A peer-reviewed version of this preprint was published in PeerJ on 22 August 2018.

View the peer-reviewed version (peerj.com/articles/5457), which is the preferred citable publication unless you specifically need to cite this preprint.

Hengl T, Walsh MG, Sanderman J, Wheeler I, Harrison SP, Prentice IC. 2018. Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. PeerJ 6:e5457 https://doi.org/10.7717/peerj.5457

PeerJ Preprints NOT PEER-REVIEWED

Global mapping of potential natural vegetation: an assessment of Machine Learning algorithms for estimating land potential

Tomislav Hengl Corresp., [1] , **Markus G Walsh** [2] , **Jonathan Sanderman** [3] , **Ichsani Wheeler** [4] , **Sandy P Harrison** [5] , **Iain C Prentice** [6]

[1] Envirometrix Ltd, Wageningen, The Netherlands
[2] The Earth Institute, Columbia University, New York, United States
[3] Woods Hole Research Center, Falmouth, United States
[4] Envirometrix, Wageningen, The Netherlands
[5] University of Reading, Reading, United Kingdom
[6] Department of Life Sciences and Grantham Institute - Climate Change and the Environment, Imperial College London, London, United Kingdom

Corresponding Author: Tomislav Hengl
Email address: tom.hengl@gmail.com

Potential Natural Vegetation (PNV) is the vegetation cover in equilibrium with climate, that would exist at a given location non-impacted by human activities. PNV is useful for raising public awareness about land degradation and for estimating land potential. This paper presents results of assessing Machine Learning Algorithms (MLA) for operational mapping of Potential Natural Vegetation (PNV). The following MLA were considered: neural networks (nnet package), random forest (ranger), gradient boosting (gmb), K-nearest neighborhood (class) and cubist. Three case studies were considered: (1) global distribution of biomes based on the BIOME 6000 data set (8057 modern pollen-based site reconstructions), (2) distribution of forest tree taxa in Europe based on detailed occurrence records (1,546,435 ground observations), and (3) global monthly Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) values (30,301 randomly-sampled points). A stack of 160 global maps representing biophysical conditions over land, including atmospheric, climatic, relief and lithologic variables, were used as explanatory variables. The overall results show that random forest gives the overall best performance. The highest accuracy for predicting BIOME 6000 classes (20) was estimated at 68% (33% with spatial Cross Validation) with the most important predictors being total annual precipitation, monthly temperatures and bioclimatic layers. Predicting forest tree species (73) resulted in mapping accuracy of 25%, with the most important predictors being monthly cloud fraction, mean annual and monthly temperatures and elevation. Regression models for FAPAR (monthly images) gave an R-square of 90% with most important predictors being total annual precipitation, monthly cloud fraction, CHELSA bioclimatic layers and month of the year, respectively.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018

PeerJ Preprints NOT PEER-REVIEWED

Further developments of PNV mapping could include using GBIF records to map global distribution of plant species at different taxonomic levels. This methodology could also be extended to dynamic modeling of PNV, so that future climate scenarios can be incorporated. Global maps of biomes, FAPAR and tree species at 1 km spatial resolution are available for download via http://dx.doi.org/10.7910/DVN/QQHCIK.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018

PeerJ Preprints NOT PEER-REVIEWED

Global Mapping of Potential Natural Vegetation: An Assessment of Machine Learning Algorithms for Estimating Land Potential

Tomislav Hengl[1], Markus G. Walsh[2], Jonathan Sanderman[3], Ichsani Wheeler[1], Sandy P. Harrison[4], and I. Colin Prentice[5]

[1]**Envirometrix Ltd., Wageningen, the Netherlands**
[2]**The Earth Institute, Columbia University, USA / Selian Agricultural Research Inst., Arusha, Tanzania**
[3]**Woods Hole Research Center, MA USA**
[4]**School of Archeology, Geography and Environmental Science, University of Reading, UK**
[5]**AXA Chair of Biosphere and Climate Impacts, Grand Challenges in Ecosystem and the Environment, Department of Life Sciences and Grantham Institute — Climate Change and the Environment, Imperial College London, UK**

Corresponding author:
Tomislav Hengl[1]

Email address: tom.hengl@envirometrix.net

ABSTRACT

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018

Potential Natural Vegetation (PNV) is the vegetation cover in equilibrium with climate, that would exist at a given location non-impacted by human activities. PNV is useful for raising public awareness about land degradation and for estimating land potential. This paper presents results of assessing Machine Learning Algorithms (MLA) for operational mapping of Potential Natural Vegetation (PNV). The following MLA were considered: neural networks (nnet package), random forest (ranger), gradient boosting (gmb), K-nearest neighborhood (class) and cubist. Three case studies were considered: (1) global distribution of biomes based on the BIOME 6000 data set (8057 modern pollen-based site reconstructions), (2) distribution of forest tree taxa in Europe based on detailed occurrence records (1,546,435 ground observations), and (3) global monthly Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) values (30,301 randomly-sampled points). A stack of 160 global maps representing biophysical conditions over land, including atmospheric, climatic, relief and lithologic variables, were used as explanatory variables. The overall results show that random forest gives the overall best performance. The highest accuracy for predicting BIOME 6000 classes (20) was estimated at 68 % (33 % with spatial Cross Validation) with the most important predictors being total annual precipitation, monthly temperatures and bioclimatic layers. Predicting forest tree species (73) resulted in mapping accuracy of 25 %, with the most important predictors being monthly cloud fraction, mean annual and monthly temperatures and elevation. Regression models for FAPAR (monthly images) gave an R-square of 90 % with most important predictors being total annual precipitation, monthly cloud fraction, CHELSA bioclimatic layers and month of the year, respectively. Further developments of PNV mapping could include using GBIF records to map global distribution of plant species at different taxonomic levels. This methodology could also be extended to dynamic modeling of PNV, so that future climate scenarios can be incorporated. Global maps of biomes, FAPAR and tree species at 1 km spatial resolution are available for download via http://dx.doi.org/10.7910/DVN/QQHCIK.

Submitted to PeerJ on 26th of March 2018

INTRODUCTION

Potential Natural Vegetation (PNV) is the *"vegetation cover in equilibrium with climate, that would exist at a given location non-impacted by human activities"* (Levavasseur et al., 2012). It is a hypothetical vegetation state assuming natural (undisturbed) physical conditions, a reference status of vegetation assuming no degradation and/or no unusual ecological disturbances. PNV is especially useful for raising public awareness about land degradation and for estimating land potential (Herrick et al., 2013). For example, Omernik (1987) details PNV maps for USA; Bohn et al. (2007) provides maps for EU; Carnahan (1989) for Australia; Marinova et al. (2017) maps PNV for the Eastern Mediterranean–Black Sea–Caspian-Corridor; and maps of PNV for Latin America are available in Marchant et al. (2009). Regarding specific tree species, San-Miguel-Ayanz et al. (2016) provide habitat suitability maps for the main forest tree species in Europe, based on environmental variables, especially bioclimatic variables such as average temperature of the coldest month, precipitation of the driest month and similar. Potapov et al. (2011) have generated a global map of potential forest cover at 1 km resolution (publicly available from http://globalforestwatch.org/map/). Erb et al. (2017) produced a global map of potential biomass stocks by reversing the current managed land use systems to natural vegetation. Levavasseur et al. (2012) and Tian et al. (2016) predict global PNV classes using environmental covariates such as climatic images and landform parameters. Griscom et al. (2017) have recently produced a global reforestation map at 1 km resolution.

A common limitation of prior maps is their coarse spatial resolution (about 25 km) limiting the use of these maps for operational planning (e.g. Marchant et al. (2009); Levavasseur et al. (2012) and Tian et al. (2016)). In addition, comparisons of multiple overlapping sources of PNV maps shows that they rarely agree with one another since they do not share the same mapping criteria and, traditionally, emphasize regionally-specific botanical groupings rather than functional classifications. Limitations of maps based on field surveys of PNV (e.g., Bohn et al. (2007)) are the assumptions about the controls on vegetation distribution applied to the extrapolation of a limited number of field surveys.

Here we provide an update of comparable global PNV maps produced by Potapov et al. (2011); Levavasseur et al. (2012); Tian et al. (2016) and Erb et al. (2017). We explore the possibility of increasing the mapping accuracy using up-to-date maps of climate, atmosphere dynamics, landform and lithology, and state-of-the-art machine learning methods. Our final aim is to produce PNV maps that are both more detailed, richer in information, based on objective reproducible methods; and potentially more usable for global modeling and awareness raising projects. We focus on improving the spatial detail, thematic accuracy and reproducibility of maps, at the cost of increasing the total computing load, but also consider automation of the prediction process so that the maps can be rapidly updated as the new ground data arrives. Our modeling follows three phases:

(a) model selection: we compare possible models of interest for PNV mapping and choose the optimal spatial prediction framework based on the cross-validation results,

(b) model assessment: we assess the uncertainty of predictions per vegetation class and try to determine objectively the limitations of the mapping products for wider uses, and

(c) prediction: we use the best performing models to produce spatial predictions, then visually assess maps and if necessary repeat steps a–c.

METHODS AND MATERIALS

Theory

PNV is the hypothetical vegetation cover that would be present if the vegetation were in equilibrium with environmental controls, including climatic factors and disturbance, and not subject to human management. When considering PNV, one needs to distinguish between potential *"natural'* and potential *"managed"* vegetation, and *"actual"* natural and *"actual"* managed vegetation. PNV also changes through time, therefore it can refer to the current state of environment or some past historic state (Fig. 1b). However, parts of the world that have not been subject to human disturbance/management can provide clearer information about the vegetation cover in historic times which can serve as a guide to PNV. A major limitation of modeling PNV is that we unfortunately do not have equally detailed information about the status of vegetation and environment across historic periods. For instance, about half of the Earth's mature tropical forests have disappeared in the last 150 years and original habitats have reduced to 10 % (Hansen et al., 2013). Given that climates have changed and few areas are truly *"human impact free"*, even undisturbed historic vegetation only represents the expression of PNV for a given set of climate conditions at a specific time.

Regardless of the hypothetical nature of PNV, the concept (both as a classification and as a regression type problem) is still a helpful yardstick against which land cover change can be quantitatively measured and land restoration designs could be planned. Indeed Erb et al. (2017) have estimated that almost half of the standing global vegetation biomass carbon stocks have been lost, almost equally due to land cover change (e.g. tree cover to cropland) and management effects within land cover types (e.g. croplands managed at lower biomass carbon stocks than tree covered areas). PNV maps can help quantify the the differences (both deficit and surplus) in biomass stocks caused by the current land management system more objectively. For efforts aimed at land restoration, such information could also be a valuable input into the redesign of land management systems.

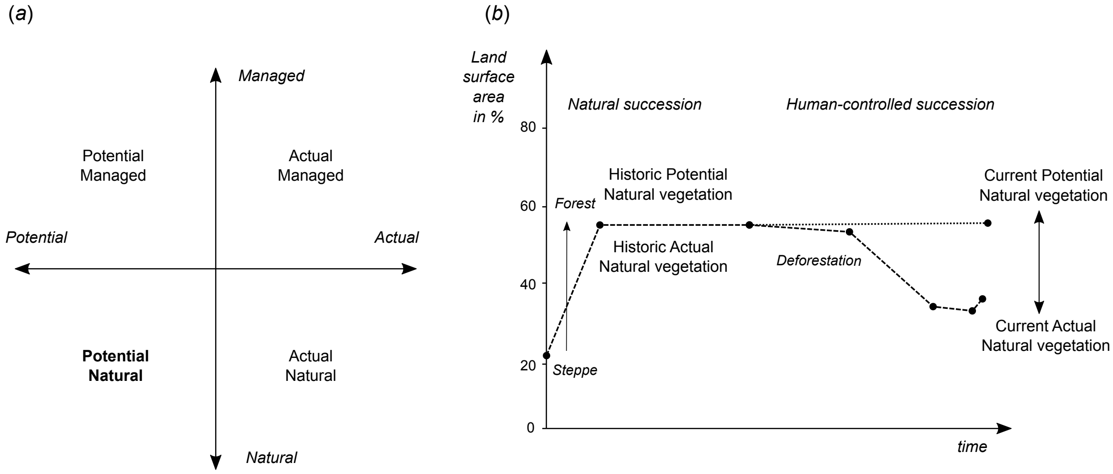


Figure 1. Schematic explanation of differences between (a) potential and actual natural/managed vegetation, and (b) current and historic vegetation. This paper focuses on providing estimates of Potential Natural Vegetation (PNV) using current state biophysical conditions.

Mapping PNV is a special case of species distribution modeling (Elith and Leathwick, 2009; Hijmans and Elith, 2018): at the core of PNV mapping is statistical modeling of relationship between species (or natural association of species or communities) and a list of predictors i.e. biotic and abiotic site factors (Elith and Leathwick, 2009). The difference between mapping actual distribution of species and mapping PNV is that PNV involves extrapolating the model to the whole land mask, assuming a hypothetical distribution under a specific set of undisturbed bioclimatic and/or biophysical conditions:

$$\Pr(Y) = f\,(\text{Relief}, \text{BioClimate}, \text{Lithology}) \tag{1}$$

where Y is the target variable, which could be vegetation types or plant species with finite number of states $Y \in \{1,2,\ldots,k\}$ and/or vegetation properties. PNV mapping can be considered as a *classification-type* or *regression-type* problem depending on whether we map factors such as vegetation types or continuous vegetation properties such as biomass or leaf area index.

