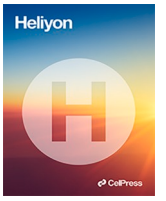
Heliyon 10 (2024) e38419

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Modelling *Eucalyptus globulus* spatial distribution in the upper Blue Nile basin using multi spectral Sentinel-2 and environmental data

Abdurohman Yimam [a,b,*], Asnake Mekuriaw [a], Dessie Assefa [c,d], Woldeamlak Bewket [a]

[a] *Department of Geography and Environmental Studies, Addis Ababa University, Addis Ababa, Ethiopia*
[b] *Department of Geography and Environmental Studies, Debre Birhan University, Debre Birhan, Ethiopia*
[c] *Department of Natural Resources Management, Bahir Dar University, Bahir Dar, Ethiopia*
[d] *Institute of Forest Ecology, University of Natural Resources and Life Sciences, Vienna, Austria*

A R T I C L E I N F O

Keywords:
Variance inflation factor
Random forest
Support vector machine
Boosted regression tree

A B S T R A C T

Eucalyptus plantations are widespread in the highlands of northern Ethiopia. The species has been used for centuries for various purposes. However, there are controversies surrounding the species with excessive soil nutrient and water consumption. Modelling the spatial distribution of the species is fundamental to understand its ecological and hydrological effects in the region for policy inputs. Therefore, the purpose of this study is to develop a model for mapping the spatial distribution of *Eucalyptus globulus*. We used the spectral bands of Sentinel-2 data, vegetation indices, and environmental data as predictor variables and three machine learning algorithms (Random Forest, Support Vector Machine, and Boosted Regression Trees) to model the current distribution of *Eucalyptus globulus*. Eleven of the twenty-five predictor variables were filtered using a variance inflation factor (VIF). 419 in situ georeferenced data points were used for training, and validating the models. The area under the curve (AUC), kappa statistic (K), true skill statistic (TSS), Root Mean Squared Error and coefficient of determination (R^2) were used to validate the models' performance. The model validation metrics confirmed the highest performance of Random Forest. The prediction map of Random Forest revealed that *Eucalyptus globulus* was fairly detected in non-*Eucalyptus globulus* woody vegetation ($R^2 = 0.86$, $P < 0.001$; RMSE = 0.31). We found that the Green Normalized Difference Vegetation Index and environmental variables, such as elevation and distance from the road, were the most important predictor variables in explaining the distribution of *Eucalyptus globulus*. Our findings demonstrate that machine learning algorithms with Sentinel-2 spectral bands and vegetation indices compounded with environmental data can effectively model the spatial distribution of *Eucalyptus globulus*.

1. Introduction

Eucalyptus is one of the most widely planted species in the world [1]. Zhang et al. [2] estimated that the global areal coverage of *Eucalyptus* plantations is about 23 million hectares (ha). The species was introduced to East Africa in the late 19th and early 20th centuries. The largest area of *Eucalyptus* plantations in the region was found in Ethiopia, covering approximately 895,000 ha in 2011

* Corresponding author. Department of Geography and Environmental Studies, Addis Ababa University, Addis Ababa, Ethiopia.
E-mail address: abdurohman2015@gmail.com (A. Yimam).

https://doi.org/10.1016/j.heliyon.2024.e38419
Received 10 August 2023; Received in revised form 23 September 2024; Accepted 24 September 2024
Available online 25 September 2024

[3].There are about 55 different *Eucalyptus* subspecies in Ethiopia, although farmers favor *Eucalyptus globulus* and *Eucalyptus camaldulensis* [4,5]

Currently, *Eucalyptus* has become the dominant plantation in the northern parts of Ethiopia and is the main source of fuel wood and building materials compared to other tree species [6]. Although *Eucalyptus* plantations are very popular among landowners and farmers, there is growing concern about their potential negative effects on water and soil resources [7–9]. The holistic impacts of the species on the environment, such as the extensive use of soil water [8,10] and the high consumption of soil nutrients compounded with its allelopathic effect [11] has led to controversies among academics, researchers, and policymakers. However, so far only a few investigations have been made to map the spatial distribution of *Eucalyptus* plantations– as an aggregate without developing a single-tree species inventory approach– in the upper Blue Nile basin [12–14]. In addition, there is no effective automated method that is able to detect and map *Eucalyptus* trees expansion over large areas [15,16]

Therefore, it is important to map and quantify the spatial distribution of *Eucalyptus* plantations to provide reliable information for decision makers and land use planners. Furthermore, accurate data on the current spatial distribution of the *Eucalyptus* plantations can provide a foundation for analysis and modelling efforts to examine the ecological and hydrological impacts of the species in the region. Thus, the single-tree species inventory approach has become the starting point for this endeavor to model the current spatial distribution of *Eucalyptus globulus* (*E. globulus*) plantations in the study area.

Species distribution models (SDMs) are used to model and predict the geographic distributions of species using the occurrence of species and their environmental variables to [15,16]. Since the modelling is based on the statistical relationship between species and environment, there are many statistical methods available, ranging from multiple linear regressions to sophisticated machine learning algorithms [17,18]. Machine-learning (ML) algorithms, which are non-parametric, have emerged since 1995 as alternatives to other classifier algorithms [19,20], as they tend to generate higher accuracy than conventional parametric classifiers [21–23].

ML algorithms have the ability to learn complex patterns and have high generalization capacity [24–28]. Further, the unique characteristics of ML algorithms, such as independence of the data statistical distribution, makes the algorithms to incorporate data from different sensors, auxiliary data and even categorical variables [22,29]. Therefore, our study employed three non-parametric ML algorithms, namely Random Forests (RF), Support Vector Machine (SVM), and Boosted Regression, which are powerful methods for modelling forest cover [30,31]. Moreover, these methods produce better overall accuracy than alternative machine-learning classifiers [30–32].

Over the past few decades, the development of Remote Sensing (RS) technologies and the use of empirical models have made it

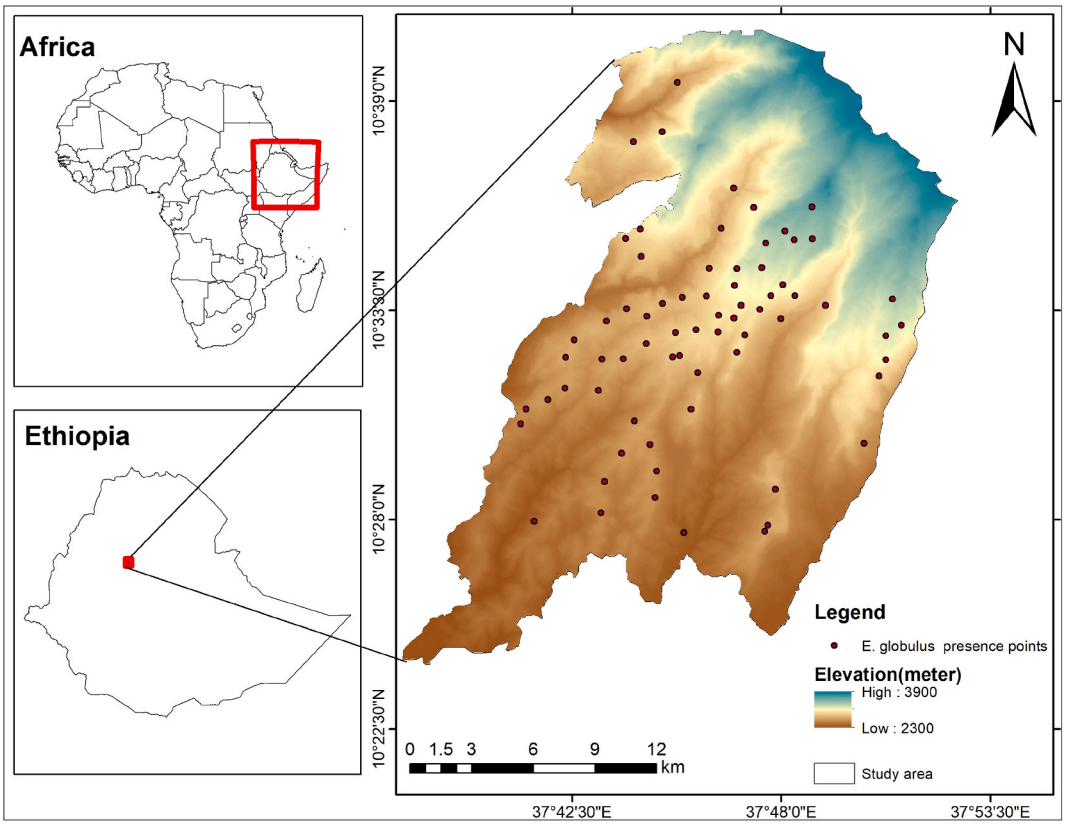


Fig. 1. Location map of the study area (right), map of Ethiopia including the location of the study area (upper left) and map of Africa including location of Ethiopia (lower left).

possible to extract information accurately and effectively model the distribution of tree species [30,33,34]. However, producing accurate maps at the species level using medium-resolution satellites, such as Landsat, is challenging for forest monitoring and conservation planning [35], particularly in heterogeneous environments, such as Ethiopia [36].

