distribution tasks, with RF clearly outperforming every other algorithm in realized distribution tasks. Overall, GLM and RF always scored the lowest logloss values, from two to three times lower than all the other algorithms in some cases.

Table 2. Average logloss for the compared algorithms and for the subset of seven target species. In bold are highlighted the best performing learners for each task.

Species	Distribution	ANN	C5.0	GBT	GLM	KNN	RF	CART
A. alba	Potential	0.242±0.024	0.053±0.009	0.097±0.003	**0.027±0.003**	0.120±0.021	0.033±0.006	0.063±0.008
C. sativa	Potential	0.210±0.019	0.118±0.020	0.128±0.003	0.058±0.006	0.197±0.033	**0.057±0.008**	0.132±0.011
F. sylvatica	Potential	0.516±0.027	0.114±0.011	0.138±0.003	**0.055±0.004**	0.108±0.012	0.060±0.004	0.184±0.014
P. abies	Potential	0.390±0.017	0.176±0.010	0.199±0.005	0.144±0.006	0.314±0.022	**0.114±0.005**	0.292±0.009
P. halepensis	Potential	0.220±0.023	0.053±0.012	0.092±0.003	**0.019±0.002**	0.075±0.015	0.023±0.003	0.070±0.013
P. sylvestris	Potential	0.655±0.041	0.370±0.017	0.358±0.005	0.318±0.008	0.569±0.030	**0.232±0.006**	0.430±0.013
Q. robur	Potential	0.422±0.019	0.117±0.012	0.144±0.004	**0.065±0.004**	0.152±0.026	0.068±0.007	0.154±0.010
A. alba	Realized	0.383±0.030	0.145±0.017	0.140±0.007	0.106±0.009	0.245±0.059	**0.059±0.009**	0.151±0.021
C. sativa	Realized	0.316±0.025	0.175±0.031	0.161±0.010	0.118±0.017	0.351±0.065	**0.077±0.010**	0.173±0.017
F. sylvatica	Realized	0.654±0.055	0.118±0.010	0.147±0.005	0.129±0.007	0.209±0.039	**0.057±0.011**	0.200±0.025
P. abies	Realized	0.666±0.053	0.180±0.011	0.208±0.009	0.177±0.009	0.467±0.042	**0.125±0.008**	0.280±0.035
P. halepensis	Realized	0.292±0.034	0.065±0.014	0.104±0.004	**0.025±0.005**	0.074±0.022	**0.025±0.005**	0.084±0.013
P. sylvestris	Realized	0.656±0.043	0.451±0.013	0.473±0.009	0.478±0.009	0.776±0.052	**0.304±0.006**	0.549±0.012
Q. robur	Realized	0.642±0.049	0.105±0.016	0.133±0.005	0.091±0.011	0.200±0.057	**0.053±0.008**	0.200±0.030

The absolute difference between values scored by GLM and RF is lower than when RF had the advantage over GLM. This indicates a high reliability of RF performances even when other models outperform it. The ANN scored the highest logloss values in all tasks, so it was immediately excluded from the pool of level 0 models to choose from. It was time consuming to find a common hyperparameter range well suited for different tasks, since neural networks are often extremely situation-dependent. After a preliminary selection, we used the range shown in Table S3: despite that, our results remained inferior to those obtained with the other learners. On top of that, the *mlr* implementation of neural networks, based on the *deepnet* R package (Rong and Rong, 2014), doesn't allow the use of ReLU (rectified linear activation function) as an activation function, which would have been beneficial for our purposes. Based on logloss performances, we selected RF and GLM as the first two components of the ensemble. Based on similar values of logloss (within one standard deviation of the average performance) scored by C5.0, GBT, KNN and CART, we used computational costs to choose the third component model (Table 3).

Table 3. Hyperparameter tuning, cross validation and prediction time for each model and distribution task. Time values are reported in seconds. Tests were conducted in a parallel computing setup on a CPU server running 2 x Intel(R) Xeon(R) Gold 6248R - 3.00GHz (96 threads) with 504 GB RAM.

Distribution	Process	ANN	C5.0	GBT	GLM	KNN	RF	CART
Potential	Tuning	561.2	310.7	527.2	448.9	2433.6	104.4	576.5
Potential	Cross validation	57.2	44.7	192.5	620.4	356.9	240.2	26.5
Potential	Prediction	24.1	91.4	21.9	14.8	19272.9	35.5	15.4
Realized	Tuning	1031.6	964.2	688.4	859.1	12321.9	396.6	1650.2
Realized	Cross validation	119.2	165.4	290.8	1372.9	1445.1	805.5	114.9
Realized	Prediction	26.1	198.1	27.2	17.3	$>$ 1 day	52.4	17.4
	Total	**1819.2**	**1774.5**	**1748.0**	**1455.5**	**$>$ 1 day**	**1634.6**	**2400.9**

KNN was excluded due to computing time values being from one to two order of magnitude higher than the ones scored by the other models. Even though CART scores very low values in cross validation and prediction time in both potential and realized tasks, tuning time is the second highest, just behind KNN. C5.0 is faster than GBT in the whole potential workflow (446.8 seconds against 741.6) but slower in the realized workflow (1327.7 seconds against 1006.4). Considering both workflows, GBT proved to be faster and more consistent in cross validation and prediction time, showing an increase in tuning time of just 30% with double the amount of training data (see Table S2).