The primary assumptions we make when applying a PNV model to the training data are:

1. The ecological gradients captured in training data reflect only natural ecological gradients and not human controls such as land use systems, civil engineering constructions, or one-off major disturbance events such as volcanic eruptions, floods, or tsunamis.

2. Remote sensing data such NDVI reflect human-altered vegetation patterns and ought not be used as covariates in PNV mapping (Leong and Roderick, 2015).

3. The training data are representative of the study area, especially considering the feature space (ecological gradients) of the study area.

Assuming a log-linear relationship between ecological gradients and target variables, PNV classes can be modeled using a multinomial log-linear model:

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i} \tag{2}$$

where $f(k,i)$ is the linear predictor function, β are the regression coefficients associated with the mth explanatory variable and the kth outcome. An efficient implementation of the multinomial logistic regression is the `multinom` function from the R package nnet (Venables and Ripley, 2002). The output of prediction produced using `multinom` are k probability maps (0–100 %) such that all predictions at each site sum up to 1:

$$\sum_{k=1}^{K} \Pr(Y_i = k) = 1 \tag{3}$$

In this paper, all predictions models are used in the *"probability"* mode i.e. to derive probability maps per class.

Note that a PNV spatial prediction model divides geographic space among all possible states given the training points. It is therefore necessary, for Eq.(1), that all possible states of Y are represented with training data so that the model can be applied over the whole spatial domain of interest. If not all of the states are known, then the space will be artificially filled-in with classes occupying similar ecological niches and which can lead to prediction bias. In other words, similar to species distribution modeling of individual species, both presence and absence data play an equally important role for model calibration (Elith and Leathwick, 2009).

Input data: training points

We consider three ground-truth data sets for model calibration:

1. an expanded version of the BIOME 6000 DB data set representing site-based reconstructions from surface pollen samples of major vegetation types or biomes (http://dx.doi.org/10.17864/1947.99),

PeerJ Preprints NOT PEER-REVIEWED

2. EU Forest (Mauri et al., 2017) and GBIF (Global Biodiversity Information Facilities) occurrence records of the 76 main forest tree taxa in Europe (http://dx.doi.org/10.15468/dl.fhucwx),
3. Long-term Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) monthly images derived using a time-series of Copernicus Global Land products (https://land.copernicus.eu),

BIOME 6000 and EU Forest and GBIF occurrences are point data sets, while FAPAR are remote sensing images at relatively fine spatial resolution (250 m), from which we sample large number of values (ca $100,000$) using random sampling after masking for areas of natural vegetation.

BIOME 6000

The BIOME 6000 data set (http://dx.doi.org/10.17864/1947.99) includes vegetation reconstructions from modern pollen samples, preserved in lake and bog sediments and from moss polsters, soil and other surface deposits. The use of pollen data to reconstruct PNV relies on the fact that although modern pollen samples may contain markers of land use, the predominant pollen types found in any one sample are those of the regional vegetation within a radius on the order of 10–30 km around the sampling site. Even if forests have fragmented, these fragments continue to produce and disperse pollen grains, and the composition of the pollen assemblage provides information on tree taxa that are still present.

The BIOME 6000 data set is an amalgamation of multiple data sets. BIOME 6000 initially produced maps for individual regions: Europe, Africa and the Arabian Peninsula, the Former Soviet Union and Mongolia and China. Additional regions were subsequently added including Beringia, western North America, Canada and the eastern United States and Japan, and the data for northern Eurasia, China and southern Europe and Africa were also updated. These regional compilations were summarized in Prentice and Jolly (2000). Subsequent regional updates include China (Harrison et al., 2001), the circum-Artic region (Bigelow et al., 2003), Australia (Pickett et al., 2004) and South America (Marchant et al., 2009). Additionally, we have also combined these data with pollen-based vegetation reconstructions from the Eastern Mediterranean-Black Sea-Caspian Corridor (EMBSeCBIO) region (Marinova et al., 2017) available from http://dx.doi.org/10.17864/1947.109, to produce a more complete and up-to-date compilation of the BIOME 6000.

Some sites in the BIOME 6000 data set have multiple reconstructions based on multiple nearby modern pollen samples (up to 30), which provides a useful measure of the reconstruction uncertainty, but could lead to modeling bias because the number of modern samples varies between sites. To reduce these unwanted effects, we use only the most frequently reconstructed biome at each site and for the sites with two equally common reconstructions (ca. 900) we use both observations.

The number of biomes differentiated varies from region to region, and some biomes were only reconstructed in specific regions where they are particularly characteristic, although they may occur elsewhere. Furthermore, some biomes that can be recognized on the modern landscape were never reconstructed in the BIOME 6000 data set (e.g. cushion forb tundra) — either because of the sample distribution or because the characteristic plant-functional types were also spread amongst other biomes. Simplified or *“megabiome”* classifications (e.g. Harrison and Bartlein (2012)) involve a substantial loss

PeerJ Preprints NOT PEER-REVIEWED

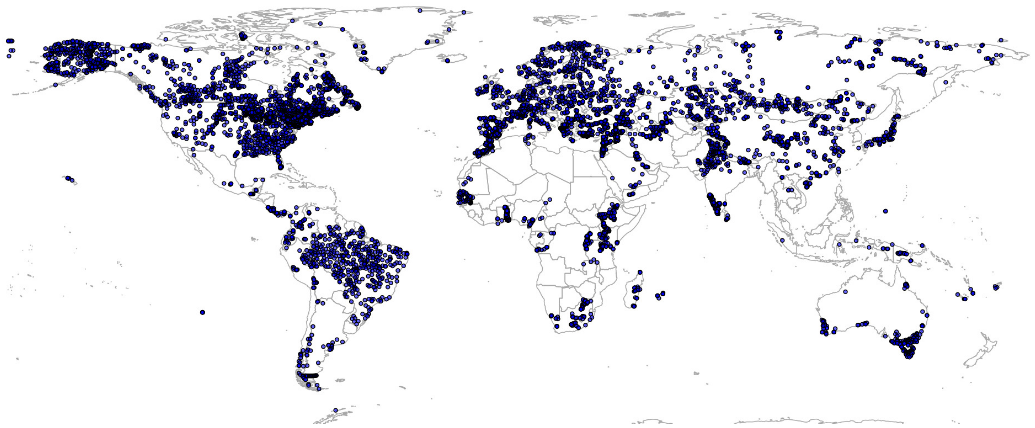


Figure 2. Spatial distribution of BIOME 6000 training points. A total of 8057 unique locations are shown on the map.

of information. We have therefore created a new standardization of the classification scheme (see further Table 1; the final scheme has 20 globally applicable and distinctive biomes) which preserve the maximum number of distinct biomes that were reconstructed as present in multiple regions.

There are relatively few data vegetation reconstructions for tropical South America, which could lead to extrapolation problems and omission of important PNV classes in Latin America, but also potentially in tropical parts of Africa and Asia. To reduce under-representation of tropics, we have added 350 randomly simulated points based on the RADAM Brazil natural vegetation polygon map at high spatial detail (Radam Vegetação SIRGAS map) (Veloso et al., 1992) obtained from `ftp://geoftp.ibge.gov.br/`. Before generating the pseudo-observations for Brazil, we have translated SIRGAS map legends to match the BIOME 6000 classes. This translation is also available via the project's github repository. This gave a total of 8057 unique individual locations represented in the combined data set i.e. a total of 8959 training observations (Fig. 2).

We have mapped the distribution of biomes for all land pixels, with the exception of water bodies, barren land and permanent ice areas. Barren land and permanent ice areas were masked out using the ESA's global land cover maps for the period 2000–2015 (`https://www.esa-landcover-cci.org`) and the long-term FAPAR images, both available at relatively fine resolution of 300 m. We only mask out pixels that are permanent ice/barren ground and have a FAPAR = 0 throughout the period 2000–2015.

European Forest Tree occurrence records

For mapping PNV distribution of forest tree taxa (note: most of these are individual species, but some are only recognised at sub-genus or genus level) in Europe we have merged two point data sets: EU Forest (Mauri et al., 2017) (588,983 records covering 242 species) and GBIF occurrence records of the main forest tree taxa in Europe. The GBIF Occurrence data was downloaded on 23rd January 2017 (`http://dx.doi.org/10.15468/dl.fhucwx`). We focus on modeling just the 76 forest tree taxa indicated in the European Atlas of Forest Tree Species (San-Miguel-Ayanz et al., 2016).

Global GBIF occurrence data can be obtained by using the rgbif package, in which case the only important parameter is the `taxonKey` (e.g. *"Betula spp."* corresponds to GBIF taxon key 2875008). After

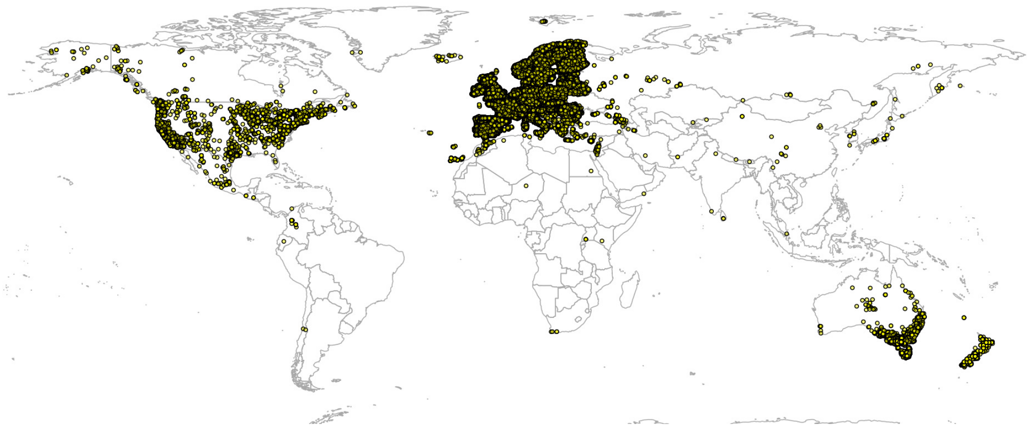


Figure 3. Merge of EU Forest (Mauri et al., 2017) and GBIF occurrence records used to build models to predict PNV for the 76 forest tree taxa. Total of 1,546,435 shown on the map.

the bulk data download (which gives about 4 million occurrences), we imported all points and then subset occurrences based on the list of taxon keys and and coordinate uncertainty (<2 km positional error). This gave a total of 1,546,435 training points from which about 2/3 are GBIF points (Fig. 3). We assume in further analysis that the EU Forest point locations and representativeness are more trustworthy, hence we assign 4× higher weights to these points than to the GBIF points.

Certain forest tree species (*Chamaecyparis lawsoniana*, *Eucalyptus globulus* and *Pseudotsuga menziesii*), that are shown in the European Atlas of Forest Tree Species are introduced i.e. planted and do not generally propagate naturally. Hence, they were removed from the list of target forest tree species. We retained, however, three species (*Ailnthus altissima*, *Picea sitchensis* and *Robinia pseudoacacia*) that are not native but are extensively naturalized. The total number of target forest tree taxa was 73.

We built predictive models for European forest tree taxa using information on their global distribution, but only generate predictions for Europe. In other words, we use a global compilation for model training to increase the precision of the definition of the ecological niche of each taxon, but then predict only for Europe as the selection of taxa is based on the European Atlas of Forest Tree Species (San-Miguel-Ayanz et al., 2016).

FAPAR

Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) monthly images for 2014–2017 were obtained from `https://land.copernicus.eu`. From a total of 142 images downloaded from `https://land.copernicus.eu` we derived minimum, median and maximum value of FAPAR per month (12) using the 95 % probability interval using the data.table package (`http://r-datatable.com`). For regression modeling we only show results of predictions of median values of FAPAR; predictions of minimum and maximum FAPAR can be obtained from the data repository.