Recently, the operation of the Sentinel-2A/B satellites have played significant contribution for successful forest classification and monitoring by providing high spatial, spectral and temporal resolution data [37,38]. Abdi [30] and Immitzer et al. [39] have shown the potential of Sentinel-2's red-edge spectral bands, the shortwave infrared bands, and also the blue, green and near-infrared bands [40] for forest species classification. Several studies using species distribution models have found that spectral bands and vegetation indices provide meaningful information for understanding the distribution of species [31,41–43]. Further, the environmental variables comprised of climatic factors, topographic features, and soil properties influenced the distributions of woody plant species. Hościło et al. [44] pointed out that environmental factors, along with multi-spectral Sentinel-2 data, have been shown to provide significant results in forest species identification.

To the best of our knowledge, there are no studies that develop an automated method that is able to detect and map spatial distribution of *E. globulus* in the upper Blue Nile basin of Ethiopia using Sentinel-2 spectral bands, vegetation indices, and environmental variables. Therefore, the current study was undertaken with the objectives of 1) assessing the current spatial distribution and cover of *E. globulus* by applying robust machine learning algorithms; 2) identifying the dominant predictor variables that explain the current distribution of *E. globulus*; and 3) examining the performance of the machine learning algorithms in modeling the spatial distribution of *E. globulus* plantations in the upper Blue Nile basin of Ethiopia.

2. Materials and methods

2.1. Description of the study area

The study was conducted in the Sinan District, which is located in the upper Blue Nile basin of Ethiopia. Geographically, it lies between 10°20′35″ to 10°50′38″ North and 37°35′10″ to 37°52′20″ East (Fig. 1). The study area covers 436 km^2, with altitudes ranging between 2300 and 3900 m above sea level. The study area is characterized by diverse topography and climate [45]. It has cool humid (75 %), cool sub-humid (23 %), and cool (2 %) agroclimatic zones. The rainfall pattern is unimodal, with the total annual rainfall ranging from 840 to 1266 mm with a mean annual of 1070 mm [46]. The mean annual air temperature is 16.6 °C which ranges from 11.3 °C to 25.8 °C.

The major trees and shrubs vegetation in the cool sub-humid (*Woinadega*) and cool humid (*Dega*) agroecological zones are

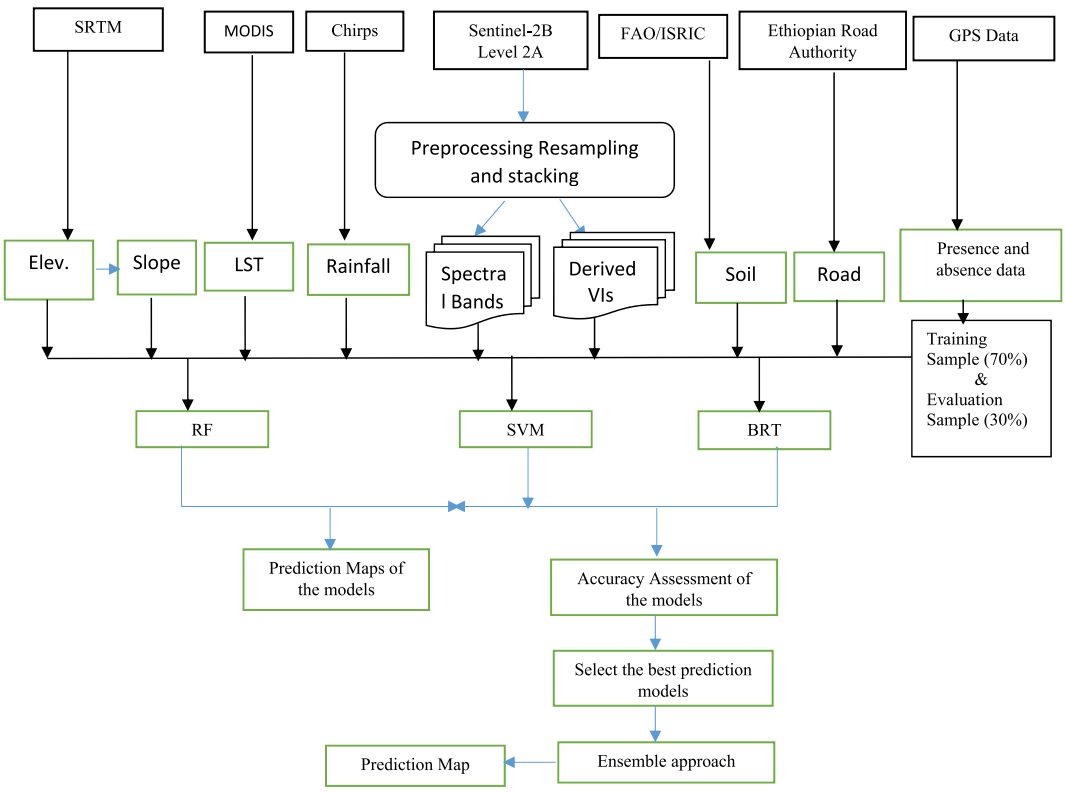


Fig. 2. Flowchart of the methodology for the prediction of the current distribution of *E. globulus*.

Eucalyptus globulus, *Cupressus lusitanica*, *Acacia abyssinica*, *Juniperus procera*, *Acacia lahai*, *Arundinaria alpine*, *Acacia decurrens* and *Rosa abyssinica*. In the cool (*wurich*) agroecological zone, *Lobelia rhynchopetalum*, *Erica arborea*, and *Euryops pinifolius* were dominant [47].

Mixed farming is the mainstay of the household economy in the study area and is intensively carried out by those who have farmland and livestock. The main crops grown include cereals (*Hordeum vulgare*, wheat, and *Eragrostis tef*), pulses (*horse bean*, *Cicer arietinum*, and *Lens culinaris*), fruits, and vegetables (*Malus domestica*, *Brassica oleracea*, and *Allium cepa*) [45].

The main soil types in the study area include luvisols, leptosols, and Alisols [47]. These soil types originate from volcanic sources, specifically Mio-Pliocene shield volcano lavas and, at lower elevations, Oligocene flood basalts [48]. Alisols can be found in the midland area at altitudes ranging from 2400 to 2800 m above sea level. Leptosols and luvisols, on the other hand, are located in hilly and mountainous highlands, with altitudes ranging from 2800 to 3800 m [47].

Even though the current land use type of the study area has been dominated by agricultural land, in the last two decades, there has been a dramatic change in the land use system due to the change in tree cover [45]. In particular, there is an exclusive expansion of *E. globulus* over farmlands and grazing lands in the study area [45,47].

2.2. Methodology

Potential predictor variables were identified based on literature reviews and expert knowledge to model the current distribution of *E. globulus* in the upper Blue Nile region of Ethiopia. As shown below (Fig. 2), the identified variables were derived from different sources. In situ georeferenced data of the training presence and absence point data (70 % of the collected data) were used by the selected ML methods (e.g., RF, SVM, and BRT) to identify the most important predictor variables and model the current distribution of *E. globulus*. The performance of the models was assessed using the presence and absence of georeferenced point data (30 % of the collected data) by employing the AUC, Kappa, and TSS evaluation metrics.

2.3. Spatial datasets

2.3.1. Ground truthing data collection

The major land use/land cover (LULC) classes of the study area that were used as inputs for ground truthing were *Eucalyptus* plantations, other woody vegetation, cropland, grazing land, and settlement. LULC types were classified by considering the environmental setting of the study area where:

Cropland: Land used for cultivation of annual or perennial crops.

Gazing land: Areas used for communal or private grazing and browsing land.

Eucalyptus Plantation: Areas exclusively used for the plantation of *E. globulus*.

Woody Vegetation: Areas covered by bushes, shrubs, and trees other than *E. globulus*.

Settlement: Areas occupied by any means of shelter, including backyards and road infrastructure.

The species presence and absence data were collected using a stratified sampling technique (considering the area coverage of each LULC class) [49]. Thus, using the Garmin Handheld Global Positioning System (GPS), 218, 89, 45, 37, and 30 georeferenced data points were collected from cropland, *Eucalyptus* plantation, woody vegetation, grazing land, and settlement from December 2022 to January 2023, respectively. Because auto-correlated datasets introduce biases in spatial distribution models at the calibration and validation stages [50,51], the in situ georeferenced data points were spatially filtered using the spatially rarefy occurrence data function of the SDM toolbox in ArcGIS 10.8. This function helped increase the models' ability to predict spatially independent data. Therefore, 70 % and 30 % of the georeferenced data were used to train and validate the models, respectively.

2.4. Images acquisition and preprocessing

The Sentinel-2B Level-2A product (available free of charge) was used in this study (https://sentinel.esa.int/web/sentinel). To obtain a cloud-free image and minimize confusion between active field crops and the *Eucalyptus* species, Sentinel-2B captured after

Table 1
Sentinel-2 spectral bands used to model the current distribution of *E. globulus*.