3.2 Variable importance

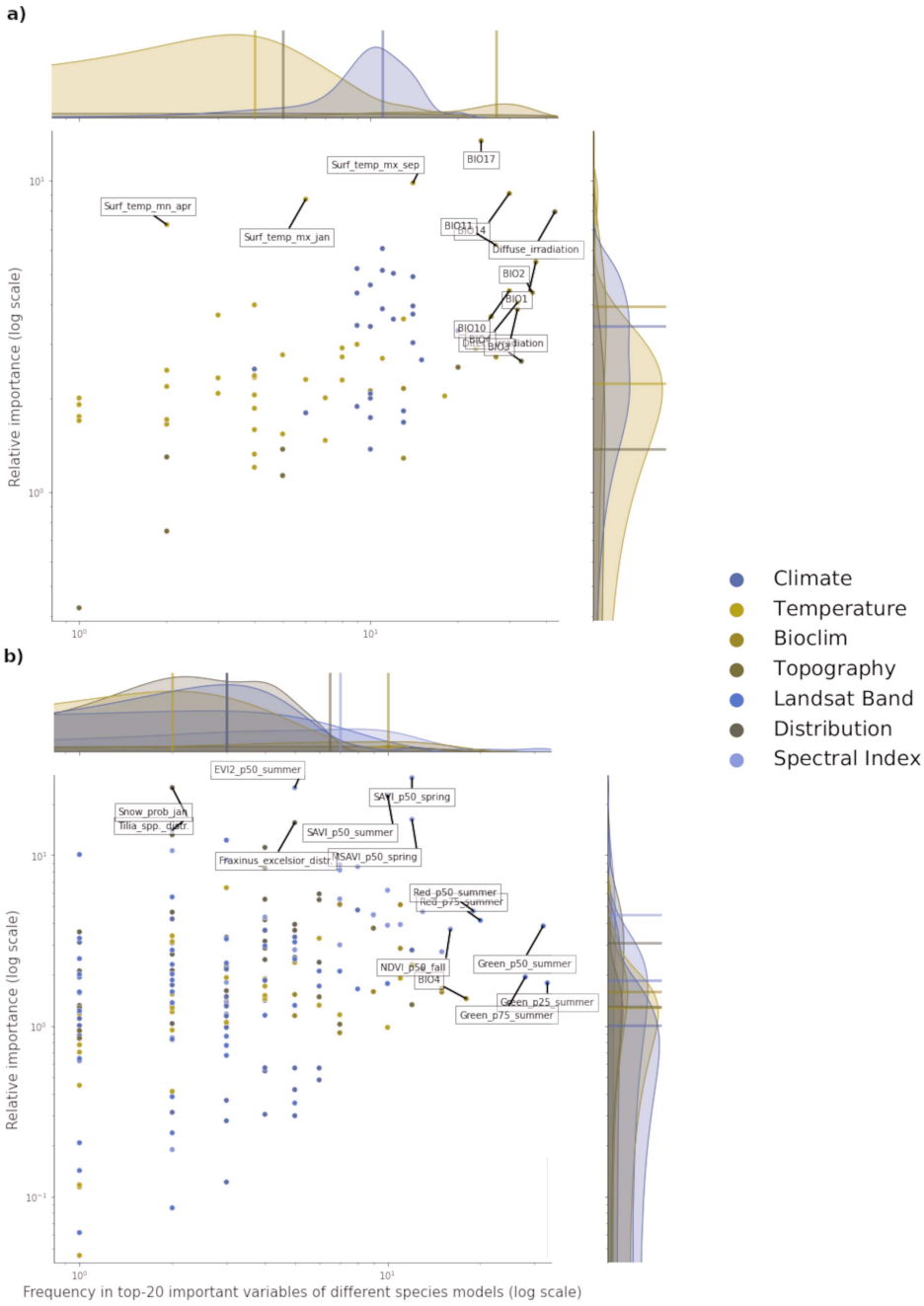

Figure 3. Relative variable importance vs frequency of the variables of the top–20 most important across the component models and all species for potential **(a)** and realized **(b)** distribution. Each plot can be divided in four quadrants, from the top left clockwise: variables with high relative importance but low frequency (i.e. important for one or few species), variables with high importance and high frequency (i.e. important for all species), variables with low importance and high frequency (i.e. they occured often but were not important) and variables with low importance and low frequency. Labeled dots are variables that recorded high values of relative variable importance or frequency.

Of all the features used in both potential and realized distribution, 60 are considered important for

both tasks. For potential distribution, diffuse irradiation, precipitation of the driest quarter (BIO17) and precipitation of the driest month (BIO14) were the most important and most frequent predictors across all component models and species (see Fig. 3). The density distributions per macroclass help understanding how the Bioclim macro-class was the one with on average both most important and most frequent variables. Other variables are more species-specific: the minimum surface temperature of April records the highest absolute value in relative importance but it was important for only one species (*Q. robur*, see Fig. S7 in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.7). The Temperature macro-class accounts the highest numbers of predictors, but the values recorded in both variable importance and frequency are the lowest among all the macro-classes. The Climate macro-class had the largest variety in predictors and variables in this class are homogeneously spread out across all the species in both variable importance and frequency.

For realized distribution, the summer aggregates of Landsat green (25th, 50th and 75th quantiles) were the three most important and most frequent variables across all models and species, closely followed by the summer aggregates of Landsat red and fall aggregates of NDVI (Fig. 3). While the Spectral Index macroclass clearly outperformed the other ones in relative importance, the Bioclim class scored as the most frequent across all the species. The distribution maps scored the highest values for variable importance (distribution of the *F. excelsior* and the *Tilia spp.*) but they were species-specific.

Overall, the component models show more differences in variable importance in the potential distribution models than in the realized ones. On average, RF and GBT selected the same variables in the top–10 but not always in the same order, while GLM tended to choose completely different variables (i.e. spectral indices for realized distributions and wind speed for potential distribution). This suggests how the ensemble models tend to use a wider proportion of the feature space than single models. This tendency is most apparent in the potential distribution models. In the realized distribution models, the component models agree in selecting the top-10 most important variables predictors from Landsat bands or Spectral indices. RF and GBT considered on average the Landsat bands as the most important, while GLM selected the spectral indices more often.

3.3 Accuracy assessment

Fig. 4 shows that on average the ensemble model outperformed all component models in both potential and realized distributions. AUC values seem to be overoptimistic and with low variability for all algorithms and distributions, with the largest interquartile range (IQR) being GLM - potential, going from 0.97 to 0.99. Values for TSS and R^2_{logloss} seem to be more conservative, with the ensemble still having the highest average (TSS = 0.898, 0.874 and R^2_{logloss} = 0.857, 0.839, respectively, for potential and realized distribution) values, and lowest IQR (TSS = 0.85 - 0.96, 0.82 - 0.92 and R^2_{logloss} = 0.82 - 0.93, 0.81 - 0.89, respectively, for potential and realized distribution).

While results from our modeling framework proved GLM and RF being the best models in both potential and realized tasks (see Table 2), GBT achieved overall better performances than both algorithms. In general, the models for potential distribution achieved better predictive performances than those for realized distribution; however potential distribution has greater IQRs as well as larger outliers.

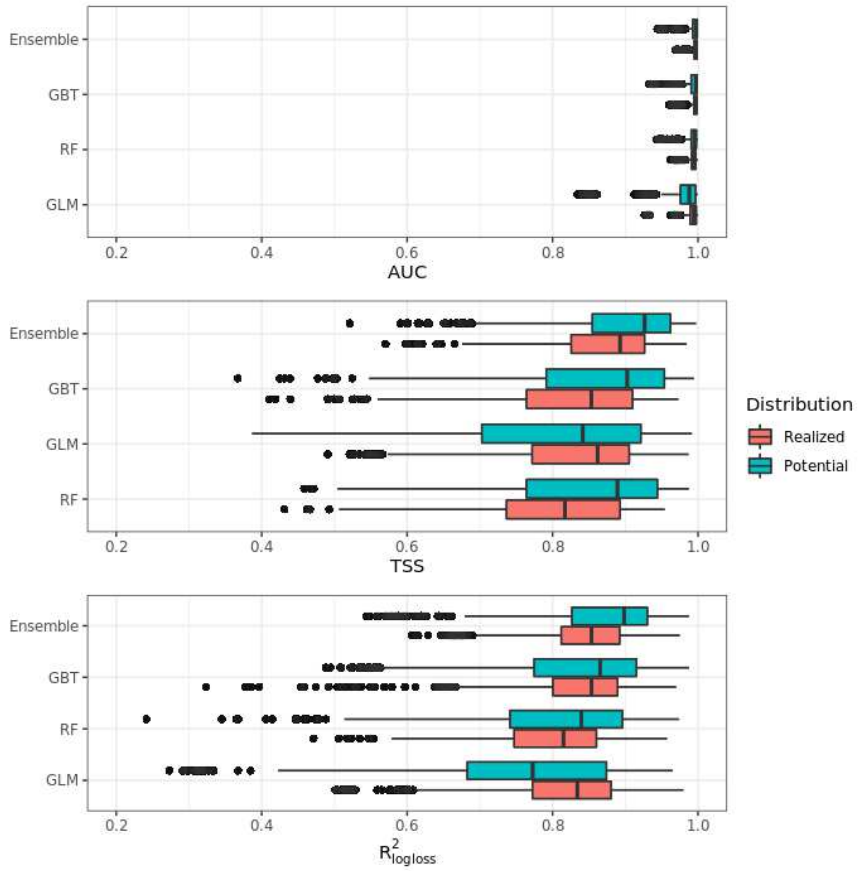

Figure 4. Aggregated results of the accuracy assessment per model and distribution expressed using AUC, TSS and $R^2_{logloss}$

Fig. 5 shows the performances of the ensemble model per species and distribution. Results for the component models are available in Table S4 and S5 (found in https://zenodo.org/record/6516728/preview/Supplementary_material.pdf#subsection.0.8). Following the trend shown in Fig. 4, differences in performances are minimal if we look at the AUC results, while they grow significantly if we look at the TSS and $R^2_{logloss}$ ones. We see that for both potential and realized distribution, models for *Q. suber* achieved the best performances (TSS = 0.968, 0.959 and $R^2_{logloss}$ = 0.952, 0.949, respectively, for potential and realized distribution), while *P. sylvestris* (TSS = 0.731, 0.785 and $R^2_{logloss}$ = 0.585, 0.670) and *P. nigra* (TSS = 0.658 0.686 and $R^2_{logloss}$ = 0.623, 0.664) achieved the worst.

Furthermore, while *Q. suber* model has overall best performances in all metrics, in both potential and realized distribution AUC grades *P. sylvestris* as the worst and *P. nigra* as the second worst; the opposite is true for TSS scores. For $R^2_{logloss}$, *P. sylvestris* scored as the worst in potential distribution and second worst in realized distribution, with the opposite happening for realized distribution.

