DESERT 2022, 27(1): 167-181
DOI: 10.22059/JDESERT.2022.88514

RESEARCH PAPER

Mapping spatial patterns of plant species based on machine-learning and regression models

Hamidreza Keshtkar[1,*], Paria Pourmohammad[1]

[1] Department of Arid and Mountainous Regions Reclamation, Faculty of Natural Resources, University of Tehran, Karaj, Iran

Received: 25 July 2021, Revised: 23 December 2021, Accepted: 6 April 2022

Abstract
Various statistical techniques have been used for species distribution modeling that attempt to predict the occurrence of a given species with respect to environmental conditions. The current study was conducted to compare the performance of three regression-based models (multivariate adaptive regression splines, generalized additive models, and generalized linear models) with three machine-learning algorithms (random forest, artificial neural networks, and generalized boosted models). Also in this study, three sets of explanatory variables (climate-only, topography-only and combined topography-climate) for each species (i.e. *Achillea millefolium, Festuca rupicola,* and *Centaurea jacea)* were quantified and the effect of the interaction of the predictor variables with the modeling approaches on determining the accuracy of the predictions was tested. Model accuracy was evaluated using the area under the curve (AUC) of the receiver operating characteristics and true skill statistics (TSS). It was found that regression-based approaches, especially generalized additive model, performed better than those of machine-learning. The results showed that the topography-climate variables were the most important for mapping potentially suitable habitats of target species. The response curves associated with these variables indicate that there are ecological thresholds for favorable growth of all plant species studied.
Keywords: plant distribution; suitable habitats; explanatory variable; Data Mining.

Introduction

Spatial species distribution and the relationship between species and environmental factors have been studied for several years (Guisan and Zimmerman, 2000; Norberg et al., 2019). Linking environmental variables with the physiological tolerance threshold of species has made it possible to model the effect and consequences of environmental change on species and ecological systems (Naghipour et al., 2021). To implement such schemes, a proper understanding of the relationship between the species and environment is required. This understanding is usually achieved through theoretical and statistical methods which relate the environmental variables to the emergence of species. In addition, the relationships between species and the environment should be related to the structural characteristics of habitats using GIS data.

Ecologists commonly assume that the ranges of current geographic species represent the characteristics of the species' habitats which support or limit their presence in a specific location. Accordingly, a range shift can be justified by measuring changes in the bioclimatic envelope (a set of biological and physical conditions suitable for the development and establishment of a particular species). In fact, species are under the influence of environmental change (Bateman et al., 2013).

* Corresponding author e-mail: Hkeshtkar@ut.ac.ir

Predicting geographical species distributions via statistical models has become essential in several aspects of biogeography, ecology and biology. Species distribution models (SDMs) are among the most appropriate methods to predict the impact of ecological factors on species distribution (Dirnböck et al., 2011; Naimi and Araújo, 2016). Using these models, researchers can predict a probability of existence species in a location where no occurrence data is known. SDMs are valuable tools for the evaluation and protection of regions degrading and losing their biodiversity due to various factors (Kosanic et al., 2018). Robertson (2003) suggested that the prediction provided by each model may present different conceptions of the potential distribution and biology of the target species. Despite that, it is essential to perfectly understand restrictions and ambiguities embedded in species distribution modeling to produce suitable and precise models (Zimmermann et al., 2010; Kumar, 2012; Zomer et al., 2015; Akhter et al., 2017).

The number of predicting techniques used to model the distribution of plant species has increased considerably in recent years (Guisan et al., 2013) and some studies have been compered the performance of models for terrestrial species (e.g., Cianci et al., 2015; Duan et al., 2014). Currently, however, it is not clear how species distribution models (SDMs) vary in their ability to predict the spatial distributions of species (Oppel et al., 2012) and which techniques yield the most reliable predictions for distribution of plant species.

Recently, regression and machine learning techniques have been used more than other methods. Most of the regression models used to predict the geographic species distribution presents the highly interpretable and meaningful results. These models are usually restricted to binary data organizations that have a precise and regular sampling strategy; generalized additive models (GAM) is one such modeling method that has a specifically forceful performance when modeling species presence/absence data (Lehmann et al., 2002). Machine learning techniques include a variety of non-parametric methods able to compute regression or classification tasks using available information. These methods show some benefits with reference to statistical methods: they are capable of handling non-linear relationships among predictors, able to deal with complicated relationships among predictors that can occur in big data sets and capable of managing complicated and noise data (Thuiller et al., 2016).