Bands	Spectral region	Spatial Resolution(m)	Wavelength (μm)
2	Blue	10	0.458–0.523
3	Green	10	0.543–0.578
4	Red	10	0.650–0.680
8	Near Infrared	10	0.698–0.713
5	Red Edge	20[a]	0.733–0.748
6	Red Edge	20[a]	0.773–0.793
7	Red Edge	20[a]	0.785–0.900
8A	Near Infrared	20[a]	0.855–0.875
11	Shortwave Infrared	20[a]	1.565–1.655
12	Shortwave Infrared	20[a]	2.100–2.280

[a] 20 m spectral resolution of the bands were resample to 10 m.

field crop harvest and during the dry season (specifically on January 21, 2023) was selected and downloaded from the Copernicus Open Access Hub (https://scihub.copernicus.eu/). Sentinel-2B bands: 2, 3, 4, and 8 (10 m resolution) and 5, 6, 7, 8a, 11, and 12 (20 m resolution) were resampled to 10 m using nearest neighbours in SNAP 9.0 software (Table 1) and used as explanatory variables.

In addition, vegetation indices (Table 2) derived from the sentinel 2B image were used as explanatory variables to model the spatial distribution of *E. globulus*.

Further, environmental variables such as elevation, rainfall, Land Surface Temperature (LST), and soil pH were used as predictor variables (Table 3) to model the spatial distribution of *E. globulus*. Among environmental variables, climatic factors, such as precipitation and temperature, are key determinants of plant and animal species distribution [61]. In addition, topographic features such as slope, aspect, and elevation control the distribution of tree species as they affect the local climate [62] and soil conditions, such as the acidity levels of the soil [63].

Hence, precipitation satellite-based rainfall products from the monthly Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS), the Shuttle Radar Topographic Mission (SRTM), and the Moderate Resolution Image Spectroradiometer (MODIS) were used to obtain precipitation data, elevation and slope datasets, and Land Surface Temperature (LST) data for the study area, respectively. To obtain average precipitation and LST data for the study area, ten years of monthly CHIRPS and MODIS data (2011–2021) were used. Soil pH data with a spatial resolution of 250 m were downloaded from the open-access global soil database (https://www.isric.org/). During field observations, we observed *Eucalyptus* plantations closer to road access. Thus, distance from the road was considered a potential explanatory variable and was extrapolated from the Road Map of Ethiopia (acquired from the Ethiopian Road Authority).

All datasets were resampled with the highest spatial resolution (10 m) and Sentinel-2 data and were kept in raster format with the same reference systems using ArcGIS 10.8. Finally, modelling was performed using the SDM package in the R 4.1 software [64].

2.5. Selection of predictor variables

Initially, 25 predictor variables (Table 1) were used to explain the current distribution of *E. globulus*. To assess the predictive power of the variables, four cases with different sets of predictor (explanatory) variables were used to model and evaluate the performance of the prediction models: Case 1 included only the spectral bands (Table 1), Case 2 considered only the vegetation indices (Table 2), Case 3 included the environmental variables only (Table 3), and Case 4 (final model) included the filtered predictor variables from three cases which did not have collinearity problems.

Hence, in the final model, the predictor variables were tested for multicollinearity, which otherwise caused instability in the parameter estimation and biases in the prediction [65]. Therefore, using variance inflation factors (VIF) (eq. (1)), the predictor variables were tested for multicollinearity using the SDM packages in R 4.1 software. VIF is one of the most precise methods for examining collinearity between predictor variables, as it evaluates how a predictor variable can be explained by another predictor variable [66].

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1}$$

where VIF_i refers to the variance inflation factor for the *i*th predictor and R_i^2 is *the* value of the coefficient of determination obtained by regressing the *i*th predictor on the other predictors

As a rule of thumb, a predictor variable with a VIF coefficient greater than 10 is a sign of collinearity [67]. Hence, in this study, out of the 25 predictor variables, 14 predictor variables were excluded because they had a VIF coefficient greater than 10.

2.6. Machine learning algorithms

ML algorithms manage complex data with large sets of variables and a high-dimensional feature space [22] because they are highly robust and accurate compared to conventional classification methods, such as maximum likelihood classifiers, generalized linear models, and logistic regression [21,23,54,55]. Thus, in this study, three machine-learning algorithms were used to model the current distribution of *E. globulus*. Random Forest (RF) is one of the ML algorithms developed from Classification and Regression Trees (CART)

Table 2
Vegetation indices used as to model the current distribution of *E. globulus*.

Vegetation Indices	Formula	Sources/references
Atmospherically Resistant Vegetation Index (ARVI)	(NIR – (2 * RED) + BLUE)/(NIR + (2 * RED) + BLUE)	[52]
Green Normalized Difference Vegetation Index(GNDVI)	(NIR – GREEN)/(NIR + GREEN)	[53]
Transformed Normalized Difference Vegetation Index (TNDVI)	sqrt((NIR – RED)/(NIR + RED) + 0.5)	[54]
Normalized Difference Vegetation Index(NDVI)	(NIR – RED)/(NIR + RED)	[55]
Normalized Difference Red Edge (NDRE)	(Red-edge3-Red-edge2)/((Red-edge3-Red-edge2)	[56]
The Modified Chlorophyll Absorption in Reflectance Index(MCARI)	(Red-Edge - R) - 0.2 * (Red-Edge - G) * (Red-Edge/Red)	[57]
The Green Leaf Index (GLI)	((Green - Red) + (Green- Blue))/((2 * Green) + (Blue + Red))	[58]
Green Red Vegetation Index (GRVI)	(Green - Red)/(Green + Red)	[59]
Transformed Soil Adjusted Vegetation Index(TSAVI)	((NIR – R)/(NIR + R + L))* (1+L)	[60]

Table 3
Environmental variables used to model the current distribution of *E. globulus*.

Environmental Variables	Description	Source/References
Elevation	Shuttle Radar Topography Mission digital elevation model (30 m spatial resolution)	USGS
Slope	Derived from Elevation	SRTM, USGS
Rainfall	Mean annual rainfall	CHRIPS
LST	Land surface Temperature	MODIS
Soil pH	Acidity level of the soil	ISRIC/FAO
Dist_ Road	Distance from the road	Ethiopian Road Authority

[68]. It is built on multiple CART and the prediction is an ensemble of classifiers. Boosted Regression Tree (BRT) as RF follows an ensemble approach using a sequential decision tree approach to improve the accuracy of the model. However, unlike RF, BRT is built using input data that are assessed in subsequent trees, and weekly modelled data in the previous trees have a chance to be selected in the next new tree [69]. Support Vector Machine (SVM) is another successful non-parametric method used for forest species classification [70–72]. SVM undertakes an iterative process until it finds the optimum minimization or hyperplane boundary that can separate the classification into a predefined set of classes [28]. Further details on RF, BRT, and SVM are found in Refs. [68,73–75], respectively.

2.7. Model evaluation

Since each evaluation metric has its own unique aspect to quantify the predicted performance of the model, employing more than one metric provides an overall assessment of the accuracy of the models [76]. Hence, in this study, five evaluation metrics were used to assess the performance of the models: area under the curve (AUC), kappa statistic (K), true skill statistic (TSS), root mean square error (RMSE) and coefficient of determination (R^2). A receiver operating characteristic (ROC) curve plots the false positive rate (1-specificity) on the x-axis against the true positive rate (sensitivity values) on the y-axis [77]. In our study, sensitivity (eq. (2)) represents the proportion of correctly predicted cells of presence of *E. globulus* while specificity (eq. (3)) quantifies pixels correctly identified as absence of *E. globulus*. Sensitivity and Specificity were calculated as:

$$\text{Sensitivity} = \frac{TN}{TN + FN} \tag{2}$$

where TP (True Positives) refers to correctly identified records of the presence of *E. globulus* and FN (False Negative) donates to the records of the absence of *E. globulus* but in fact *E. globulus* is found.

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{3}$$

where TN (True Negative) refers to correctly identified records of the absence of *E. globulus* and FP (False Positive) donates to the records of the presence of *E. globulus* but in fact *E. globulus* is absent.

The values of the area under the ROC curve (AUC) range between 0 and 1, where an AUC value closer to 1 indicates better discrimination performance and less than 0.5 value implies worse predictive discrimination. An alternative metric to the AUC is the True Skill Statistic, which combines correctly predicted presence and absence observations [78]. Thus, TSS (eq. (4)) considers both omission and commission errors, which offers a reasonably better model performance evaluation. True Skill Statistic (TSS) was calculated as:

$$TSS = \text{Sensitivity} + \text{specificity} - 1 \tag{4}$$

The kappa statistic (K) was used as the other evaluation metric. It was computed from the cross-tabulation of the predicted and observed values of the presence and absence observations (eq. (5)). It ranges from-1 to +1, where +1 indicates perfect agreement between observed values of the presence and absence observation, and values of zero or less indicate a performance that is no better than random [79].The formula for Cohen's Kappa statistics can be formulated as:

$$K = \frac{2(TP\text{x}TN - FN\text{x}FP)}{(TP + FP)\text{x}(FP + TN) + (TP + FN) + (FN + TN)} \tag{5}$$

where TP, TN, FP and FN are defined in the questions 2 and 3.