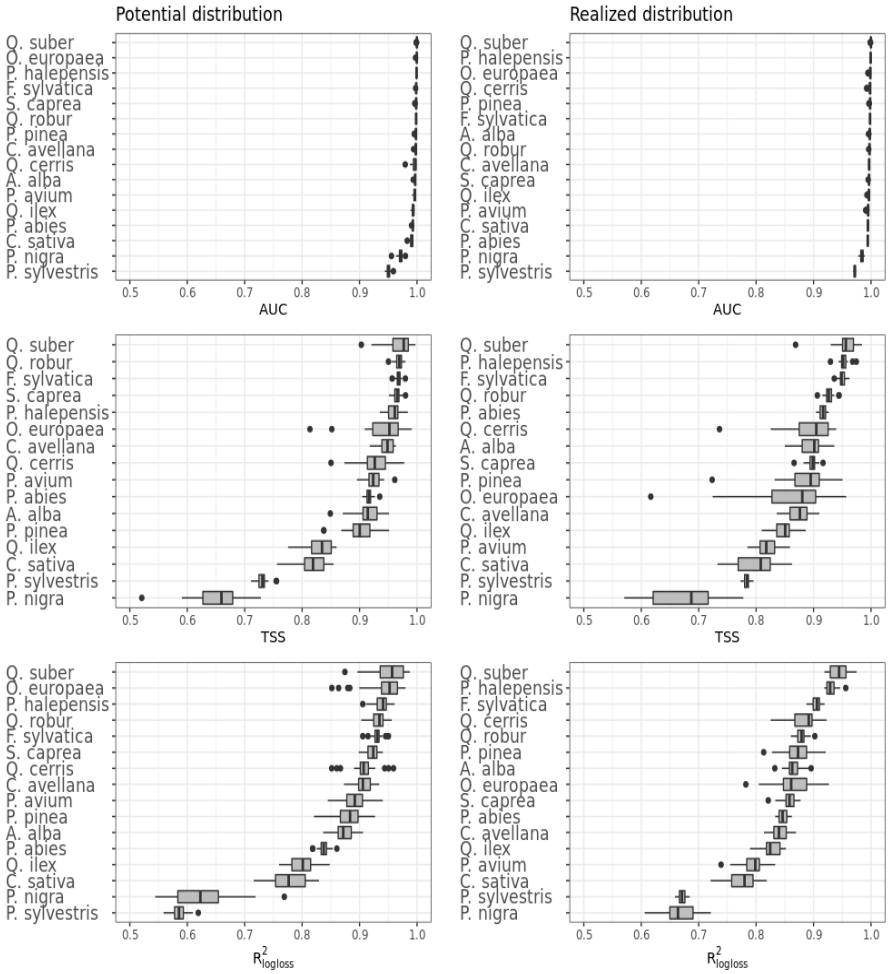

Figure 5. Results of the accuracy assessment per model and distribution for the ensemble model only expressed using AUC, TSS and $R^2_{logloss}$

3.4 Influence of high resolution on predictive performances

Fig. 6 shows that ensemble models for realized distribution including Landsat data consistently outperformed models without Landsat data. In all metrics, values scored by the Landsat models show higher median and average values than the ones without Landsat and lower IQR range. The same trend is shown by all metrics: like in the previous cases shown in section 3.3, AUC values are high, reaching 1 for some species, and with low IQR. TSS and $R^2_{logloss}$ show a larger IQR and lower median and average values than AUC: given the large differences in values scored by the models in TSS and $R^2_{logloss}$, these two metrics proved to be more helpful to discriminate model performances across multiple species. On average, including Landsat data increases TSS performances by 6.5% and $R^2_{logloss}$ by 7.5%, while when comparing median values the increase in performances is higher in TSS (+7.5%) and lower in $R^2_{logloss}$ (+7%).

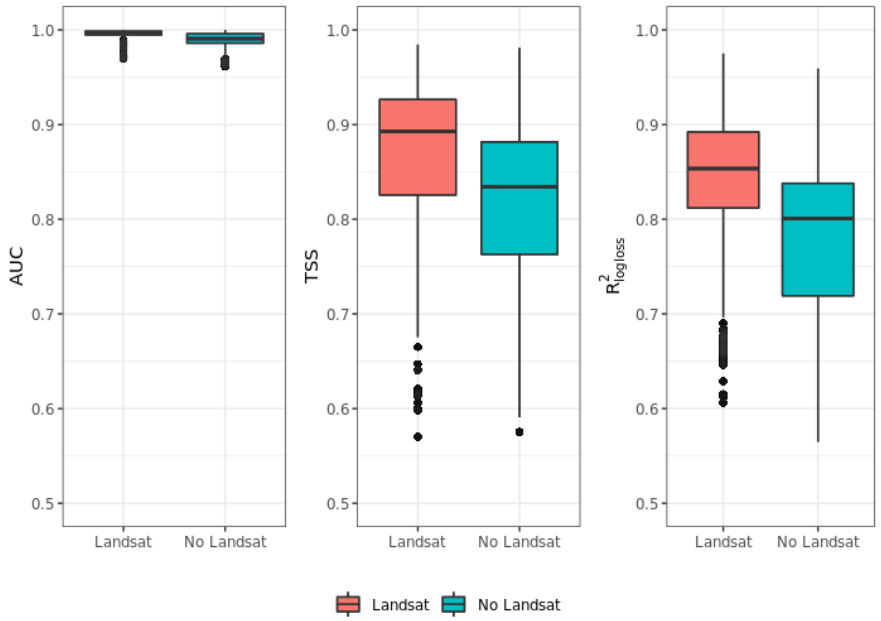

Figure 6. Aggregated results of the accuracy assessment for modeling realized distribution with and without the Landsat bands and spectral indices expressing using AUC, TSS and $R^2_{logloss}$

4 DISCUSSION

4.1 Modeling framework

Combining models using the ensemble approach is thought to reduce model uncertainty and increase its robustness in modeling species distributions (Araújo and New, 2007). We used ensemble with stacked generalization as ensemble approach, which has not been tested yet for species distribution modeling. We also trained the models in a spatio-temporal framework, expecting the models to generalize better when predicting in a temporal window not included in the training data.

Part of the intent of the paper was to provide a robust reproducible framework to model species distributions based on ensemble ML. Hao et al. (2020) used a similar methodological framework to the one used in this study. They modelled the distribution of 13 species of the genus *Eucalyptus* in South Australia and tested performances of ensemble model against individual models; they used mean and weighted average as ensemble strategies. They also tested cross validation versus spatial cross validation for model performances. The study doesn't specify which type of distribution was modelled: according to the definition provided in our study, we can compare their results with our potential distribution results. Their results show how spatial cross validation performances were more conservative than non spatial cross validation ones when compared with performances on independent validation sets. This supports and reinforces our use of spatial cross validation as a validation strategy for the modeling framework. Ensemble models performed well but were outperformed by untuned individual models and by a tuned GBT. There was also no clear advantage in predictive performances when using different ensemble strategies. This is in contrast with our results, where the ensemble based on stacking outperformed even tuned component models (13 cases out of 16), performed as good as the best component model (2 out 16)

and in just one case performed worse (tuned GBT was better than the ensemble). However, this is true only when comparing results from the Logloss and $R^2_{logloss}$: AUC and TSS both show that the ensemble outperformed or performed as good as tuned GBT in all cases and never performed worse. Given the very few occurrences in which the ensemble performed worse, this may be an indication of stacking being a better ensemble strategy when modeling species distribution. Valavi et al. (2021) reported an ensemble of tuned individual models as outperforming all other ML and regression based algorithms when benchmarking model performances on potential distribution of 225 different species; their results also show nonparametric techniques outperforming traditional regression methods. Among SDM studies focused on testing and comparing different SDM methodologies, the study from (Valavi et al., 2021) is also one of the few reporting computation time for all the models: this is a metric seldomly reported, but relevant when considering the optimal trade-off between accuracy and time, a well-known issue in the ML field (Hosseinzadeh et al., 2021).

In our cross validation estimates, AUC proved the least useful of the performance metrics used in this study, with low variability in AUC scores among different species and distributions despite the difference in predictor variables used and amount and source (i.e. LUCAS, other tree species) of absence data. A general trend shown by the other metrics is mostly picked up (i.e. ensemble superior to base models, models for *Q. suber* being the most accurate), but AUC scores are too similar to ascertain critical problems or possible artifacts in our models. Both TSS and the $R^2_{logloss}$ provided more useful metrics, showing models for *P. sylvestris* and *P. nigra* performing poorly compared to other models in both potential and realized distribution. Our results seem to agree with the ones from Chakraborty et al. (2021), who predicted current and future potential distribution of tree species over Europe using an ensemble framework based on averaging. We compared our results only with species present in both studies (*A. alba, F. sylvatica, P. abies, P. sylvestris, Q. robur*): final model AUC values are all $\geq$0.94 on both test set and external validation set, while TSS values start from 0.80.