We model median and upper 95 % FAPAR values as a function of the same covariate layers used in all three case studies. For model training we use ca. 30,000 randomly sampled points (Simple Random Sampling) exclusively from protected area as shown in the World Database on Protected Areas (WDPA) data set (`http://protectedplanet.net`) and the Intact Forest Landscapes (IFL) data set for 2000 and

PeerJ Preprints NOT PEER-REVIEWED

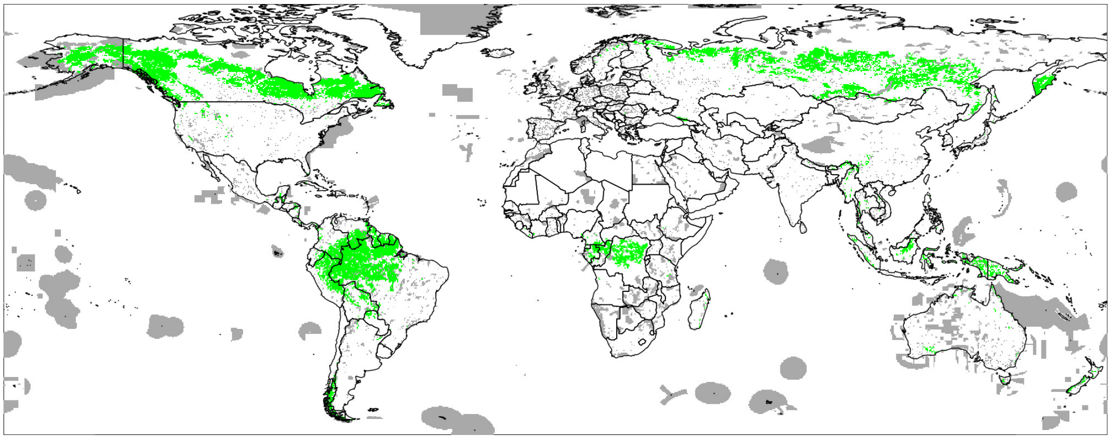


Figure 4. World's Protected Areas (dark gray) based on http://protectedplanet.net and Intact Forest Landscapes for year 2000 (green) based on http://intactforests.org. These maps were used to randomly select some 30,000 training points to predict potential FAPAR under PNV.

2013 (Potapov et al., 2008) Fig. 4). We use about 3× more training points from the IFL 2013 areas than from the WDPA and IFL 2000 masks to emphasize ecological conditions of intact vegetation.

The prediction model for FAPAR under PNV is in the form of:

```
R> FAPAR ~ cm + X1m + X2m + X3 + ... + Xp
```

where `X1m` is the covariate with monthly values (for example precipitation, day-time and night-time temperatures etc), `X3` is the environmental covariates that does not vary through year (e.g. lithology or DEM derivatives), and `cm` is the cosine of the month number:

$$c_m = \cos(\mu/12 \cdot 2 \cdot \pi) \tag{4}$$

where μ is the month number 1–12. The total number of training observations used to build models is in fact 180,483 (each training site is represented up to 12 times).

For PNV FAPAR mapping we have masked out all water bodies including lakes and rivers, following the ESA's global land cover maps for the period 2000–2015 (https://www.esa-landcover-cci.org) and permanent ice/barren ground.

Input data: environmental covariates

For modeling purposes, we use a stack of 160 spatially explicit co-variate data layers that represent ecological gradients essential for growth and survival of plants:

- DEM derivatives quantifying various landscape metrics and hydrological processes: slope, curvature, topographic index, topographic openness, valley depth and multi-resolution valley bottom index; all derived using the SAGA GIS (Conrad et al., 2015);
- Mean, minimum and maximum monthly temperatures derived as a mean between WorldClim v2 (http://worldclim.org/version2) and CHELSA climate (Karger et al., 2017).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018

- Mean monthly precipitation images derived as a weighted average between the WorldClim v2, CHELSA climate and Global Precipitation Measurement Integrated Multi-satellitE Retrievals for GPM (IMERG) rainfall product.
- CHELSA Bioclimatic layers downloaded from http://chelsa-climate.org/, including: annual mean temperature, mean diurnal temperature range, isothermality (day-to-night temperature oscillations relative to the summer-to-winter oscillations), temperature seasonality (standard deviation of monthly temperature averages), maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation amount, precipitation of wettest month, precipitation of driest month, precipitation of wettest quarter, precipitation of driest quarter (Karger et al., 2017);
- European Space Agency's CCI-LC snow probability monthly averages based on MODIS snow products MOD10A2 downloaded from http://maps.elie.ucl.ac.be/CCI/viewer/index.php;
- USGS Global Ecophysiography landform classification and lithological map at 250 m resolution obtained from http://rmgsc.cr.usgs.gov/outgoing/ecosystems/Global/ and based on Global Lithological Map (GLiM) (Hartmann and Moosdorf, 2012);
- MODIS Cloud fraction monthly images obtained from http://www.earthenv.org/cloud (Wilson and Jetz, 2016);
- Global Water Table Depth in meters based on Fan et al. (2013);
- NASA's monthly MODIS Precipitable Water Vapor images (MYDAL2_M_SKY_WV data set at http://neo.sci.gsfc.nasa.gov);
- Potential wetlands GIEMS map (Fluet-Chouinard et al., 2015);
- Global Surface Water dynamics images: occurrence probability, surface water change, and water maximum extent; downloaded from https://global-surface-water.appspot.com/download (Pekel et al., 2016);
- Density of earthquakes based on the USGS Earthquake Archives (http://earthquake.usgs.gov/earthquakes/);

Some CHELSA bioclimatic layers contained too many missing pixels or artifacts (e.g. mean temperature of wettest quarter, mean temperature of driest quarter, precipitation seasonality, precipitation of warmest quarter and precipitation of coldest quarter) and were hence not used for further modeling to avoid propagating those artifacts to final predictions.

All original layers have been resampled to the same grid at spatial resolution of 1/120 decimal degrees (about 1 km) covering latitudes between -62.0 and 87.37. We do not map Antarctica as this continent is dominantly covered with permanent ice and there are no training points. We limit all analysis to 1 km i.e. 1/120 degrees in geographical coordinates, to avoid too high computational load, even though many of environmental covariates are also available at finer resolutions.

We use the same stack of covariates for mapping global distribution of biomes, FAPAR and forest tree species in Europe, to be able to compare model performance and investigate whether the most important covariates differ between the three case studies.

Machine Learning Algorithms (MLA) examined

We examine predictive performance of the following MLA's:

- Neural networks (Venables and Ripley, 2002),
- Random forest (Breiman, 2001; Cutler et al., 2007; Biau and Scornet, 2016; Hengl et al., 2018),
- Generalized Boosted Regression Models (Friedman, 2002),
- K-nearest neighbors (Venables and Ripley, 2002),

Neural networks are available from several packages in R. Here we use the nnet package (Ripley and Venables, 2017) also described in Venables and Ripley (2002). Random forest is efficiently implemented in the ranger package (Wright and Ziegler, 2016) and can be used to to process large data sets. Generalized Boosted Regression Models are available via the gbm package (Ridgeway, 2017). The K-nearest Neighbour Regression is available via the class package i.e. the `knn` function (Venables and Ripley, 2002). Of these four algorithms, the K-nearest neighbors is computationally least intensive and results in relatively simple models, while random forest is computationally most intensive and results in large models. However, a limitation of the K-nearest neighbors approach is that it does not handle high dimensional data in comparison to random forest or neural nets.

We also test using the same packages to fit models for regression-type problems (e.g. modeling of FAPAR), with the exception of the class package i.e. the `knn` function which can only be used for classification problems. For modeling FAPAR we used instead also the Cubist approach, available via the Cubist package (Kuhn et al., 2014), and the Extreme Gradient Boosting approach available via the xgboost package (Chen and Guestrin, 2016).

The caret package has many more MLA of interest for classification and regression problems than presented here, but many are not all optimized for large data sets and hence also not applicable for large data sets ($\gg$ 1000 observations with $\gg$ 100 covariates).

Model selection

For model fitting and model selection we use the caret package implementation for automated evaluation of models. When comparing performance of the models we look at classification accuracy based on cross-validation with refitting implemented in the caret package via the setting (Kuhn, 2008; Kuhn and Johnson, 2013):

```
R> ctrl <- trainControl(method="repeatedcv", number=5, repeats=2)
```

which translates as: models are refit 5 times using 80 % of the data and that predictions derived from the fitted models are compared with the remaining observations; this process is then repeated two times to produce stable results. The reported accuracy is the map accuracy (0–100 %) and/or Root Mean Square Error (RMSE) derived using all merged cross-validations (Kuhn, 2008; Kuhn and Johnson, 2013). Since most of data sets are fairly large and model fitting can take hours, even in a High Performance Computing environment, we limit the number of repetitions to 2.

For FAPAR (regression modeling) and selection of the final prediction model we use the same repeated cross-validation as implemented via the caret package. This is in principle similar to evaluation of the classification accuracy, except the comparison criterion is RMSE.

All analyses were run on a High Performance Computing Amazon ec2 server with 64 cores and 256 GiB RAM. Total computing time to produce all outputs is about 12 hours of optimized computing (or about 600 CPU hours). 1 km data can be processed with 2 degree tiles, which usually leads to some 5000 tiles to represent the land mask. All processing steps and preparation of input and output maps are fully documented at https://github.com/envirometrix/PNVmaps. All output maps are available for download via http://dx.doi.org/10.7910/DVN/QQHCIK under the Open Database License (ODbL).

Performance of classification algorithms

For biomes and tree species we use the map purity (0–100 %) and kappa metrics for the dominant (hard) classes at cross-validation points as the key measure of predictive performance (Kuhn and Johnson, 2013). For each class we also provide predicted probabilities, which can be used to model transition zones and correlation between classes. For the predicted probabilities of class occurrences (0–1) we derived the True Positive Rate (TPR) and the Area Under the receiver operating characteristic Curve (AUC) as implemented in the ROCR package (Sing et al., 2005, 2016). TPR values range from 0 to 1 where 1 indicates a perfect match to the class positives in ground data and TPR values < 0.5 can be considered poor. Values of AUC close to 1 show high prediction performance, while values around 0.5 and below are considered poor. TPR and AUC provide probably a more informative measure of the mapping accuracy than overall mapping accuracy / kappa, as they also allow detection of problematic classes.

We also use Scaled Shannon Entropy Index, which can be derived using the per-class probability maps (Shannon, 1949; Borda, 2011):

$$\mathsf{EI}_s(x) = -\sum_{i=1}^{b} P_i(x) \cdot \log_b P_i(x) = \frac{-\sum_{i=1}^{b} P_i(x) \cdot \log P_i(x)}{-b \cdot b^{-1} \cdot \log b^{-1}} \tag{5}$$

where b is the total number of possible classes and P is probability of class i. The Scaled Shannon Entropy Index (EI_s) is in the range from 0–1, where 0 indicates a perfect classification and 1 (or 100 %) indicates maximum confusion. Scaled Shannon Entropy Index should not be confused with classification accuracy assessment: $\mathsf{EI}_s < 60\,\%$ already indicates relatively low confusion between classes i.e. high accuracy, while mapping error of 60 % would still be considered a relatively poor classification accuracy result.

For the biomes data set, where spatial clustering of points is significant, we also use repeated spatial cross-validation as implemented in the mlr package (Bischl et al., 2016):

```
R> learner.rf = makeLearner("classif.ranger", predict.type = "prob")
R> resampling = makeResampleDesc("RepCV", fold = 5, reps = 5)
```

It has been shown that spatial autocorrelation in data and serious spatial clustering in training points can lead to somewhat biased estimate of the actual accuracy (Brenning, 2012). Solution to this problem is to apply spatial partitioning so that possible bias due to spatial proximity is minimized.

We also compare results of modeling potential distribution of tree species in Europe with the habitat type maps of Europe produced independently by San-Miguel-Ayanz et al. (2016) and Brus et al. (2012). This comparison is visually based only.

Performance of regression algorithms

For testing spatial predictability of FAPAR we use the root mean squared error (RMSE) and mean error (ME):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{m} [\hat{y}(\mathbf{s}_j) - y(\mathbf{s}_j)]^2}{n}}$$

$$\mathrm{ME} = \frac{\sum_{j=1}^{m} [\hat{y}(\mathbf{s}_j) - y(\mathbf{s}_j)]}{n}$$

where $\hat{y}(\mathbf{s}_j)$ is the predicted value of y at the cross-validation location, and m is total number of cross-validation points. We also report amount of variation explained by the model (R^2) derived as:

$$R^2 = \left[1 - \frac{SSE}{SST} \right] \times 100\% \tag{6}$$

where SSE is the sum of squared errors at cross-validation points and SST is the total sum of squares. A coefficient of determination close to 1 indicates a perfect model.

RESULTS

Global maps of biomes

Results show that a relatively accurate model of PNV can be produced from the BIOME 6000 data set using the existing stack of covariates at 1 km spatial resolution. Results of cross-validation show that random forest (RF) model is the best performing method and distinctively better than other approaches (Fig. 5). The choice of the random forest `mtry` parameter has little impact on overall accuracy, most likely because there is a high overlap in covariate maps so that even with smaller `mtry` bagging the performance is relatively similar. The best prediction accuracy from the four methods used for mapping global biomes is about 68 %. The predicted biome classes are available in Fig. 6.

The most important covariates for the random forest model are: total annual precipitation, monthly temperatures, CHELSA bioclimatic layers, atmospheric water vapor images and monthly precipitation. Landform parameters and lithology are not amongst the top 20 most important predictors. The decline in variable importance is, however, gradual — even lower ranked covariates might still affect the accuracy of predictions.

