Received: 9 September 2021 | Accepted: 13 January 2022

DOI: 10.1111/gcb.16110

Global Change Biology WILEY

RESEARCH ARTICLE

Identifying climate thresholds for dominant natural vegetation types at the global scale using machine learning: Average climate versus extremes

Rita Beigaitė[1] | Hui Tang[2,3] | Anders Bryn[2] | Olav Skarpaas[2] | Frode Stordal[3] | Jarle W. Bjerke[4] | Indrė Žliobaitė[1,5]

[1]Department of Computer Science, University of Helsinki, Helsinki, Finland

[2]Natural History Museum, University of Oslo, Oslo, Norway

[3]Department of Geosciences, University of Oslo, Oslo, Norway

[4]Norwegian Institute for Nature Research, FRAM – High North Research Centre for Climate and the Environment, Tromsø, Norway

[5]Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

Correspondence
Rita Beigaitė, Department of Computer Science, University of Helsinki, Helsinki, Finland.
Email: rita.beigaite@helsinki.fi

Funding information
Norges Forskningsråd, Grant/Award Number: 294948; Universitetet i Oslo, Grant/Award Number: UiO/GEO1039; Academy of Finland, Grant/Award Number: 314803

Abstract

The global distribution of vegetation is largely determined by climatic conditions and feeds back into the climate system. To predict future vegetation changes in response to climate change, it is crucial to identify and understand key patterns and processes that couple vegetation and climate. Dynamic global vegetation models (DGVMs) have been widely applied to describe the distribution of vegetation types and their future dynamics in response to climate change. As a process-based approach, it partly relies on hard-coded climate thresholds to constrain the distribution of vegetation. What thresholds to implement in DGVMs and how to replace them with more process-based descriptions remain among the major challenges. In this study, we employ machine learning using decision trees to extract large-scale relationships between the global distribution of vegetation and climatic characteristics from remotely sensed vegetation and climate data. We analyse how the dominant vegetation types are linked to climate extremes as compared to seasonally or annually averaged climatic conditions. The results show that climate extremes allow us to describe the distribution and eco-climatological space of the vegetation types more accurately than the averaged climate variables, especially those types which occupy small territories in a relatively homogeneous ecological space. Future predicted vegetation changes using both climate extremes and averaged climate variables are less prominent than that predicted by averaged climate variables and are in better agreement with those of DGVMs, further indicating the importance of climate extremes in determining geographic distributions of different vegetation types. We found that the temperature thresholds for vegetation types (e.g. grass and open shrubland) in cold environments vary with moisture conditions. The coldest daily maximum temperature (extreme cold day) is particularly important for separating many different vegetation types. These findings highlight the need for a more explicit representation of the impacts of climate extremes on vegetation in DGVMs.

KEYWORDS
climate extremes, climate thresholds, decision trees, DGVMs, machine learning, vegetation distribution

1 | INTRODUCTION

Variation in climate is the major factor determining the distribution of vegetation around the world (Adams, 2009). As the world is facing climate change, large-scale future dynamics in vegetation distribution are expected, which in turn may exert strong biophysical and biochemical feedback on the climate (Pearson et al., 2013; Sitch et al., 2008). Predicting future vegetation distribution in response to climate change, however, is particularly challenging, requiring a detailed understanding of how vegetation distribution on a large scale is linked to climate. Historically, von Humboldt and Bonpland (1807) started this process by presenting the first zonal vegetation maps based on climate gradients in the high Andes, but the first quantitative classification of world climate was presented by Wladimir Köppen (Köppen, 1900; Kottek et al., 2006), in which he delineated vegetation zones by mean rainfall and monthly temperature. Other well-known attempts to classify the climatic life zones were made by Whittaker (1962) and Holdridge (1967). While insightful, these schemes did not have sufficient resolution for predicting local vegetation in many parts of the world (Adams, 2009).

In recent decades, efforts to understand and predict large-scale vegetation distributions under different climate conditions (past, present and future) have been made mainly with two bottom-up approaches. One is statistical modelling of the relationship between climate and species distribution or plant functional traits, and usage of the assembly of species or plant functional traits to predict vegetation distribution at the community or biome level (e.g. Conradi et al., 2020; Yang et al., 2019). The other is process-based vegetation modelling of large-scale vegetation distribution, such as dynamic global vegetation models (DGVMs) (e.g. Hickler et al., 2012; Ito et al., 2020; Scheiter et al., 2020; Sitch et al., 2008). DGVMs can be coupled with Earth system models (ESMs) (Fisher & Koven, 2020), thus being essential tools for predicting vegetation distribution changes and feedbacks with the climate system. Various processes have been parameterized in DGVMs to describe the large-scale dynamics of major vegetation types (referred to as plant functional types, PFTs; see Wullschleger et al., 2014), such as photosynthesis, phenology, carbon allocation, recruitment, mortality and fire disturbance (Lasslop et al., 2020). Ideally, the distribution or dominance of different PFTs should emerge from the competitions among PFTs for light, water and nutrients if the above-mentioned processes are adequately described in the model. However, in reality, simple and hard-coded climate thresholds have had to be implemented in DGVMs for various vegetation processes for which detailed descriptions are lacking, such as survival, establishment or mortality, so as to faithfully represent the geographic distribution of different PFTs (see Table 1). These hard-coded climate thresholds are one group of the most uncertain parameters in DGVMs (Forkel et al., 2019; Horvath et al., 2021; Song & Zeng, 2014; Zhu et al., 2018). They may lead to unrealistically strong and fast response of vegetation to climate changes in DGVMs, hampering their application to ESMs for the future projections (Masson-Delmotte et al., 2021).

TABLE 1 Climatic thresholds used for describing vegetation dynamics (e.g. survival and establishment) in LPJml (from Schaphoff et al., 2018). Similar climate thresholds have also been adopted by other DGVMs such as LPJ-GUESS (Miller & Smith, 2012), CLM-DGVM (Levis et al., 2004), ORCHIDEE-DGVM (Krinner et al., 2005), SDGVM (Cramer et al., 2001) and SEIB-DGVM (Sato & Ise, 2012). Here, T_{cmin} is minimum coldest monthly mean temperature, T_{cmax} is maximum coldest monthly mean temperature, GDD_{min} is minimum growing degree days (at or above 5°C)

Vegetation types	T_{cmin}	T_{cmax}	GDD_{min}
Tropical broadleaved evergreen tree	15.5	—	—
Tropical broadleaved raingreen tree	15.5	—	—
Temperate needle-leaved evergreen tree	−2	22	900
Temperate broadleaved evergreen tree	3	18.8	1200
Temperate broadleaved summergreen tree	−17.7	15.5	1200
Boreal needle-leaved evergreen tree	−32.5	−2	600
Boreal broadleaved summergreen tree	—	−2	350
Boreal needle-leaved summergreen tree	−46.5	−5.4	350
Tropical herbaceous	7	—	—
Temperate herbaceous	−39	15.5	—
Polar herbaceous	—	−2.6	—

Several recent studies have started the task of improving these hard-coded thresholds, from different perspectives (e.g. Horvath et al., 2021; Liu et al., 2018a). However, the data sources vary in resolution and quality, and only average climate thresholds are often employed in model test beds. Contemporaneously, it has been reported that climate extremes, that is, which statistically deviate from the average climate records and occurring at daily or submonthly scales, can have large impacts on biome ranges and vegetation dynamics (Julio Camarero et al., 2015; Li et al., 2018b; Shao et al., 2021; Ummenhofer & Meehl, 2017). For instance, drought can cause a decrease in dominant grass species (Li et al., 2018a), since in arid or semi-arid grassland, water is the most limiting resource for plant (Robinson et al., 2013; Yan et al., 2015). Findings of O'sullivan et al. (2017) suggest that during heatwave events combined with drought,

the upper canopy leaf metabolism may be at substantially increased risk. Phoenix and Bjerke (2016) and Treharne et al. (2020) remark that extreme weather events and winter warming can contribute to damage-induced declining vegetation productivity (browning) in the Arctic. Woodward (1990) emphasized that geographical plant distribution is influenced by low temperature extremes, for example, regulating the survival of different functional types of trees globally (Woodward et al., 2004). Plants adapted to tolerating cold in winters rarely thrive or reproduce during dormancy, and its reversal is not triggered by declining temperatures in winter or warming in spring respectively (Harrison et al., 2010). Whereas some tropical plants can be damaged by chilling temperatures (Graham & Patterson, 1982), boreal evergreen needleleaf trees can be damaged and die because of extreme warming spells when the soil is frozen (winter warming and spring drought; Song et al. (2021)) or because of extreme cold winter temperatures even at the trailing edge (Julio Camarero et al., 2015). Dahl (1998) found rough correlations with temperatures of the coldest and warmest months and the distribution of a large number of plant species in northern Europe, and related these to ecophysiological limitations such as frost tolerance and drought stress. According to Zimmermann et al. (2009), the predictive performance of species distribution models increases when mean climatic predictors are complemented by climate extremes. A changing climate influences the duration, frequency, intensity, timing and spatial extent of climate extremes (Seneviratne et al., 2012). For instance, daily temperature and precipitation extremes, in particular, have been observed to increase in frequency and intensity due to global warming (Ummenhofer & Meehl, 2017) with distinct spatial pattern from average climate changes. How climate extremes will affect vegetation distribution in the future remains largely unknown.

Machine learning techniques have become increasingly popular in the biogeosciences (Reichstein et al., 2019). Models built upon observational data offer the potential to combine a higher resolution while keeping investigations at the largest possible scales. Machine learning has been used in a variety of studies: in forest ecology (Liu et al., 2018b), rare species distribution modelling (Mi et al., 2017), calibration of aquatic microfossil proxies (Salonen et al., 2016), mapping fractional cover of an invasive plant species in a dryland ecosystem (Shiferaw et al., 2019), forest type classification (Chatterjee et al., 2016), land cover classification from remote sensing images (Abdi, 2020; Ge et al., 2020; Talukdar et al., 2020) and global mapping of potential natural vegetation (Hengl et al., 2018). In this study, we employed a decision tree approach from machine learning (Breiman et al., 1984) to explore available climate and vegetation data, and to systematically re-examine long-lasting and reappearing scientific questions regarding climate–vegetation relations. This approach enabled us to analyse whether any novel climate thresholds affecting the large-scale distribution of vegetation types could be detected, particularly climate extreme thresholds that have been overlooked in previous studies.

Decision tree models are easily interpretable, that is, it is easy to extract decision rules and trace why a certain classification is made. We trained decision tree models with the present-day global climate and vegetation data, and further tested their ability to predict natural dominant vegetation types from climatic variables. Here, the term 'dominant vegetation type' refers to a vegetation type which occupies most of the natural space in a given territory. Decision trees can provide boundary conditions for the distribution of each dominant vegetation type. To the best of our knowledge, no attempts have yet been made to use machine learning for understanding threshold conditions that govern and separate dominant vegetation types at a global scale.

We first investigated the added value of including climate extremes in the decision tree induction to demonstrate the importance of climate extremes in shaping the present-day vegetation distribution. We then applied the decision tree models to future climate scenarios and compared the results with those from DGVMs and other approaches to further demonstrate the importance of climate extremes in predicting dominant vegetation changes in the future. These results are expected to inform process-based models, such as DGVMs, to further improve their parameterization of the climate thresholds of different processes for each vegetation type rather than be used as a purely empirical approach.

2 | MATERIALS AND METHODS

2.1 | Data sources and variables

In this study, to illustrate the workflow of the method and analysis (Figure 1), as an instance, we chose MODIS (Friedl & Sulla-Menashe, 2015; Friedl et al., 2010) land cover product (MCD12C1, https://doi.org/10.5067/MODIS/MCD12C1.006), in the year 2001. This product has been produced by the data providers primarily based on supervised learning classifications of MODIS Terra and Aqua reflectance data. One of the main reasons for choosing this product was that climate data were not involved in their classification algorithm, with the exception of land surface data derived from the same satellite product (Friedl et al., 2010). Rather than blending vegetation classes from several sources, we chose a single data product for our main scenario to ensure consistency of treatment. The data product includes three different land cover classification schemes. In this study, the International Geosphere-Biosphere Programme (IGBP) classification scheme was used. The definition of the 17 land cover types in the IGBP scheme can be found in Strahler et al. (1999). The original data set had a resolution of 0.05 × 0.05 degrees. We first regridded it to 10 × 10 min grids and then resampled to 50 × 50 km grids in line with the climatic variables used in the study. MODIS land cover data provided fractions of each land cover type for a given grid cell. We extracted the dominant vegetation type variable by assigning each observation a class label of the vegetation type which had the highest fraction in a given grid cell. Since we aimed to model natural vegetation, the grid cells which had 100% human activity cover (land cover types: urban & built-up, cropland, cropland & natural vegetation mosaic), water or a combination of both were eliminated. The 13 natural

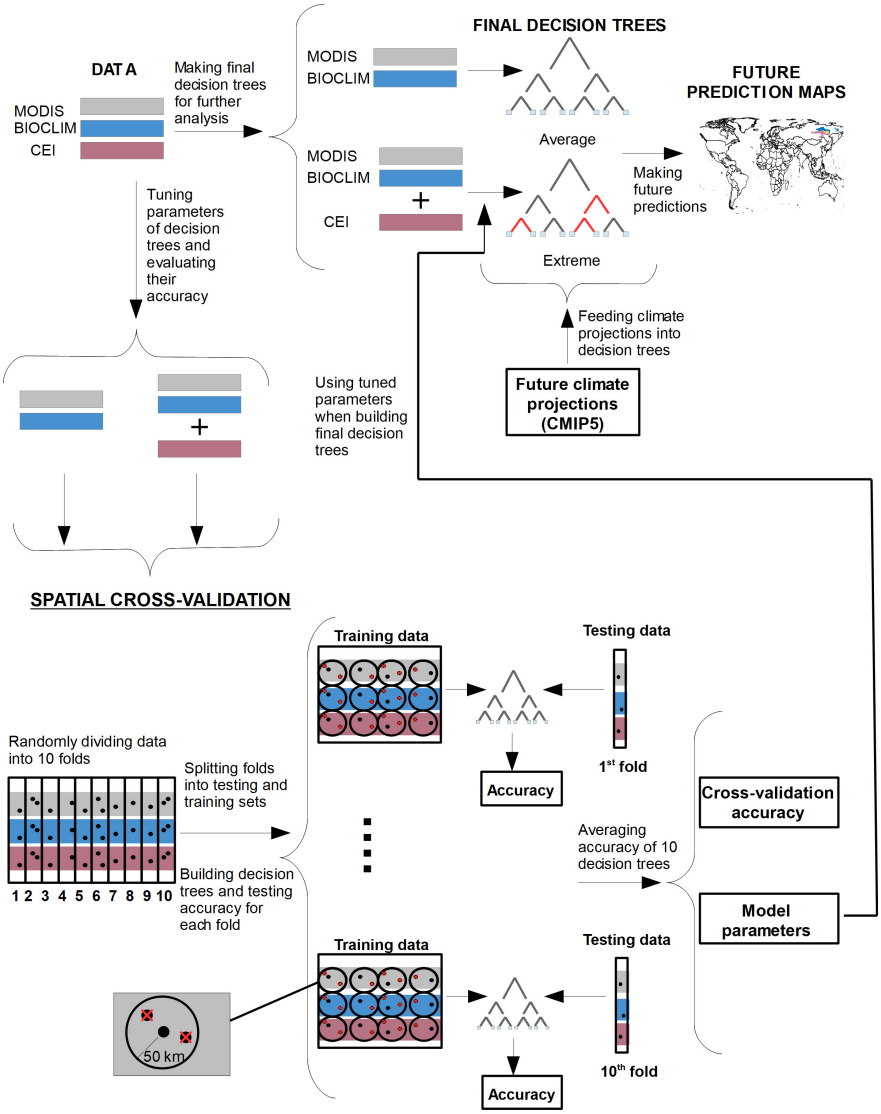


FIGURE 1 Decision tree modelling process

vegetation types used in this study are listed in Table 2. Trying to keep as many as possible observations in Europe, North America and India, where high proportions of the landscapes are dominated by intensive land use types, we made an assumption that the relative proportion of natural land cover types would remain unchanged despite human activity. For example, if the land cover of a certain area consists of 50% cropland, 30% mixed forest and 20% deciduous broadleaf forest (DBF), we assumed that the natural and dominant vegetation type for that area is mixed forest.