RF and GLM are the best component models to map both potential and realized distributions when trained on a data sample, but GBT often outperforms RF or even the ensemble when tuned and trained on the whole dataset. In general, differences in predictive performances between the ensemble and the component models are also higher in potential distribution than in realized distribution. The list of variable importance per component model, species and task may give an insight to this: in the potential tasks, the component models use different parts of the feature space before the predictions are combined by the meta-learner. All the models select as most important variables for the task different predictors. For realized distribution tasks, the models all agree in selecting either Landsat bands or spectral indices as most important variables, resulting in predictions that are highly correlated and with less variance between the models.

Ensemble modeling is known to perform best when there is a high diversity between the base models and no or negative correlation between their outputs (Zhou, 2019). The introduction of Landsat bands and spectral indices in general greatly increased the predictive performances of the models for realized distribution compared to potential distribution models. However, this also homogenized predictions, which makes the second condition reported above not always respected. We separately compared the repeated spatial cross validation performances of ensemble and component models excluding the Landsat bands

and spectral indices. In this case, the ensemble never performed worse than the best component model. In general, if the ensemble provides predictive performances as good as or worse than the best component model, the best component model must be preferred (Zhang and Ma, 2012). However, ensemble models can still provide more advantages than individual models since they reduce model uncertainty and are more robust towards extrapolation (Mehra et al., 2019).

High resolution or hyperspectral data have not been used so far for SDMs but are extremely popular in tree species classification: the usage of such predictors has consistently increased over the years the predictive performances of ML tree species classifiers (Deur et al., 2020). Such data is not always available on a large spatial and temporal scale, so studies including these predictors usually cover a limited area compared to the one covered by our study. Bridging this gap may help having operational continental scale species distribution maps. Similarly, ML methods have mostly been used for tree species classification, where predictor variables such as temperature or precipitation are seldomly included and environmental variables not throughly considered, rather than for SDM. Despite that, we found in literature several studies which agree with our results: when classifying five (three broadleaves and two conifers) forest tree species in Portugal, Łoś et al. (2021) found out that GBT outperformed RF and KNN, reaching accuracy values $\geq$90% using Sentinel-2 reflectance bands only. Wessel et al. (2018) similarly reached high level of accuracy using only Sentinel-2 bands in German forests, but their best results were achieved using an object-based multitemporal Support Vector Machine (SVM) classifier: despite SVM being a very powerful ML method, it was was purposefully excluded from this study due to its computation intensity and lack of parallelization.

In our study, the ANN tested performed poorly, mostly due to the limited implementation options in the R environment of this method. Contrary to our results, Raczko and Zagajewski (2017) found ANN outperforming RF and SVM when including hyperspectral data; however, they also showed that ANN seemed to be the algorithm which predictions were strictly dependent from the dataset used, while RF and SVM showed more stable performances: the extreme sensibility of ANN to perturbations is a well known issue in the ML field (Colbrook et al., 2022). This issue is partially solved by Convolutional Neural Networks (CNNs), which have achieved considerable results when applied for SDM purposes, even when compared with ML methods such as GBT and RF. CNNs also showed to be particularly promising when commonly used remotely sensed predictor variables such as LiDAR (Light Detection and Ranging) and hyperspectral high resolution data are available (Zhang et al., 2020; Fricker et al., 2019). In some cases, CNNs have even outperformed tuned ML methods given their ability to grasp how local landscape structure affects prediction of species occurrence, in contrast with more conventional ML methods which cannot acknowledge the influence of environmental structure in local landscapes (Deneu et al., 2021; Sothe et al., 2020). The ML framework presented in this study could greatly benefit from the inclusion of CNNs.

4.2 Species distributions

Our cross-validation accuracy assessment results indicate high predictive performances for all species, in both potential and realized distributions. In the case of mapping potential distribution, diffuse irradiation and precipitation of the driest quarter (BIO17) are the most important predictors overall. These results are

partially in contrast with Dyderski et al. (2018), who modelled current and future potential distribution of 12 tree species over Europe. We compared our results only with species present in both studies (*A. alba, F. sylvatica, P. abies, P. sylvestris, Q. robur*): in their case, temperature-related bioclimatic variables (BIO1, BIO5, BIO7 and BIO10) were more important than precipitation-related bioclimatic variables. Few peer-reviewed studies have reported on the importance of predictors other than bioclimatic ones in shaping species' potential distributions. We found that, on average, each component model considers two or more predictors from the Bioclim macro-class among the top-10 most important variables to predict the potential distribution. Previous findings in literature have shown the importance of bioclimatic variables when modeling species distributions (Fourcade et al., 2018), but this may also be a consequence of bioclimatic variables and elevation being the most employed, if not the only, predictors in numerous SDM studies (Fois et al., 2018). Bucklin et al. (2015) compared the influence of different sets of environmental predictors on model performances, but the list of predictors used in the study included human influenced factors, so their results cannot be used to assess the driving factors for potential distributions. Even if our results show the bioclimatic variables as the most important predictors for potential distributions, further studies in this direction may be needed. The scale of the study may affect the importance of predictor variables: on a large scale, distribution may be influenced by macro environmental factors, while at a local scale, other environmental factors may limit distribution more significantly. Walthert and Meier (2017) and Weigel et al. (2019) proved that soil properties are more important than either bioclimatic or only climatic variables when modeling potential tree species distribution at, respectively, country and regional scale.

Variable importance confirms that Earth Observation layers such as the 25th, 50th and 75th quantile summer aggregates for the Landsat green and red band and the 50th quantile fall aggregates of NDVI are overall the most important layers for mapping realized distribution of species. The inclusion of Landsat data and derived spectral indices increases predictive performances and contains more detailed information on species distribution ranges. Importance of NDVI is well known since it is one of the most used proxies in vegetation studies such as biodiversity estimation (Madonsela et al., 2017; He et al., 2009), net primary productivity (Schloss et al., 1999) and land degradation (Easdale et al., 2018), phenology (Fawcett et al., 2021) and species composition changes (Wang et al., 2021). NDVI incorporates information from the red and the near-infrared (NIR) portion of the electromagnetic spectrum. Vegetation's behavior in this portion of the spectrum has long been used in vegetation mapping to distinguish between coniferous and deciduous tree species (Hoffer, 1984). The green band, although usually less important than the red and NIR band, has already proved useful in vegetation mapping to classify forest types (Gao et al., 2015), predict forest variables (stem volume, diameter and tree height) at species level (Astola et al., 2019) and forest biomass at community level (Nandy et al., 2017).

Comparing our results with chorological maps from the European Atlas of Forest Tree Species (San-Miguel-Ayanz et al., 2016), we can see that in general both potential and realized distribution correctly capture the species ranges. Overall, potential distribution maps show homogeneous patterns of high probability values for all target species, while realized distribution maps show very fragmented patterns. The realized distribution model helps discriminating the presence or absence of the species due to biotic or other external factors. A high geographical overlap between probability maps of realized distribution

of different species may reflect co-existence within the same forest stands and could help in clearly define forest communities.

However, our results have to be interpreted and analyzed carefully: contrary to process based models, correlative models describe the patterns, not the mechanisms, in the association between species occurrences and predictor variables; for this reason, correlative SDMs risk overlooking potentially important driving factors determining species distributions since they cannot distinguish between direct and indirect effects (Sirén et al., 2022). A known issue in this sense is the masking effect of abiotic factors on competition and predation: SDMs could estimate abiotic predictors as the most important for species abundance, even in those cases when distribution is strongly affected by competition (Godsoe et al., 2017) or when biotic interactions are strictly correlated with abiotic factors (Filazzola et al., 2020).