The detailed cross-validation results show that the only difficult classes to predict is prostrate dwarf shrub tundra (Table 1). The TPR value for most class probabilities ranges from 0.83 to 0.94 indicating relatively high match with ground data. The Scaled Shannon Entropy Index map (Fig. 7) shows that the zones of highest confusion between classes can be found in Afghanistan, Nepal, mountainous parts of the USA and Mexico, parts of Angola and Zambia. The map of the SSEI is comparable to the confusion map

PeerJ Preprints NOT PEER-REVIEWED

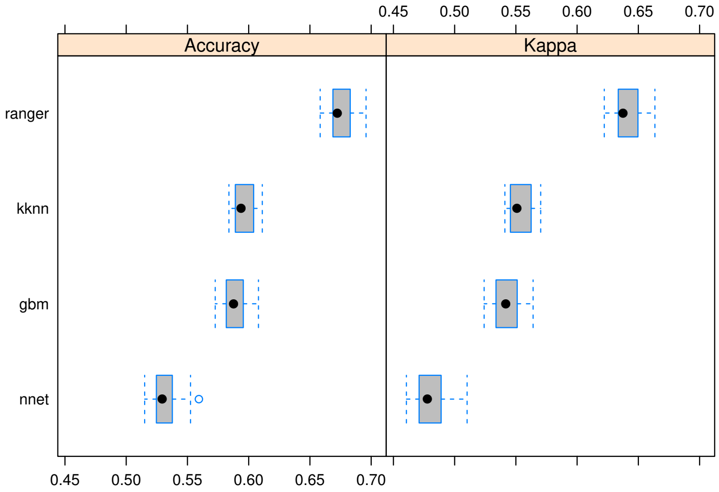


Figure 5. Predictive performance of the target machine learning algorithms for mapping global distribution of biomes ($N = 8653$; spatial distribution of training points is available in Fig. 2). ranger = random forest, kkn = K-nearest neighbors, gbm = Generalized Boosted Regression Models, nnet = Neural networks.

Table 1. Summary results of cross-validation for mapping global distribution of biomes (20 classes). Classification accuracy for predicted class probabilities based on 5–fold cross-validation with refitting. ME = *"Mean Error"*, TPR = *"True Positive Rate"*, AUC = *"Area Under Curve"*, N = *"Number of occurrences"*.

Biome class	**ME**	**TPR**	**AUC**	**N**
cold deciduous forest	-0.01	0.89	0.96	201
cold evergreen needleleaf forest	0.01	0.87	0.98	892
cool evergreen needleleaf forest	-0.07	0.87	0.93	201
cool mixed forest	0.01	0.86	0.97	1549
cool temperate rainforest	0.01	0.92	0.99	95
desert	0.00	0.89	0.96	330
erect dwarf shrub tundra	-0.01	0.89	0.98	145
graminoid and forb tundra	-0.03	0.83	0.91	128
low and high shrub tundra	-0.01	0.88	0.98	393
prostrate dwarf shrub tundra	-0.02	**0.54**	0.90	11
steppe	0.01	0.87	0.94	889
temperate deciduous broadleaf forest	-0.01	0.84	0.94	961
temperate evergreen needleleaf open woodland	0.01	0.92	0.97	307
temperate sclerophyll woodland and shrubland	0.00	0.94	0.99	154
tropical deciduous broadleaf forest and woodland	0.01	0.86	0.97	215
tropical evergreen broadleaf forest	0.00	0.87	0.99	333
tropical savanna	0.01	0.89	0.99	291
tropical semi evergreen broadleaf.forest	-0.05	0.87	0.98	160
warm temperate evergreen and mixed forest	0.01	0.85	0.96	985
xerophytic woods scrub	-0.02	0.88	0.95	388

produced by Levavasseur et al. (2012), except in our case the Rocky Mountains in USA and mountains chains in South America show somewhat higher confusion. Many of the areas with high confusion index happen because the prediction model has problems distinguishing between closely-related biomes such as

PeerJ Preprints NOT PEER-REVIEWED

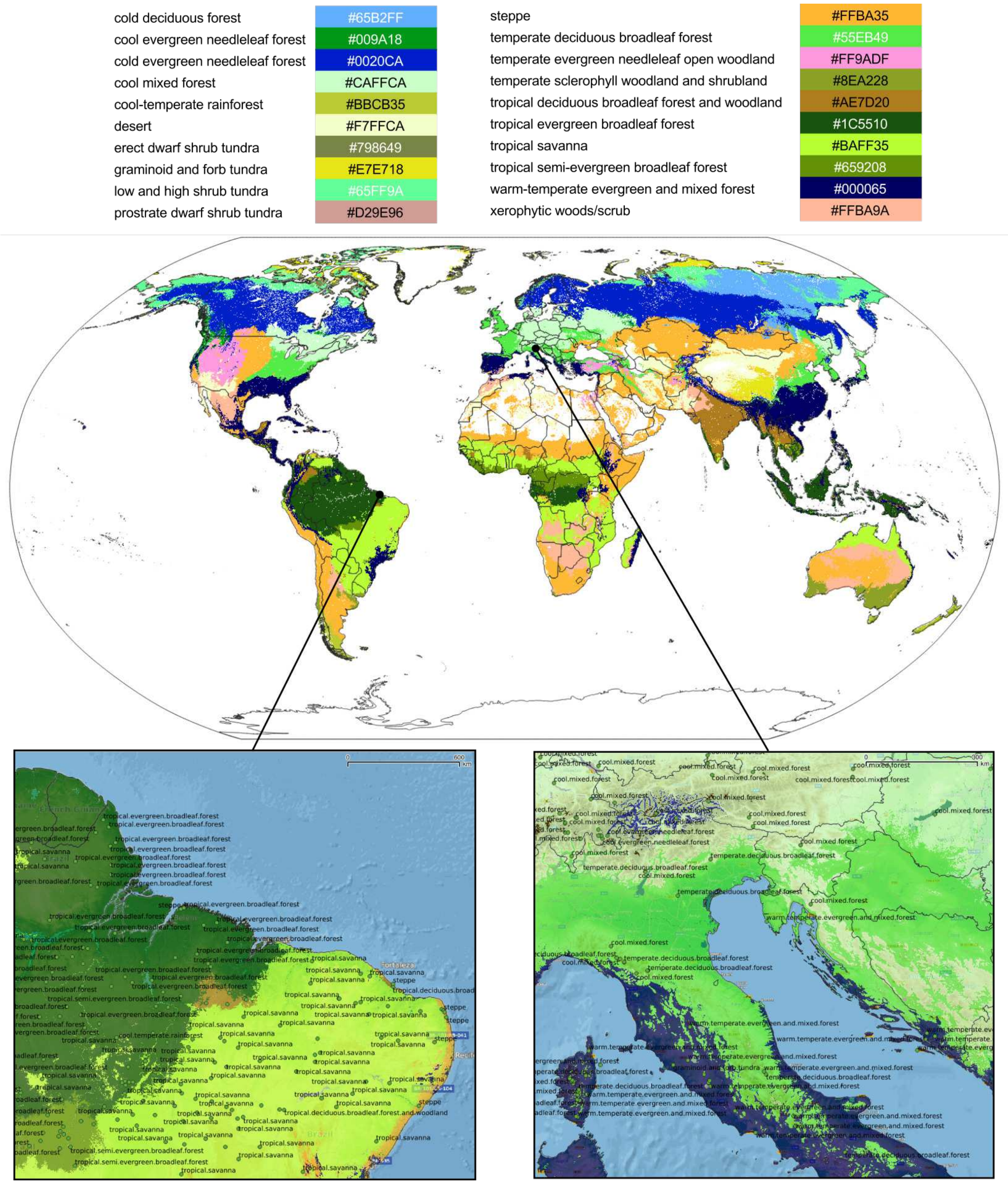

cold deciduous forest #65B2FF
cool evergreen needleleaf forest #009A18
cold evergreen needleleaf forest #0020CA
cool mixed forest #CAFFCA
cool-temperate rainforest #BBCB35
desert #F7FFCA
erect dwarf shrub tundra #798649
graminoid and forb tundra #E7E718
low and high shrub tundra #65FF9A
prostrate dwarf shrub tundra #D29E96
steppe #FFBA35
temperate deciduous broadleaf forest #55EB49
temperate evergreen needleleaf open woodland #FF9ADF
temperate sclerophyll woodland and shrubland #8EA228
tropical deciduous broadleaf forest and woodland #AE7D20
tropical evergreen broadleaf forest #1C5510
tropical savanna #BAFF35
tropical semi-evergreen broadleaf forest #659208
warm-temperate evergreen and mixed forest #000065
xerophytic woods/scrub #FFBA9A
0
600 km
0
300 km

Figure 6. Predicted PNV distribution for global biomes with a zoom in on areas in Brazil and Europe. Labels indicates training points from the BIOME 6000 data set (Fig. 2).

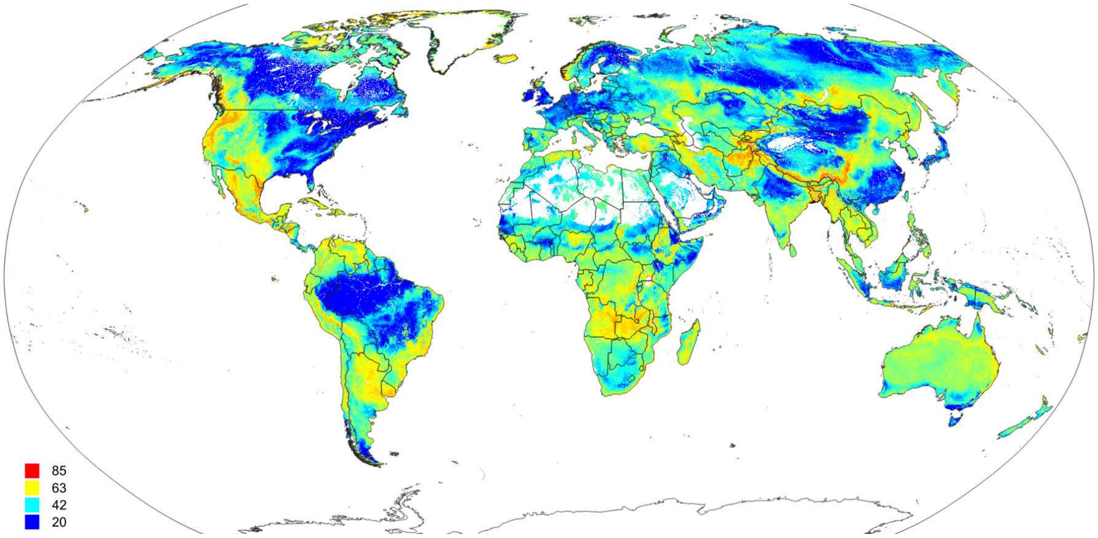


Figure 7. Scaled Shannon Entropy Index (SSEI) derived using predicted probabilities for 20 biomes (classes) based on Eq.(5). High values of SSEI (red color) indicate high confusion between classes.

the *"cold evergreen needleleaf forest"* and *"cool evergreen needleleaf forest"* (e.g. Scotland).

Results of the spatial Cross-Validation, as implemented in the mlr package (Bischl et al., 2016), further indicate that the spatial clustering of points does have a large effect on the mapping accuracy: spatial CV drops to 0.33 and weighted kappa to 0.45. This likely happens due to high spatial clustering of the biome points, but obviously biomes are also spatially highly autocorrelated.

European forest tree species

The results of 5–fold cross validation with re-fitting at each fold, confirms that random forest is also the best prediction method for the forest taxa data set (Fig. 8). The overall mapping accuracy is significantly lower than for biomes, but this reduction in accuracy is to be expected as many of these taxa occur in communities, resulting in natural overlap of forest tree taxa distribution. The mapping accuracy of individual taxa, however, can be relatively high with TPR values of between 0.16–0.90 and an average value of around 0.69 (Table 2). The final maps (Fig. 9) show a relatively good match with ground data, meaning that with the exception of some species of rarer occurrence (*Picea omorika*, *Cupressus sempervirens*, *Prunus mahaleb*), the species probability distribution maps are relatively accurate.

Table 2. Results of cross-validation for the forest tree taxa (San-Miguel-Ayanz et al., 2016). Classification accuracy for predicted class probabilities based on 5–fold cross-validation. ME = *"Mean Error"*, TPR = *"True Positive Rate"*, AUC = *"Area Under Curve"*, N = *"Number of occurrences"*. Taxa with less than < 50 observations were omitted from analysis.