We used two sets of climatic variables: BIOCLIM variables from WorldClim 2 (Fick & Hijmans, 2017 downloaded from https://www.worldclim.org/data/worldclim21.html, last access: 9 September 2021) and climate extreme indices (CEIs) from CLIMDEX (Sillmann et al., 2013a, 2013b downloaded from https://climate-modelling.canada.ca/climatemodeldata/climdex/, last access: 9 September 2021). They are listed in Table 3. BIOCLIM variables were derived from multiyear averaged monthly temperature and rainfall data during 1970–2000 and have been widely used in species distribution modelling as well as other ecological modelling techniques (Galbrun et al., 2018). They represent annual means (e.g. BIO1 and BIO12), seasonality (e.g. BIO4, BIO7 and BIO15) and also limiting environmental factors on a monthly scale (e.g. BIO5, BIO6, BIO13 and BIO14). In this study, we downloaded original BIOCLIM data at a spatial resolution of 10 × 10 min and resampled them into 50 × 50 km grids using nearest neighbour interpolation for decision tree mining. Unlike the BIOCLIM variables, the CEIs better represent extreme conditions on a daily scale (e.g. minimum value of daily maximum temperature (TXn), maximum length of dry spell (CDD, consecutive dry days)). CEI indices were derived from the ERA-Interim reanalysis data set covering the period from 1979 to 2010. They are averaged over the entire 32-year period. Multiyear average of extreme indices is a common practice to show the averaged extreme conditions in the past and future (Seneviratne & Hauser, 2020; Sillmann et al., 2013b). The original resolution of the data set was 1.5 × 1.5 degrees. To match the BIOCLIM variables, CEIs were first interpolated onto 10 × 10 min grids by conservative interpolation and then resampled to 50 × 50 km grids using nearest neighbour interpolation. It has been documented that CEIs derived from ERA-Interim can reliably reproduce observed extremes (Donat et al., 2014).

TABLE 2 Natural vegetation types of MODIS data set used in modelling

Name	Description	Prevalence (%)
Evergreen needleleaf forest (ENF)	Dominated by evergreen conifer trees (canopy > 2 m). Tree cover > 60%	2.67
Evergreen broadleaf forest (EBF)	Dominated by evergreen broadleaf and palmate trees (canopy > 2 m). Tree cover > 60%	11.45
Deciduous needleleaf forest (DNF)	Dominated by deciduous needleleaf (larch) trees (canopy > 2 m). Tree cover > 60%	1.04
Deciduous broadleaf forest (DBF)	Dominated by deciduous broadleaf trees (canopy > 2 m). Tree cover > 60%	1.45
Mixed forest (MF)	Dominated by neither deciduous nor evergreen (40%–60% of each) tree type (canopy > 2 m). Tree cover > 60%	7.04
Closed shrubland	Dominated by woody perennials (1–2 m height). Tree cover > 60%	0.46
Open shrubland	Dominated by woody perennials (1–2 m height) 10%–60% cover	17.61
Woody savanna	Tree cover 30%–60% (canopy > 2 m)	10.76
Savanna	Tree cover 10%–30% (canopy > 2 m)	9.31
Grassland	Dominated by herbaceous annuals (<2 m). Tree cover < 10%	16.49
Permanent wetland	Permanently inundated lands with 30%–60% water cover and >10% vegetation cover	0.90
Permanent snow and ice (snow and ice)	At least 60% of area is covered by snow and ice for at least 10 months of the year	2.59
Barren	At least 60% of area is non-vegetated barren (sand, rock, soil) areas with <10% vegetation cover	18.25

For future projections with decision trees, BIOCLIM (http://www.worldclim.com/cmip5_10m, last access: 9 September 2021) and climate extreme variables (https://crd-data-donnees-rdc.ec.gc.ca/CCCMA/products/CLIMDEX/CMIP5/, last access: 9 September 2021) (Sillmann et al., 2013a) derived from the future climate projections of the Coupled Model Intercomparison Project Phase 5 (CMIP5) are employed. Three different future scenarios, that is, RCP (Representative Concentration Pathway) 2.6, 4.5 and 8.5, were used (Seneviratne et al., 2012). These are greenhouse gas concentration trajectories projected by the Intergovernmental Panel on Climate Change covering a wide range of possible changes in future anthropogenic greenhouse gas emissions under different socio-economic assumptions. More specifically, RCP2.6 is a low-emission pathway that would keep atmospheric carbon dioxide (CO_2) concentration similar to the present day and global temperature rise below 2°C by 2100, while RCP4.5 and RCP8.5 are the intermediate and high emission pathways that will lead to the rise of atmospheric CO_2 concentration to about 600 ppm and 1200 ppm by 2100 respectively. All the data are based on the ensemble mean of 11 models participating in CMIP5 and are averaged over two specific time periods, that is, 2041–2060 and 2061–2080.

2.2 | Machine learning procedure: Decision tree modelling of current vegetation

To model the complex associations between climate and the global distribution of dominant vegetation types while keeping the model itself transparent and interpretable, we used a decision tree approach (Breiman et al., 1984), also known as classification trees or regression trees (and conceptually unrelated to hierarchical clustering). A tree-structured predictive model allows us to reach reasonably high accuracy and extract the climatic thresholds responsible for the separation of different vegetation types. To achieve state-of-the-art accuracy, one could use tree-based methods such as random forests (Breiman, 2001) or XGBoost (Chen & Guestrin, 2016), which employ ensembles of decision trees. However, as we focused on extraction of the threshold values, we used a single tree model, which is more transparent for interpretation and has a lower risk of overfitting the data.

Similar to standard statistical approaches such as linear regression, building a decision tree model requires matching observations of climatic variables and vegetation types, a so-called training data set. A decision tree model was built iteratively by first splitting the training data set based on the climate variable that is the most informative regarding vegetation classes, then on the next most informative variable and so on until the observations at the end leaf nodes are well classified according to a selected fitness criterion. Each separation (split) into the tree leaves is not necessarily homogeneous and a small share of the observations will inevitably be misclassified.

We used R 4.0.5 suite (R Core Team, 2021), that is, the rpart (v4.1-15; Therneau & Atkinson, 2021) and the caret (v6.0-86; Kuhn et al., 2008) packages, for fitting the decision trees. Within the rpart package, decision trees are built using the classification and regression tree (CART) algorithm (Breiman et al., 1984). As a splitting criterion, we tried out the Gini index (James et al., 2013) and the information criterion (Maindonald & Braun, 2013) but chose to proceed with the Gini index, since it provided an accuracy similar to the information criterion but had lower computational time. In order to keep models simple and easy to interpret as well as prevent potential overfitting, we regulated the complexity parameter (Maindonald & Braun, 2013), which indirectly controls the number of splits by imposing a relative cost for each split. The splitting process stops when the increase in cost of complexity surpasses the reduction in relative prediction error. Based on the visual elbow

TABLE 3 Variables of BIOCLIM and CLIMDEX data sets used in modelling

ID	Description	Units
BIO1	Annual mean temperature	°C
BIO2	Mean diurnal range (mean of monthly (max temp − min temp))	°C
BIO3	Isothermality	Percent
BIO5	Maximum temperature of the warmest month	°C
BIO6	Minimum temperature of the coldest month	°C
BIO8	Mean temperature of the wettest quarter	°C
BIO9	Mean temperature of the driest quarter	°C
BIO10	Mean temperature of the warmest quarter	°C
BIO11	Mean temperature of the coldest quarter	°C
BIO12	Annual precipitation	mm
BIO13	Precipitation of the wettest month	mm
BIO14	Precipitation of the driest month	mm
BIO16	Precipitation of the wettest quarter	mm
BIO17	Precipitation of the driest quarter	mm
BIO18	Precipitation of the warmest quarter	mm
BIO19	Precipitation of the coldest quarter	mm
FD	Number of frost days: annual count when TN (daily minimum) < 0°C	days
SU	Number of summer days: annual count of days when TX (daily maximum temperature) > 25°C	days
ID	Number of icing days: annual count of days when TX (daily maximum temperature) < 0°C	days
TR	Number of tropical nights: annual count of days when TN (daily minimum temperature) > 20°C	days
GSL	Growing season length: annual (1 January to 31 December in the northern hemisphere (NH), 1 July to 30 June in the southern hemisphere (SH)) count between first span of at least 6 days with TG (daily mean temperature) > 5°C and first span after 1st of July (1st of January in SH) of 6 days with TG < 5°C	days
TXx	Monthly maximum value of daily maximum temperature	°C
TNx	Monthly maximum value of daily minimum temperature	°C
TXn	Monthly minimum value of daily maximum temperature	°C
TNn	Monthly minimum value of daily minimum temperature	°C
Tn10p	Cool nights: percentage of days when TN < 10th percentile	percent
Tx10p	Cool days: percentage of days when TX < 10th percentile	percent
Tn90p	Warm nights: percentage of days when TN > 90th percentile	percent
Tx90p	Warm days: percentage of days when TX > 90th percentile	percent
WSDI	Warm spell duration index: annual count of days with at least six consecutive days when TX > 90th percentile	days
CSDI	Cold spell duration index: annual count of days with at least six consecutive days when TN < 10th percentile	days
DTR	Diurnal temperature range: monthly mean value of difference between Tx and Tn	°C
Rx1day	Monthly maximum consecutive 1-day precipitation	mm
Rx5day	Monthly maximum consecutive 5-day precipitation	mm
SDII	Simple precipitation intensity index: annual total precipitation divided by the number of wet days (defined as PRCP ≥ 1.0 mm) in the year	mm/day
R10mm	Number of heavy precipitation days: annual count of days when PRCP ≥ 10 mm	days
R20mm	Number of very heavy precipitation days: annual count of days when PRCP ≥ 20 mm	days
R1mm	Number of wet days: annual count of days when PRCP ≥ 1 mm	days
CDD	Maximum length of dry spell: maximum number of consecutive days with RR (daily precipitation amount) < 1 mm	days
CWD	Maximum length of wet spell: maximum number of consecutive days with RR ≥ 1 mm	days
R95p	Very wet days precipitation: annual total PRCP when RR > 95th percentile	mm
R99p	Extremely wet days precipitation: annual total PRCP when RR > 99th percentile	mm
PRCPTOT	Annual total precipitation on wet days (RR ≥ 1 mm)	mm

method (Clarke et al., 2009), we set the complexity parameter to a minimum value at an intersection with the point where the relative error stops decreasing significantly.

The primary performance measure for assessing the quality of resulting decision trees was classification accuracy. The classification accuracy is the ratio between correct predictions and the total number of predictions, that is, the fraction of observations correctly classified (Han et al., 2011). In addition, for gaining more insight, we calculated the precision and recall of each vegetation type in a one-versus-all setting. Precision is the ratio between the true positives and the sum of the true positives and false positives, while the recall is the ratio between the true positives and the sum of the true positives and false negatives. Here, true positive means that an observation was assigned the correct class label, true negative means that it was correctly classified as some other class and false positive means that the observation was incorrectly classified. Precision shows what fraction of positive identifications for a class was actually correct and recall shows what fraction of class examples was classified to the right class (Han et al., 2011).

For testing the prediction accuracy of our decision tree models, we used 10-fold cross-validation (Fushiki, 2011). To account for spatial non-independence of observations, we used a spatial variant of cross-validation instead of the regular variant. Spatial cross-validation helps to avoid underestimation of the predictive error due to ignoring the spatial structure of the data. Spatial cross-validation was implemented using distance-based buffers around hold-out points (Le Rest et al., 2014; Roberts et al., 2017). The data set was randomly divided into 10 subsets. Nine subsets were used for training and one subset for testing. We repeated this 10 times, each time using a different subset for model testing. In addition, during each turn, points, which were within 50 km radius around any of the training subset points, were removed from the training data and were not used either for testing or for training.

After selecting the decision tree complexity parameter, which allowed us to achieve the lowest cross-validation error while keeping the model simple, we fitted the final tree models to the whole data set. Models produced during the cross-validation procedure were only used for tuning the parameters and assessing the performance (prediction accuracy) of the models, whereas models fitted on the whole data set were used for further analysis.

In order to evaluate to what extent climate extremes can help to improve the prediction accuracy, two global decision trees were built. One used only BIOCLIM variables to predict current global vegetation distribution, and the other used both BIOCLIM and CEI variables to predict current global vegetation distribution. To further demonstrate the robustness of the decision tree results, several decision trees using different input data, for example, climate and vegetation data at different spatial resolutions were also used. More detailed decision trees for predicting regional vegetation distribution, such as in boreal and Arctic regions, have also been used. They are not very different to the global decision tree results and thus are only shown in the supplementary materials.