Selection of environmental factors to apply as predictors is a major challenge in SDMs (Araujo and Guisan, 2006). Selection of predictors with direct effects on species distribution is the best solution to this problem (Austin, 2007). In some areas, it is not possible to include different types of predictor factors because of the limitation in availability of data (Bucklin et al., 2015). Because climate is a factor affecting species distribution, one subclass of SDMs comprises only climate (hereafter climate-only) predictors (Thuiller et al., 2016). Climate-only SDMs are essential to guiding future conservation efforts (Elith and Leathwick, 2009; Xu et al., 2021), although the lack of sufficient information for determination of climate range in species distribution have caused some scientists to criticize climate-only models (Beale et al., 2008). If non-climatic variables are used along with climatic information, this problem could be solved (Austin and Van Niel, 2011). Although studies have incorporated climatic and topographic variables in modeling, few have examined the climate-only and topography-only models versus combined models.

The objective of the current study is to evaluate the performance of a number of presence-absence distribution models using spatial distribution data of three plant species. Specifically, three regression methods and three machine-learning methods were compared. Considering that differences in predictive accuracy between methods depends upon the explanatory variables (Bucklin et al., 2015; Oppel et al., 2012; Rafiee et al., 2020; Rajpoot et al., 2020), three sets of explanatory variables (climate-only, topographic-only and combined topographic-climate) were quantified for each species and the differences among predictor variables interacting with the modeling approaches were tested to determine the accuracy of the predictions. The tolerance

range of the plant species to environmental change was investigated and the limiting factors and ecological drivers of the systems modeled were elaborated. Such knowledge could assist in the selection of predictors for practical SDM applications and provide information on which modeling techniques are the most useful for a group of species.

Materials and Methods

Study areas

The area under study is in Free State of Thuringia and covers 6900 Km^2. The average elevation is 486 m.a.s.l, with minimum and maximum altitudes of 114 m and 982 m , respectively. The climatic condition is of the continental type, with the mean annual precipitation of 604 mm, and the mean annual temperature of 8.6 °C (based on monthly data of 18 meteorological stations from 1960–2010). The soil parent material is mainly calcareous.

Species and data preparation

In this study, three native non-woody species were selected: 1) *Achillea millefolium millefolium (A. millefolium)*, 2) *Festuca rupicola (F. rupicola)*, and 3) *Centaurea jacea (C. jacea). A. millefolium* is an herbaceous and perennial plant in the family Asteraceae. The selected species comprise a balanced composition of occurrence frequencies, so that *F. rupicola*, *A. millefolium*, and *C. jacea* represent very common, relatively frequent, and relatively rare conditions, respectively. A total of 201 plots were available (Fig. 1), and the studied plant species (i.e. *F. rupicola, A. millefolium, and C. jacea*) occurred in 144, 102, and 58 plots, respectively. Field studies were done to record the occurrence points between 2013 and 2014.

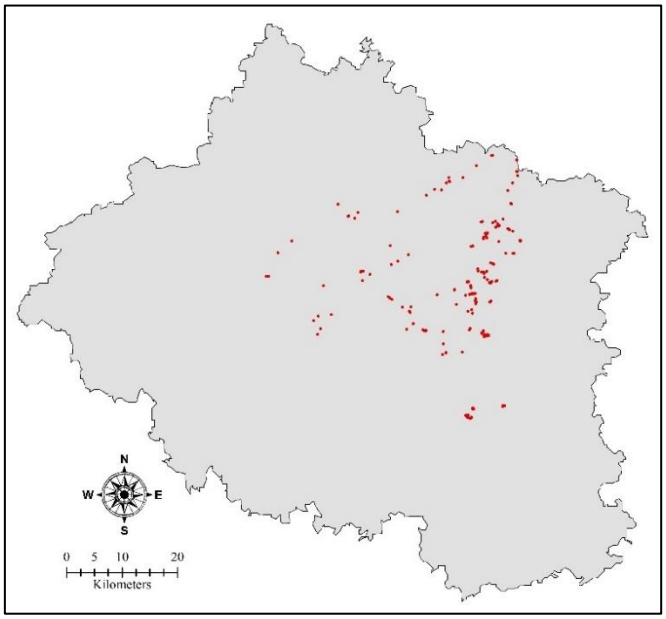

Figure 1. Spatial location of plots in study area

Environmental predictors

The set of quantitative topographic and climatic predictors selected covered the basic physiological requirements (i.e. nutrients, energy and water) of species. In total, 11 predictor variables (six climatic and five topographic) were calculated at 25-m in spatial resolution (Table 1). The environmental variables used are described in Zimmermann and Kienast (1999) and Parviainen et al. (2008) and are only briefly discussed here. Digital elevation model (DEM)

was employed to extract all topographic factors. The Topographic Wetness Index (TWI) can provide suitable information about the local relative differences in moisture conditions (Balazy et al., 2019). The topographic position index (TPI), or difference from mean elevation (DIFF), is a useful measure that increasingly is used to express the exposure of a central point in space as compared to the surrounding terrain (Wilson and Gallant, 2000).