Species name	GBIF taxon ID	ME	TPR	AUC	N
Abies alba	2685484	-0.01	0.77	0.92	16,150
Acer campestre	3189863	-0.01	0.65	0.83	19,819
Acer platanoides	3189846	-0.02	0.68	0.82	30,801
Acer pseudoplatanus	3189870	-0.01	0.69	0.79	65,039
Aesculus hippocastanum	3189815	-0.01	0.59	0.85	8,088
Ailanthus altissima	3190653	0.04	0.69	0.92	1,576
Alnus cordata	2876607	0.05	0.73	0.95	904

Continued on next page

PeerJ Preprints NOT PEER-REVIEWED

Table 2 – *Continued from previous page*

Species name	GBIF taxon ID	ME	TPR	AUC	N
Alnus glutinosa	2876213	0.00	0.71	0.77	91,292
Alnus incana	2876388	-0.03	0.76	0.95	6,873
Betula spp.	2875008	-0.03	0.63	0.83	7,313
Carpinus betulus	2875818	0.00	0.75	0.89	22,765
Carpinus orientalis	2875780	0.07	**0.21**	0.92	284
Castanea sativa	5333294	0.00	0.74	0.91	13,049
Celtis australis	2984492	-0.01	0.54	0.92	594
Cornus mas	3082263	0.03	0.51	0.90	827
Cornus sanguinea	3082234	-0.03	0.59	0.82	8,837
Corylus avellana	2875979	-0.02	0.67	0.76	48,140
Cupressus sempervirens	2684030	-0.04	**0.21**	0.70	284
Euonymus europaeus	3169131	-0.02	0.61	0.83	12,119
Fagus sylvatica	2882316	0.00	0.73	0.81	89,044
Frangula alnus	3039454	-0.02	0.71	0.86	26,873
Fraxinus angustifolia	7325877	-0.05	0.63	0.94	1,757
Fraxinus excelsior	3172358	0.00	0.67	0.74	91,111
Fraxinus ornus	3172347	0.02	0.86	0.99	2,765
Ilex aquifolium	5414222	-0.01	0.66	0.82	26,873
Juglans regia	3054368	-0.03	0.60	0.89	3,643
Juniperus communis	2684709	-0.03	0.71	0.86	21,189
Juniperus oxycedrus	2684451	-0.07	0.71	0.97	1,705
Juniperus phoenicea	2684640	-0.07	0.74	0.98	1,137
Juniperus thurifera	2684528	-0.03	0.87	0.99	1,886
Larix decidua	2686212	-0.01	0.71	0.89	15,581
Olea europaea	5415040	0.00	0.90	0.99	7,080
Ostrya carpinifolia	5332305	0.06	0.90	0.99	1,809
Picea abies	5284884	0.02	0.76	0.86	122,713
Picea sitchensis	5284827	0.05	0.80	0.96	13,023
Pinus cembra	5285134	-0.01	0.77	0.96	853
Pinus halepensis and Pinus brutia	5285604	0.03	0.86	0.99	16,951
Pinus mugo	5285385	0.00	0.85	0.98	6,667
Pinus nigra	5284809	0.01	0.79	0.93	13,540
Pinus pinaster	5285565	0.01	0.86	0.98	17,080
Pinus pinea	5285165	-0.04	0.85	0.99	4,910
Pinus sylvestris	5285637	0.02	0.78	0.85	153,928
Populus alba	3040233	-0.01	0.54	0.86	4,522
Populus nigra	3040227	-0.01	0.65	0.89	5,478
Populus tremula	3040249	-0.02	0.66	0.74	44,057
Prunus avium	3020791	-0.01	0.63	0.77	25,711
Prunus cerasifera	3021730	0.00	0.73	0.94	3,928
Prunus mahaleb	3022789	-0.01	**0.31**	0.75	517
Prunus padus	3021037	-0.03	0.63	0.78	21,705
Prunus spinosa	3023221	-0.01	0.69	0.81	31,783
Quercus cerris	2880580	0.00	0.80	0.97	4,109
Quercus ilex	2879098	0.02	0.85	0.99	22,972
Quercus pubescens	2881283	0.01	0.86	0.98	9,096
Quercus pyrenaica	2878826	0.00	0.88	0.99	6,253
Quercus robur and Quercus petraea	2878688	0.01	0.69	0.76	141,938
Quercus suber	2879411	-0.04	0.86	0.99	5,504
Robinia pseudoacacia	5352251	0.01	0.71	0.90	13,411
Salix alba	5372513	0.02	0.72	0.90	11,938
Salix caprea	5372952	-0.03	0.68	0.78	40,879
Sambucus nigra	2888728	0.00	0.70	0.81	44,961
Sorbus aria	3012680	-0.01	0.59	0.87	5,426
Sorbus aucuparia	3012167	-0.01	0.70	0.76	86,977
Sorbus domestica	3013206	-0.04	**0.48**	0.87	801
Sorbus torminalis	3012567	-0.03	0.62	0.92	2,558
Taxus baccata	5284517	-0.02	0.58	0.82	8,062
Tilia spp.	3152041	-0.02	0.50	0.82	4,393
Ulmus spp.	2984510	-0.03	0.64	0.92	5,426
Tilia spp.	3152041	0.00	0.58	0.85	4,522
Ulmus spp.	2984510	-0.02	0.69	0.91	5,375

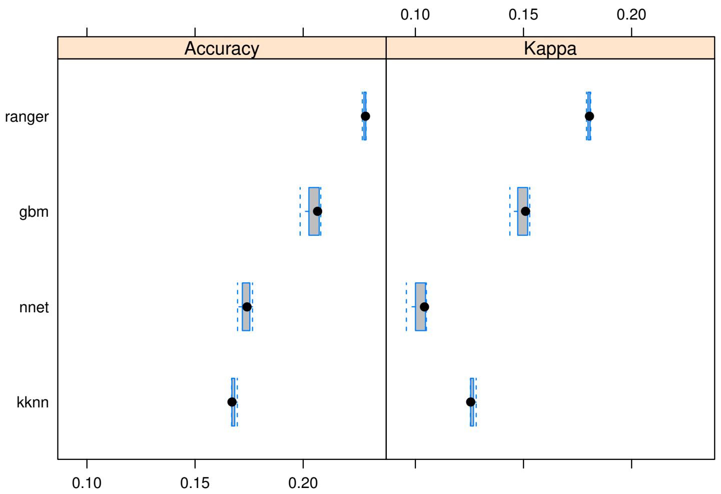


Figure 8. Predictive performance of the target machine learning algorithms for mapping forest tree species (N =1.5 million distribution of training points is available in Fig. 3). ranger = random forest, gbm = Generalized Boosted Regression Models, nnet = Neural networks, kkn = K-nearest neighbors.

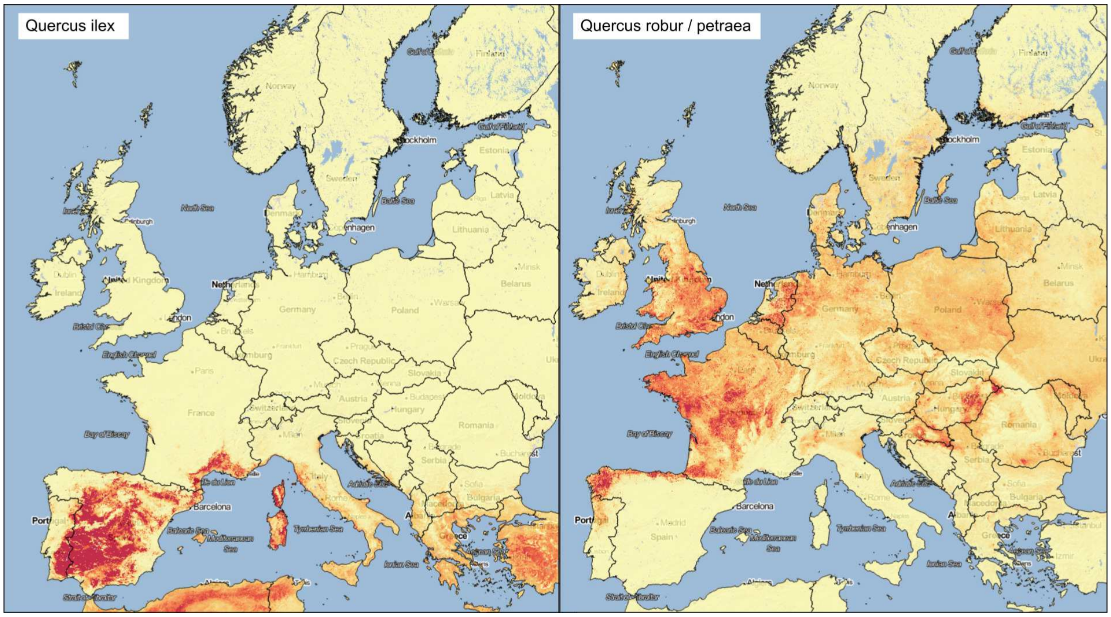


Figure 9. Examples of predicted PNV distributions (probabilities) for European forest tree species.

The most important predictors in the random forest model for forest tree taxa are mean annual daily temperature, and other monthly temperatures, elevation, CHELSA bioclimatic images, monthly precipitation and MODIS cloud fraction images. Covariates for lithology and landform classification did not feature in the top 20 predictors. It could be that the Global Lithological Map (GLiM) (Hartmann and Moosdorf, 2012), which was used to represent changes in lithology, is too general for this scale of work.

Fig. 10 illustrates differences between the map of actual distribution of *Fagus sylvatica*, generated by

PeerJ Preprints NOT PEER-REVIEWED

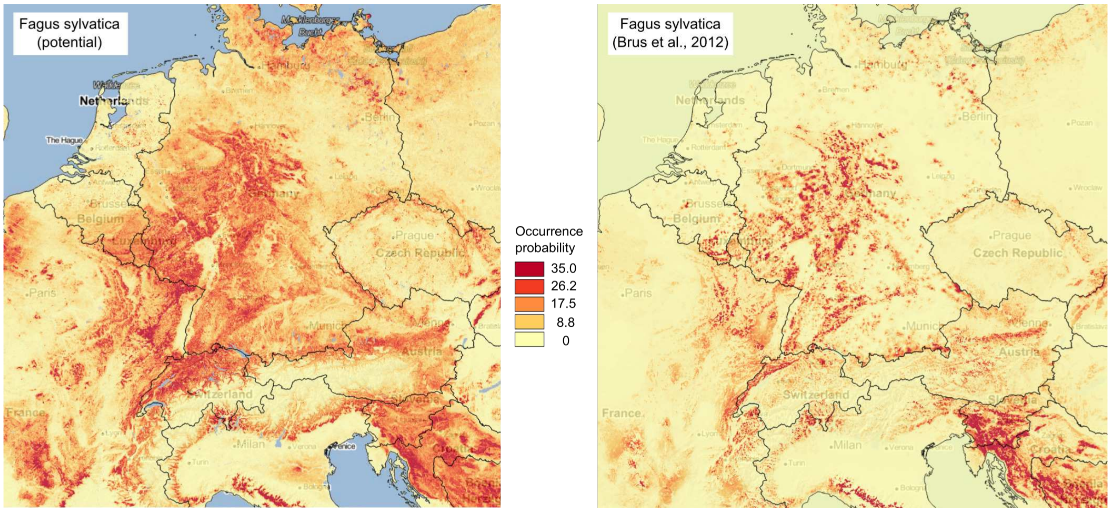


Figure 10. Comparison between predicted PNV distribution for *Fagus sylvatica* based on our results, and based on the maps generated by Brus et al. (2012) i.e. showing mapping of the actual distribution of the tree species.

Brus et al. (2012), and our predictions. In this case potential for extending habitat of *Fagus sylvatica* is significant especially over parts of France and Germany.

Correlation analysis (predicted distribution maps) indicates that many forest species are positively correlated especially *Fagus sylvatica* and *Abies alba* and *Populus nigra* and *Salix alba*. High overlap between species probability maps reflects co-existence within communities, and thus provides a way of objectively defining forest communities. Table 2 indicates that even if individual taxa are highly correlated, they can still be mapped reasonably accurately and often with a TPR above 0.80.

PeerJ Preprints NOT PEER-REVIEWED

Global monthly FAPAR

The random forest approach also provides the best preditcions of potential FAPAR (Fig. 11). The models for FAPAR are highly significant with R-squared around 90 % and RMSE at ±24 (original values in the range 0–232 where 235 corresponds to FAPAR=100 %) for the most accurate model based on 5–fold Leave-Location-Out cross-validation. However, unlike with biomes and forest species distributions, the performance of the regression-tree Cubist model shows equal performance with random forest. The most important covariates for predicting FAPAR are total annual precipitation, MODIS cloud fraction images, CHELSA bioclimatic images, and monthly precipitation images. The caret package further suggest that `mtry` parameter for the Random Forest needs to be set higher than the default values for modeling FAPAR. Setting up `mtry` >25 helps reduce the RMSE by about 7–8 %.

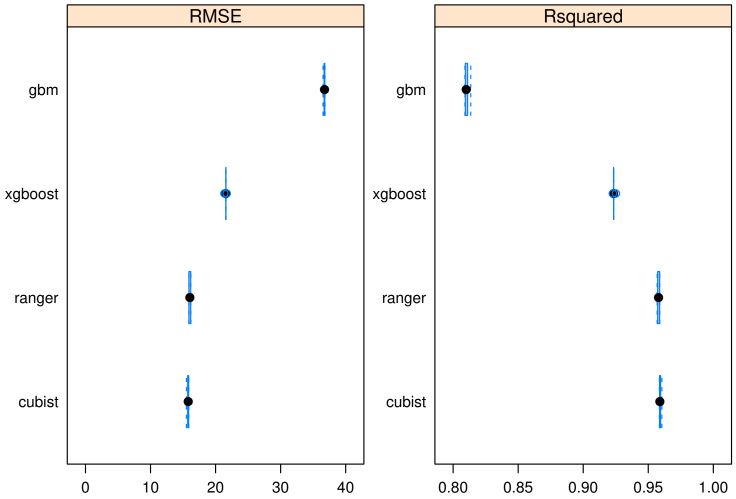


Figure 11. Predictive performance of four machine learning algorithms for mapping global distribution of FAPAR ($N = 180,990$). gbm = Generalized Boosted Regression Models, xgboost = Extreme Gradient Boosting, ranger = random forest, cubist = Cubist Regresion Models.