2.3 | Future vegetation projection with DGVMs

To further explore the importance of incorporating climate extremes in understanding vegetation distribution, the two global decision trees built with current climate and vegetation data were employed to predict vegetation changes in the future (2060–2080) using climate projections for different future scenarios (i.e. RCP2.6, RCP4.5, RCP8.5) from different climate models (see Section 2.1). The results were compared with the vegetation changes predicted by a process-based DGVM under the same future climate forcing. The DGVM results are from the Inter-Sectoral Impact Model Intercomparison Project 2b (ISIMIP2b) (Frieler et al., 2017; Warszawski et al., 2014). Among the DGVMs contributing to ISIMIP2b, only one (the Lund-Potsdam-Jena DGVM with managed Land (LPJmL)) provides changes in the vegetation cover fraction for both RCP2.6 and RCP8.5 (downloaded from: https://esg.pik-potsdam.de/search/isimip/, last access: 2020); therefore, it was used in the following analysis of this study. LPJmL is one of the state-of-the-art DGVMs (Schaphoff et al., 2018) and has been widely used for projecting future vegetation changes. It includes the potential drivers and their interactions for future vegetation changes (e.g. climate, land use and CO_2) (e.g. Boit et al., 2016). But, to be more comparable with the decision tree model (which future projections do not consider the effect of land use and CO_2), the simulations of LPJml with CO_2 and land use, fixed at year 2005 levels for the RCPs, are used. LPJml was run at 0.5×0.5 degree resolution with a fire module but no nitrogen limitation. It has competitions among PFTs for light and water; thus, the boundaries for the dominance of different PFTs can emerge from these processes. The difference of the future vegetation projections between the decision tree and DGVM approaches can provide useful insights on the uncertainty when we use different methods (pure statistical vs. process-based model) to predict future vegetation changes and the potential issues with using the decision tree approach. For instance, DGVM's future projections represent transient changes and hence are expected to be much smaller than that from the decision trees which represent equilibrium responses of vegetation to climate.

2.4 | Comparison of decision trees built on alternative land cover data products

The thresholds in decision tree rules are optimized to separate the underlying classes. Therefore, they can be different when the model is trained on different land cover schemes, which reflect different perspectives of land cover experts towards vegetation types (Ullerud et al., 2018). The perfect land cover data set does not exist and global maps have inaccuracies as well as varying definitions of vegetation classes (Hua et al., 2018). Often even experts standing on the ground at a place would not agree upon a precise definition of a vegetation type. Blending several schemes to one's taste carries extra risks. To ensure objectivity of model training, we resorted to working with externally defined schemes, one scheme at a time.

Comparing two decision trees built to classify different targets is a challenging task. Solutions exist in cases when we have additional data coming from the same domain (Perner, 2013). However, in the case of land cover products, the classes are defined in different ways and often are not equivalent.

Nonetheless, we can show that decision tree rules can equivalently describe conceptually similar vegetation classes. For this analysis, we built a decision tree using ESA CCI LC land cover classification scheme (Poulter et al., 2015), which is of the same year as the MODIS data used in this article. We included both BIOCLIM and CEI variables in the modelling. To assure that the results are robust, we integrated analyses from different schemes in the following way: We analysed which predictions were made by the ESA tree model for each leaf of the MODIS decision tree. That is, what vegetation types were predicted by the ESA tree in the locations where the MODIS tree indicated one vegetation type.

3 | RESULTS

3.1 | Decision trees: Extremes versus averages

The decision tree (Figure 2) for classification of all MODIS vegetation types using only climate averages (BIOCLIM variables) as input data produced informative and reasonably accurate results. The cross-validated accuracy of this model was 65%. It significantly exceeded the 15% accuracy of a baseline majority class model in which all observations are predicted to have a presence of the biggest class. In addition, it exceeded a 49% accuracy of a baseline majority class model in which observations of the same latitude were assigned a label of the biggest class in that latitude.

The decision tree using BIOCLIM and CEI variables is illustrated in Figure 3. The accuracy of this tree reached 67%. Both decision trees start the splitting based on the BIO12 variable (i.e. annual precipitation). If this variable is <152 mm in a grid cell, the grid cell is assigned the vegetation type barren. If BIO12 is greater than or equal to 1584 mm in a grid cell, it is assigned the vegetation type evergreen broadleaf forest (EBF).

Prediction maps of the present-day vegetation distribution by both decision trees are provided in Figure 4. We can see that the decision trees divide some of the MODIS classes into several leaves (subclasses) which are clustered in distinct territories.

In both of the decision trees, two of the smallest classes, permanent wetland and closed shrubland, are not separated into leaves and are thus not predicted by the tree. Another smaller class deciduous broadleaf forest (DBF) is separated by both trees into a leaf. However, this leaf only represents DBF in the northern latitudes and not in the tropical climate zones. In the decision tree with only BIOCLIM variables, the evergreen needleleaf forest (ENF) class is not separated into a leaf within the restriction of the complexity

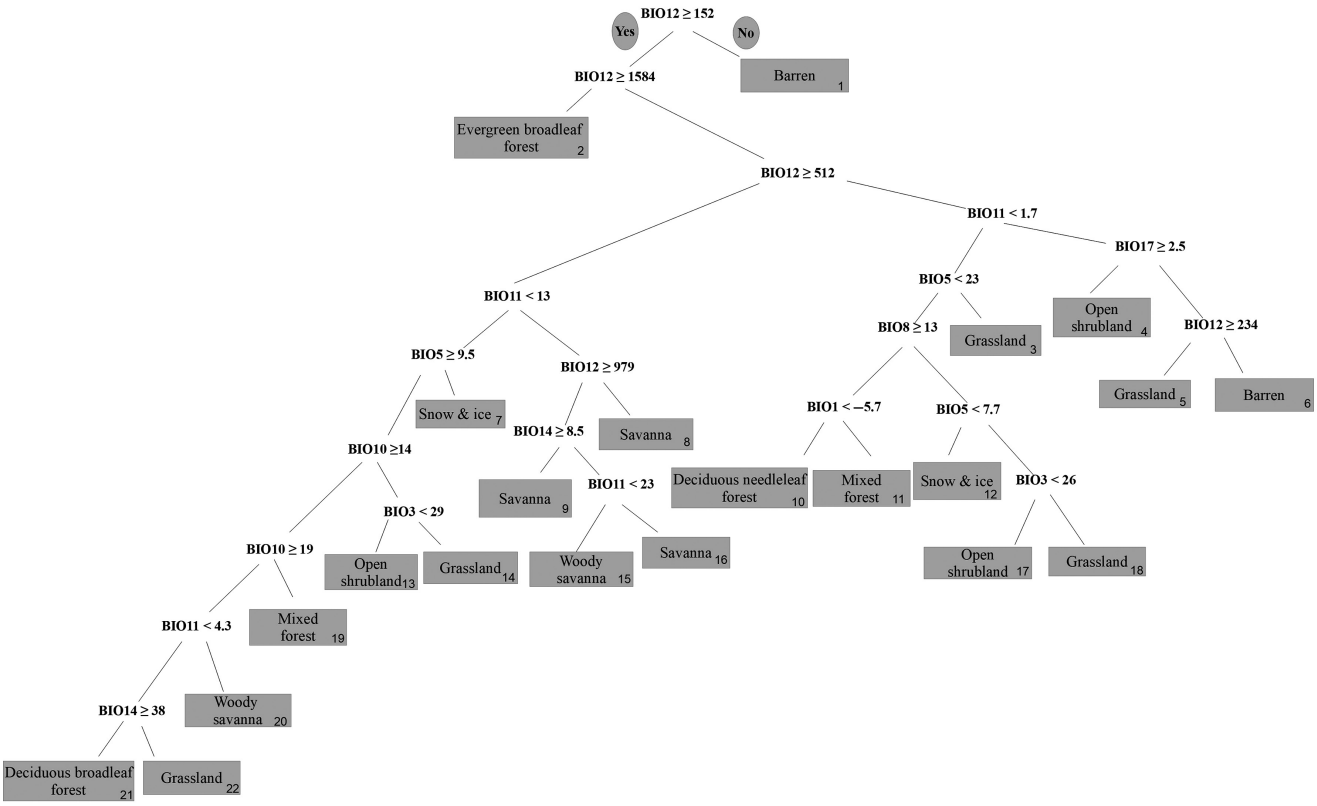


FIGURE 2 Decision tree with only climatic averages from BIOCLIM data set. Numbers in the lower right corners are an arbitrary referencing system

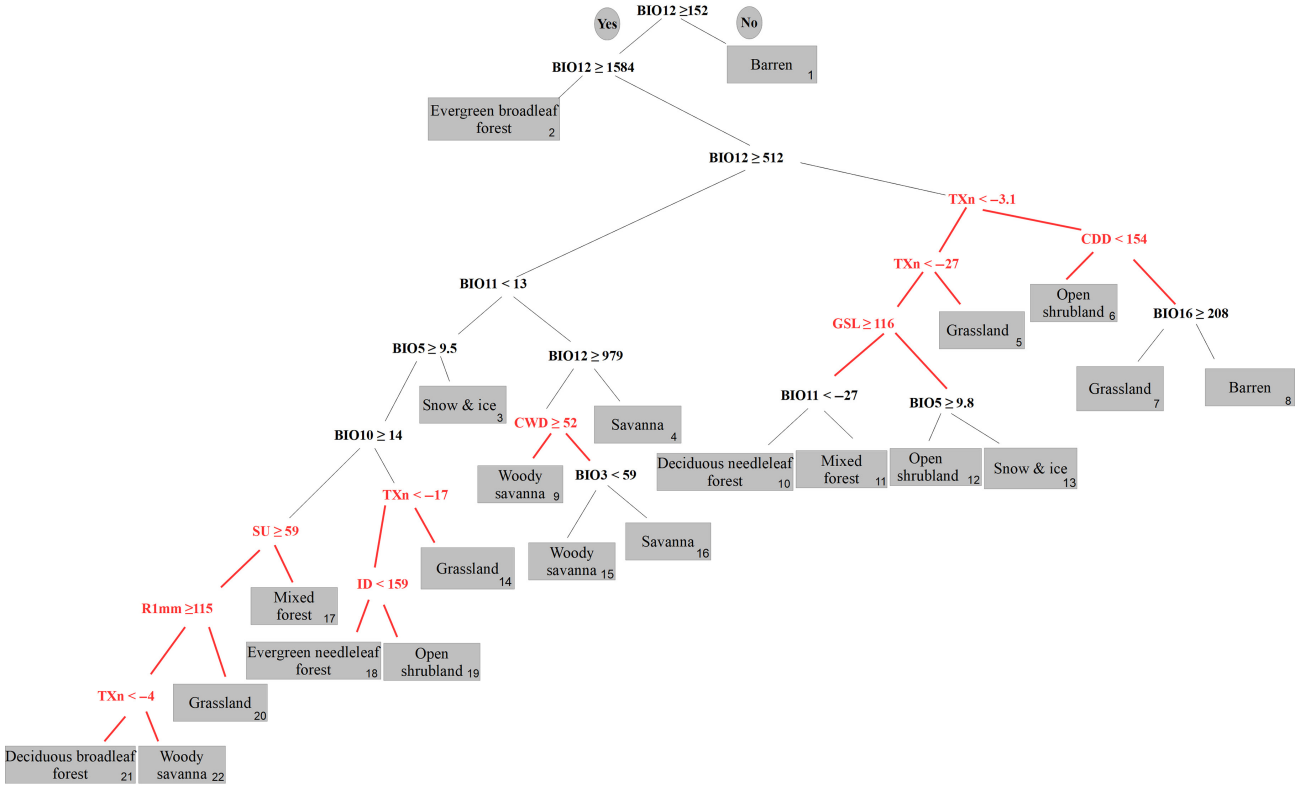


FIGURE 3 Decision tree with both climatic averages from the BIOCLIM data set and climate extremes from the CLIMDEX data set. Splits made using climate extremes are highlighted in red. Numbers in the lower right corners are an arbitrary referencing system

parameter for the decision tree. When climate extremes are included in the training data set, this type of forest is separated from open shrubland and is thus predicted by the decision tree.

Precision and recall of each class of the two global decision trees are reported in Table 4 and the confusion matrices are provided in the supplementary materials. The most accurately classified types are barren ground, snow and ice as well as EBF. Barren grid cells have a slightly higher recall value in the decision tree with CEI variables compared to that without CEI variables. In the predictions, it is most often confused with open shrubland and grassland. Snow and ice grid cells have a higher recall value in the decision tree with CEI variables but a higher precision value in the tree with only BIOCLIM variables. EBF is classified equally well in both decision trees. EBF is most often confused with woody savanna.

The vegetation type with the least prediction accuracy is ENF. Out of all the grid cells in which ENF is dominant, only 27% were identified as ENF by the extremes decision tree. This class is most often falsely predicted to be mixed forest. DBF has the second lowest recall value. Only 35% of grid cells, where DBF is dominant, are correctly assigned with the DBF class in the tree with both BIOCLIM and CEI variables (36% in the tree with only BIOCLIM). However, the precision values are quite high for this type, meaning that other vegetation types rather than DBF are less often assigned with the DBF label. DBF is most often falsely predicted to be mixed forest or savanna. For grassland, both the recall and precision values are higher in the decision tree with CEIs. Grassland is most often confused with open shrubland and savanna.

Deciduous needleleaf forest (DNF) has a slightly higher recall value in the decision tree with only BIOCLIM variables. However, the precision value for this type is much higher in the decision tree with CEI variables. This means that the extremes tree has more distinctive threshold values for separating DNF, and other classes are less likely to be falsely predicted as DNF. In the decision tree with CEI variables, DNF is mainly confused with mixed forest and open shrubland.