The fourth evaluation metric used in the study was the root mean square error (RMSE), which was calculated using observed and predicted sample points. Equation (6) can be used to calculate the RMSE.

$$RMSE = \frac{\sum_{i=1}^{n} (Observed_i - predicted_i)}{n} \tag{6}$$

where n refers to the number of sample point.

2.8. Model validation

A bootstrap replication approach was used to validate the prediction models. This involved ten-fold cross-validation, where the validation process was repeated ten times with different calibration batches and the average results of the evaluation metrics were recorded. Furthermore, the models were tested outside the study area to validate their replicability. Hence, using the same procedure as it was used to calibrate the models, model validation was performed in *Machakel* district using 209 georeferenced data which were collected using a Handheld Global Positioning System (GPS), 51 of the georeferenced data from *E. globulus*, and 158 from non-eucalyptus plantation areas (cropland, grazing land, other woody vegetation, and settlements and towns). In the district, areas that were dominantly covered by E. globulus and had similar agroecological zones as the study area were selected for validation.

2.9. Statistical analyses and mapping

The modelling and validation were performed using the SDM package in the open-source software R, version 4.1. Multicollinearity within the predictor variables was evaluated using the variance inflation factors in the USDM R package. The importance of the explanatory variables was also examined using correlation metrics in the SDM package.

The prediction maps were produced in ArcGIS 10.8 using the threshold value (maximum sensitivity + specificity), which is one of the recommended thresholds in SDM [80]. The area above the threshold value was considered as the presence of *E. globulus* species; otherwise, it was considered the absence of the species. The coefficient of determination was analyzed in the open-source software R version 4.1, using the predicted probability values with the calibrated reference data of the presence and absence of *E. globulus* and non-*E. globulus* woody vegetation. In addition, ensemble analysis was employed using the weighted average of the best predictors of each model in the SDM package in R.

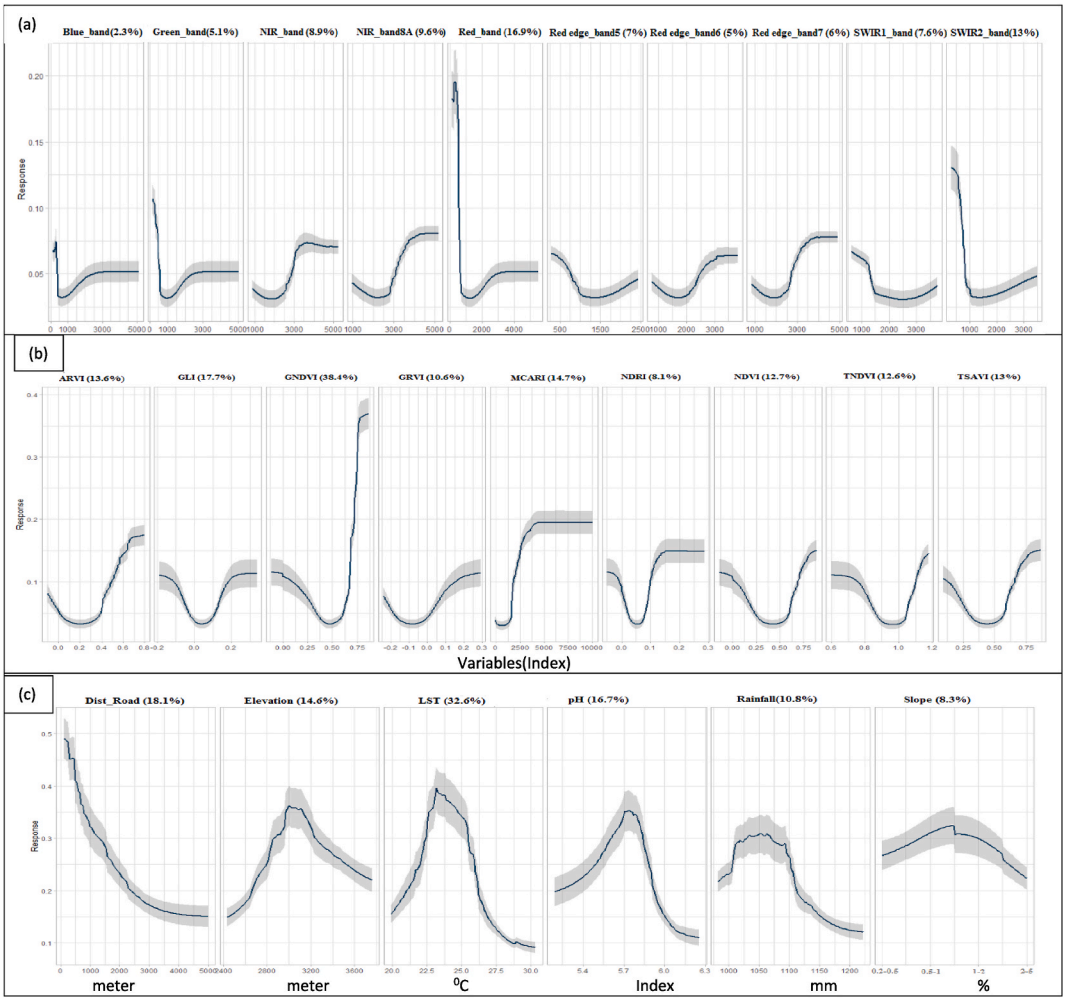


Fig. 3. The response curve of the predictor variables and their relative importance values (the number in brackets next to each variables name): Spectral Bands (a), Vegetation Indices (b) and Environmental variables(c).

3. Results

3.1. Relative importance of the predictor variables

In this study, we applied three ML algorithms, RF, SVM, and BRT, to model the current distribution of *E. globulus* in the study area. Fig. 3 shows the initial 25 predictor variables, their relative importance values, and their response curves. The relative importance values of the spectral bands (Fig. 3a) of Sentinel 2 B revealed that the red band(Band_4), SWIR2(Band_12), and NIR_band8A (Band_8A) were the three most important predictor variables, accounting for 17 %, 13 %, 9 %, and 10 % relative importance values, respectively.

Among the vegetation indices, the Green Normalized Difference Vegetation Index (GNDVI) (38.4 %) was the most significant predictor variable, followed by the Modified Chlorophyll Absorption Reflectance Index (MCARI) (14.7 %) and the Atmospherically Resistant Vegetation Index (ARVI) (13.6). LST (32.4 %), Dist_Road (18.1) and soil pH (16.7 %) were ranked as the highest among the environmental predictor variables in explaining the current distribution of *E. globulus* (Fig. 3c).

In the final model, 11 predictor variables from the initial twenty-five variables were filtered to model the current distribution of *E. globulus* as 14 predictor variables had collinearity problems. The variable importance value of the final model revealed that the vegetation indices were the most important predictor variables in determining the current distribution of *E. globulus* (Fig. 4). GNDVI was by far the most significant predictor variable, with 41 % relative importance followed by NDRE (13 %).

3.2. Models evaluation and validation

Table 4 shows the accuracy assessment of the model performance used in this study: AUC, kappa coefficient, and TSS methods. The accuracy assessment showed some variation within the models and in the evaluation of the statistical techniques. In the case where spectral bands were used as the predictor variables, the performances of the models (RF, SVM and BRT) were above 0.90 for AUC and TSS metrics but the value of the kappa coefficient for SVM and BRT models was 0.88 and 0.85, respectively. The models showed excellent performance in the case where the vegetation indices were used as predictor variables in the three evaluation metrics. However, the performance of TSS and the kappa coefficient were relatively poor when the environmental variables were treated independently as predictor variables.

In the final model analysis, where the 11 selected predictor variables were considered, the performance of the models was fairly accurate in determining the current distribution of *Eucalyptus globulus*, where the AUC, kappa coefficient, and TSS metrics were greater than 0.95, almost in all cases. The AUC statistics of the final model of the receiver operating characteristic (ROC) curve is 0.99 for the RF, SVM, and BRT algorithms(Fig. 5)

The validation results (Table 5) clearly show that the models performed well in mapping the distribution of *E. globulus* in the selected areas of the *Machakel* district of Ethiopia. The kappa coefficient, AUC, and TSS were 0.98, 0.96, and 0.94 values for RF, respectively. In addition, the predictor variables employed showed a similar magnitude of relative variable importance to the calibration finding. The importance value of the predictor variables (Fig. 6) revealed that GNDVI (29 %) emerged as one of the most significant predictor variables in determining the spatial distribution of E. globulus in Machakel district followed by distance to the road (20 %).