Fig. 12 depicts an example of actual vs predicted (PNV based) FAPAR for February in the urban area around São Paulo, where lower actual FAPAR reflects the removal of natural vegetation. Even larger differences between the potential and actual FAPAR are observed in parts of Africa (Fig. 13), likely reflecting land degradation and destruction of vegetation cover. In areas of intensive agricultural production (e.g. Western Australia and Midwest USA), actual FAPAR can be much higher than potential FAPAR under potential natural vegetation in a given month. However this is often a temporal effect, as when PNV FAPAR is aggregated over the whole year, most places modified by human management show actual FAPAR is lower than potential. In Western Australian cropping zones for example, crop fields have higher FAPAR during the winter growing season, but since the fields are bare for most of the year, aggregated annual PNV FAPAR is higher overall. Whilst this pattern may hold for rain-fed agriculture, in intensively irrigated areas the FAPAR of the managed vegetation can be much higher than of the PNV over the whole year, especially in arid and semi-arid areas (e.g. Nile Delta). This supplemental irrigation, plus the fact that total annual precipitation is the most important covariate, indicates that water availability/use

PeerJ Preprints NOT PEER-REVIEWED

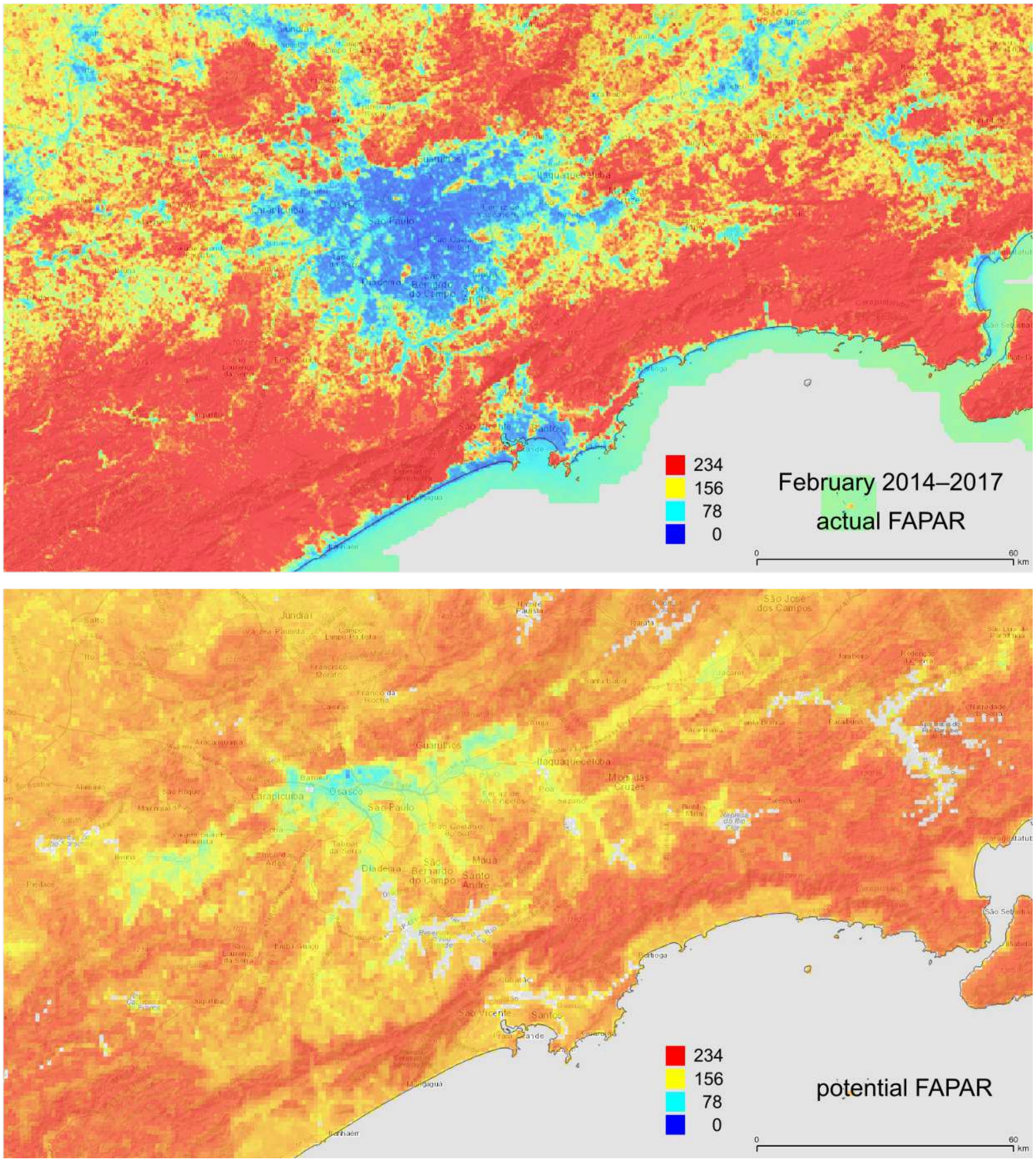


Figure 12. Actual (250 m resolution) and predicted (1 km resolution) FAPAR values for February based on the PNV samples: a zoom in area around the city of São Paulo in Brazil.

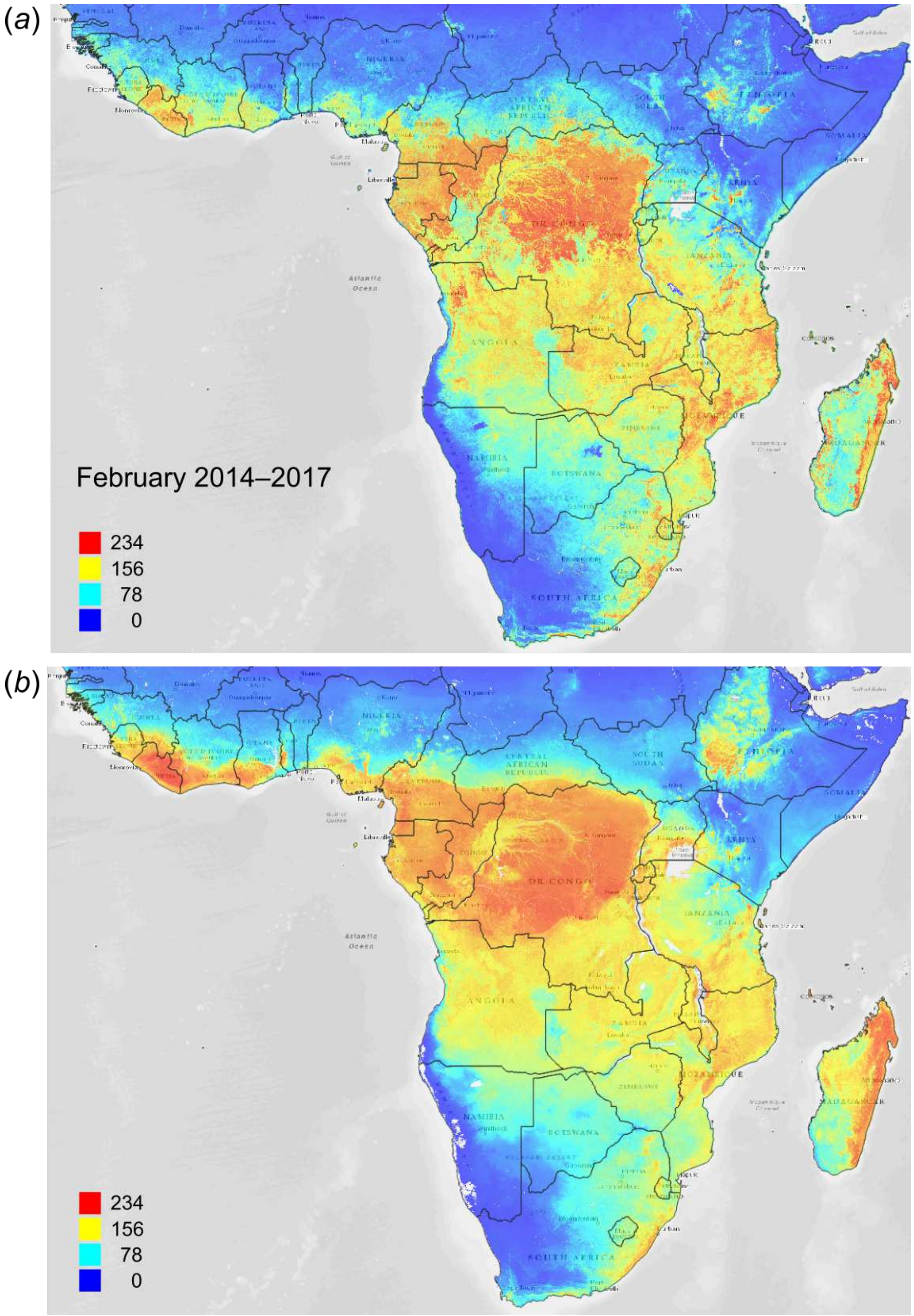


Figure 13. FAPAR values for Subsaharan Africa: (a) actual (250 m resolution) and (b) predicted (1 km resolution) potential FAPAR values for February.

PeerJ Preprints NOT PEER-REVIEWED

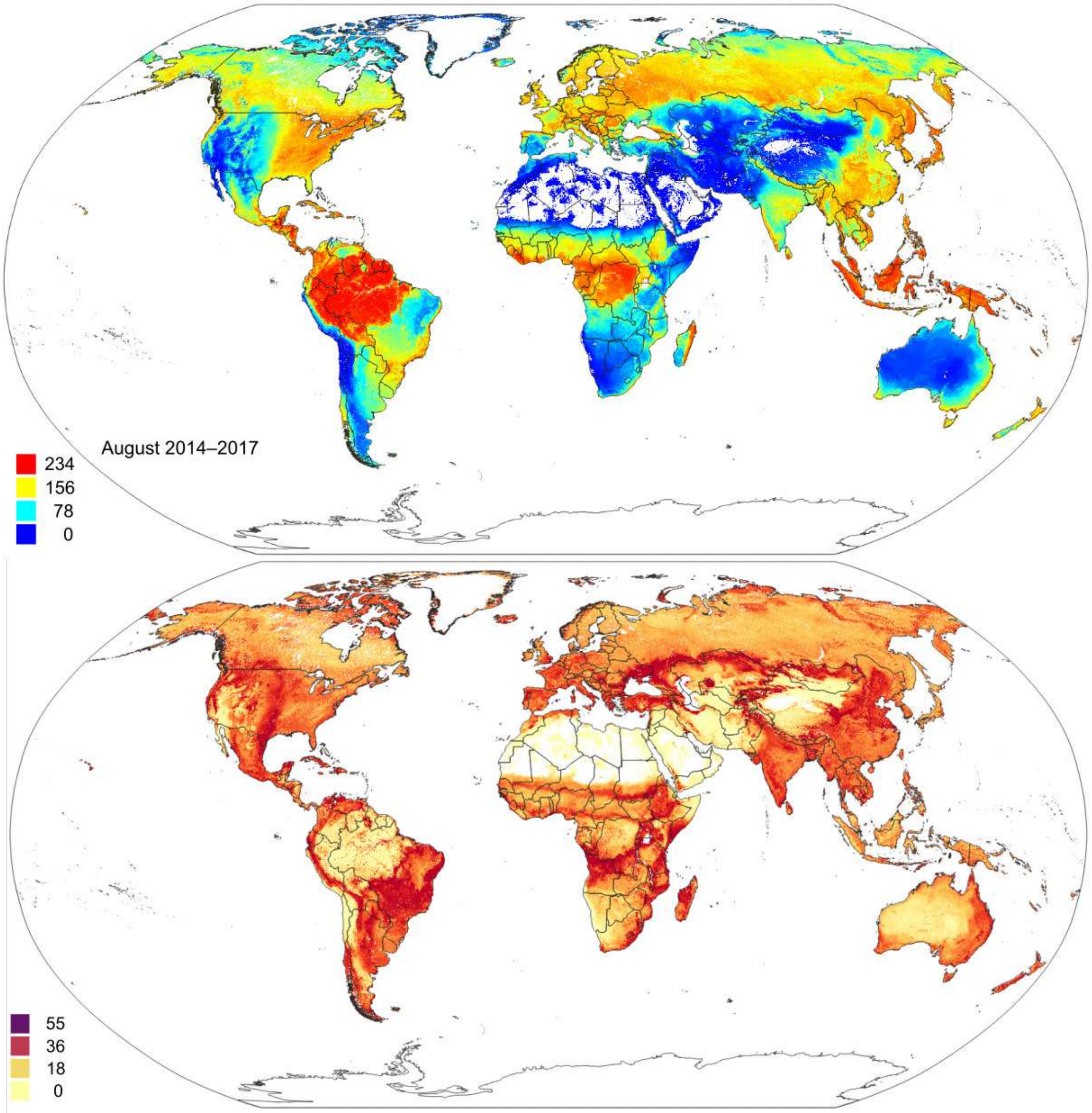


Figure 14. Predicted global FAPAR values for August (above) and standard deviation of the prediction error for the map above (below). To convert to percent divide by 253.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018

efficiency is likely the main driver of FAPAR beyond natural conditions.