3.2 | Thresholds of dominant vegetation types

The thresholds determining the dominance of each vegetation type in the decision trees are summarized in Tables 5 and 6. From the results, we can see that annual precipitation (BIO12) is essential for the dominance of EBF (≥1584 mm) and barren ground (≤152 mm) in both decision trees. The separation of other types of vegetation requires consideration of both precipitation and temperature thresholds. Vegetation types covering a wide range of climate conditions, such as mixed forest, grassland, open shrubland, woody savanna and savanna, rely on different combinations of temperature and precipitation thresholds to effectively separate them from each other under distinct climate conditions, such as warm dry, warm wet, cold dry and cold wet. The most active temperature-related BIOCLIM variables in the decision tree are BIO11 (mean temperature of coldest quarter), BIO5 (maximum temperature of warmest month), BIO10 (mean temperature of warmest quarter) and BIO3 (isothermality,

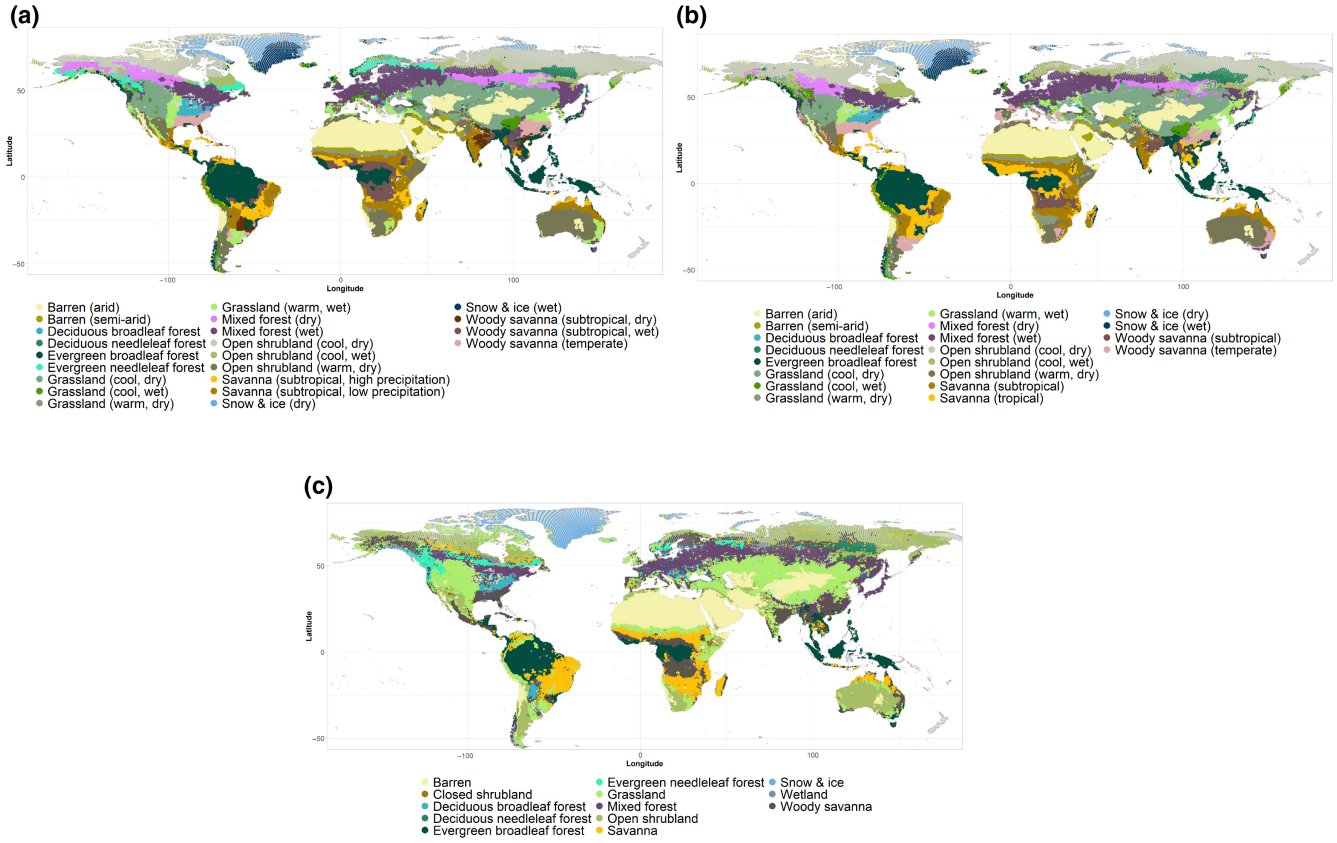


FIGURE 4 Distribution of MODIS vegetation types. (a) Predictions by decision tree with extremes. (b) Predictions by decision tree without extremes. (c) Present-day MODIS vegetation types (after correcting for the land use)

TABLE 4 Precision and recall of each class in the decision trees

MODIS class	Recall % (extreme tree)	Recall % (average tree)	Precision % (average tree)	Precision % (average tree)
Evergreen needleleaf forest	27	0	35	—
Evergreen broadleaf forest	85	85	72	72
Deciduous needleleaf forest	65	68	82	56
Deciduous broadleaf forest	35	36	65	68
Mixed forest	68	56	54	56
Closed shrubland	0	0	—	—
Open shrubland	78	79	73	66
Woody savanna	36	34	52	51
Savanna	63	67	52	46
Grassland	62	57	70	68
Permanent wetland	0	0	—	—
Permanent snow and ice	80	78	85	93
Barren	89	87	88	88

TABLE 5 Thresholds extracted from the decision tree of averages. Symbol ∧ indicates a logical conjunction

Climatic variables →									
Main vegetation types ↓	**BIO12**	**BIO11**	**BIO5**	**BIO8**	**BIO1**	**BIO10**	**BIO14**	**BIO3**	**BIO17**
Evergreen broadleaf forest	≥ 1584	–	–	–	–	–	–	–	–
Evergreen needleleaf forest	–	–	–	–	–	–	–	–	–
Deciduous needleleaf forest	≥ 152 ∧ < 512	< 1.7	< 23	≥ 13	< − 5.7	–	–	–	–
Deciduous broadleaf forest (temperate)	≥ 512 ∧ < 1584	< 4.3	≥ 9.5	–	–	≥ 19	≥ 38	–	–
Mixed forest (wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	–	–	≥ 14 ∧ < 19	–	–	–
Mixed forest (dry)	≥ 152 ∧ < 512	< 1.7	< 23	≥ 13	> − 5.7	–	–	–	–
Grassland (warm, wet)	≥ 512 ∧ < 1584	< 4.3	≥ 9.5	–	–	≥ 19	< 38	–	–
Grassland (cool, wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	–	–	< 14	–	≥ 29	–
Grassland (cool, dry)	≥ 152 ∧ < 512	< 1.7	≥ 23	–	–	–	–	–	–
	≥ 152 ∧ < 512	< 1.7	≥ 7.7 ∧ < 23	< 13	–	–	–	≥ 26	–
Grassland (warm, dry)	≥ 234 ∧ < 512	≥ 1.7	–	–	–	–	–	–	< 2.5
Open shrubland (cool, wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	–	–	< 14	–	< 29	–
Open shrubland (cool, dry)	≥ 152 ∧ < 512	< 1.7	≥ 7.7 ∧ < 23	–	< 13	–	–	< 26	–
Open shrubland (warm, dry)	≥ 152 ∧ < 512	≥ 1.7	–	–	–	–	–	–	≥ 2.5
Woody savanna (temperate)	≥ 512 ∧ < 1584	≥ 4.3 ∧ < 13	≥ 9.5	–	–	≥ 19	–	–	–
Woody savanna (subtropical)	≥ 979 ∧ < 1584	≥ 13 ∧ < 23	–	–	–	–	< 8.5	–	–
Savanna (subtropical)	≥ 512 ∧ < 979	≥ 13							
Savanna (tropical)	≥ 979 ∧ < 1584	≥ 13	–	–	–	–	≥ 8.5	–	–
	≥ 979 ∧ < 1584	≥ 23	–	–	–	–	< 8.5	–	–
Barren (arid)	< 152	–	–	–	–	–	–	–	–
Barren (semi-arid)	≥ 152 ∧ < 234	≥ 1.7	–	–	–	–	–	–	< 2.5
Snow and ice (wet)	≥ 512 ∧ < 1584	< 13	< 9.5	–	–	–	–	–	–
Snow and ice (dry)	≥ 152 ∧ < 512	< 1.7	< 7.7	< 13	–	–	–	–	–

TABLE 6 Thresholds extracted from the decision tree of extremes. Symbol ∧ indicates a logical conjunction

Climatic variables →													
Main vegetation types ↓	**BIO12**	**BIO11**	**BIO5**	**BIO10**	**BIO16**	**BIO3**	**TXn**	**ID**	**SU**	**R1mm**	**GSL**	**CDD**	**CWD**
Evergreen broadleaf forest	≥ 1584	–	–	–	–	–	–	–	–	–	–	–	–
Evergreen needleleaf forest	≥ 512 ∧ < 1584	< 13	≥ 9.5	< 14	–	–	< – 17	< 159	–	–	–	–	–
Deciduous needleleaf forest	≥ 152 ∧ < 512	< – 27	–	–	–	–	< – 27	–	–	–	≥ 116	–	–
Deciduous broadleaf forest (temperate)	≥ 512 ∧ < 1584	< 13	≥ 9.5	≥ 14	–	–	< – 4	–	≥ 59	≥ 115	–	–	–
Mixed forest (wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	≥ 14	–	–	–	–	< 59	–	–	–	–
Mixed forest (dry)	≥ 152 ∧ < 512	> – 27	–	–	–	–	< – 27	–	–	–	≥ 116	–	–
Grassland (warm, wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	≥ 14	–	–	–	–	≥ 59	< 115	–	–	–
Grassland (cool, wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	< 14	–	–	≥ – 17	–	–	–	–	–	–
Grassland (cool, dry)	≥ 152 ∧ < 512	–	–	–	–	–	≥ – 27 ∧ < – 3.1	–	–	–	–	–	–
Grassland (warm, dry)	≥ 152 ∧ < 512	–	–	–	≥ 208	–	≥ – 3.1	–	–	–	–	≥ 154	–
Open shrubland (cool, wet)	≥ 512 ∧ < 1584	< 13	≥ 9.5	< 14	–	–	< – 17	≥ 159	–	–	–	–	–
Open shrubland (cool, dry)	≥ 152 ∧ < 512	–	≥ 9.8	–	–	–	< – 27	–	–	–	< 116	–	–
Open shrubland (warm, dry)	≥ 152 ∧ < 512	–	–	–	–	–	≥ – 3.1	–	–	–	–	< 154	–
Woody savanna (temperate)	≥ 512 ∧ < 1584	< 13	≥ 9.5	≥ 14	–	–	≥ – 4	–	≥ 59	≥ 115	–	–	–
Woody savanna (subtropical, wet)	≥ 979 ∧ < 1584	≥ 13	–	–	–	–	–	–	–	–	–	–	≥ 52
Woody savanna (subtropical, dry)	≥ 979 ∧ < 1584	≥ 13	–	–	–	< 59	–	–	–	–	–	–	< 52
Savanna (subtropical, low precipitation)	≥ 512 ∧ < 979	≥ 13	–	–	–	–	–	–	–	–	–	–	–
Savanna (subtropical, high precipitation)	≥ 979 ∧ < 1584	≥ 13	–	–	–	≥ 59	–	–	–	–	–	–	< 52
Barren (arid)	< 152	–	–	–	–	–	–	–	–	–	–	–	–
Barren (semi-arid)	≥ 152 ∧ < 512	–	–	–	≥ 208	–	≥ – 3.1	–	–	–	–	≥ 154	–
Snow and ice (wet)	≥ 512 ∧ < 1584	< 13	< 9.5	–	–	–	–	–	–	–	–	–	–
Snow and ice (dry)	≥ 152 ∧ < 512	–	< 9.8	–	–	–	< – 27	–	–	–	< 116	–	–

i.e. the ratio of mean diurnal range to temperature annual range in percent). The most often picked precipitation-related BIOCLIM variables other than BIO12 are BIO14 (precipitation of driest month) and BIO17 (precipitation of driest quarter), which are particularly used to separate grassland (drier) from DBF under warm wet conditions, and open shrubland (drier) from grassland under warm dry conditions.

When CEIs are used in the decision tree, the variable most often picked is TXn (minimum value of daily maximum temperature), highlighting the importance of extreme cold conditions in limiting the distribution of different vegetation types. The number of icing days (ID) is also found to be a critical threshold for the dominance of open shrubland (ID ≥ 159 days) and ENF (ID < 159 days) in the boreal region. The maximum duration of a dry spell (CDD) is effective in separating open shrubland (CDD < 154 days) from grassland/barren (CDD ≥ 154 days) under warm and dry conditions.

In the resulting trees, the temperature thresholds for the dominance of a vegetation type in a cold environment vary with the moisture conditions. For instance, the dominance of open shrubland requires BIO3 to be smaller (larger in case of grassland) than 26% in dry climate conditions but 29% in wet climate conditions. Similarly, the dominance of snow and ice requires BIO5 to be <7.2°C in dry climate conditions but 9.5°C in wet climate conditions. This highlights the importance of applying different temperature thresholds (rather than a uniform temperature threshold) according to the living environment of the vegetation type to depict its distribution accurately.

We note that even though the MODIS land cover data set does not distinguish tropical, temperate and boreal biomes, we can separate them with the decision tree. For example, grasslands are separated into several leaves. Looking at the threshold values leading up to these leaves, we can notice that such separation is distinguishing grasslands from tropical, temperate and boreal zones respectively (Tables 5 and 6).

3.3 | Projections using decision trees

The total occupied territory of each vegetation type is projected to change in different future scenarios (Figure 5). It is visible that projected changes increase in magnitude from RCP2.6 to RCP8.5. Based on both decision trees, areas dominated by barren ground, snow and ice and mixed forest are predicted to shrink (Figure 5a,b). However, the shrinkage of mixed forest is predicted to be of lower magnitude by the decision tree with CEI variables than by the decision tree without CEI variables. The latter tree predicts a much greater expansion of grassland in all scenarios of the future, while the decision tree with CEI variables suggests a relatively small change in the areas dominated by grassland in RCP8.5 and even a decline in the area dominated by

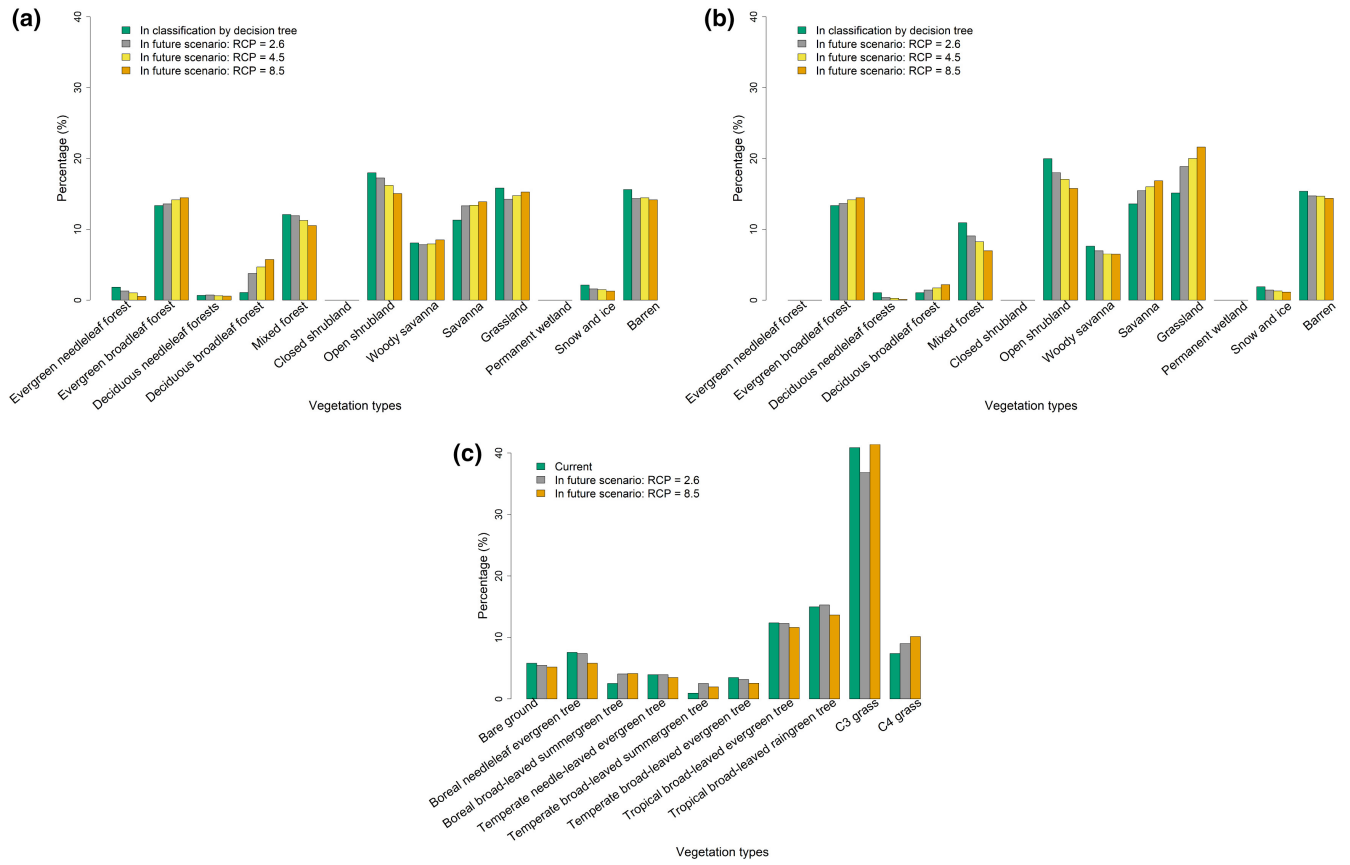


FIGURE 5 Change in total occupied territory for each vegetation type and representative concentration pathway (RCP) scenario. (a) Decision tree predictions with extremes. (b) Decision tree predictions without extremes. (c) Dynamic global vegetation model (LPJmL) predictions without carbon dioxide changes for RCP2.6 and RCP8.5, ensemble mean

grassland in RCP2.6. This is in better agreement with the future projection by the DGVM (including both C3 and C4) (Figure 5c).