3.3. Spatial distribution of E. globulus

The current distribution of *E. globulus* was predicted using the RF model (Fig. 7a) as one of the best predictors and ensemble approach (Fig. 7b). The threshold values of 0.46 and 0.4 were used to produce the presence and absence maps of the *E. globulus* for the RF model and the ensemble approach, respectively. The results showed that the species was highly concentrated in the central part of

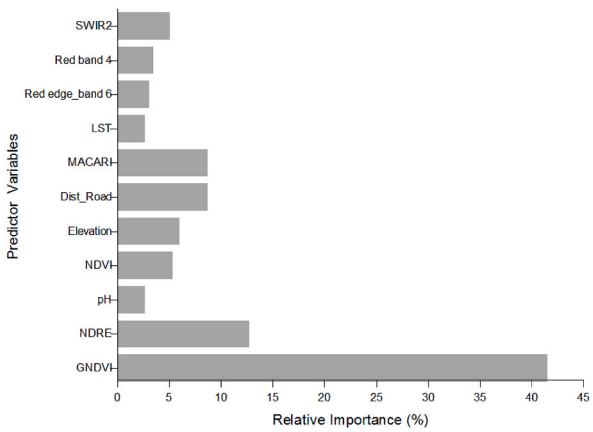


Fig. 4. Relative importance (%) of 11 predictor variables for explaining *E. globulus* distribution in final model.

Table 4
Model mean performance using the test dataset of the study area.

Predictor Variables	Model	AUC	Kappa	TSS
Spectral Bands	RF	0.99	0.93	0.96
	SVM	0.98	0.88	0.93
	BRT	0.98	0.85	0.94
Vegetation Indices	RF	0.99	0.95	0.97
	SVM	0.98	0.96	0.94
	BRT	0.97	0.90	0.97
Environmental Variables	RF	0.93	0.64	0.76
	SVM	0.86	0.47	0.65
	BRT	0.88	0.63	0.71
Selected Variables from the three groups	RF	0.99	0.97	0.97
	SVM	0.99	0.96	0.95
	BRT	0.99	0.93	0.97

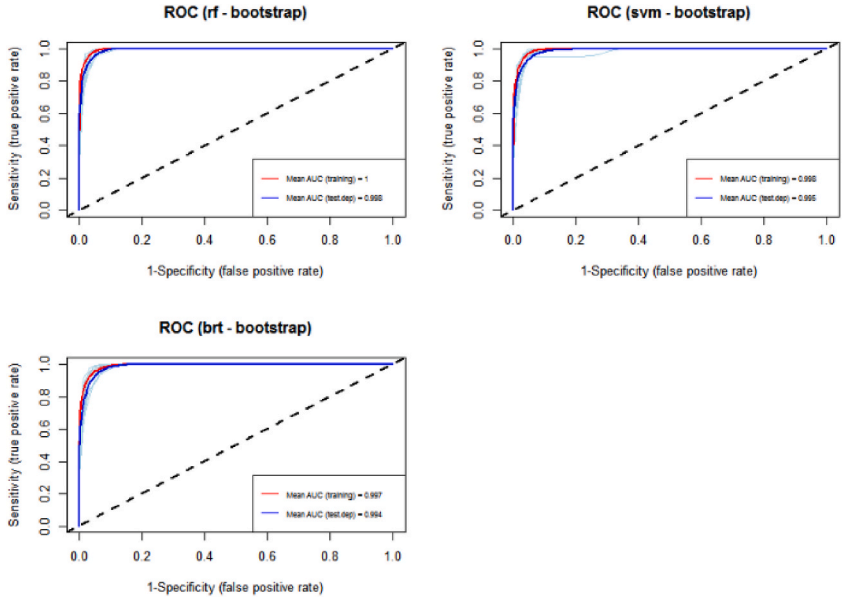


Fig. 5. Receiver operator characteristics (ROC) curve of RF, SVM and BRT.

Table 5
Model mean performance using test dataset in the validation area (Machakel district).

Predictor Variables	Model	AUC	Kappa	TSS
Selected Predictor Variables	RF	0.98	0.93	0.98
	SVM	0.96	0.88	0.97
	BRT	0.94	0.85	0.91

the study area (Fig. 7). In addition, some patches of this species were found in the southeastern part of the study area. The total area covered by *E. globulus* was 3930 ha (9.5 %) and 3536 ha (8.6 %) of the study area, as predicted by the RF model and the ensemble approach, respectively.

The analysis of the RF model prediction map (Fig. 8) shows that there is a strong significant relationship between the predicted probability and calibration reference data points of the presence and absence of *E. globulus* and non- *E. globulus* woody vegetation ($R^2 = 0.86$, $p < 0.001$, RMSE = 0.31). *E. globulus* was fairly identified from the non-*E. globulus* woody vegetation in the study area. The georeferenced data points used in the model calibration discriminated for the presence of *E. globulus* and non-*E. globulus* woody vegetation, with few overlapping reference points.

4. Discussion

In this study, we used Sentinel-2B spectral bands, vegetation indices (VIs), and environmental variables (Table 1) to fix the SDMs.

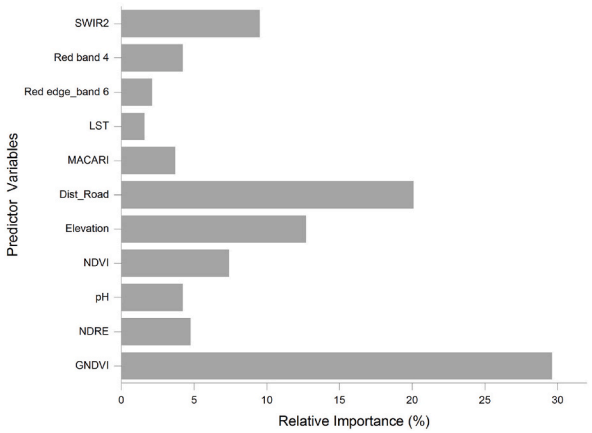


Fig. 6. Relative importance (%) of predictor variables for explaining *E. globulus* distribution in the validation district (Machakel district).

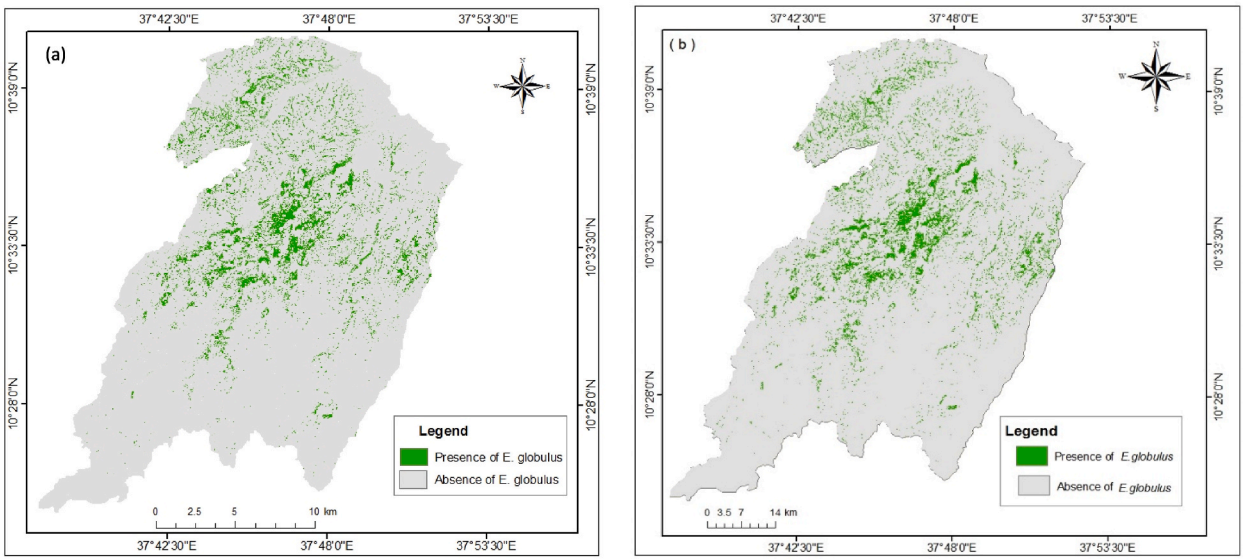


Fig. 7. *E. globulus* distribution maps predicted by the RF model (a) and ensemble approach (b).

Specifically, we applied RF, SVM, and BRT machine learning algorithms to detect and map the spatial distribution of *E. globulus* in the Upper Blue Nile basin of Ethiopia.

The accuracy assessment results of the models varied among the four cases and among the ML algorithms. In the first three cases, where the predictor variables (spectral bands, vegetation indices, and environmental variables) were treated independently, spectral bands and vegetation indices achieved high accuracy in the three evaluation metrics (AUC, Kappa, and TSS) employed in the study. When the environmental predictor variables were considered independently, the kappa coefficient values derived from RF, BRT, and SVM were 0.64, 0.63, and 0.47, respectively. Hence, the performance of the models with respect to the kappa metric is relatively poor.