Maps of the standard deviation (s.d.) of the prediction error (Fig. 14) as derived in the ranger package by using the quantreg setting (Meinshausen, 2006) provide useful information about model quality i.e. where collection of additional points would maximize model improvement and which additional covariates could be considered. For example, the highest prediction errors for FAPAR for month August occur in the transition areas between tropical forest and savanna areas, and in various biome transition zones in Asia.

DISCUSSION

Accuracy and reliability of produced PNV maps

Our results of modeling potential spatial distribution of global biomes, potential FAPAR and European forest tree taxa, show that relatively accurate maps of PNV can be produced using existing data and publicly available environmental grids. In the case of the biomes and forest tree taxa case studies, random forest outperforms neural networks, gradient boosting and similar MLA's. However, random forest and Cubist models perform equally well in the case of FAPAR. Accuracy assessment results of our work indicate improvement in product accuracy in terms of higher spatial detail and smaller classification error than the mapping products of Levavasseur et al. (2012) and Tian et al. (2016).

Precipitation, temperature maps and bioclimatic images are consistently the most important covariates in all three case studies. Lithology/parent material are not indicated as significantly important covariates in any of the case studies. This may be because the existing lithologic map (Hartmann and Moosdorf, 2012) is not detailed enough, and/or it may also reflect the fact that differences in lithology/parent material are important at finer resolutions/scales than those mapped here. Landform and lithology/parent material covariates may be important at local scales, but globally vegetation distribution seems to be dominated by climate. This is not surprising since nutrient availability is also partially controlled by climate and partially by the vegetation itself. Upon visualization of the mapping products however, it was noticed that the influence of topography is visible, especially in the maps of European forest tree taxa, suggesting that DEM derivatives are still important for mapping PNV.

Further improvements in prediction accuracy of global biome may be limited due to:

1. BIOME reconstructions representing the vegetation of an area around the site rather than at the exact point location, since the source of the pollen is on the order of 10–30 km around the site.

2. The ambiguity of reconstructions for about 10 % of the sites, so that maximum accuracy of any prediction technique may not exceed 90 % without additional observation data.

3. The fact that the BIOME reconstruction accuracy is known to be lower at ecotonal boundaries and in mountainous areas because of pollen transport issues, particularly the long-distance transport of tree pollen.

4. The BIOME data set is compiled from many regional reconstructions and all harmonization was done a posteriori, which may have introduced additional noise in the data.

So far, we did not explore opportunities for combining multiple MLA models based on validation data i.e. for doing ensemble predictions, model averages or model stacks. Stacking models can improve upon individual best techniques even up to >30 %, with the additional costs being higher computation loads (Michailidis, 2017). In our case, the extensive computational load from derivation of models and product predictions had already obtained improved accuracies, making increasing computing loads further still a matter of diminishing returns.

Our models of PNV FAPAR are based on simulated point data and accuracy of how well models represent natural vegetation areas are dependent on the representativeness of the http://protectedplanet.

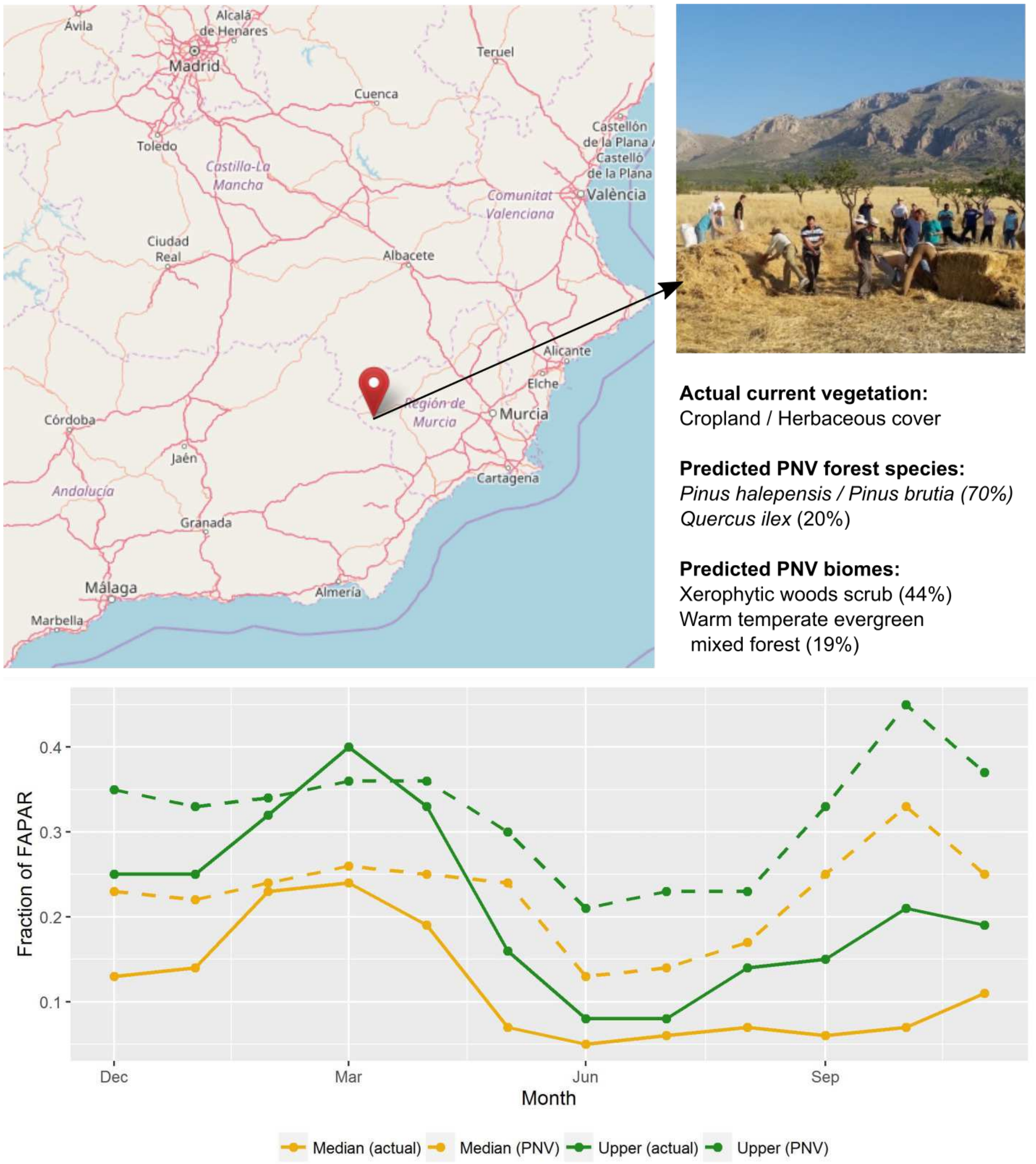
net and http://intactforests.org data. Also, many of the world's biomes such as the Mediterranean region and similar, have sustained high levels of human impact in the past and are perhaps under-represented in the http://protectedplanet.net data set. Nevertheless, our cross-validation results (Leave-Location-Out method) indicate a good match between training and validation points.

It would be further useful to explore what the performance of the models we used would be if we removed whole continents in the cross-validation process, or at least larger countries such as USA, China, Brazil, Australia, India and/or the South African Republic. For biomes, spatial Cross Validation showed a significant drop in accuracy; taking out some larger country from model training will likely also make difference. We did not explore effects of spatial proximity on mapping forest species and FAPAR as these are very dense point data sets. In addition, FAPAR training points were generated using simple random sampling, so spatial clustering should be non-existent.

Possible uses of the produced PNV maps

Newbold et al. (2016) have argued that many terrestrial biomes today have transgressed safe limits for biodiversity, with grasslands being most affected, and tundra and boreal forests least affected. *"Slowing or reversing the global loss of local biodiversity will require preserving the remaining areas of natural (primary) vegetation and, so far as possible, restoring human-used lands to natural."* (Newbold et al., 2016) Roughly half of the difference of around 466 billion tonnes of carbon can be attributed to the clearing of forests and woodlands, mostly for agricultural purposes (Erb et al., 2017). The other half of biomass carbon stock losses is derived from the management effects within a land cover class (Erb et al., 2017). The expansion of agriculture will probably continue in the coming years, leading to decreased biodiversity and soil degradation (Mauser et al., 2015; Molotoks et al., 2017). On the other hand, Griscom et al. (2017) identify reforestation (e.g. biomass restoration) as the largest natural pathway to hold global warming below 2 °C. In that context, accurate maps of PNV could become increasingly useful for assessing the level of land degradation/biomass shortfall against the potential of a site. Such information can also inform selection of optimal steps towards restoring biomass stocks in managed vegetation in ways that better reflect the PNV FAPAR in those areas.

Other uses of PNV maps include assessing the land potential i.e. land use efficiency given the difference between actual and potential vegetation. Consider for example a location in southern Spain called *"Altiplano Estepario"*, which has been identified by the Commonland company (http://commonland.com) and partners as a landscape restoration site. Fig. 15 shows results of a spatial query for this location and values of our PNV and PNV FAPAR predictions, in comparison to the actual land cover and actual FAPAR images. The figure shows that the actual FAPAR is as good as PNV FAPAR in February and March but that differences are large in the summer months. Overall, the median and upper FAPAR for this specific location are only 51 % of the PNV FAPAR, so we can say that this site is currently operating at 51 % of the predicted FAPAR capability under PNV. This comparison should also consider that our estimates of FAPAR come with an RMSE of ±0.085. Furthermore, as landscape restoration efforts have recently begun on this site — this work suggests that it ought to be possible to: (a) identify priority areas of PNV FAPAR shortfall, (b) use this information to inform in part the type of restoration strategies used, and (c) monitor the progress of restoration efforts in monthly time steps over several decades. Such



Actual current vegetation:
Cropland / Herbaceous cover

Predicted PNV forest species:
Pinus halepensis / Pinus brutia (70%)
Quercus ilex (20%)

Predicted PNV biomes:
Xerophytic woods scrub (44%)
Warm temperate evergreen
mixed forest (19%)

Figure 15. Example of comparison between the actual land cover and actual FAPAR curves and our predicted potential natural vegetation (PNV) and predicted PNV FAPAR curves. According to our results, this location in southern Spain (latitude=37.957332, longitude=-2.163181) currently utilizes 51 % of the predicted FAPAR capability under PNV, indicating a substantive short fall in on-site photosynthetically active biomass. Background map source: OpenStreetMap; photograph source: Commonland.

practical measurement, monitoring and verification efforts are required to mobilize further investment in this emerging sector.

Our PNV maps could also be used to estimate soil carbon sequestration and/or evapotranspiration potential, and gains in net primary productivity assuming return of natural vegetation (Fig. 16). Further more, by combining various estimates of potential natural and managed vegetation, one could design the optimal use of land both regionally and globally. Herrick et al. (2013), for example, provide a theoretical

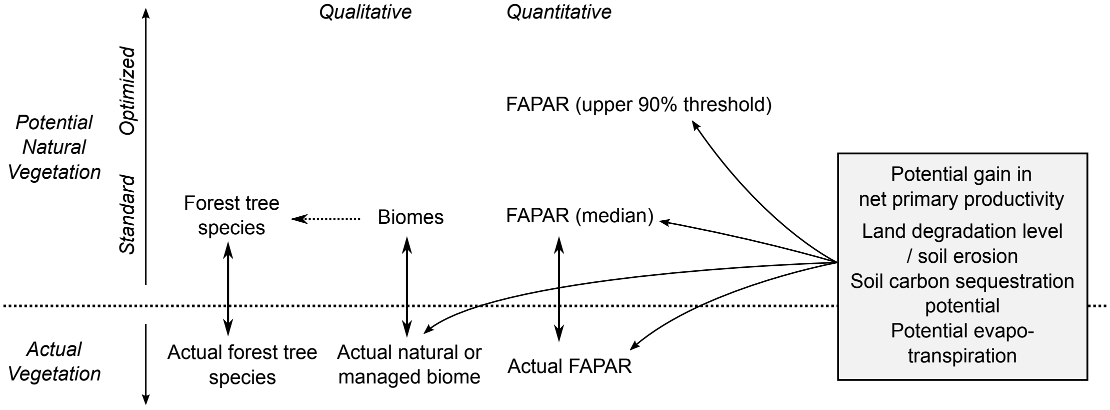

Qualitative
Quantitative
Potential Natural Vegetation
Optimized
Standard
Actual Vegetation
FAPAR (upper 90% threshold)
Forest tree species
Biomes
FAPAR (median)
Actual forest tree species
Actual natural or managed biome
Actual FAPAR
Potential gain in net primary productivity
Land degradation level / soil erosion
Soil carbon sequestration potential
Potential evapo-transpiration

Figure 16. Some possible uses of maps of Potential Natural Vegetation.

framework for estimating land potential productivity which could theoretically connect all land owners in the world to share local and regional knowledge.