Figure 6 illustrates the spatial distribution of the places susceptible to a change in dominant vegetation type in scenario RCP8.5. The changes for RCP2.6 and RCP4.5 are provided in the supplementary materials. Approximately 30% of the grid cells exhibit a change in dominant vegetation type as predicted by both decision trees (Figure 6a,b). The areas susceptible to a shift in dominant vegetation type are largely over the boreal and Arctic regions (Figure 6a,b). There are also some regions, such as the periphery of the tropical rainforest in Africa and South America, the northern India, central and southern China and the coastal area of Australia, that show a change in dominant vegetation type. The spatial pattern is generally consistent with the prediction by the DGVM (Figure 6c) but exhibits a large overestimation for the boreal and Arctic regions compared to the results from the DGVM.

Compared with the decision tree with only BIOCLIM variables, the decision tree with both BIOCLIM and CEI variables predicts less extensive changes in the dominant vegetation type over the boreal and Arctic zone, and therefore agrees more with the DGVM results.

We further analysed how the spatial distribution of each individual vegetation type will change in the future scenarios. As an example, Figure 7 illustrates the predicted RCP8.5 changes to grassland. The results for other vegetation types can be found in the supplementary materials. Figure 7 shows that temperate grassland is predicted to expand greatly to the boreal region by the decision tree with only BIOCLIM variables (Figure 7b), while the expansion of temperate grassland towards the north is very limited in the prediction using the decision tree with CEI variables (Figure 7a). The latter mainly predicts a cover of different forest types in the locations where grassland is projected to expand by the decision tree with only BIOCLIM variables (Figure 7d).

Prediction of the decision tree with CEI variables is in better agreement with the prediction by the DGVM (Figure 7c). We attribute this to a possibly slower change of extreme variables in the future scenario. For example, TXn values are projected to increase in many locations. However, such increase is not yet large enough to reach the threshold value which separates grassland from DNF, mixed forest, ENF and open shrubland.

Both decision trees predict the loss of territories dominated by grassland towards the southern part, which is similar to the DGVM. Since the definition of vegetation types in the DGVM (i.e. PFTs) does not exactly match that used in our decision trees, it is impossible to provide a more detailed and quantitative comparison of the results between the two methods. Nevertheless, it is clearly shown that the decision tree with extremes has a better potential to reproduce the results predicted by the DGVM than the decision tree without extremes.

3.4 | Comparison of the MODIS and ESA CCI LC decision trees

The decision tree (Figure 8) built using the ESA CCI LC scheme vegetation types reached the same accuracy of 67% as the one of the MODIS decision tree (Figure 3). Even though these two trees look distinct, we can identify several similarities. Both trees make the two first splits on the BIO12 variable with very similar threshold values and distinguish barren ground (bare soil) as well as evergreen broadleaf forest (tree broadleaf evergreen) types first. The BIO5 variable is used to separate

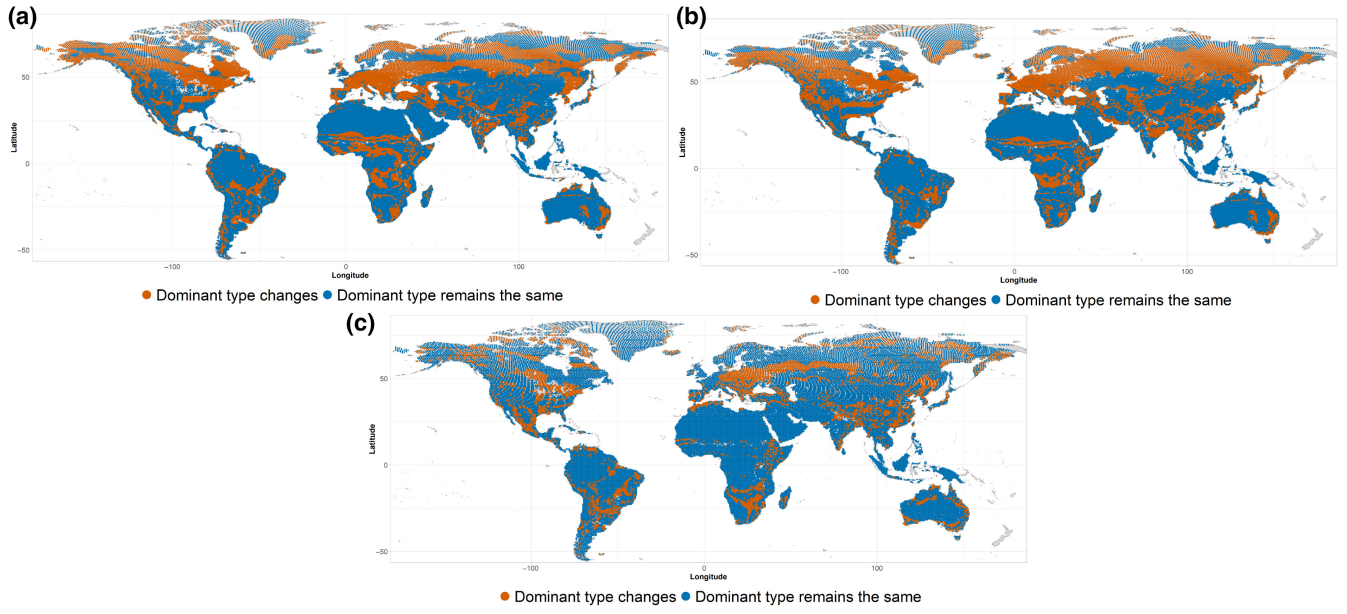


FIGURE 6 Global map of where changes are identified comparing predictions of the decision trees and future projections when the representative concentration pathway is 8.5. (a) Decision tree predictions with extremes. (b) Decision tree predictions without extremes. (c) Dynamic global vegetation model (LPJmL) predictions without carbon dioxide changes for RCP2.6 and RCP8.5, ensemble mean

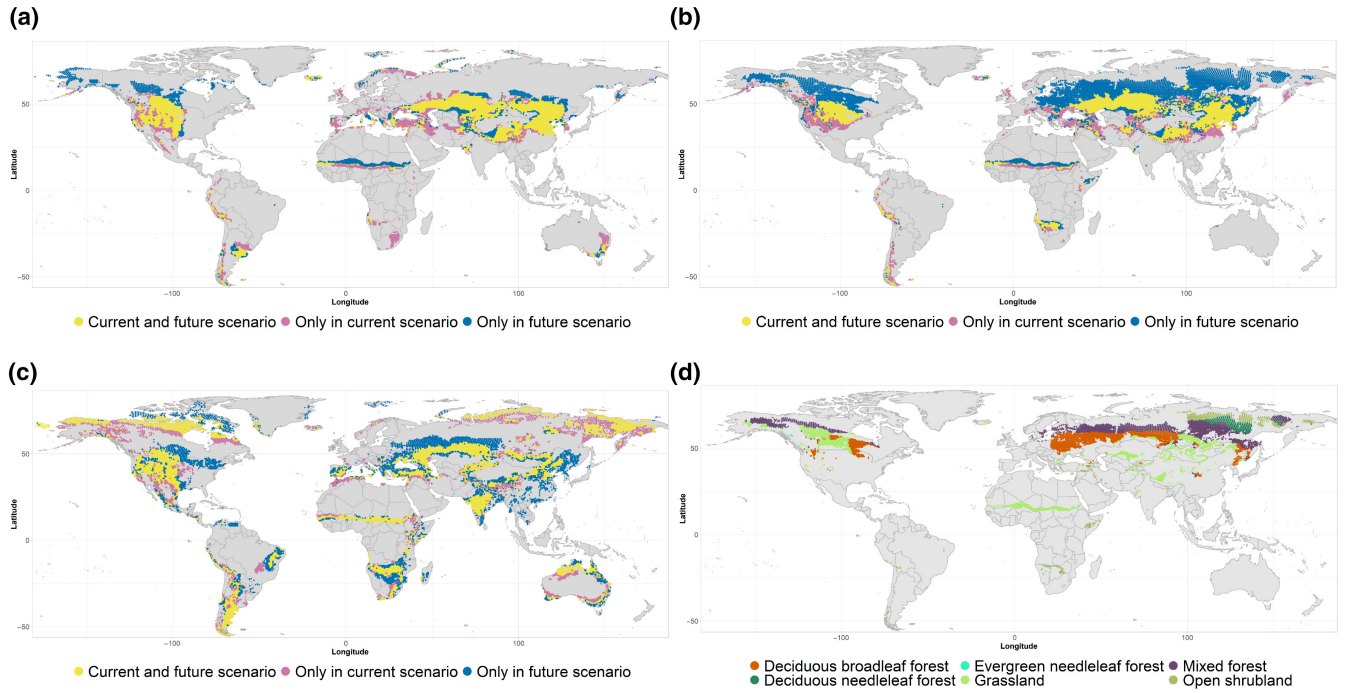
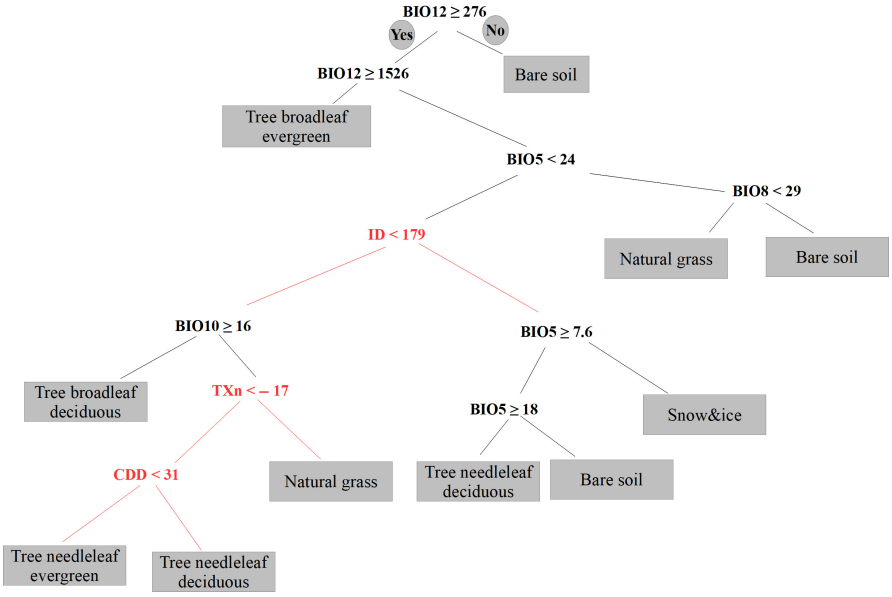


FIGURE 7 Predicted change in grassland under the representative concentration pathway 8.5. (a) Decision tree predictions with extremes. (b) Decision tree predictions without extremes. (c) Dynamic global vegetation model (LPJmL) prediction without carbon dioxide changes for RCP2.6 and RCP8.5 (both C3 and C4). (d) Vegetation types which are predicted in the future scenario by the extremes decision tree in the locations where grassland is predictd to expand by the decision tree without extremes

FIGURE 8 Decision tree using ESA CCI LC land cover product data with both climatic averages from the BIOCLIM data set and climate extremes from the CLIMDEX data set. Splits made using climate extremes are highlighted in red



the snow and ice class from other vegetation types. The TXn threshold with the identical value of −17 is used in both trees to separate grassland from evergreen needleleaf forest, while the BIO10 variable separates grassland from deciduous broadleaf forest. The same group of variables, apart from BIO8, is selected as important in both decision trees.