Table 1. List of environmental variables which tested for multi-collinearity

Variables	Unit	Details
Topography		
Slope*	Degrees	Slope inclination
Aspect*	Degrees	The compass direction that a slope face
Elevation*	m	The elevation of a geographic locations
Topographic wetness index*	m	It quantifies the role of topography for redistributing water in the landscape
Topographic position index	Unitless	Identification of topographic features at various spatial scales
Climate		
Mean annual temperature	°C	Average of annual temperature
Mean summer temperature*	°C	Average temperature from April to September
Sum annual precipitation	mm	Sum of annual precipitation
Sum summer precipitation*	mm	Sum of precipitation from April to September
Summer solar radiation*	$kJ \times m^{-2} \times day^{-1}$	Sum of monthly average of daily global solar radiation from April to September
Soil moisture index*	mm	Difference between precipitation and potential evapotranspiration

Temperature and precipitation data (1961-2010) were collected from the German Meteorological Service (https://www.dwd.de). Since topography strongly affects temperature and precipitation (Sartz, 1972), DEM was used as a co-variable to implement co-kriging interpolation method throughout the study area. Soil moisture variable was computed as the monthly difference between precipitation and potential evapotranspiration.

Multi-collinearity analysis

In this study, the Pearson correlation coefficient was run to address the multicollinearity problem between explanatory variables (Shrestha, 2020). Accordingly, variables that showed high correlation with other predictors (>|0.7|) were left out of the model. Finally, 8 environmental predictors were kept for model calibration (marked with '*' in Table 1), out of the original 11 variables.

Calibration of statistical models

Six predictive models (three regression methods and three machine-learning methods) were run and compared to predict species distributions, which are known to provide good predictions: (1) generalized linear models (GLMs; McCullagh and Nelder, 1989); (2) generalized additive model (GAM; Hastie and Tibshirani, 1990); (3) multivariate adaptive regression splines (MARS; Friedman, 1991); (4) generalized boosted models (GBMs, also known as boosted regression trees (BRT); Ridgeway, 1999); (5) random forest (RF; Breiman, 2001); (6) artificial neural networks (ANNs; Ripley, 1996). By relating the independent and dependent variables,

all the models mentioned here can specify at what probability percentage a pixel will be hosting the target species.

GLMs and GAMs were fitted for each species with a binomial variance and a logit transformation. In both models, the selection of significant variables was done with an Akaike information criterion-based stepwise method (Akaike, 1998) in forward and backward directions. To calibration the GBMs a maximum number of 2500 trees and internal 3-fold cross-validation procedure was used. MARS models were calibrated using a maximum interaction degree equal to 2. For RF model, we set 500 for the number of trees to grow (ntree), and for the number of input variables (mtry) we used the default value, which is the square root of variables' number. For ANN model, the model optimized the number hidden layer (size) and the weight decay (decay) factor by cross validation based on area under the curve (AUC) of the receiver operating characteristic (ROC). To optimizing, the model test different values for "size" and "decay", respectively, (2, 4, 6, 8) and (0.001, 0.01, 0.05, 0.1), and the one given the best AUC will be selected. All models were run in R (3.6) software using the biomod2 package (Thuiller et al., 2013). For each species, above models were fitted using three different sets of explanatory variables: (1) Topographic variable only (hereafter abbreviated ''Topo''; Table 1); (2) climate only (abbreviated ''Clim''; Table 1); and (3) Topo + Clim variables (abbreviated ''ALL'').

Assessment of model performance

Since there was no independent data to assess the predictive ability of the model, repeated data-splitting procedure was used. We used the formula presented by Huberty (1994) to determine the optimal ratio of training and testing data. This formula is limited to presence/absence models, and the ratio of required testing data is based on $[1 + (p - 1)^{½}]^{-1}$, where p is the number of predictor variables. Accordingly, training of the model was performed using 70% of random data samples (presences and pseudo-absences data) and remaining 30% was used to validation the model through AUC and TSS indices (Allouche et al., 2006).