ML methods strongly depend from high quality datasets, so a considerable effort was spent in creating two different presence-absence datasets for each target species, one for potential and one for realized distribution: while the same geographical extent was used for both datasets, different rules were used to select true absence (LUCAS dataset) or pseudo-absence (other tree species occurrences) points. This modeling choice may be one of the causes of cross validation estimates for potential distribution being higher than for realized distribution. Restricting the study area from which true and pseudo absence points are collected reduces the applicability of the models for predictive purposes (Pearson and Dawson, 2003), with unpredictable effects on future projections (over- or under-prediction, see Thuiller et al. (2004)). On the other end, no spatial constraint leads to unwanted situations where larger scale differences rather than local ones are picked, leading to the infamous SDM case of "there-are-no-polar-bears-in-the-Sahara" (Lobo et al., 2010). Chefaoui and Lobo (2008) proved how best hypotheses for potential distribution are obtained using absence points that are placed farther apart than the ones needed for the best hypotheses on realized distribution, hence using the same study area for both distributions may have lead to an overstimation of presence in potential distribution. Furthermore, Lobo et al. (2010) proved that for realized distribution best practice would be to avoid the absences from nearest localities due to possible contamination with methodological absences; while this was not an issue when considering absences coming from the LUCAS dataset, using information of other tree species occurrences as pseudo-absence may have affected our models. The criteria used to select pseudo-absence occurrences in SDM represent a big challenge in SDM, such that the topic has been the focus of multiple studies in the last decade (Iturbide et al., 2018a,b; Senay et al., 2013).

4.3 High resolution contributions: is finer always better?

Bioclimatic variables available only at coarse spatial resolution were used as predictor variables in both potential and realized distribution. The Landsat bands and the spectral indices were not the only high resolution layers used in this study: terrain and terrain-derived predictors were also included at 30 m resolution. However, despite the terrain data high resolution, the tree species potential distribution patterns mostly reflect the original spatial resolution of the bioclimatic variables. Thus, climate influences species distribution at the European scale. Even though this might indicate that mapping potential distributions at high resolution may not be necessary, it can still be useful for different case studies. For example, comparing the difference, and hence mapping the gap, between potential and realized distribution at the

same fine scale may prove to be an invaluable tool for both forest managers and conservation planners that work on the local level.

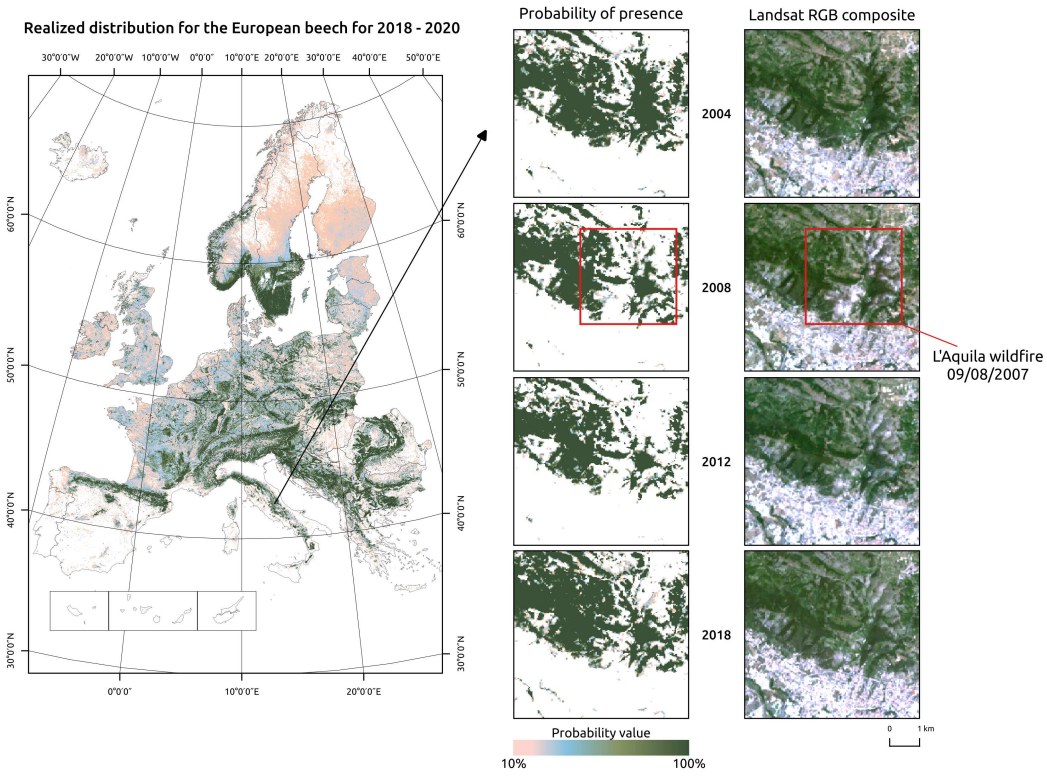

Figure 7. Realized distribution of *Fagus sylvatica* for the period 2018-2020. Detailed insets show a region around L'Aquila city, in Central Italy. The *Fagus sylvatica* forest on the northern outskirts of the city was affected by a serious wildfire in 2007. The realized distribution maps can be used to track compositional changes through time.

Potential distribution maps can be used to identify suitable areas for species in reforestation and restoration programs; realized distribution maps can inform the forest managers on the presence or absence of said species in those areas at a particular point in space and time (Fig. 7). By removing the biotic factors that limit the presence of the species in a potential reforestation site, using multiple distribution maps and including expert knowledge on species synecology, structurally complex forest stands could be planned and developed in a much more informed and data-driven way. A similar approach could be used by conservation planners. Potential distribution is modelled by studying the relationship between a species and the environmental conditions found in its native range, where the species is at equilibrium (Jiménez-Valverde et al., 2011). Invasive species are usually more abundant and productive in the introduced range than in their native ranges (Hierro et al., 2005). This is due to the absence of biotic factors that normally limit species distribution in their native range in the introduced range. Thus, a species that occupies only 10% of its potential distribution in its native range may end up occupying a bigger percentage of it in the introduced range. Estimation of potential distribution in the introduced range that depends only on environmental factors are conservative by definition, potential distribution maps may provide a good indication to conservation planners of how much the invasive species could

spread in the introduced range.

For realized distribution, including high resolution predictor variables in the model not only increases predictive performances but also lowers overall and local values of uncertainty. For forest management purposes, a large, consistent, standardized, long-term and high resolution image collection such as the one provided by the Landsat program can help extending in space and time information on tree species presence, composition and abundance. A spatial resolution of 30 m is particularly well suited for NFI applications: Strickland et al. (2020) derived probability maps of forest tree species for a 25 years time period (1985–2010) using yearly Landsat composites to extend missing information from the Canadian NFI and estimating changes in forest cover, species composition and forest disturbances. The increasing availability of even higher-spatial resolution satellite data from the European Copernicus program (i.e. Sentinel 1 and 2) and commercial providers (i.e. Planet) can potentially further enhance predictions by including more data and a better spatial matching of in-situ and satellite-derived information.

5 CONCLUSION

In this paper we have developed, tested and reported a methodological framework for predicting potential and realized distributions of 16 forest tree species using ensemble ML and analysis-ready EO data. In general, our ensemble models achieved better predictive performances than individual models when modeling both potential and realized distribution, while performing as good as the best individual model in the worst cases. Bioclimatic variables proved in general to be the most important and frequent predictors for potential distribution across Europe, mainly through precipitation-related predictors (BIO17 and BIO14) even at high resolution (i.e. on a local scale), while reflectance-based covariates were the most important predictors of the realized distributions. Overall, realized distribution proved to be more complex to map accurately than potential distribution and, among the species analyzed, distributions of specialist species proved easier to classify than pioneer species. In general, the ensemble and component models achieved better predictive performances for the potential distributions than for the realized distributions as judged from the cross-validation estimates. Our results indicate a consistent increase in predictive performances for realized distribution when adding high resolution data, especially Landsat data at 30 m resolution and spectral indices to the list of predictors. Significant findings of our work include: (**a**) distribution mapping for forest tree species can be efficiently automated to the level of full automation, but this assumes high quality / artifact free training points with a homogenous distribution of occurrence and absence points whenever possible; (**b**) complexity of ML methods can be significantly reduced by implementing efficient hyperparameter tuning and feature selection; (**c**) analysis-ready, high resolution reflectance time-series layers are maybe cumbersome to prepare and gap-fill for clouds and artifacts, but overall come as the most important inputs for maximizing predictive performances of realized tree species distribution.

We have released the maps and the code under open data / open source licenses to enable other similar research and to help speed up land restoration and reforestation projects in Europe. The code is publicly available in our GitLab repository at https://gitlab.com/geoharmonizer_inea/spatial-layers/-/tree/master/veg_mapping, while the datasets and predictions of tree species are available as Cloud-Optimized GeoTIFFs on Zenodo (see https://doi.org/10.5281/zenodo.