Table 7 represents how the observations in the leaves of the MODIS decision tree are predicted in the decision tree based on ESA CCI data set. For example, in the locations where the MODIS tree predicts the dominant vegetation type to be barren ground or evergreen broadleaf forest, the ESA model also predicts corresponding types, that is, bare soil and broadleaf evergreen tree, accordingly. In those locations where the MODIS tree predicts deciduous needleleaf forest, the ESA tree predicts the corresponding type tree needleleaf deciduous in 83% of observations, tree needleleaf evergreen in 8% of observations and bare soil in 7% of observations. The maps of mismatches are provided in the supplementary materials. Overall, the

TABLE 7 Distribution of the ESA CCI LC decision tree predictions in the leaves of the MODIS tree with climatic extremes. Bold text represents corresponding vegetation types in both classification schemes or conceptually similar classes to the one of the leaves of the MODIS decision tree. Vegetation types which comprise <1% are not listed. A number in the brackets indicates the number of the leaf in the MODIS tree (Figure 3)

Leaves of MODIS decision tree	Predictions of ESA CCI LC tree
(1) Barren	**100% Bare soil**
(2) Evergreen broadleaf forest	**100% Tree broadleaf evergreen**
(3) Snow and ice	**90% Snow and ice**; 5% Bare soil; 4% Grass
(4) Savanna	**90% Grass**; 9% Bare soil
(5) Grassland	**54% Grass**; 31% Bare soil; 5% Tree broadleaf deciduous; 4% Tree needleleaf deciduous
(6) Open shrubland	55% Bare soil; **45% Grass**
(7) Grassland	**64% Grass**; 36% Bare soil
(8) Barren	**94% Bare soil**; 6% Grass
(9) Woody savanna	85% Grass; 14% Tree broadleaf evergreen
(10) Deciduous needleleaf forest	**83% Tree needleleaf deciduous**; 8% Tree needleleaf evergreen; 7% Bare soil
(11) Mixed forest	**55% Tree needleleaf evergreen**; **23% Tree broadleaf deciduous**; **10% Tree needleleaf deciduous**; 7% Grass; 5% Bare soil
(12) Open shrubland	77% Bare soil; 20% Tree needleleaf deciduous; 3% Tree needleleaf evergreen
(13) Snow and ice	71% Bare soil; **29% Snow and ice**
(14) Grassland	**98% Grass**; 2% Tree broadleaf evergreen
(15) Woody savanna	87% Grass; 10% Tree broadleaf evergreen; 3% Bare soil
(16) Savanna	**93% Grass**; 6% Tree broadleaf evergreen
(17) Mixed forest	**40% Tree broadleaf deciduous**; 33% Grass; 23% **Tree needleleaf evergreen**; **21% Tree needleleaf deciduous**
(18) Evergreen needleleaf forest	**99% Tree needleleaf evergreen**
(19) Open shrubland	35% Tree needleleaf evergreen; 35% Bare soil; 30% Tree needleleaf deciduous
(20) Grassland	**96% Grass**; 4% Tree broadleaf deciduous
(21) Deciduous broadleaf forest	95% Grass; **5% Tree broadleaf deciduous**
(22) Woody savanna	89% Grass; 8% Tree broadleaf evergreen; 3% Tree broadleaf deciduous

share of mismatches is relatively small and makes good sense given the large differences in the definitions of the two land cover schemes.

4 | DISCUSSION

4.1 | Climate thresholds in shaping vegetation distributions

The thresholds identified by the decision trees for separating the dominance of different vegetation types are generally consistent with our ecological understanding of the vegetation types. For instance, the dominance of DBF–mixed forest–ENF is primarily separated by temperature thresholds, while the dominance of DBF–grassland–savanna–woody savanna is primarily determined by precipitation thresholds (Figures 1 and 2; Tables 5 and 6). These thresholds also share many similarities with those used in traditional climate/vegetation/biome classifications (Conradi et al., 2020; Holdridge, 1967; Kottek et al., 2006; Whittaker, 1962). For example, in the Köppen classification, temperature in the coldest month (similar to BIO11) is broadly used for the separation of the major climate types (Kottek et al., 2006). Temperature in the warmest month (similar to BIO5) is also used for defining a snow/polar climate. Annual precipitation (BIO12) is also used for separating tropical evergreen forest from barren ground. Such similarities further support the close association of the Köppen classification with biome distribution (Rohli et al., 2015). In the Holdridge life zone, the classification of rainforest is independent of temperature as long as the annual mean precipitation is over 1000 mm. This is in line with our results. In the Whittaker biome classification, the precipitation thresholds for separating tropical forest, savanna and desert/barren are roughly 1500 mm and 500 mm, which is close to what we found in the decision tree (1584 mm and 512 mm) (Figure 3; Tables 5 and 6).

The branches of the later splits of the decision trees extract more specific ecological constraints of different vegetation types under different climate conditions, which can hardly be formalized otherwise. One interesting example of such constraints is how the relative dominance of grassland and open shrubland is determined by temperature in the cold environment. In general, both decision trees (Figures 2 and 3) indicate that open shrubland is more abundant than grassland at a colder temperature (e.g. BIO3 < 29%, TXn < −17). This is in line with previous studies showing that shrubs have higher cold tolerance than grasses (Venn et al., 2013), although the evolution of cold acclimation within grasses probably came alongside the diversification of this plant group (e.g. Humphreys & Linder, 2013; Schubert et al., 2019; Vidal Jr et al., 2021). In addition, the temperature threshold for the dominance of grassland is lower in a dry climate (BIO3 ≥ 26%, TXn ≥ −27°C) than in a wet climate (BIO3 > 29%, TXn ≥ −17°C), implying a higher (lower) cold tolerance of grass in dry (wet) climate conditions (Table 6). The climatic tolerance of different plants and vegetation types varies globally (e.g. Lancaster & Humphreys, 2020). The dependence of a plant's cold tolerance on moisture conditions has been found in previous studies

(e.g. Geange et al., 2021; Sierra-Almeida et al., 2016), and the higher cold tolerance of a plant under drier conditions can be attributed to the presence of less tissue water to freeze, thus reducing the probability of ice nucleation and tissue damage in cold conditions (Sierra-Almeida & Cavieres, 2010).

From our results, the importance of climate extremes rather than average climate in determining the dominance of a vegetation type is highlighted in both decision trees (Figures 2 and 3). In the decision tree with only BIOCLIM variables, the variables depicting long-term monthly/seasonal extremes of temperature and precipitation are mostly selected (e.g. BIO5, BIO11 and BIO14), while in the decision tree with CEI variables included, the variables depicting extremes on a daily scale are widely picked, such as TXn, CDD and ID. Despite the difference in the prediction accuracy of the two decision trees being rather small, our interpretation is that the climate extreme variables (i.e. CEIs) can be particularly useful to more effectively separate vegetation classes (Figure 3). This is in line with the understanding of the bioclimatic control on the ecophysiological traits of different PFTs in previous studies (e.g. Harrison et al., 2010, and references therein). For instance, the decision tree with CEIs separates ENF into a separate leaf, whereas the decision tree with only BIOCLIM variables is not able to separate this type within the same complexity limit. The precision in predicting DNF, woody savanna and open shrubland is also improved when CEIs are included in the decision tree (Table 4), indicating the importance of climate extremes (on a daily scale), such as cold, drought and freezing events, in limiting the distribution of these vegetation types. In particular, ENF is found to be more vulnerable to the duration of the daily maximum below 0°C in winter season, that is, icing conditions (ID < 159 days) compared to shrubland. This is probably related to higher thermal demands (i.e. length of season) of larger trees compared to smaller shrubs (Körner, 2012), as well as challenges with frost drought, which is well known from many ENF tree species (Mayr et al., 2006). Frost drought will damage evergreen trees when the ground is frozen due to long periods of icing days, but the ambient temperature is above 0°C during the photoperiod (Huang et al., 2020), so that photosynthesis is activated when water is unavailable. ID may also be related to freeze–thaw cycles (at least in some areas, where winters are characterised by mild subfreezing temperature), and hence, possibly related to frost damage and top breaks. It is noted that the inclusion of more climate extreme characteristics not only improves the accuracy of the decision tree model for depicting present-day vegetation but also produces more reasonable spatial changes for the future that are more in line with DGVMs (Figure 6), further supporting the importance of climate extremes in determining the spatial range of different vegetation types, for example, the role of TXn in expansion of temperate grassland (Figure 7).

4.2 | Implication for improving DGVMs and predicting vegetation in the future

DGVMs have been a major modelling tool for describing and understanding large-scale vegetation distribution and its changes. The fidelity of DGVMs, however, suffers from large uncertainties in their parameterization of the vegetation processes, including the climate thresholds for the key processes critical for vegetation distribution, such as establishment, survival and mortality (Fisher et al., 2015; Forkel et al., 2019; Horvath et al., 2021; Masson-Delmotte et al., 2021). There have been various ways to derive the climate thresholds for these processes. They can be derived from the biogeographic limits of certain species or vegetation types retrieved from observation or statistical models (e.g. Horvath et al., 2021). They can also be derived from the ecophysiological climate tolerance of certain species or groups of species (e.g. Geange et al., 2021). Often, the reference sources for the parameters are neither comprehensive, up to date nor necessarily consistent with each other.

We argue that the climate thresholds derived from the decision tree mining of land cover and climate data may provide a valuable source for a more systematic and consistent parameterization of the climate thresholds required by DGVMs. One approach could be to directly apply those thresholds as a priori constraints (climate envelopes) to where they are needed in DGVMs (e.g. mortality). Another approach would be to further explore and improve processes in DGVMs to allow the prediction of biome boundaries directly from plant physiological traits via their competitive interactions, and thus better representation of the threshold response of vegetation to the climate variables, especially the climate extremes as found in our results (e.g. Fisher et al., 2015, 2018).

As shown in Table 1, the climate thresholds employed in DGVMs are mostly monthly mean variables and they are static. Our decision tree results, however, emphasize the importance of using climate extremes, especially extremes on a daily scale, in defining the climate thresholds of different vegetation types. This is in agreement with the recent study by Forkel et al. (2019), which found that the performance of DGVMs can be improved by incorporating CEIs in parameterizing the mortality of different vegetation types. How vegetation responds to cold and drought extremes in DGVMs can, in particular, be essential for depicting the spatial distribution of certain vegetation types and their changes, such as savanna, boreal forest and Arctic tundra. As implied by our decision tree results, the tolerance of vegetation to climate extremes may vary with average climate (e.g. higher tolerance to cold extremes under dry conditions). It is, therefore, critical for the DGVMs to implement varying climate thresholds as a function of mean climate conditions rather than hard-coded static climate thresholds for the whole globe, which is commonly done in the DGVMs. Such improvement in the DGVMs is expected to have a large impact on the future projections of vegetation changes, as (1) including climate extremes offer more refinement than average climate in characterizing changes in future climate, for example, it can well be that an increase in average precipitation could be accompanied by an increase in CDD (Seneviratne & Hauser, 2020), and (2) the ability of vegetation to tolerate climate extremes may change with the average climate in the future, for example, increasing mean temperature can reduce the tolerance of plants to cold extremes (Sierra-Almeida & Cavieres, 2010).

It has been argued that climate extremes may have contributed to the reduced plant growth in the Arctic in the past decades, a phenomenon referred to as Arctic browning (Phoenix & Bjerke, 2016). This contradicts to the effects of the average climate changes, which lead to an overall enhanced growth of plants in the Arctic (referred to as Arctic greening) (Myers-Smith et al., 2020). Our results further support the role of climate extremes in limiting vegetation changes in the boreal region in response to global warming (Figures 5–7). It is thus expected that better description of the vegetation response to climate extremes may help in reducing the sensitivity of Arctic vegetation to global warming in DGVMs, and hence, alleviate the strong positive vegetation–climate feedback found in the ESM coupled with DGVM (Zhang et al., 2018).

4.3 | Limitations and uncertainties with climatic thresholds from decision tree mining

The reliability of the climatic thresholds derived from the decision tree mining to inform DGVM parameterization is affected by both the accuracy of the decision tree model and the uncertainties of the climate and vegetation data sets used to build the decision tree. The global vegetation cover data set used in this study (i.e. MODIS land cover data set) is one of the most validated and used vegetation products (Friedl et al., 2010; Grekousis et al., 2015), but it has also shown various biases for different land cover types in different regions, in particular in high-latitude regions (Liang et al., 2019). The land cover types used in MODIS can also add another layer of uncertainty as different vegetation classification schemes may have different biases and uncertainties (Grekousis et al., 2015). To further evaluate the influence of the uncertainties in vegetation cover data sets and their classification schemes, we have also employed the ESA CCI LC land cover data set. The results show that although the exact values of the climate thresholds or decision rules for different vegetation types might differ, the climate variables (either BIOCLIM or CEI) selected for separating the same vegetation types are quite similar. Analysing similarities of predictions of the ESA and MODIS trees, we see that predicted vegetation types are mostly conceptually similar in both trees. This means that similarly defined vegetation types can be described by an equivalent set of rules.

To address the uncertainties regarding the resolution of the climate and vegetation data set used for decision tree mining, we have performed decision tree mining at different spatial resolutions from 10 × 10 min to 1.5 × 1.5 degrees, and the results do not change much, with the exception that the exact threshold values differ slightly (see supplementary materials).

While assigning dominant vegetation type labels according to the land cover data set, we made an assumption that the proportion of natural land cover types would remain unchanged despite human activity. This assumption comes with uncertainty, since humans tend to occupy the most productive land. However, even in heavily human-modified areas, fragments of the original vegetation often survive (Adams, 2009) and form the basis for potential natural vegetation. Our initial exploratory experiments with observations, which have dominant vegetation type with 40% or more occupancy in a grid cell, showed that including observations where assigned dominant vegetation type is unclear does not change the accuracy of the produced models significantly. This means that even if not all grid cells follow our assumption, this should not significantly bias the results.

As to the accuracy of the decision tree mining, some of the vegetation types, such as grassland or woody savanna, can be separated into more than one leaf with reasonable accuracy. This is due to the fact that these types can occur in different climatic zones. DBF, which dominates in both northern latitudes and tropical climate zones, is only separated for the northern latitudes. For this reason, tropical DBF is never predicted with the correct label, and thus, the accuracy of the DBF class is one of the lowest. In addition, many of the DBFs appear in territories highly impacted by humans and the initial label given of DBF dominance can be misleading. Furthermore, DBF can be part of mixed forest and can be already misclassified in the MODIS data set or exist in climatic conditions very similar to some mixed forest. The low accuracy of the decision tree for permanent wetland and closed shrubland is primarily due to their low coverage (dominant in <1% of the grid cells of the MODIS land cover data set). In addition, these types can occur in various climatic zones. Permanent wetland is often more related to topography characteristics than climate (Branton & Robinson, 2020), and currently, its coverage is not accurately estimated (Mahdavi et al., 2018). Therefore, the number of observations in each of the different climatic zones is too small and the observations are not distinct enough from other vegetation types to be separated by the tree.

We note that barren ground and EBF are two of the biggest classes and they can be well distinguished only by a lack or large quantity of precipitation respectively. Therefore, these classes are the first to be separated by the tree on the precipitation amount variable. All observations with a very high precipitation threshold are given the EBF label as this class has the most observations. This leads to very inaccurate predictions for some places (e.g. part of Greenland) in the northern latitudes in the future scenarios which have as high precipitation as EBF (see supplementary materials). One way to avoid such bias would be to engineer an extra binary feature combining both temperature and precipitation (see supplementary materials). This feature could indicate, for instance, whether the climate in a grid cell is very humid and warm or not. However, in this case, we need to manually select the thresholds which indicate high precipitation and warm temperature respectively. Thus, rather than extracting the climatic thresholds from the model, we would be manually encoding them into the decision tree. To achieve higher accuracy of the decision tree mining and the climatic thresholds for a specific vegetation type, a regional decision tree can be applied. We note that applying the decision tree to historical or future scenarios can provide additional validation of the robustness of the climate thresholds derived from the decision tree mining for each vegetation type. According to the results for the future scenarios (Figures 5–7), the climate thresholds derived from the decision tree using both BIOCLIM and

CEI variables seem to be more realistic and reliable for most of the vegetation types, such as grassland and DNF.