The TSS evaluation value varies from 0 to 1, where a value of 0 can be interpreted as random predictions and value 1.0 indicates a perfect agreement (Franklin, 2009). The AUC value varies from 0 to 1, where a value below than 0.5 interprets that predictions are no better than random, values of 0.5–0.7 indicate low predictions, 0.7-0.9 indicate useful predictions, and >0.9 indicate excellent predictions (Franklin, 2009; Eskildsen et al., 2013). This index is calculated as specificity (proportion of correctly predicted presences) + sensitivity (proportion of correctly predicted absences)-1 (Franklin, 2009). The resampling technique (data-splitting) was repeated 25 times for the models and the evaluation metrics averaged. We weighted the presence and pseudo-absence data in the modeling procedure so that both gave prevalence of 0.5. This equal prevalence prevents the model bias towards over-prediction of either presences or pseudo-absences data (Isabelle et al., 2014).

For the final calibration of models, all of the data were used to the implementation of spatial projections. The predictive maps were developed for the target species after calibration of the models. Although continuous predictions need to be converted to a binary map (i.e. a species is either predicted present or absent), we used threshold classification according to the minimization of the absolute difference between sensitivity and specificity (Liu et al., 2005). To test whether the probability of the occurrence values for each species were predicted by the predictive models and whether the three sets of explanatory variables differed from each other, we used the non-parametric Wilcoxon's signed-rank test (Phillips et al., 2009).

Results

Multi-collinearity among variables

All of the variables that were used for model calibration had correlation values <0.7. Table 1 shows variables that were kept for modeling. All in all, one topography (Topographic position index) and two climate variables (Mean annual temperature and sum annual precipitation) were deleted from the study. Topography and climate variables showed almost no correlation with each other.

Efficiency of distribution models

The best model was selected based on the AUC and TSS measures. The mean AUC of the six models ranged from 0.64 (ANN) to 0.94 (GAM and MARS) and for the TSS index from 0.42 (ANN) to 0.80 (GAM). All the three species were accurately classified by all the models except the ANN model, which was found to classify inadequately (Fig. 2). Table 2 presents the results of the different statistical techniques constructed with subsets of environmental variables which were applied to the *F. rupicola, A. millefolium* and *C. jacea* data sets. The Wilcoxon signed-rank test showed that there were no statistical differences between the performance of the regression models (GLM, MARS and GAM; $p > 0.05$; Table 3).

Table 2. Mean evaluation values of TSS (true skill statistics) and AUC (the area under the receiver-operated characteristic curve) of six modeling techniques for predicting the distribution of three plant species based on three set of explanatory variables. TOPO= topography-only variables, CLIM= climate-only variables, All= topo-climate variables. See Fig. 2 for the technique's abbreviations

Models	TOPO		CLIM		ALL	
	TSS	AUC	TSS	AUC	TSS	AUC
C. jacea						
RF	0.62	0.78	0.61	0.76	0.66	0.84
ANN	0.42	0.65	0.47	0.67	0.55	0.73
GBM	0.61	0.74	0.68	0.84	0.67	0.85
GAM	0.65	0.78	0.70	0.87	0.71	0.90
GLM	0.67	0.81	0.71	0.84	0.73	0.92
MARS	0.63	0.76	0.67	0.79	0.70	0.87
A. millefolium						
RF	0.63	0.77	0.65	0.80	0.69	0.84
ANN	0.57	0.69	0.57	0.72	0.61	0.74
GBM	0.62	0.75	0.68	0.80	0.72	0.89
GAM	0.71	0.78	0.73	0.85	0.74	0.93
GLM	0.66	0.76	0.71	0.84	0.69	0.83
MARS	0.68	0.84	0.69	0.79	0.71	0.85
F. rupicola						
RF	0.65	0.79	0.70	0.81	0.71	0.83
ANN	0.49	0.64	0.53	0.68	0.56	0.73
GBM	0.68	0.82	0.69	0.87	0.72	0.86
GAM	0.72	0.87	0.75	0.93	0.80	0.94
GLM	0.73	0.83	0.76	0.89	0.75	0.89
MARS	0.71	0.89	0.76	0.92	0.79	0.94

With the *F. rupicola* data set, the best performance was achieved by the MARS (AUC = 0.92; TSS = 0.76), although the GAM results were very similar. For *C. jacea*, the best performance was obtained by the GLM (AUC = 0.92; TSS = 0.73). The best projection was carried out by

the GAM (AUC = 0.93; TSS = 0.74) for *A. millefolium.* The results illustrate that the GAM method presented significantly more accurate predictions than all the machine-learning algorithms (Wilcoxon signed rank test; $p < 0.05$; Table 3 and Fig. 2).