In the final model analysis, the 11 predictor variables were filtered to run the models. All three ML algorithms achieved a higher level of accuracy. In our study, the greatest accuracy was achieved by RF, with AUC values of 0.99 and 0.97 for both TSS and kappa coefficient (Table 4). Analysis of the RF prediction map (Fig. 8) revealed that *E. globulus* was fairly detected from the non-*E. globulus* woody vegetation with a strong significant relationship, that is between the calibration georeferenced data and the predicted probability of the presence and absence of *E. globulus* and non-*E. globulus* woody vegetation. The ensemble architecture of the RF was responsible for the greater accuracy of its performance [30]. The work of López-Sánchez et al. [81] also revealed that Random Forest was the best method for modelling the suitable niche of *E. globulus* in northern Spain. Furthermore, a study on the classification of some *Eucalyptus* species in Australian native forests found that RF, SVM, and BRT were the optimal algorithms, scoring higher classification accuracies than conventional classification algorithms [82]. Similarly, a study on natural forest mapping in the Andes [83] revealed that RF and SVM accurately classified mountain forests in the Andes. Ahmed et al. [31] compared the performance of RF, BRT, and SVM with other regression and profile methods for mapping invasive Prosopis in Ethiopia and found that RF and BRT performed better

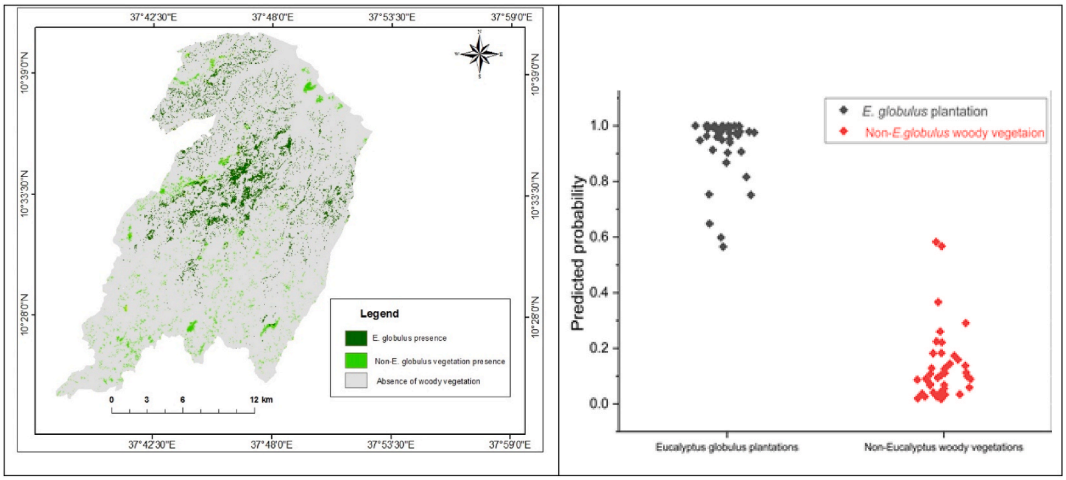


Fig. 8. *E. globulus* and non-*E. globulus* woody vegetation distribution map (left) predicted by RF model and scatterplot of the predicted probability(y-axis) and calibration reference data (x-axis) of *E. globulus* and non-*E. globulus* woody vegetation (right).

than regression and profile methods.

Regarding the relative importance of the predictor variables, the red band and shortwave infrared (SWIR2) bands were the most important predictor variables among the spectral bands of Sentinel-2B imagery (Fig. 3a) The finding of our study agreed with the studies on forest species mapping with the Sentinel-2 data [27,84–86]. They revealed that the red, red-edge, and shortwave infrared bands were the most significant spectral bands for forest species mapping and classification. The work of [87] also revealed that Sentinel-2 SWIR bands (Bands 11 and 12) could differentiate tree species by detecting variability in water content within the tree species.

Besides the spectral bands of Sentinel-2, VIs derived from the spectral bands of Sentinel-2 data were examined independently as predictor variables for modelling *E. globulus*. Among the VIs, our final model demonstrated that the GNDVI emerged as the most significant predictor variable, followed by the NDRE. The NIR and Green bands in both indices were more significant in detecting the spectral response of the E. globules. Similarly, de Oliveira et al. [88] found that the VIs derived from the NIR and Green bands provided greater accuracy in the spectral response of Eucalyptus species. Further, the use of the green band in GNDVI makes it less saturated [89] and more responsive to chlorophyll variability than NDVI [90]. Recently, Taddeo et al. [91] revealed that GNDV outperformed other VIs as predictor variables for wetland plant species.

Further, five potential environmental variables (elevation, distance from the road, LST, soil pH, and slope) were considered to model the spatial distribution of *E. globulus*. Although the values of the kappa coefficient and TSS metrics for the predictor variables were not satisfactory like AUC, the response curve (Fig. 3c) revealed that the probability of the presence of *E. globulus* decreased further away from the road. Since *E. globulus* has become a commercial crop in the area, many *Eucalyptus* plantation sites are closer to road access. Moreover, the response curve showed that the probability of the presence of the species decreased above 3000 m a.s.l. The analysis of the current distribution of *E. globulus* (Fig. 7) revealed that more than 93 % of the species are found within 2600–3500 m above sea level. Altitude most likely worked as a proxy for local edapho-climatic conditions, where low temperatures may define the highest altitudinal limit for *E. gobulus* plantations [92,93]. Steeper slopes, which frequently have a thinner soil layer, could limit species distribution.

Further, the analysis (Fig. 7) showed that more 98 % of the *E. globulus* are found within 900–1100 mm mean annual rainfall. Most studies found that the species can endure rainfall between 900 and 1400 mm, with a dry season lasting up to three months, if sufficient soil moisture is present [94,95]. According to Kirkpatrick [93], excessive or insufficient water may limit the occurrence of *E. globulus*. The finding of Pohjonen and Pukkala [96] complemented with our analysis that the annual rainfall between 1000 and 1200 mm is the ideal rainfall zone for the productivity of *E. globulus* in the highlands of Ethiopia. Furthermore, Alemneh et al. [14] found that the dramatic expansion of *Eucalyptus* plantations, particularly in mid agro ecological zones (2400–3800 m a.s.l) of the upper Blue Nile, was due to climatic conditions that were neither too cold nor affected by frequent drought, limiting species growth. *E. globulus* is very flexible to different climatic conditions, the most suitable climatic condition for the species is a mild and temperate climate [83,84]. Though *Eucalyptus globulus* is very adaptable and has a high level of ecological plasticity [97], the species requires annual precipitation of more than 900 mm and a mean annual temperature of 14 ± 4 °C [92,94,96].

5. Conclusion

Our study demonstrated that the three ML algorithms (RF, SVM, and BRT) can effectively map the spatial distribution of *E. globulus* from Sentinel-2 spectral bands, vegetation indices, and environmental variables. Although the three ML algorithms produced similar accuracy, the highest accuracy was achieved by Random Forest because of its capability to combine several predicting trees and its computational efficiency. The results showed that among the 11 predator variables, GNDVI emerged as the single most significant

predictor variable to explain the current distribution of *E. globulus* in the study area. Furthermore, the environmental variables give confidence to understand the suitable niche of the species and effectively detect the current distribution of the E. globulus, which otherwise could not be. Accordingly, more than 90 % of the current distribution of *E. globulus* is found in the mid-agro-ecological zone, which has mild and temperate climatic characteristics. The study showed that considering other potential predictor variables, such as environmental variables, other than the spectral bands and derived vegetation indices, increases the accuracy of the model performance in predicting *E. globulus* distribution accurately. Furthermore, the validation results attested to the replicability of the model outside of the study area. Thus, this would make it possible to model *E. globulus* spatial distribution on a large scale, contributing to an improvement in the management processes of tree plantations and providing up-to-date information for policymakers and land resource conservation at the local and regional scales.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Abdurohman Yimam: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Asnake Mekuriaw:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Dessie Assefa:** Writing – review & editing, Supervision. **Woldeamlak Bewket:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Acknowledgments

We would like to thank Debre Birhan University and Addis Ababa University for allowing this doctoral study. The full cost of this study was covered by Addis Ababa University Thematic Research Project (Eucalyptus plantations in the Ethiopian highlands: Extent of coverage and its effects on the environment and rural livelihoods).