5 | CONCLUSIONS

In this study, we employed decision tree induction to explore the global linkage between vegetation and climate. Important climate thresholds for the dominance of different vegetation types have been identified. Among them, the thresholds of climate extremes (e.g. extreme cold or drought) have been found to be essential for the dominance of certain vegetation types such as evergreen needleleaf forest, deciduous needleleaf forest, grassland, open shrubland and savanna in both the present day and the future. Moreover, the climate thresholds for a vegetation type, such as its cold tolerance, may vary with environmental conditions (e.g. moisture). All these aspects of vegetation response to climate have not been fully considered in DGVMs. This highlights the need for further improvements of DGVMs for representing the threshold response of different vegetation types to climate extremes in order to provide a better projection of future vegetation changes for Earth system models.

Decision tree modelling proved to be a powerful tool to separate the land cover types into more detailed subtypes and to generate and update our understanding of the relationship between climate and vegetation distribution from emerging big climate and vegetation data sets in a coherent way. Nonetheless, we do not advise employing the decision trees for vegetation prediction stand alone, but rather coupling them with expert knowledge to critically assess the biological significance and implications of the identified thresholds. To facilitate the use of decision tree mining in exploring potential climate thresholds for the vegetation types in different regions and their application to the parameterization of DGVMs, a reproducible workflow for the decision tree mining using global climate data and remotely sensed land cover data is provided in R. We note that the decision tree built using this workflow can also be applied to quickly generate a reasonable first guess of large-scale vegetation distribution in equilibrium with the climate in past or future periods when DGVM results are not available. However, it has to be applied with caution, as some environmental variables that are critical for vegetation are not considered in the current decision tree model, such as CO_2 concentrations.

ACKNOWLEDGEMENTS

Research leading to these results is funded by the Academy of Finland (grant no. 314803 to IŽ). This work forms a contribution to LATICE, which is a Strategic Research Initiative funded by the Faculty of Mathematics and Natural Sciences at the University of Oslo (grant no. UiO/GEO103920). It is also part of the EMERALD project (grant no. 294948) funded by the Research Council of Norway.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in WorldClim—Global Climate Data repository at https://www.worldclim.org/data/worldclim21.html and at http://www.worldclim.com/cmip5_10m; in Canadian Centre for Climate Modelling and Analysis repository at https://crd-data-donnees-rdc.ec.gc.ca/CCCMA/products/CLIMDEX/CMIP5/ and at https://climate-modelling.canada.ca/climatemodeldata/climdex/; in Moderate Resolution Imaging Spectroradiometer repository at https://doi.org/10.5067/MODIS/MCD12C1.006. The code for decision tree mining will be available in a GitHub repository (https://github.com/ritabei/Dominant-natural-vegetation).

ORCID

Rita Beigaitė https://orcid.org/0000-0003-3308-4493
Hui Tang https://orcid.org/0000-0002-8745-3859
Anders Bryn https://orcid.org/0000-0003-4712-8266
Olav Skarpaas https://orcid.org/0000-0001-9727-1672
Frode Stordal https://orcid.org/0000-0002-5190-6473
Jarle W. Bjerke https://orcid.org/0000-0003-2721-1492
Indrė Žliobaitė https://orcid.org/0000-0003-2427-5407

REFERENCES

Abdi, A. M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience & Remote Sensing*, *57*, 1–20. https://doi.org/10.1080/15481603.2019.1650447

Adams, J. (2009). *Vegetation-climate interaction: How plants make the global environment*. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-00881-8

Boit, A., Sakschewski, B., Boysen, L., Cano-Crespo, A., Clement, J., Garcia-alaniz, N., Kok, K., Kolb, M., Langerwisch, F., Rammig, A., Sachse, R., Van Eupe, M., Von Bloh, W., Zemp, D. C., & Thonicke, K. (2016). Large-scale impact of climate change vs. land-use change on future biome shifts in Latin America. *Global Change Biology*, *22*, 3689–3701. https://doi.org/10.1111/gcb.13355

Branton, C., & Robinson, D. T. (2020). Quantifying topographic characteristics of wetlandscapes. *Wetlands*, *40*, 433–449. https://doi.org/10.1007/s13157-019-01187-2

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press. https://doi.org/10.1201/9781315139470

Chatterjee, S., Ghosh, S., Dawn, S., Hore, S., & Dey, N. (2016). Forest type classification: A hybrid nn-ga model based approach. In *Information systems design and intelligent applications* (pp. 227–236). Springer. https://doi.org/10.1007/978-81-322-2757-1_23

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Clarke, B., Fokoue, E., & Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media.

Conradi, T., Slingsby, J. A., Midgley, G. F., Nottebrock, H., Schweiger, A. H., & Higgins, S. I. (2020). An operational definition of the biome for global change research. *New Phytologist*, *227*, 1294–1306. https://doi.org/10.1111/nph.16580

Cramer, W., Bondeau, A., Woodward, F. I., Prentice, I. C., Betts, R. A., Brovkin, V., Cox, P. M., Fisher, V., Foley, J. A., Friend, A. D. et al.

(2001). Global response of terrestrial ecosystem structure and function to CO_2 and climate change: Results from six dynamic global vegetation models. *Global Change Biology*, *7*, 357–373. https://doi.org/10.1046/j.1365-2486.2001.00383.x

Dahl, E. (1998). The phytogeography of northern Europe (British Isles, fennoscandia and adjacent areas). https://doi.org/10.1017/CBO9780511565182

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., & Zwiers, F. W. (2014). Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets. *Journal of Climate*, *27*, 5019–5035. https://doi.org/10.1175/JCLI-D-13-00405.1

Fick, S. E., & Hijmans, R. J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*, 4302–4315. https://doi.org/10.1002/jgra.50188

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*, e2018MS001453. https://doi.org/10.1029/2018MS001453

Fisher, R. A., Koven, C. D., Anderegg, W. R. L., Christoffersen, B. O., Dietze, M. C., Farrior, C. E., Holm, J. A., Hurtt, G. C., Knox, R. G., Lawrence, P. J., Lichstein, J. W., Longo, M., Matheny, A. M., Medvigy, D., Muller-Landau, H. C., Powell, T. L., Serbin, S. P., Sato, H., Shuman, J. K., … Moorcroft, P. R. (2018). Vegetation demographics in earth system models: A review of progress and priorities. *Global Change Biology*, *24*, 35–54. https://doi.org/10.1111/gcb.13910

Fisher, R. A., Muszala, S., Verteinstein, M., Lawrence, P., Xu, C., McDowell, N. G., Knox, R. G., Koven, C., Holm, J., Rogers, B. M., Spessa, A., Lawrence, D., & Bonan, G. (2015). Taking off the training wheels: The properties of a dynamic vegetation model without climate envelopes, CLM4. *Geoscientific Model Development*, *8*, 3593–3619. https://doi.org/10.5194/gmd-8-3593-2015

Forkel, M., Drüke, M., Thurner, M., Dorigo, W., Schaphoff, S., Thonicke, K., von Bloh, W., & Carvalhais, N. (2019). Constraining modelled global vegetation dynamics and carbon turnover using multiple satellite observations. *Scientific Reports*, *9*, 1–12. https://doi.org/10.1038/s41598-019-55187-7

Friedl, M., & Sulla-Menashe, D. (2015). Mcd12c1 modis/terra+ aqua land cover type yearly l3 global 0.05 deg cmg v006. *NASA EOSDIS Land Processes DAAC*.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. (2010). Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, *114*, 168–182. https://doi.org/10.1016/j.rse.2009.08.016

Frieler, K., Lange, S., Piontek, F., Reyer, C. P., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K. et al. (2017). Assessing the impacts of 1.5 c global warming–simulation protocol of the inter-sectoral impact model intercomparison project (isimip2b). *Geoscientific Model Development*, *10*, 4321–4345. https://doi.org/10.5194/gmd-10-4321-2017

Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, *21*, 137–146. https://doi.org/10.1007/s11222-009-9153-8

Galbrun, E., Tang, H., Fortelius, M., & Žliobait, I. (2018). Computational biomes: The ecometrics of large mammal teeth. *Palaeontologia Electronica*, *21*, 1–31. https://doi.org/10.26879/786

Ge, G., Shi, Z., Zhu, Y., Yang, X., & Hao, Y. (2020). Land use/cover classification in an arid desert-oasis mosaic landscape of china using remote sensed imagery: Performance assessment of four machine learning algorithms. *Global Ecology and Conservation*, *22*, e00971. https://doi.org/10.1016/j.gecco.2020.e00971

Geange, S. R., Arnold, P. A., Catling, A. A., Coast, O., Cook, A. M., Gowland, K. M., Leigh, A., Notarnicola, R. F., Posch, B. C., Venn, S. E., Zhu, L., & Nicotra, A. B. (2021). The thermal tolerance of photosynthetic tissues: A global systematic review and agenda for future research. *New Phytologist*, *229*, 2497–2513. https://doi.org/10.1111/nph.17052

Graham, D., & Patterson, B. D. (1982). Responses of plants to low, nonfreezing temperatures: Proteins, metabolism, and acclimation. *Annual Review of Plant Physiology*, *33*, 347–372. https://doi.org/10.1146/annurev.pp.33.060182.002023

Grekousis, G., Mountrakis, G., & Kavouras, M. (2015). An overview of 21 global and 43 regional land-cover mapping products. *International Journal of Remote Sensing*, *36*, 5309–5335. https://doi.org/10.1080/01431161.2015.1093195

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*, 83–124. https://doi.org/10.1073/pnas.0901643106

Harrison, S. P., Prentice, I. C., Barboni, D., Kohfeld, K. E., Ni, J., & Sutra, J.-P. (2010). Ecophysiological and bioclimatic foundations for a global plant functional classification. *Journal of Vegetation Science*, *21*, 300–317. https://doi.org/10.1111/j.1654-1103.2009.01144.x

Hengl, T., Walsh, M. G., Sanderman, J., Wheeler, I., Harrison, S. P., & Prentice, I. C. (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

Hickler, T., Vohland, K., Feehan, J., Miller, P. A., Smith, B., Costa, L., Giesecke, T., Fronzek, S., Carter, T. R., Cramer, W., Kühn, I., & Sykes, M. T. (2012). Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. *Global Ecology and Biogeography*, *21*, 50–63. https://doi.org/10.1111/j.1466-8238.2010.00613.x

Holdridge, L. R. (1967). *Life zone ecology*.

Horvath, P., Tang, H., Halvorsen, R., Stordal, F., Tallaksen, L. M., Berntsen, T. K., & Bryn, A. (2021). Improving the representation of high-latitude vegetation distribution in dynamic global vegetation models. *Biogeosciences*, *18*, 95–112. https://doi.org/10.5194/bg-18-95-2021

Hua, T., Zhao, W., Liu, Y., Wang, S., & Yang, S. (2018). Spatial consistency assessments for global land-cover datasets: A comparison among GLC2000, CCI LC, MCD12, GLOBCOVER and GLCNMO. *Remote Sensing*, *10*, 1846. https://doi.org/10.3390/rs10111846

Huang, J.-G., Ma, Q., Rossi, S., Biondi, F., Deslauriers, A., Fonti, P., Liang, E., Mäkinen, H., Oberhuber, W., Rathgeber, C. B. K., Tognetti, R., Treml, V., Yang, B., Zhang, J.-L., Antonucci, S., Bergeron, Y., Camarero, J. J., Campelo, F., Čufar, K., … Ziaco, E. (2020). Photoperiod and temperature as dominant environmental drivers triggering secondary growth resumption in northern hemisphere conifers. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 20645–20652. https://doi.org/10.1073/pnas.2007058117

Humphreys, A. M., & Linder, H. P. (2013). Evidence for recent evolution of cold tolerance in grasses suggests current distribution is not limited by (low) temperature. *New Phytologist*, *198*, 1261–1273. https://doi.org/10.1111/nph.12244

Ito, A., Reyer, C. P. O., Gädeke, A., Ciais, P., Chang, J., Chen, M., François, L., Forrest, M., Hickler, T., Ostberg, S., Shi, H., Thiery, W., & Tian, H. (2020). Pronounced and unavoidable impacts of low-end global warming on northern high-latitude land ecosystems. *Environmental Research Letters*, *15*, 044006. https://doi.org/10.1088/1748-9326/ab702b

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Julio Camarero, J., Gazol, A., Sancho-Benages, S., & Sangüesa-Barreda, G. (2015). Know your limits? Climate extremes impact the range of scots pine in unexpected places. *Annals of Botany*, *116*, 917–927. https://doi.org/10.1093/aob/mcv124

Köppen, W. (1900). Versuch einer klassifikation der klimate, vorzugsweise nach ihren beziehungen zur pflanzenwelt. *Geographische Zeitschrift*, *6*, 593–611. https://www.jstor.org/stable/27803924

Körner, C. (2012). Treelines will be understood once the functional difference between a tree and a shrub is. *Ambio*, *41*, 197–206. https://doi.org/10.1007/s13280-012-0313-2

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., & Prentice, I. C. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, *19*, https://doi.org/10.1029/2003GB002199

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lancaster, L. T., & Humphreys, A. M. (2020). Global variation in the thermal tolerances of plants. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 13580–13587. https://doi.org/10.1073/pnas.1918162117

Lasslop, G., Hantson, S., Harrison, S. P., Bachelet, D., Burton, C., Forkel, M., Forrest, M., Li, F., Melton, J. R., Yue, C., Archibald, S., Scheiter, S., Arneth, A., Hickler, T., & Sitch, S. (2020). Global ecosystems and fire: Multi-model assessment of fire-induced tree-cover and carbon storage reduction. *Global Change Biology*, *26*, 5027–5041. https://doi.org/10.1111/gcb.15160

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, *23*, 811–820. https://doi.org/10.1111/geb.12161

Levis, S., Bonan, G., Vertenstein, M., & Oleson, K. (2004). The community land model's dynamic global vegetation model (CLM-DGVM): Technical description and user's guide. *NCAR Tech. Note TN-459+IA*, 50. https://doi.org/10.5065/D6P26W36

Li, C., Leal Filho, W., Wang, J., Yin, J., Fedoruk, M., Bao, G., Bao, Y., Yin, S., Yu, S., & Hu, R. (2018). An assessment of the impacts of climate extremes on the vegetation in Mongolian Plateau: Using a scenarios-based analysis to support regional adaptation and mitigation options. *Ecological Indicators*, *95*, 805–814. https://doi.org/10.1016/j.ecolind.2018.08.031

Li, C., Wang, J., Hu, R., Yin, S., Bao, Y., & Ayal, D. Y. (2018). Relationship between vegetation change and extreme climate indices on the Inner Mongolia Plateau, China, from 1982 to 2013. *Ecological Indicators*, *89*, 101–109. https://doi.org/10.1016/j.ecolind.2018.01.066

Liang, L., Liu, Q., Liu, G., Li, H., & Huang, C. (2019). Accuracy evaluation and consistency analysis of four global land cover products in the arctic region. *Remote Sensing*, *11*, 1396. https://doi.org/10.3390/rs11121396

Liu, Z., Ballantyne, A. P., Poulter, B., Anderegg, W. R., Li, W., Bastos, A., & Ciais, P. (2018). Precipitation thresholds regulate net carbon exchange at the continental scale. *Nature Communications*, *9*, 1–10. https://doi.org/10.1038/s41467-018-05948-1

Liu, Z., Peng, C., Work, T., Candau, J.-N., DesRochers, A., & Kneeshaw, D. (2018). Application of machine-learning methods in forest ecology: Recent progress and future challenges. *Environmental Reviews*, *26*, 339–350. https://doi.org/10.1139/er-2018-0034

Mahdavi, S., Salehi, B., Granger, J., Amani, M., Brisco, B., & Huang, W. (2018). Remote sensing for wetland classification: A comprehensive review. *Giscience & Remote Sensing*, *55*, 623–658. https://doi.org/10.1080/15481603.2017.1419602

Maindonald, J., & Braun, J. (2013). *Data analysis and graphics using R: An example-based approach* (Vol. 10). Cambridge University Press. https://doi.org/10.1017/CBO9781139194648

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., & Zhou, B. (2021). IPCC, 2021: Climate change 2021: The physical science basis. contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change.