5818021) and can be displayed in 2D and 3D using the compare tool on the Open Data Science Europe viewer (see Fig. 8).

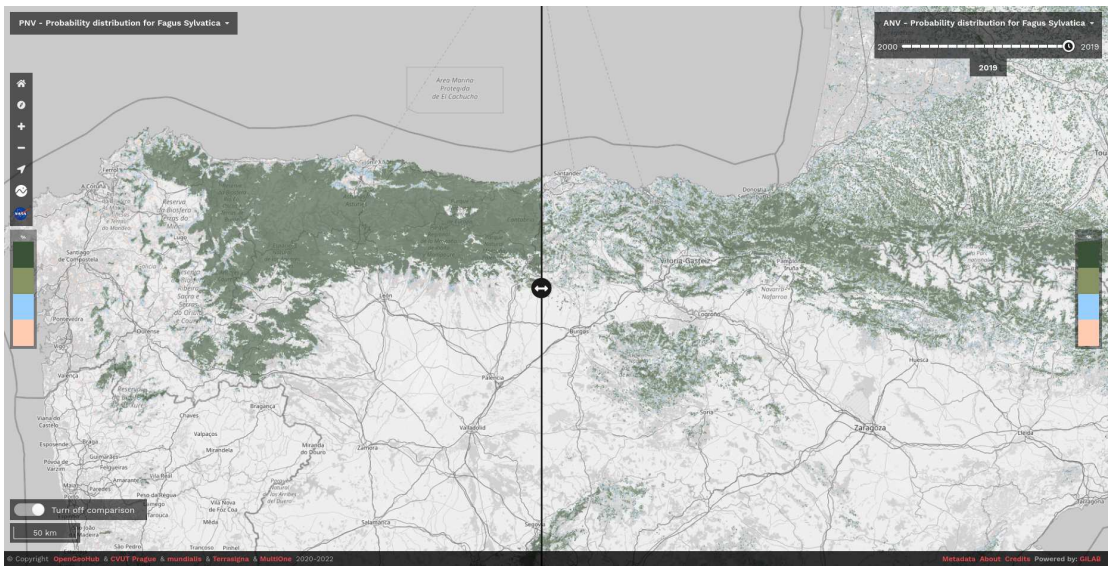

Figure 8. Difference between potential and realized distribution for *Fagus sylvatica* in Northern Spain for the period 2018–2020 visualized using slider in the Open Data Science Europe viewer (https://maps.opendatascience.eu). ©Copyright OpenGeoHub & CVUT Prague & mundialis & Terrasigna & MultiOne 2020–2022.

Even though we achieved high values of predictive performances, we still recognize many future areas of improvements. Given the importance of Landsat data for the results of this study, using a larger and higher resolution stack of reflectance-based predictors could help to improve precision of the predictions. A good example in this direction would be fusing all EO data currently available such as Harmonized Landsat Sentinel-2 (HLS) (Claverie et al., 2018) and eventually Sentinel 1 datasets. Hyper-spectral images (i.e. from future hyper-spectral missions such as ENMAP, see https://www.enmap.org/) are also proving to be useful to discriminate between different tree species and especially those that grow under dominant species (Fricker et al., 2019; Shen and Cao, 2017). As any ML-derived product, our predictions would benefit from having more and better quality data on tree species, in particular those that come from NFI plots: it is now crucial to have such data freely available to monitor processes such as species compositional changes, niche shifts, forest regrowth and degradation, as recently stated by Nabuurs et al. (2022). Exploring more sophisticated and different ML algorithms such as Deep Learning (DL) techniques (Lakshminarayanan et al., 2016) to our ensemble framework is also another area of improvement given the wide variety of applications these methods possess and the results obtained in comparison with other conventional ML algorithms (Choe et al., 2021; Deneu et al., 2021; Anand et al., 2021).

European forest dynamics, even though some recent results indicate increased mortality in European forests (Popkin, 2021; Senf et al., 2021, 2018), are probably among the least troubling in comparison to other continents. Our methodological framework could potentially be implemented at a global scale, and

possibly through Google Earth Engine (GEE) (van den Hoogen et al., 2021) or through the European Space Agency's OpenEO platform (https://openeo.cloud/) to produce high resolution (10–30 m) predictions of forest dynamics. Globally, there are many more tree species which are more important for forest management and monitoring. For example, South America as a whole has 4 times the amount of tree species present in Europe and 50% of all tree species on Earth (Cazzolla Gatti et al., 2022); in Brazil, it has been estimated that about 220 tree species cover most of the land and represent over 95% of the biomass (i.e. so called *"hyper-dominant species"* (Draper et al., 2021)). Scaling up the approach described in this paper to help producing objective predictions, to assist with monitoring forest dynamics and to support reforestation efforts globally is part of our next objectives.

ACKNOWLEDGEMENTS

We are grateful to the GiLAB company from Belgrade, Serbia for their support with processing and publishing produced data via the opendatascience.eu data portal. We are also grateful to the Geo-harmonizer project partners CVUT Prague, mundialis, Terrasigna & MultiOne for helping with processing all LUCAS ground observations and quality control.

CONFLICT OF INTEREST

The authors confirm no conflict of interest.

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). spthin: an r package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5):541–545.

Anand, A., Pandey, M. K., Srivastava, P. K., Gupta, A., and Khan, M. L. (2021). Integrating multi-sensors data for species distribution mapping using deep learning and envelope models. *Remote Sensing*, 13(16):3284.

Andrewartha, H. G. and Birch, L. C. (1954). *The distribution and abundance of animals.* The University of Chicago Press.

Araújo, M. B. and New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1):42–47.

Astola, H., Häme, T., Sirro, L., Molinier, M., and Kilpi, J. (2019). Comparison of Sentinel-2 and Landsat 8 imagery for forest variable prediction in boreal region. *Remote Sensing of Environment*, 223:257–273.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M. (2016). mlr: Machine learning in r. *Journal of Machine Learning Research*, 17(170):1–5.

Bonannella, C., Hengl, T., Heisig, J., Leal Parente, L., Wright, M., Herold, M., and de Bruin, S. (2022). Supplementary material for "Forest tree species distribution for Europe 2000 - 2020: mapping potential and realized distributions using spatiotemporal Machine Learning".

Breiman, L. (2001a). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Brown, K. E., Bhuiyan, F. A., and Talbert, D. A. (2020). Uncertainty quantification in multimodal ensembles of deep learners. In *The Thirty-Third International Flairs Conference*.

Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romanach, S. S., Speroterra, C., and Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and distributions*, 21(1):23–35.

Büttner, G., Steenmans, C., Bossard, M., Feranec, J., and Kolár, J. (1998). The European CORINE land cover database. *International Archives of Photogrammetry and Remote Sensing*, 32:633–638.

Cazzolla Gatti, R., Reich, P. B., Gamarra, J. G., Crowther, T., Hui, C., Morera, A., Bastin, J.-F., De-Miguel, S., Nabuurs, G.-J., Svenning, J.-C., et al. (2022). The number of tree species on earth. *Proceedings of the National Academy of Sciences*, 119(6):e2115329119.

Chakraborty, D., Móricz, N., Rasztovits, E., Dobor, L., and Schueler, S. (2021). Provisioning forest and conservation science with high-resolution maps of potential distribution of major european tree species under climate change. *Annals of Forest Science*, 78(2):1–18.

Chefaoui, R. M. and Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological modelling*, 210(4):478–486.

Choe, H., Chi, J., and Thorne, J. H. (2021). Mapping potential plant species richness over large areas with deep learning, modis, and species distribution models. *Remote Sensing*, 13(13):2490.

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., and Justice, C. (2018). The harmonized landsat and sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219:145–161.

Colbrook, M. J., Antun, V., and Hansen, A. C. (2022). The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and smale's 18th problem. *Proceedings of the National Academy of Sciences*, 119(12):e2107151119.

d'Andrimont, R., Verhegghen, A., Meroni, M., Lemoine, G., Strobl, P., Eiselt, B., Yordanov, M., Martinez-Sanchez, L., and van der Velde, M. (2021). Lucas copernicus 2018: Earth-observation-relevant in situ data on land cover and use throughout the european union. *Earth System Science Data*, 13(3):1119–1133.

de Rigo, D., Houston Durrant, T., Caudullo, G., and Barredo, J. I. (2016). European forests: an ecological overview. In *European Atlas of Forest Tree Species*, pages 24 – 31. Publication Office of the European Union, Luxembourg.