Maps of PNV for European tree species could also be used as a supplement to the distribution and ecology of tree species produced by San-Miguel-Ayanz et al. (2016) and Brus et al. (2012). PNV type analysis could be made even more quantitative so that even predictions of dendrometric properties of tree species could be produced using similar frameworks. Also, similar PNV mapping algorithms could be used to map the potential canopy height based on the previously estimated map of the global canopy height (Simard et al., 2011).

Technical limitations and further challenges

Running Machine Learning Algorithms on larger and larger data is computational demanding; however, by using fully parallelized implementation of random forest in the ranger package, we were able to produce spatial predictions within days. Model fitting and prediction using EU Forest and GBIF data (1.5 million training points) was, however, very memory and time consuming and is not recommended for systems with <126 GiB RAM. In our case, model fitting took several hours even with full parallelization, and final models were >10 GiB in size. Prediction of probabilities took additional 5–6 hours with the current computational set-up. In the future, scalable cloud computing could be used to overcome some of these computational limits.

With enough computing capacity, one could theoretically use all 160 million records of distribution of plant species currently available via GBIF (Meyer et al., 2016) and from other national inventories, and map global distribution of each forest tree species. In Europe the list is very short; globally this list could be quite long (e.g. 60,000 species). The primary problems of using GBIF for PNV mapping will remain however, as these are primarily due to high clustering of points and under-representation of often inaccessible areas with very high biodiversity (Yesson et al., 2007; Meyer et al., 2016). GBIF records have been shown in the past to give biased results (Escribano et al., 2016), so that spatial prediction methods that account for high spatial clustering, i.e. bias in training point representation in both space and time; would need to be developed further to minimize such effects.

CONCLUSIONS

Although PNV is a hypothetical concept, ground-truth observations can be used to cross-validate PNV models and produce an objective estimate of accuracy. As the prediction accuracy becomes more significant, the reliability of the PNV maps increases. Our analyses show that the highest accuracy for predicting 20 biome classes is about 68 % (33 % with spatial Cross Validation) with the most important predictors being total annual precipitation, monthly temperatures and bioclimatic layers. Predictions of 73 forest tree species had a mapping accuracy of 25 % and with average TPR of 0.69, with the most important predictors being mean annual and monthly temperatures, elevation and monthly cloud fraction. Regression models for FAPAR (monthly images) were most accurate with R-square of 90 % (Leave-Location-Out CV) and with the most important predictors being total annual precipitation, MODIS cloud fraction images, CHELSA bioclimatic layers and month of the year, respectively. Machine learning can be successfully used to model vegetation distribution, and is especially applicable when the training data sets consists of large number of observations and large number of covariates. Extending the coverage of observations of natural and managed vegetation, including through making new ground observations, will allow regular improvements of such PNV maps.

ACKNOWLEDGMENTS

Funding was provided by The Nature Conservancy and the Doris Duke Charitable Foundation. SPH acknowledges support from the ERC-funded project GC2.0 (Global Change 2.0: Unlocking the past for a clearer future, grant number 694481) and from JPI-Belmont Forum via NERC for the project "PAleao-Constraints on Monsoon Evolution and Dynamics (PACMEDY)". This research is a contribution to the AXA Chair Programme in Biosphere and Climate Impacts and the Imperial College initiative on Grand Challenges in Ecosystems and the Environment (ICP). Authors are grateful to Karger et al. (2017) for maintaining the CHELSA Climate images, US agencies NASA and USGS for distributing high resolution images of Earth's atmosphere and the European Copernicus Land program. We are grateful to Mauri et al. (2017) for sharing the EU-Forest — a high-resolution tree occurrence dataset for Europe. We are also grateful to the Open Source software developers of the packages ranger, xgboost, caret, raster, GDAL, SAGA GIS and similar, and without which this work would have not be possible.

REFERENCES

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227.

Bigelow, N. H., Brubaker, L. B., Edwards, M. E., Harrison, S. P., Prentice, I. C., Anderson, P. M., Andreev, A. A., Bartlein, P. J., Christensen, T. R., Cramer, W., et al. (2003). Climate change and arctic ecosystems: 1. vegetation changes north of 55 n between the last glacial maximum, mid-holocene, and present. *Journal of Geophysical Research: Atmospheres*, 108(D19).

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.

Bohn, U., Zazanashvili, N., and Nakhutsrishvili, G. (2007). The map of the natural vegetation of europe

PeerJ Preprints NOT PEER-REVIEWED

and its application in the caucasus ecoregion. *Bulletin of the Georgian National Academy of Sciences*, 175:112–121.

Borda, M. (2011). *Fundamentals in Information Theory and Coding*. Springer Berlin Heidelberg.

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

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

Brus, D., Hengeveld, G., Walvoort, D., Goedhart, P., Heidema, A., Nabuurs, G., and Gunia, K. (2012). Statistical mapping of tree species over europe. *European Journal of Forest Research*, 131(1):145–157.

Carnahan, J. (1989). *Australia natural vegetation: Australia's vegetation in the 1780's*. Australian Surveying and Land Information Group, Dept. of Administrative Services, Queensland, AU.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. (2015). System for automated geoscientific analyses (saga) v. 2.1.4. *Geoscientific Model Development*, 8(7):1991–2007.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1):677.

Erb, K.-H., Kastner, T., Plutzar, C., Bais, A. L. S., Carvalhais, N., Fetzel, T., Gingrich, S., Haberl, H., Lauk, C., Niedertscheider, M., et al. (2017). Unexpectedly large impact of forest management and grazing on global vegetation biomass. *Nature*.

Escribano, N., Ariño, A. H., and Galicia, D. (2016). Biodiversity data obsolescence and land uses changes. *PeerJ*, 4:e2743.

Fan, Y., Li, H., and Miguez-Macho, G. (2013). Global patterns of groundwater table depth. *Science*, 339(6122):940–943.

Fluet-Chouinard, E., Lehner, B., Rebelo, L.-M., Papa, F., and Hamilton, S. K. (2015). Development of a global inundation map at high spatial resolution from topographic downscaling of coarse-scale remote sensing data. *Remote Sensing of Environment*, 158:348–361.

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

Griscom, B. W., Adams, J., Ellis, P. W., Houghton, R. A., Lomax, G., Miteva, D. A., Schlesinger, W. H., Shoch, D., Siikamäki, J. V., Smith, P., Woodbury, P., Zganjar, C., Blackman, A., Campari, J., Conant, R. T., Delgado, C., Elias, P., Gopalakrishna, T., Hamsik, M. R., Herrero, M., Kiesecker, J., Landis, E., Laestadius, L., Leavitt, S. M., Minnemeyer, S., Polasky, S., Potapov, P., Putz, F. E., Sanderman, J., Silvius, M., Wollenberg, E., and Fargione, J. (2017). Natural climate solutions. *Proceedings of the National Academy of Sciences*, 114(44):11645–11650.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., Thau, D., Stehman,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018

PeerJ Preprints NOT PEER-REVIEWED

S., Goetz, S., Loveland, T., et al. (2013). High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853.

Harrison, S., Yu, G., Takahara, H., and Prentice, I. (2001). Plant diversity and palaeovegetation in east asia. *Nature*, 413:129–130.

Harrison, S. P. and Bartlein, P. (2012). Chapter 14 — records from the past, lessons for the future: What the palaeorecord implies about mechanisms of global change. In Henderson-Sellers, A. and McGuffie, K., editors, *The Future of the World's Climate (Second Edition)*, pages 403 – 436. Elsevier, Boston, second edition edition.

Hartmann, J. and Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems*, 13(12):n/a–n/a.

Hengl, T., Nussbaum, M., Wright, M. N., and Heuvelink, G. B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ Preprints*, 6:e26693v1.

Herrick, J. E., Urama, K. C., Karl, J. W., Boos, J., Johnson, M.-V. V., Shepherd, K. D., Hempel, J., Bestelmeyer, B. T., Davies, J., Guerra, J. L., et al. (2013). The global land-potential knowledge system (landpks): Supporting evidence-based, site-specific land use and management through cloud computing, mobile applications, and crowdsourcing. *Journal of Soil and Water Conservation*, 68(1):5A–12A.

Hijmans, R. J. and Elith, J. (2018). *Species distribution modeling with R*. Environmental Science and Policy, University of California.

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*, 4:170122.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(1):1–26.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., and Quinlan, R. (2014). *Cubist: rule-and instance-based regression modeling*. R package version 0.0.

Leong, M. and Roderick, G. K. (2015). Remote sensing captures varying temporal patterns of vegetation between human-altered and natural landscapes. *PeerJ*, 3:e1141.

Levavasseur, G., Vrac, M., Roche, D., and Paillard, D. (2012). Statistical modelling of a new global potential vegetation distribution. *Environmental Research Letters*, 7(4):044019.

Marchant, R., Cleef, A., Harrison, S., Hooghiemstra, H., Markgraf, V., Van Boxel, J., Ager, T., Almeida, L., Anderson, R., Baied, C., et al. (2009). Pollen-based biome reconstructions for latin america at 0, 6000 and 18 000 radiocarbon years ago. *Climate of the Past*, 5:725–767.

Marinova, E., Harrison, S. P., Bragg, F., Connor, S., Laet, V., Leroy, S. A., Mudie, P., Atanassova, J., Bozilova, E., Caner, H., et al. (2017). Pollen-derived biomes in the eastern mediterranean–black sea–caspian-corridor. *Journal of Biogeography*.

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

Mauser, W., Klepper, G., Zabel, F., Delzeit, R., Hank, T., Putzenlechner, B., and Calzadilla, A. (2015). Global biomass production potentials exceed expected future demand without the need for cropland expansion. *Nature communications*, 6:8946.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Meyer, C., Weigelt, P., and Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, 19(8):992–1006.

Michailidis, M. (2017). *Investigating machine learning methods in recommender systems*. PhD thesis, UCL (University College London).

Molotoks, A., Kuhnert, M., Dawson, T. P., and Smith, P. (2017). Global Hotspots of Conflict Risk between Food Security and Biodiversity Conservation. *Land*, 6(4).

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L., Hoskins, A. J., Lysenko, I., Phillips, H. R., et al. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? a global assessment. *Science*, 353(6296):288–291.

Omernik, J. M. (1987). Ecoregions of the conterminous united states. *Annals of the Association of American geographers*, 77(1):118–125.

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*.

Pickett, E. J., Harrison, S. P., Hope, G., Harle, K., Dodson, J. R., Peter Kershaw, A., Colin Prentice, I., Backhouse, J., Colhoun, E. A., D'Costa, D., et al. (2004). Pollen-based reconstructions of biome distributions for australia, southeast asia and the pacific (seapac region) at 0, 6000 and 18,000 14c yr bp. *Journal of Biogeography*, 31(9):1381–1444.

Potapov, P., Laestadius, L., and Minnemeyer, S. (2011). *Global Map of Potential Forest Cover*. World Resources Institute.

Potapov, P., Yaroshenko, A., Turubanova, S., Dubinin, M., Laestadius, L., Thies, C., Aksenov, D., Egorov, A., Yesipova, Y., Glushkov, I., et al. (2008). Mapping the world's intact forest landscapes by remote sensing. *Ecology and Society*, 13(2).

Prentice, I. C. and Jolly, D. (2000). Mid-holocene and glacial-maximum vegetation geography of the northern continents and africa. *Journal of Biogeography*, 27(3):507–519.

Ridgeway, G. (2017). *gbm: generalized boosted regression models*. R package version 1.6-3.1.

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

San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., and Mauri, A. (2016). *European Atlas of forest tree species*. European Commission, Joint Research Centre.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Simard, M., Pinto, N., Fisher, J. B., and Baccini, A. (2011). Mapping forest canopy height globally with spaceborne lidar. *Journal of Geophysical Research: Biogeosciences*, 116(G4):NA.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2016). *ROCR: Visualizing the Performance of Scoring Classifiers*. R package version 1.0.7.

Tian, H., Lu, C., Ciais, P., Michalak, A. M., Canadell, J. G., Saikawa, E., Huntzinger, D. N., Gurney,

PeerJ Preprints NOT PEER-REVIEWED

K. R., Sitch, S., Zhang, B., et al. (2016). The terrestrial biosphere as a net source of greenhouse gases to the atmosphere. *Nature*, 531(7593):225–228.

Veloso, H. P., Oliveira-Filho, L., Vaz, A., Lima, M., Marquete, R., and Brazao, J. (1992). Manual técnico da vegetação brasileira.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer-Verlag, New York, 4th edition.

Wilson, A. M. and Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLOS Biology*, 14(3):1–20.

Wright, M. N. and Ziegler, A. (2016). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, page 18.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., and Culham, A. (2007). How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, 2(11):e1124.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26811v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018