References

[1] W. Chen, Y. Zou, Y. Dang, T. Sakai, Spatial distribution and dynamic change monitoring of Eucalyptus plantations in China during 1994–2013, Trees Struct. Funct. 36 (1) (2022).
[2] Y. Zhang, X. Wang, Geographical spatial distribution and productivity dynamic change of eucalyptus plantations in China, Sci. Rep. 11 (1) (2021) 19764.
[3] M. Lemenih, H. Kassa, Re-greening Ethiopia: history, challenges and lessons, Forests 5 (7) (2014) 1717–1730.
[4] M. Zewdie, Temporal Changes of Biomass Production, Soil Properties and Ground Flora in Eucalyptus Globulus Plantations in the Central Highlands of Ethiopia, vol. 2008, 2008.
[5] H. Zegeye, Environmental and socio-economic implications of Eucalyptus in Ethiopia, Ethiop Inst Agric Res2010. (2010) 184–205.
[6] A.B. Madalcho, B. Lemma, M.M. Mena, B.B. Badesso, Is the expansion of Eucalyptus tree a curse or an opportunity? Implications from a dispute on the tree's ecological and economic impact in Ethiopia: a review, J. Ecol. Nat. Environ. 11 (2019) 75–83.
[7] T. Chanie, A.S. Collick, E. Adgo, C.J. Lehmann, T.S. Steenhuis, Eco-hydrological impacts of Eucalyptus in the semi humid Ethiopian highlands: the lake tana plain, J. Hydrol. Hydromechanics 61 (1) (2013) 21–29b.
[8] D. Jaleta, B. Mbilinyi, H. Mahoo, M. Lemenih, Eucalyptus expansion as relieving and provocative tree in Ethiopia, Journal of Agriculture and Ecology Research International 6 (3) (2016) 1–12.
[9] S. Fikreyesus, Z. Kebebew, A. Nebiyu, N. Zeleke, S. Bogale, Allelopathic effects of Eucalyptus camaldulensis Dehnh. on germination and growth of tomato. Am-Eurasian, J Agric Environ Sci. 11 (5) (2011) 600–608.
[10] S. Feyera, E. Beck, U. Lüttge, Exotic trees as nurse-trees for the regeneration of natural tropical forests, Trees (Berl.) 16 (2002) 245–249.
[11] R.M. Wise, P.J. Dye, M.B. Gush, A comparison of the biophysical and economic water-use efficiencies of indigenous and introduced forests in South Africa, For Ecol Manage 262 (6) (2011) 906–915.
[12] M. Mengist, G. Georg, M. Sieghardt, Eucalyptus Plantations in the Highlands of Ethiopia Revisited: A Comparison of Soil Nutrient Status after the First Coppicing, Mountain Forestry Master Programme, 2011.
[13] M. Bekele, Y. Tesfaye, Z. Mohammed, S. Zewdie, Y. Tebikew, M. Brockhaus, et al., The Context of REDD+ in Ethiopia: Drivers, Agents and Institutions, vol. 127, CIFOR, 2015.
[14] T. Alemneh, B.F. Zaitchik, B. Simane, A. Ambelu, Changing patterns of tree cover in a tropical highland region and implications for food, energy, and water resources, Front. Environ. Sci. 7 (2019) 1.
[15] R.P. Anderson, E. Martínez-Meyer, M. Nakamura, M.B. Araújo, A.T. Peterson, J. Soberón, et al., Ecological Niches and Geographic Distributions (MPB-49), Princeton University Press, 2011.
[16] A. Guisan, W. Thuiller, N.E. Zimmermann, Habitat Suitability and Distribution Models: with Applications in R, Cambridge University Press, 2017.
[17] A.J. Shirk, S.A. Cushman, K.M. Waring, C.A. Wehenkel, A. Leal-Sáenz, C. Toney, et al., Southwestern white pine (Pinus strobiformis) species distribution models project a large range shift and contraction due to regional climatic changes, For Ecol Manage 411 (2018) 176–186.
[18] J. Castaño-Santamaría, C.A. López-Sánchez, J.R. Obeso, M. Barrio-Anta, Modelling and mapping beech forest distribution and site productivity under different climate change scenarios in the Cantabrian Range (North-western Spain), For Ecol Manage 450 (2019) 117488.
[19] G. Mallinis, N. Koutsias, M. Tsakiri-Strati, M. Karteris, Object-based classification using Quickbird imagery for delineating forest vegetation polygons in a Mediterranean test site, ISPRS J. Photogrammetry Remote Sens. 63 (2) (2008) 237–250.
[20] C. Zhang, F. Qiu, Mapping individual tree species in an urban forest using airborne lidar data and hyperspectral imagery, Photogramm Eng Remote Sensing 78 (10) (2012) 1079–1087.
[21] M. Pal, P.M. Mather, Support vector machines for classification in remote sensing, Int J Remote Sens 26 (5) (2005) 1007–1011.