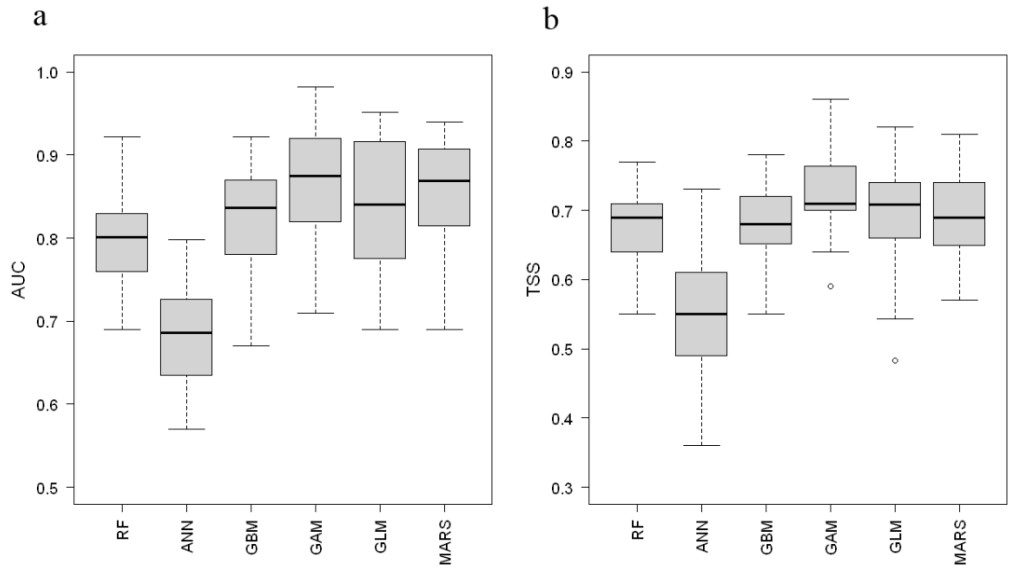

Figure 2. Comparison of AUC (a) and TSS (b) model evaluations between modeling techniques based on "All" explanatory variables. Each box-plot is built from three values (i.e. plant species). The boxes extend from the data's 1st to 3rd quartiles, box boundaries show the interquartile range and the horizontal bars in the box represent the median. RF=random forest, ANN=artificial neural networks, GBM=boosted regression trees, GAM=generalized additive models, GLM=generalized linear models, MARS=multivariate adaptive regression splines

Comparison of models fitted with different explanatory

All the models exhibited very good correctness in all three sets of the environmental variables (Table 2). Modeling with climate-only variables was significantly better than modeling with environment-only variables (Wilcoxon signed rank test; $P < 0.001$). Also, modeling using a set of climatic and environmental factors had a superior predictive ability than two other variable sets (Wilcoxon signed rank test from climate-only models; $P < 0.01$). According to these results, all further analyses examine only the models calibrated with both the climatic and topographic parameters (ALL).

The most important predictor for modeling the distribution of *F. rupicola* was the mean summer temperature (Fig. 3), followed by the sum of the summer precipitation, slope and DEM. The results show that the mean summer temperature followed by the sum summer precipitation and slope are the most important predictors for modeling the distribution of *C. jacea* and *A. millefolium.* The soil moisture index (SMI) was the least important predictor for these two plant species.

The response curves associated with these variables indicate that there may be ecological thresholds for the favorable growth of all plant species studied (Fig. 4). For example, the response curves for the best model (i.e. GAM) for *F. rupicola* indicated that the optimal value for the sum summer precipitation was 300-400 mm (Fig. 4-k). For the mean summer temperature, habitat suitability was low until the range increased to about 10°C (Fig. 4-l). It steadily increased to about 15°C, showing *F. rupicola* to have a stronger relationship with a higher summer temperature. The suitability of the habitat improved as the elevation decreased (<430 m) and the slope increased (>5°) (Fig. 4-j and i).