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., and Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 17(4):e1008856.

Deur, M., Gašparović, M., and Balenović, I. (2020). Tree species classification in mixed deciduous forests using very high spatial resolution satellite imagery and machine learning methods. *Remote Sensing*, 12(23):3926.

Domke, G. M., Oswalt, S. N., Walters, B. F., and Morin, R. S. (2020). Tree planting has the potential to increase carbon sequestration capacity of forests in the united states. *Proceedings of the National*

Academy of Sciences, 117(40):24649–24651.

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., and Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. *Advances in Water Resources*, 94:87–102.

Draper, F. C., Costa, F. R., Arellano, G., Phillips, O. L., Duque, A., Macía, M. J., Ter Steege, H., Asner, G. P., Berenguer, E., Schietti, J., et al. (2021). Amazon tree dominance across forest strata. *Nature ecology & evolution*, 5(6):757–767.

Dyderski, M. K., Paź, S., Frelich, L. E., and Jagodziński, A. M. (2018). How much does climate change threaten European forest tree species distributions? *Global Change Biology*, 24(3):1150–1163.

Easdale, M. H., Bruzzone, O., Mapfumo, P., and Tittonell, P. (2018). Phases or regimes? R evisiting NDVI trends as proxies for land degradation. *Land Degradation & Development*, 29(3):433–445.

Elith, J. and Graham, C. H. (2009). Do they? how do they? why do they differ? on finding reasons for differing performances of species distribution models. *Ecography*, 32(1):66–77.

European Commission (2021). New eu forest strategy for 2030. communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions.

EUROSTAT (2017). Land Cover/Use Statistics (LUCAS) Database.

Fawcett, D., Bennie, J., and Anderson, K. (2021). Monitoring spring phenology of individual tree crowns using drone-acquired ndvi data. *Remote Sensing in Ecology and Conservation*, 7(2):227–244.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315.

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., and Gordon, A. (2017). Metaresearch for Evaluating Reproducibility in Ecology and Evolution. *BioScience*, 67(3):282–289.

Filazzola, A., Matter, S. F., and Roland, J. (2020). Inclusion of trophic interactions increases the vulnerability of an alpine butterfly species to climate change. *Global change biology*, 26(5):2867–2877.

Fix, E. and Hodges, J. L. (1989). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238.

Fois, M., Cuena-Lombraña, A., Fenu, G., and Bacchetta, G. (2018). Using species distribution models at local scale to guide the search of poorly known species: Review, methodological issues and future directions. *Ecological Modelling*, 385:124–132.

Fourcade, Y., Besnard, A. G., and Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, 27(2):245–256.

Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in physical geography*, 19(4):474–499.

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press, Cambridge.

Fricker, G. A., Ventura, J. D., Wolf, J. A., North, M. P., Davis, F. W., and Franklin, J. (2019). A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral

imagery. *Remote Sensing*, 11(19):2326.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378. Nonlinear Methods and Data Mining.

Gao, B.-C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266.

Gao, T., Zhu, J., Zheng, X., Shang, G., Huang, L., and Wu, S. (2015). Mapping spatial distribution of larch plantations from multi-seasonal landsat-8 oli imagery and multi-scale textures using random forests. *Remote Sensing*, 7(2):1702–1720.

Gelfand, A. E. and Shirota, S. (2021). The role of odds ratios in joint species distribution modeling. *Environmental and Ecological Statistics*, 28(2):287–302.

Gobeyn, S., Mouton, A. M., Cord, A. F., Kaim, A., Volk, M., and Goethals, P. L. (2019). Evolutionary algorithms for species distribution modelling: A review in the context of machine learning. *Ecological Modelling*, 392(June 2018):179–195.

Godsoe, W., Franklin, J., and Blanchet, F. G. (2017). Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. *Ecology and evolution*, 7(2):654–664.

Gomes, C., Nocairi, H., Thomas, M., Ibanez, F., Collin, J.-F., and Saporta, G. (2012). Stacking prediction for a binary outcome. In *Compstat 2012*, pages 271–282, Limassol, Cyprus.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Gottschalk, T. K., Aue, B., Hotes, S., and Ekschmitt, K. (2011). Influence of grain size on species-habitat models. *Ecological Modelling*, 222(18):3403–3412.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan, T. J., Brotons, L., Mcdonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., and Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12):1424–1435.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., and Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162.

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., and Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853.

Hao, T., Elith, J., Guillera-Arroita, G., and Lahoz-Monfort, J. J. (2019). A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions*, 25(5):839–852.

Hao, T., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, 43(4):549–558.

Hastie, T., Qian, J., and Tay, K. (2016). An introduction to glmnet.

He, K. S., Zhang, J., and Zhang, Q. (2009). Linking variability in species composition and MODIS NDVI based on beta diversity measurements. *acta oecologica*, 35(1):14–21.

Hefley, T. J. and Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2):87–97.

Heisig, J. and Hengl, T. (2020). Harmonized Tree Species Occurrence Points for Europe. URL: https://zenodo.org/record/4061816, Dataset Version: 0.2.

Hengl, T., Leal Parente, L., Krizan, J., and Bonannella, C. (2020). Continental Europe Digital Terrain Model at 30 m resolution based on GEDI, ICESat-2, AW3D, GLO-30, EUDEM, MERIT DEM and background layers.

Hierro, J. L., Maron, J. L., and Callaway, R. M. (2005). A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *Journal of ecology*, 93(1):5–15.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 25(15):1965–1978.

Hoffer, R. (1984). Remote sensing to measure the distribution and structure of vegetation. *The Role of Terrestrial Vegetation in the Global Carbon Cycle: Measurement by Remote Sensing*, pages 131–59.

Hosseinzadeh, M., Wachal, A., Khamfroush, H., and Lucani, D. E. (2021). Optimal accuracy-time trade-off for deep learning services in edge computing systems. In *ICC 2021-IEEE International Conference on Communications*, pages 1–6. IEEE.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote sensing of environment*, 83(1-2):195–213.

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote sensing of environment*, 25(3):295–309.

Ij, H. (2018). Statistics versus machine learning. *Nat Methods*, 15(4):233.

Iturbide, M., Bedia, J., and Gutiérrez, J. M. (2018a). Background sampling and transferability of species distribution model ensembles under climate change. *Global and Planetary Change*, 166(March):19–29.

Iturbide, M., Bedia, J., and Gutiérrez, J. M. (2018b). Tackling uncertainties of species distribution model projections with package mopa. *R Journal*, 10(1).

Jiang, Z., Huete, A. R., Didan, K., and Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. *Remote sensing of Environment*, 112(10):3833–3845.

Jiménez-Valverde, A., Peterson, A. T., Soberón, J., Overton, J., Aragón, P., and Lobo, J. M. (2011). Use of niche models in invasive species risk assessments. *Biological invasions*, 13(12):2785–2797.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific data*, 4(1):1–20.

Karger, D. N., Dabaghchian, B., Lange, S., Thuiller, W., Zimmermann, N. E., and Graham, C. H. (2020). High resolution climate data for europe.

Keenan, R. J. (2015). Climate change impacts and adaptation in forest management: a review. *Annals of Forest Science*, 72(2):145–167.

Key, C. H. and Benson, N. C. (1999). The Normalized Burn Ratio (NBR): A Landsat TM radiometric measure of burn severity. *United States Geological Survey, Northern Rocky Mountain Science Center: Bozeman, MT, USA*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

Lefebvre, D., Williams, A. G., Kirk, G. J., Burgess, J., Meersmans, J., Silman, M. R., Román-Dañobeytia, F., Farfan, J., Smith, P., et al. (2021). Assessing the carbon capture potential of a reforestation project. *Scientific reports*, 11(1):1–10.

Lobo, J. M., Jiménez-Valverde, A., and Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114.

Łoś, H., Mendes, G. S., Cordeiro, D., Grosso, N., Costa, H., Benevides, P., and Caetano, M. (2021). Evaluation of xgboost and lgbm performance in tree species classification with sentinel-2 data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5803–5806. IEEE.

Madonsela, S., Cho, M. A., Ramoelo, A., and Mutanga, O. (2017). Remote sensing of species diversity using Landsat 8 spectral variables. *ISPRS Journal of Photogrammetry and Remote Sensing*, 133:116–127.