[22] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, D. Roberts, Mapping land-cover modifications over large areas: a comparison of machine learning algorithms, Remote Sens. Environ. 112 (5) (2008) 2272–2283.
[23] B. Ghimire, J. Rogan, V.R. Galiano, P. Panday, N. Neeti, An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA, GIsci Remote Sens 49 (5) (2012) 623–643.
[24] S. Moradpour, M. Entezari, S. Ayoubi, A. Karimi, S. Naimi, Digital exploration of selected heavy metals using Random Forest and a set of environmental covariates at the watershed scale, J. Hazard Mater. 455 (2023) 131609.
[25] S. Saidi, S. Ayoubi, M. Shirvani, K. Azizi, S. Zhao, Digital mapping of soil phosphorous sorption parameters (PSPs) using environmental variables and machine learning algorithms, Int J Digit Earth 16 (1) (2023) 1752–1769.
[26] M. Zeraatpisheh, S. Ayoubi, Z. Mirbagheri, M.R. Mosaddeghi, M. Xu, Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables, Geoderma Regional 27 (2021) e00440.
[27] Y. Shao, R.S. Lunetta, Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points, ISPRS J. Photogrammetry Remote Sens. 70 (2012) 78–87.
[28] G. Mountrakis, J. Im, C. Ogole, Support vector machines in remote sensing: a review, ISPRS J. Photogrammetry Remote Sens. 66 (3) (2011) 247–259.
[29] J.F. Mas, J.J. Flores, The application of artificial neural networks to the analysis of remotely sensed data, Int J Remote Sens 29 (3) (2008) 617–663.
[30] A.M. Abdi, Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data, GIsci Remote Sens 57 (1) (2020) 1–20.
[31] N. Ahmed, C. Atzberger, W. Zewdie, Species Distribution Modelling performance and its implication for Sentinel-2-based prediction of invasive Prosopis juliflora in lower Awash River basin, Ethiopia, Ecol Process 10 (1) (2021) 1–16.
[32] A.E. Maxwell, T.A. Warner, F. Fang, Implementation of machine-learning classification in remote sensing: an applied review, Int J Remote Sens 39 (9) (2018) 2784–2817.
[33] P.S. Roy, M.D. Behera, S.K. Srivastav, Satellite remote sensing: sensors, applications and techniques, Proc. Natl. Acad. Sci., India, Sect. A 87 (2017) 465–472.
[34] M. Kumar, M.P. Singh, H. Singh, P.M. Dhakate, N.H. Ravindranath, Forest working plan for the sustainable management of forest and biodiversity in India, J. Sustain. For. 39 (1) (2020) 1–22.
[35] E. Grabska, D. Frantz, K. Ostapowicz, Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the Polish Carpathians, Remote Sens. Environ. 251 (2020) 112103.
[36] V.J. Pasquarella, C.E. Holden, C.E. Woodcock, Improved mapping of forest type using spectral-temporal Landsat features, Remote Sens. Environ. 210 (2018) 193–207.
[37] N. Puletti, F. Chianucci, C. Castaldi, Use of Sentinel-2 for forest classification in Mediterranean environments, Ann Silvic Res 42 (1) (2018) 32–38.
[38] M. Szostak, P. Hawryło, D. Piela, Using of Sentinel-2 images for automation of the forest succession detection, Eur J Remote Sens 51 (1) (2018) 142–149.
[39] M. Immitzer, F. Vuolo, C. Atzberger, First experience with Sentinel-2 data for crop and tree species classifications in central Europe, Remote Sens (Basel). 8 (3) (2016) 166.
[40] W.T. Ng, P. Rima, K. Einzmann, M. Immitzer, C. Atzberger, S. Eckert, Assessing the potential of sentinel-2 and pléiades data for the detection of Prosopis and vachellia spp. in Kenya, Remote Sens (Basel). 9 (1) (2017) 74.
[41] V. Deblauwe, V. Droissart, R. Bose, B. Sonké, A. Blach-Overgaard, J. Svenning, et al., Remotely sensed temperature and precipitation data improve species distribution modelling in the tropics, Global Ecol. Biogeogr. 25 (4) (2016) 443–454.
[42] H. Shiferaw, U. Schaffner, W. Bewket, T. Alamirew, G. Zeleke, D. Teketay, et al., Modelling the current fractional cover of an invasive alien plant and drivers of its invasion in a dryland ecosystem, Sci. Rep. 9 (1) (2019) 1576.
[43] M. Amiri, M. Tarkesh, R. Jafari, G. Jetschke, Bioclimatic variables from precipitation and temperature records vs. remote sensing-based bioclimatic variables: which side can perform better in species distribution modeling? Ecol Inform 57 (2020) 101060.
[44] A. Hościło, A. Lewandowska, Mapping forest type and tree species on a regional scale using multi-temporal Sentinel-2 data, Remote Sens (Basel) 11 (8) (2019) 929.
[45] A. Tesfaw, F. Senbeta, D. Alemu, E. Teferi, Value chain analysis of Eucalyptus wood products in the blue nile highlands of northwestern Ethiopia, Sustainability 13 (22) (2021) 12819.
[46] M.T. Taye, V. Ntegeka, N.P. Ogiramoi, P. Willems, Assessment of climate change impact on hydrological extremes in two source regions of the Nile River Basin, Hydrol. Earth Syst. Sci. 15 (1) (2011) 209–222.
[47] B. Simane, B.F. Zaitchik, M. Ozdogan, Agroecosystem analysis of the choke mountain watersheds, Ethiopia, Sustainability 5 (2) (2013) 592–616.
[48] B. Kieffer, N. Arndt, H. Lapierre, F. Bastien, D. Bosch, A. Pecher, et al., Flood and shield basalts from Ethiopia: magmas from the African superswell, J. Petrol. 45 (4) (2004).
[49] L. Jiang, A fast and accurate circle detection algorithm based on random sampling, Future Generat. Comput. Syst. 123 (2021) 245–256.
[50] R.J. Hijmans, Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model, Ecology 93 (3) (2012) 679–688.
[51] R.A. Boria, L.E. Olson, S.M. Goodman, R.P. Anderson, Spatial filtering to reduce sampling bias can improve the performance of ecological niche models, Ecol Modell 275 (2014) 73–77.
[52] Y.J. Kaufman, D. Tanre, Atmospherically resistant vegetation index (ARVI) for EOS-MODIS, IEEE Trans. Geosci. Rem. Sens. 30 (2) (1992) 261–270.
[53] A.A. Gitelson, Y.J. Kaufman, M.N. Merzlyak, Use of a green channel in remote sensing of global vegetation from EOS-MODIS, Remote Sens. Environ. 58 (3) (1996) 289–298.
[54] G.M. Senseman, C.F. Bagley, S.A. Tweddale, Correlation of rangeland cover measures to satellite-imagery-derived vegetation indices, Geocarto Int. 11 (3) (1996) 29–38.
[55] C.J. Tucker, Red and photographic infrared linear combinations for monitoring vegetation, Remote Sens. Environ. 8 (2) (1979) 127–150.
[56] E.M. Barnes, T.R. Clarke, S.E. Richards, P.D. Colaizzi, J. Haberland, M. Kostrzewski, et al., Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data, in: Proceedings of the Fifth International Conference on Precision Agriculture, 2000. Bloomington, MN, USA.
[57] D. Haboudane, J.R. Miller, E. Pattey, P.J. Zarco-Tejada, I.B. Strachan, Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: modeling and validation in the context of precision agriculture, Remote Sens. Environ. 90 (3) (2004) 337–352.
[58] N. Gobron, B. Pinty, M.M. Verstraete, J.L. Widlowski, Advanced vegetation indices optimized for up-coming sensors: design, performance, and applications, IEEE Trans. Geosci. Rem. Sens. 38 (6) (2000) 2489–2505.
[59] P.J. Zarco-Tejada, J.R. Miller, T.L. Noland, G.H. Mohammed, P.H. Sampson, Scaling-up and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data, IEEE Trans. Geosci. Rem. Sens. 39 (7) (2001) 1491–1507.
[60] F. Baret, G. Guyot, Potentials and limits of vegetation indices for LAI and APAR assessment, Remote Sens. Environ. 35 (2–3) (1991) 161–173.
[61] A. Guisan, N.E. Zimmermann, Predictive habitat distribution models in ecology, Ecol Modell 135 (2–3) (2000) 147–186.
[62] C. Zhang, X. Li, L. Chen, G. Xie, C. Liu, S. Pei, Effects of topographical and edaphic factors on tree community structure and diversity of subtropical mountain forests in the Lower Lancang River Basin, Forests 7 (10) (2016) 222.
[63] F.B. Vahdati, S.S. Mehrvarz, D.C. Dey, A. Naqinezhad, Environmental factors–ecological species group relationships in the Surash lowland-mountain forests in northern Iran, Nord. J. Bot. 35 (2) (2017) 240–250.
[64] B. Naimi, M.B. Araújo, sdm: a reproducible and extensible R platform for species distribution modelling, Ecography 39 (4) (2016) 368–375.
[65] C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, et al., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (1) (2013) 27–46.
[66] D.W. Marquaridt, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, Technometrics 12 (3) (1970) 591–612.
[67] S. Chatterjee, A.S. Hadi, Regression Analysis by Example, John Wiley & Sons, 2013.
[68] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[69] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[70] S. Shataee, S. Kalbi, A. Fallah, D. Pelz, Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms, Int J Remote Sens 33 (19) (2012) 6254–6280.
[71] J. Gajardo, M. García, D. Riaño, Applications of airborne laser scanning in forest fuel assessment and fire prevention, in: Forestry Applications of Airborne Laser Scanning: Concepts and Case Studies, Springer, 2013, pp. 439–462.
[72] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann, C. Notarnicola, Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data, Remote Sens (Basel). 7 (12) (2015) 16398–16421.
[73] C. Huang, L.S. Davis, J.R.G. Townshend, An assessment of support vector machines for land cover classification, Int J Remote Sens 23 (4) (2002) 725–749.
[74] M.C. Hansen, B. Reed, A comparison of the IGBP DISCover and University of Maryland 1 km global land cover products, Int J Remote Sens 21 (6–7) (2000) 1365–1373.
[75] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[76] J. Elith, C.H. Graham, Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models, Ecography 32 (1) (2009) 66–77.
[77] A.H. Fielding, J.F. Bell, A review of methods for the assessment of prediction errors in conservation presence/absence models, Environ. Conserv. 24 (1) (1997) 38–49.
[78] M. Iturbide, J. Bedia, S. Herrera, O. del Hierro, M. Pinto, J.M. Gutiérrez, A framework for species distribution modelling with improved pseudo-absence generation, Ecol Modell 312 (2015) 166–174.
[79] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.
[80] C. Liu, P.M. Berry, T.P. Dawson, R.G. Pearson, Selecting thresholds of occurrence in the prediction of species distributions, Ecography 28 (3) (2005) 385–393.
[81] C.A. López-Sánchez, F. Castedo-Dorado, A. Cámara-Obregón, M. Barrio-Anta, Distribution of Eucalyptus globulus Labill. in northern Spain: contemporary cover, suitable habitat and potential expansion under climate change, For Ecol Manage 481 (2021) 118723.
[82] X. Shang, L.A. Chisholm, Classification of Australian native forest species using hyperspectral remote sensing and machine-learning classification algorithms, IEEE J Sel Top Appl Earth Obs Remote Sens 7 (6) (2013) 2481–2489.
[83] L.A. Vega Isuhuaylas, Y. Hirata, L.C. Ventura Santos, Torobeo N. Serrudo, Natural forest mapping in the Andes (Peru): a comparison of the performance of machine-learning algorithms, Remote Sens (Basel). 10 (5) (2018) 782.
[84] C. Bolyn, A. Michez, P. Gaucher, P. Lejeune, S. Bonnet, Forest mapping and species composition using supervised per pixel classification of Sentinel-2 imagery, Biotechnologie, Agronomie, Société et Environnement. 22 (3) (2018).
[85] M. Immitzer, M. Neuwirth, S. Böck, H. Brenner, F. Vuolo, C. Atzberger, Optimal input features for tree species classification in Central Europe based on multi-temporal Sentinel-2 data, Remote Sens (Basel). 11 (22) (2019) 2599.
[86] M. Persson, E. Lindberg, H. Reese, Tree species classification with multi-temporal Sentinel-2 data, Remote Sens (Basel). 10 (11) (2018) 1794.
[87] P. Lukeš, P. Stenberg, M. Rautiainen, M. Mottus, K.M. Vanhatalo, Optical properties of leaves and needles for boreal tree species in Europe, Remote Sensing Letters 4 (7) (2013) 667–676.
[88] B.R. de Oliveira, da Silva Aap, L.P.R. Teodoro, G.B. de Azevedo, GT. de OS. Azevedo, F.H.R. Baio, et al., Eucalyptus growth recognition using machine learning methods and spectral variables, For Ecol Manage 497 (2021) 119496.
[89] A.A. Gitelson, Y.J. Kaufman, M.N. Merzlyak, Use of a green channel in remote sensing of global vegetation from EOS-MODIS, Remote Sens. Environ. 58 (3) (1996) 289–298.
[90] A.A. Gitelson, M.N. Merzlyak, Remote sensing of chlorophyll concentration in higher plant leaves, Adv. Space Res. 22 (5) (1998) 689–692.
[91] S. Taddeo, I. Dronova, K. Harris, The potential of satellite greenness to predict plant diversity among wetland types, ecoregions, and disturbance levels, Ecol. Appl. 29 (7) (2019) e01961.
[92] C. Alegria, N. Roque, T. Albuquerque, S. Gerassis, P. Fernandez, M.M. Ribeiro, Species ecological envelopes under climate change scenarios: a case study for the main two wood-production forest species in Portugal, Forests 11 (8) (2020) 880.
[93] J.B. Kirkpatrick, Natural distribution of Eucalyptus globulus Labill, Aust. Geogr. 13 (1) (1975) 22–35.
[94] D.J. Boland, M.I.H. Brooker, G.M. Chippendale, N. Hall, B.P.M. Hyland, R.D. Johnston, et al., Forest Trees of Australia, CSIRO publishing, 2006.
[95] M.R. Jacobs, Eucalypts for Planting. FAO Forestry Series 11, Food and Agriculture Organization of the United Nations. Forestry Department, Rome, Italy, 1979.
[96] V. Pohjonen, T. Pukkala, Eucalyptus globulus in Ethiopian forestry, For Ecol Manage 36 (1) (1990) 19–31.
[97] D. Whitehead, C.L. Beadle, Physiological regulation of productivity and water use in Eucalyptus: a review, For Ecol Manage 193 (1–2) (2004) 113–140.