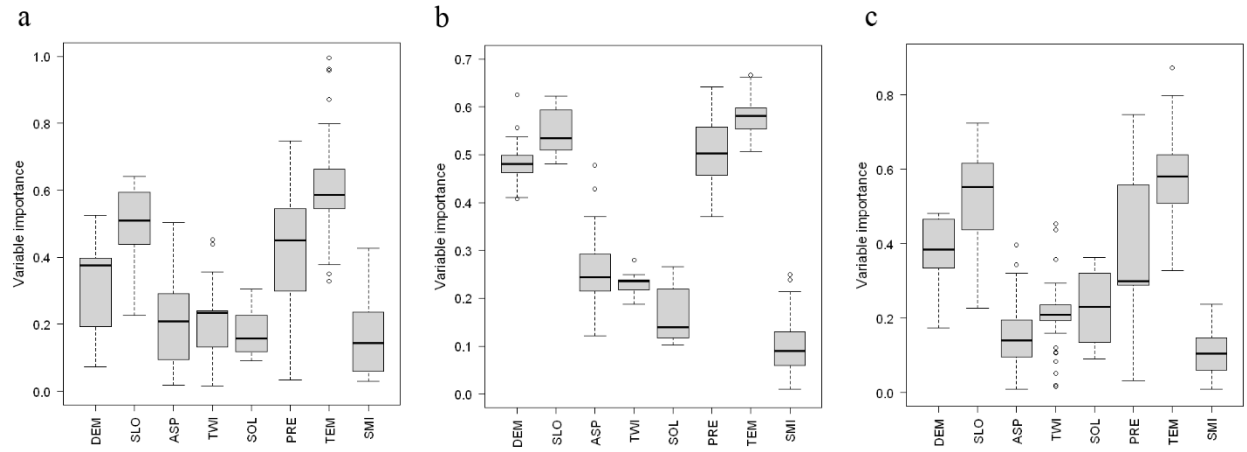

Figure 3. Importance of each predictor used in calibrated models for three species; F. rupicola (a), C. jacea (b) and A. millefolium (c). A high value (like Slope) represents a significant effect of the predictor in the model. DEM= digital elevation model; SLO= slope in degrees; ASP= aspect in degrees; TWI= topographic wetness index; SOL= sum of solar radiation for the growing season (April–September); PRE= sum precipitation over the growing season; TEM= mean temperature for the growing season; SMI= soil moisture index