Mayr, S., Hacke, U., Schmid, P., Schwienbacher, F., & Gruber, A. (2006). Frost drought in conifers at the alpine timberline: Xylem dysfunction and adaptations. *Ecology*, *87*, 3175–3185. https://doi.org/10.1890/0012-9658(2006)87[3175:fdicat]2.0.co;2

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose random forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ*, *5*, e2849. https://doi.org/10.7717/peerj.2849

Miller, P. A., & Smith, B. (2012). Modelling tundra vegetation response to recent arctic warming. *Ambio*, *41*, 281–291. https://doi.org/10.1007/s13280-012-0306-1

Myers-Smith, I. H., Kerby, J. T., Phoenix, G. K., Bjerke, J. W., Epstein, H. E., Assmann, J. J., John, C., Andreu-Hayles, L., Angers-Blondin, S., Beck, P. S. A., Berner, L. T., Bhatt, U. S., Bjorkman, A. D., Blok, D., Bryn, A., Christiansen, C. T., Cornelissen, J. H. C., Cunliffe, A. M., Elmendorf, S. C., … Wipf, S. (2020). Complexity revealed in the greening of the arctic. *Nature Climate Change*, *10*, 106–117. https://doi.org/10.1038/s41558-019-0688-1

O'sullivan, O. S., Heskel, M. A., Reich, P. B., Tjoelker, M. G., Weerasinghe, L. K., Penillard, A., Zhu, L., Egerton, J. J. G., Bloomfield, K. J., Creek, D., Bahar, N. H. A., Griffin, K. L., Hurry, V., Meir, P., Turnbull, M. H., & Atkin, O. K. (2017). Thermal limits of leaf metabolism across biomes. *Global Change Biology*, *23*, 209–223. https://doi.org/10.1111/gcb.13477

Pearson, R. G., Phillips, S. J., Loranty, M. M., Beck, P. S., Damoulas, T., Knight, S. J., & Goetz, S. J. (2013). Shifts in arctic vegetation and associated feedbacks under climate change. *Nature Climate Change*, *3*, 673–677. https://doi.org/10.1038/nclimate1858

Perner, P. (2013). How to compare and interpret two learnt decision trees from the same domain? In *2013 27th International Conference on Advanced Information Networking and Applications Workshops* (pp. 318–322). IEEE. https://doi.org/10.1109/WAINA.2013.201

Phoenix, G. K., & Bjerke, J. W. (2016). Arctic browning: Extreme events and trends reversing arctic greening. *Global Change Biology*, *22*, 2960–2962. https://doi.org/10.1111/gcb.13261

Poulter, B., MacBean, N., Hartley, A., Khlystova, I., Arino, O., Betts, R., Bontemps, S., Boettcher, M., Brockmann, C., Defourny, P., Hagemann, S., Herold, M., Kirches, G., Lamarche, C., Lederer, D., Ottlé, C., Peters, M., & Peylin, P. (2015). Plant functional type classification for earth system models: Results from the European space agency's land cover climate change initiative. *Geoscientific Model Development*, *8*, 2315–2328. https://doi.org/10.5194/gmd-8-2315-2015

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*, 195–204. https://doi.org/10.1038/s41586-019-0912-1

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., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Robinson, T. M., La Pierre, K. J., Vadeboncoeur, M. A., Byrne, K. M., Thomey, M. L., & Colby, S. E. (2013). Seasonal, not annual precipitation drives community productivity across ecosystems. *Oikos*, *122*, 727–738. https://doi.org/10.1111/j.1600-0706.2012.20655.x

Rohli, R. V., Joyner, T. A., Reynolds, S. J., & Ballinger, T. J. (2015). Overlap of global Köppen-Geiger climates, biomes, and soil orders.

Physical Geography, *36*, 158–175. https://doi.org/10.1080/02723646.2015.1016384

Salonen, J. S., Verster, A. J., Engels, S., Soininen, J., Trachsel, M., & Luoto, M. (2016). Calibrating aquatic microfossil proxies with regression-tree ensembles: Cross-validation with modern chironomid and diatom data. *The Holocene*, *26*, 1040–1048. https://doi.org/10.1177/0959683616632881

Sato, H., & Ise, T. (2012). Effect of plant dynamic processes on African vegetation responses to climate change: Analysis using the spatially explicit individual-based dynamic global vegetation model (SEIB-DGVM). *Journal of Geophysical Research: Biogeosciences*, *117*, https://doi.org/10.1029/2012JG002056

Schaphoff, S., Bloh, W. V., Rammig, A., Thonicke, K., Biemans, H., Forkel, M., Gerten, D., Heinke, J., Jägermeyr, J., Knauer, J. et al. (2018). LPJML4—A dynamic global vegetation model with managed land–part 1: Model description. *Geoscientific Model Development*, *11*, 1343–1375. https://doi.org/10.5194/gmd-11-1343-2018

Scheiter, S., Moncrieff, G. R., Pfeiffer, M., & Higgins, S. I. (2020). African biomes are most sensitive to changes in CO_2 under recent and near-future CO_2 conditions. *Biogeosciences*, *17*, 1147–1167. https://doi.org/10.5194/bg-17-1147-2020

Schubert, M., Grønvold, L., Sandve, S. R., Hvidsten, T. R., & Fjellheim, S. (2019). Evolution of cold acclimation and its role in niche transition in the temperate grass subfamily pooideae. *Plant Physiology*, *180*, 404–419. https://doi.org/10.1104/pp.18.01448

Seneviratne, S. I., & Hauser, M. (2020). Regional climate sensitivity of climate extremes in CMIP6 versus CMIP5 multimodel ensembles. *Earth's Future*, *8*, e2019EF001474. https://doi.org/10.1029/2019EF001474

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M. et al. (2012). Changes in climate extremes and their impacts on the natural physical environment. https://doi.org/10.1017/CBO9781139177245.006

Shao, H., Zhang, Y., Gu, F., Shi, C., Miao, N., & Liu, S. (2021). Impacts of climate extremes on ecosystem metrics in southwest china. *Science of the Total Environment*, *776*, 145979. https://doi.org/10.1016/j.scitotenv.2021.145979

Shiferaw, H., Bewket, W., & Eckert, S. (2019). Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. *Ecology and Evolution*, *9*, 2562–2574. https://doi.org/10.1002/ece3.4919

Sierra-Almeida, A., & Cavieres, L. A. (2010). Summer freezing resistance decreased in high-elevation plants exposed to experimental warming in the central Chilean Andes. *Oecologia*, *163*, 267–276. https://doi.org/10.1007/s00442-010-1592-6

Sierra-Almeida, A., Reyes-Bahamonde, C., & Cavieres, L. A. (2016). Drought increases the freezing resistance of high-elevation plants of the central Chilean Andes. *Oecologia*, *181*, 1011–1023. https://doi.org/10.1007/s00442-016-3622-5

Sillmann, J., Kharin, V., Zhang, X., Zwiers, F., & Bronaugh, D. (2013a). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, *118*, 1716–1733. https://doi.org/10.1002/jgrd.50203

Sillmann, J., Kharin, V. V., Zwiers, F., Zhang, X., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *Journal of Geophysical Research: Atmospheres*, *118*, 2473–2493. https://doi.org/10.1002/jgrd.50188

Sitch, S., Huntingford, C., Gedney, N., Levy, P. E., Lomas, M., Piao, S. L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C. D., Prentice, I. C., & Woodward, F. I. (2008). Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five dynamic global vegetation models (DGVMS). *Global Change Biology*, *14*, 2015–2039. https://doi.org/10.1111/j.1365-2486.2008.01626.x

Song, X., & Zeng, X. (2014). Investigation of uncertainties of establishment schemes in dynamic global vegetation models. *Advances in Atmospheric Sciences*, *31*, 85–94. https://doi.org/10.1007/s00376-013-3031-1

Song, Y., Sass-Klaassen, U., Sterck, F., Goudzwaard, L., Akhmetzyanov, L., & Poorter, L. (2021). Growth of 19 conifer species is highly sensitive to winter warming, spring frost and summer drought. *Annals of Botany*, *128*, 545–557. https://doi.org/10.1093/aob/mcab090

Strahler, A. H., Muller, J., Lucht, W., Schaaf, C., Tsang, T., Gao, F., Li, X., Lewis, P., & Barnsley, M. J. (1999). Modis BRDF/albedo product: Algorithm theoretical basis document version 5.0. *MODIS Documentation*, *23*, 42–47.

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.-A., Rahman, A. et al. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote Sensing*, *12*, 1135. https://doi.org/10.3390/rs12071135

Therneau, T., Atkinson, B., & Ripley, B. (2021). Rpart: Recursive partitioning. R Package Version, 4.1-15. http://CRAN.R-project.org/package=rpart

Treharne, R., Bjerke, J. W., Tømmervik, H., & Phoenix, G. K. (2020). Development of new metrics to assess and quantify climatic drivers of extreme event driven arctic browning. *Remote Sensing of Environment*, *243*, 111749. https://doi.org/10.1016/j.rse.2020.111749

Ullerud, H. A., Bryn, A., Halvorsen, R., & Hemsing, L. Ø. (2018). Consistency in land-cover mapping: Influence of field workers, spatial scale and classification system. *Applied Vegetation Science*, *21*, 278–288. https://doi.org/10.1111/avsc.12368

Ummenhofer, C. C., & Meehl, G. A. (2017). Extreme weather and climate events with ecological relevance: A review. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160135. https://doi.org/10.1098/rstb.2016.0135

Venn, S. E., Morgan, J. W., & Lord, J. M. (2013). Foliar freezing resistance of Australian alpine plants over the growing season. *Austral Ecology*, *38*, 152–161. https://doi.org/10.1111/j.1442-9993.2012.02387.x

Vidal Jr, J., le Roux, P. C., Johnson, S. D., & Clark, V. R. (2021). Beyond the tree-line: The C3–C4 'grass-line' can track global change in the world's grassy mountain systems. *Frontiers in Ecology and Evolution*, *861*, 760118. https://doi.org/10.3389/fevo.2021.760118

von Humboldt, A., & Bonpland, A. (1807). *Essai sur la géographie des plantes*.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The inter-sectoral impact model intercomparison project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 3228–3232. https://doi.org/10.1073/pnas.1312330110

Whittaker, R. H. (1962). Classification of natural communities. *The Botanical Review*, *28*, 1–239. https://doi.org/10.1007/BF02860872

Woodward, F. I. (1990). The impact of low temperatures in controlling the geographical distribution of plants. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *326*, 585–593. https://doi.org/10.1098/RSTB.1990.0033

Woodward, F. I., Lomas, M. R., & Kelly, C. K. (2004). Global climate and the distribution of plant biomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*, 1465–1476. https://doi.org/10.1098/rstb.2004.1525

Wullschleger, S. D., Epstein, H. E., Box, E. O., Euskirchen, E. S., Goswami, S., Iversen, C. M., Kattge, J., Norby, R. J., van Bodegom, P. M., & Xu, X. (2014). Plant functional types in earth system models: Past experiences and future directions for application of dynamic vegetation models in high-latitude ecosystems. *Annals of Botany*, *114*, 1–16. https://doi.org/10.1093/aob/mcu077

Yan, H., Liang, C., Li, Z., Liu, Z., Miao, B., He, C., & Sheng, L. (2015). Impact of precipitation patterns on biomass and species richness of annuals in a dry steppe. *PLoS One*, *10*, e0125300. https://doi.org/10.1371/journal.pone.0125300

Yang, Y., Zhao, J., Zhao, P., Wang, H., Wang, B., Su, S., Li, M., Wang, L., Zhu, Q., Pang, Z., & Peng, C. (2019). Trait-based climate change predictions of vegetation sensitivity and distribution in China. *Frontiers in Plant Science*, *10*, 908. https://doi.org/10.3389/fpls.2019.00908

Zhang, W., Miller, P. A., Jansson, C., Samuelsson, P., Mao, J., & Smith, B. (2018). Self-amplifying feedbacks accelerate greening and warming of the arctic. *Geophysical Research Letters*, *45*, 7102–7111. https://doi.org/10.1029/2018GL077830

Zhu, J., Zeng, X., Zhang, M., Dai, Y., Ji, D., Li, F., Zhang, Q., Zhang, H., & Song, X. (2018). Evaluation of the new dynamic global vegetation model in CAS-ESM. *Advances in Atmospheric Sciences*, *35*, 659–670. https://doi.org/10.1007/s00376-017-7154-7

Zimmermann, N. E., Yoccoz, N. G., Edwards, T. C., Meier, E. S., Thuiller, W., Guisan, A., Schmatz, D. R., & Pearman, P. B. (2009). Climatic extremes improve predictions of spatial patterns of tree species. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 19723–19728. https://doi.org/10.1073/pnas.0901643106

How to cite this article: Beigaitė, R., Tang, H., Bryn, A., Skarpaas, O., Stordal, F., Bjerke, J. W., & Žliobaitė, I. (2022). Identifying climate thresholds for dominant natural vegetation types at the global scale using machine learning: Average climate versus extremes. *Global Change Biology*, 00, 1–23. https://doi.org/10.1111/gcb.16110