Manzoor, S. A., Griffiths, G., and Lukac, M. (2018). Species distribution model transferability and model grain size-finer may not always be better. *Scientific Reports*, 8(1):1–9.

Martinez del Castillo, E., Zang, C. S., Buras, A., Hacket-Pain, A., Esper, J., Serrano-Notivoli, R., Hartl, C., Weigel, R., Klesse, S., Resco de Dios, V., et al. (2022). Climate-change-driven growth decline of european beech forests. *Communications biology*, 5(1):1–9.

Mauri, A., Girardello, M., Strona, G., Beck, P. S., Forzieri, G., Caudullo, G., Manca, F., and Cescatti, A. (2022). Eu-trees4f, a dataset on the future distribution of european tree species. *Scientific data*, 9(1):1–12.

Mauri, A., Strona, G., and San-Miguel-Ayanz, J. (2017). EU-Forest, a high-resolution tree occurrence dataset for Europe. *Scientific Data*, 4(1):160123.

Mehra, A., Tripathy, P., Faridi, A., and Chinmay, A. (2019). Ensemble learning approach to improve existing models. *International Journal of Innovative Science and Research Technology*, 4.

Nabuurs, G.-J., Harris, N., Sheil, D., Palahi, M., Chirici, G., Boissière, M., Fay, C., Reiche, J., and Valbuena, R. (2022). Glasgow forest declaration needs new modes of data ownership. *Nature Climate Change*, pages 1–3.

Nandy, S., Singh, R., Ghosh, S., Watham, T., Kushwaha, S. P. S., Kumar, A. S., and Dadhwal, V. K. (2017). Neural network-based modelling for forest biomass assessment. *Carbon Management*, 8(4):305–317.

Nave, L. E., Walters, B. F., Hofmeister, K., Perry, C. H., Mishra, U., Domke, G. M., and Swanston, C. (2019). The role of reforestation in carbon sequestration. *New Forests*, 50(1):115–137.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Olaya, V. (2009). Chapter 6 basic land-surface parameters. In Hengl, T. and Reuter, H. I., editors, *Geomorphometry*, volume 33 of *Developments in Soil Science*, pages 141–169. Elsevier.

Pearson, R. G. and Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of

species: are bioclimate envelope models useful? *Global ecology and biogeography*, 12(5):361–371.

Pearson, R. G., Thuiller, W., Araújo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P., and Lees, D. C. (2006). Model-based uncertainty in species range prediction. *Journal of biogeography*, 33(10):1704–1711.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422.

Pérez Chaves, P., Ruokolainen, K., and Tuomisto, H. (2018). Using remote sensing to model tree species distribution in Peruvian lowland Amazonia. *Biotropica*, 50(5):758–767.

Popkin, G. (2021). Germany's trees are dying. a fierce debate has broken out over how to respond. *Science*, 374.

Porfirio, L. L., Harris, R. M., Lefroy, E. C., Hugh, S., Gould, S. F., Lee, G., Bindoff, N. L., and Mackey, B. (2014). Improving the use of species distribution models in conservation planning and management under climate change. *PLoS ONE*, 9(11):1–21.

Potapov, P., Hansen, M. C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., and Ying, Q. (2020). Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sensing*, 12(3):426.

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

Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., and Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote sensing of environment*, 48(2):119–126.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

R Core Team (2021). R: A language and environment for statistical computing.

Raczko, E. and Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral apex images. *European Journal of Remote Sensing*, 50(1):144–154.

Ripley, B. and Venables, W. (2017). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-12.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rong, X. and Rong, M. X. (2014). Package 'deepnet'.

San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., and Mauri, A. (2016). European Atlas of Forest Tree Species.

Schloss, A., Kicklighter, D., Kaduk, J., Wittenberg, U., and Intercomparison, T. P. O. T. P. N. M. (1999). Comparing global models of terrestrial net primary productivity (NPP): comparison of NPP to climate and the Normalized Difference Vegetation Index (NDVI). *Global Change Biology*, 5(S1):25–34.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., and Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120.

Senay, S. D., Worner, S. P., and Ikeda, T. (2013). Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE*, 8(8).

Senf, C., Pflugmacher, D., Zhiqiang, Y., Sebald, J., Knorn, J., Neumann, M., Hostert, P., and Seidl, R. (2018). Canopy mortality has doubled in europe's temperate forests over the last three decades. *Nature Communications*, 9(1):1–8.

Senf, C., Sebald, J., and Seidl, R. (2021). Increasing canopy mortality affects the future demographic structure of europe's forests. *One Earth*, 4(5):749–755.

Shabani, F., Kumar, L., and Ahmadi, M. (2018). Assessing accuracy methods of species distribution models: AUC, Specificity, Sensitivity and the True Skill Statistic. *Global Journal of Human Social Science*, 18(1):6–18.

Shen, X. and Cao, L. (2017). Tree-species classification in subtropical forests using airborne hyperspectral and lidar data. *Remote Sensing*, 9(11):1180.

Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., and Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129:170–179.

Sirén, A. P., Sutherland, C. S., Karmalkar, A. V., Duveneck, M. J., and Morelli, T. L. (2022). Forecasting species distributions: Correlation does not equal causation. *Diversity and Distributions*.

Sothe, C., De Almeida, C., Schimalski, M., La Rosa, L., Castro, J., Feitosa, R., Dalponte, M., Lima, C., Liesenberg, V., Miyoshi, G., et al. (2020). Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. *GIScience & Remote Sensing*, 57(3):369–394.

Strickland, G. E. I., Luther, J. E., White, J. C., Wulder, M. A., Strickland, G. E. I., Luther, J. E., White, J. C., and Wulder, M. A. (2020). Extending Estimates of Tree and Tree Species Presence-Absence through Space and Time Using Landsat Composites. *Canadian Journal of Remote Sensing*, 46(5):567–584.

Therneau, T. M. and Atkinson, E. J. (2011). An Introduction to Recursive Partitioning Using the RPART Routines. *Mayo clinic*, 61:33.

Thuiller, W., Brotons, L., Araújo, M. B., and Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, 27(2):165–172.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, 8(2):127–150.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., and Elith, J. (2021). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 0(0):1–27.

van den Hoogen, J., Robmann, N., Routh, D., Lauber, T., van Tiel, N., Danylo, O., and Crowther, T. W. (2021). A geospatial mapping pipeline for ecologists. *BioRxiv*.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Walthert, L. and Meier, E. S. (2017). Tree species distribution in temperate forests is more influenced by soil than by climate. *Ecology and Evolution*, 7(22):9473–9484.

Wang, H., Liu, H., Huang, N., Bi, J., Ma, X., Ma, Z., Shangguan, Z., Zhao, H., Feng, Q., Liang, T., et al. (2021). Satellite-derived NDVI underestimates the advancement of alpine vegetation growth over the past three decades.

Weigel, R., Gilles, J., Klisz, M., Manthey, M., and Kreyling, J. (2019). Forest understory vegetation is more related to soil than to climate towards the cold distribution margin of european beech. *Journal of Vegetation Science*, 30(4):746–755.

Wessel, M., Brandmeier, M., Tiede, D., Seitz, R., and Straub, C. (2018). Comparison of different machine-learning algorithms for tree species classification based on sentinel data. In *PFGK18*.

Wilson, A. M. and Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS biology*, 14(3):e1002415.

Witjes, M., Parente, L., van Diemen, C. J., Hengl, T., Landa, M., Brodsky, L., Halounova, L., Krizan, J., Antonic, L., Ilie, C. M., et al. (2022). A spatiotemporal ensemble machine learning framework for generating land use/land cover time-series maps for Europe (2000–2019) based on LUCAS, CORINE and GLAD Landsat. *PeerJ*.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Zhang, B., Zhao, L., and Zhang, X. (2020). Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images. *Remote Sensing of Environment*, 247:111938.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

Zhang, J. and Li, S. (2017). A review of machine learning based species' distribution modelling. In *2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, pages 199–206. IEEE.

Zhou, Z.-H. (2019). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

Zhu, Z., Wulder, M. A., Roy, D. P., Woodcock, C. E., Hansen, M. C., Radeloff, V. C., Healey, S. P., Schaaf, C., Hostert, P., Strobl, P., et al. (2019). Benefits of the free and open Landsat data policy. *Remote Sensing of Environment*, 224:382–385.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., Yates, K. L., Zimmermann, N. E., and Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, 43(9):1261–1277.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.pdf