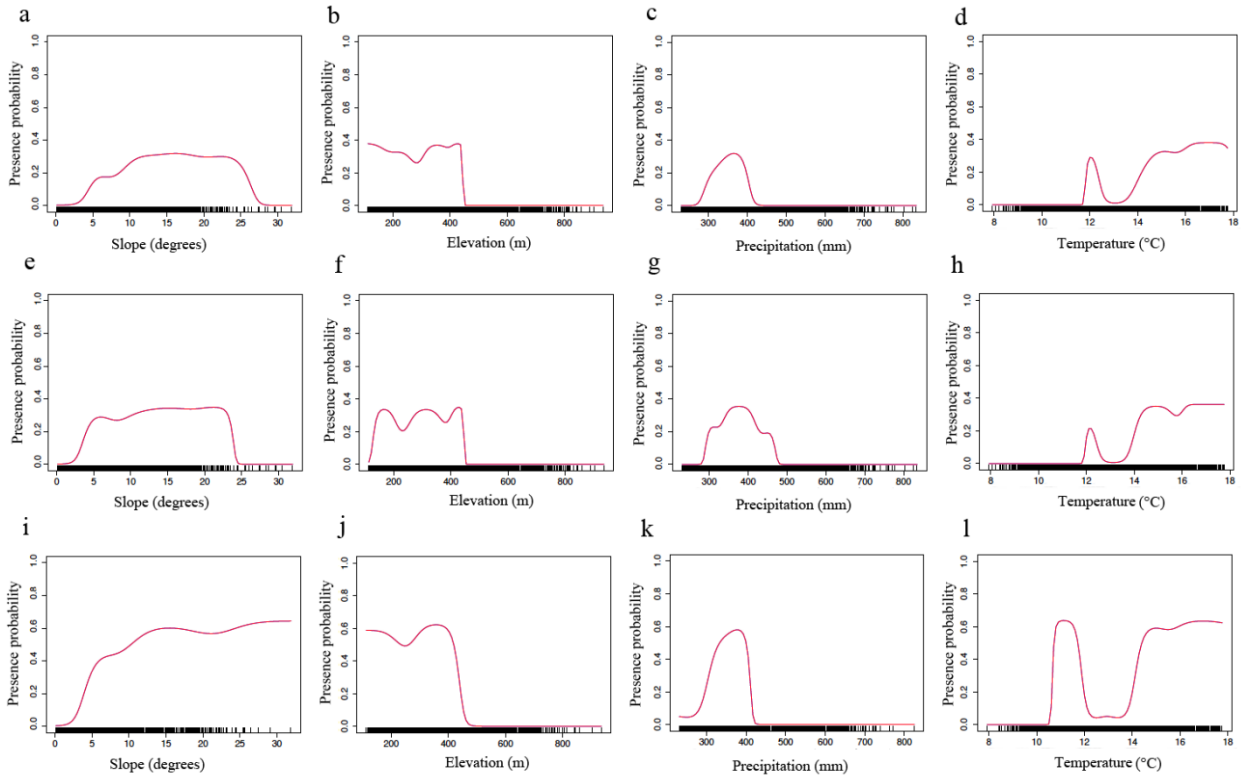

Figure 4. Response curves for A. millefolium (a-d), C. jacea (e-h) and F. rupicola (i-l) based on GAM for the four most important predictors. Frequency distribution of each predictor in study area is shown by black bands on x-axis

Assessment of models among species

An investigation of the distribution projections showed that all species obtained high assessment scores (except the ANN model for *F. rupicola* and *C. jacea*), with TSS values of 0.61 to 0.80 (Table 2). Such a TSS value means that, on average, with a probability of 80% to

90%, the model was properly able to predict the presence and absence of the species. *F. rupicola* was the species that obtained the highest assessment score (TSS-mean=0.72). *A. millefolium* obtained an average score (TSS-mean= 0.69) and *C. jacea* reached a slightly lowe r assessment score than the other species (TSS-mean=0.67).

Table 3. Statistical differences in the predictive performance of six different models for three plant species. Statistical tests of the differences among the predictive accuracies of different methods based on AUC scores were tested by Wilcoxon signed rank test (*P*-values)

Models	RF	ANN	GBM	GAM	GLM	MARS
RF	-	< 0.001	0.008	< 0.001	0.001	< 0.001
ANN		-	< 0.001	< 0.001	< 0.001	< 0.001
GBM			-	< 0.001	0.011	0.007
GAM				-	0.016	0.091
GLM					-	0.253

Discussion and Conclusion

This work presented an experimental study comparing the use of six statistical models (RF, ANN, GBM, GLM, GAM, and MARS) to predicting the spatial location of three individual plant species at a 25 m spatial resolution and investigated the impact that different sets of explanatory variables (climate-only, topography-only and topo-climatic) on model performance.

It was found in this study that the modeling method can determine spatial location of studied plant species for the purpose of conservation. Comparing the predictive power of SDMs, it was found that overall, the GAM model showed the best results (Fig. 2). This is consistent with the results of experiments performed by Leathwick et al. (2005) and Heikkinen et al. (2012) in the field of species distribution modeling. The performance of GBM and RF was acceptable but poorer than that of regression approaches when applied to the prediction of suitable habitats. The ANN modeling technique received lower evaluation scores. The MARS and GLM models, similar in performance to the GAM model, can then be considered as substitute mapping methods (Fig. 2 and Table 2). The findings of this study are in line with those obtained by Leathwick et al. (2005).

Additionally, previous studies revealed that MARS is comparable to other regression techniques (i.e. GAM and GAM) in terms of function and capability (Guisan et al., 2007; Ibáñez et al., 2014). Despite relatively similar predictive accuracies in models, the quality of the predicted distributions can vary owing to different theory and assumptions behind these models (Guisan et al., 2013; Oppel et al., 2012). In the current study, for example, the MARS and the GAM model had similar predictive efficiency (Table 2), but varying patterns could be predicted for the spatial positions of plant species (Figs. 5, 6, and 7).

The predictor factors used in this study were selected to cover a wide range of the possible ecological parameters on the distributions of the modeled species. Climate variables, particularly temperature and precipitation in the growing seasons, and topographic factors such as elevation and slope, appeared as significant determinants across all modeling techniques (Fig. 4). These variables demonstrate initial environmental factors related to the physiological requirements of plants (Pearson et al., 2002; Al-Qaddi et al., 2016). The environmental layers used in the current study was based on literature and ecological expertise. Since the same datasets were used in the induction of all predictor models, this attribute does not invalid the comparisons performed.

Results of the current study show that the ALL scenarios (topo-climate variables) are the most important variables for predicting potential suitable habitats of target species. Conversely, models based only on TOPO predictors showed lower evaluation scores. Likewise,

simultaneous incorporation of topographic and climatic variables increased the models' prediction power significantly (Wilcoxon signed rank test comparing ALL and CLIM models; $P < 0.01$). For example, under the ALL scenario, mean TSS for *F. rupicola* showed 2.7% and 6% higher performance than CLIM and TOPO scenarios, respectively. Some earlier studies confirmed that topo-climate explanatory variables (ALL scenario) strongly predict habitat distribution of species (Engler et al., 2009; Kissling et al., 2010).

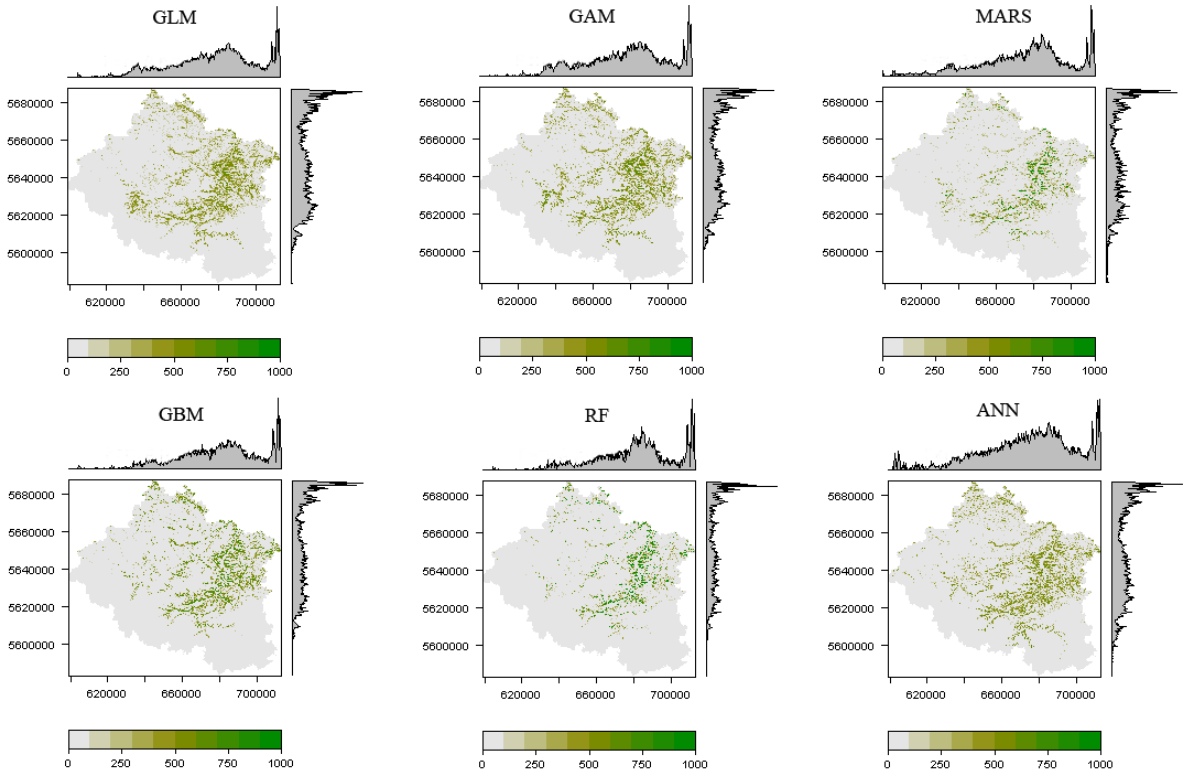

Figure 5. Predicted environmental suitability maps for F. rupicola using six predictor algorithms. Axes represent the geographical location of the region. The map legend shows the probability of the species present per pixel (0-1000). See Fig. 2 for the algorithm's abbreviations

Predictive species modeling can provide good information about the habitat of species and their method of interacting with their environment. Response curves are one of the important components in species distribution studies, because they show the tolerance range for environmental changes by species. Using these relationships between factors, we can learn about the ecological drivers of the systems that we are modeling (Holcombe et al., 2010). The main factor in the models was temperature with a superior fitness for mean summer temperatures >14°C for all plant species, almost related to elevations above 430 m a.s.l. This result could not translate into a preference for lowland regions, because higher elevations are covered by forestland in the study area while target plant species can occur only in open lands. The predictive models indicated the most probable presence of plants is in the areas having an average rainfall of 350 mm and suggest that these plants tend to be present in semi-arid regions. This feature will help them to resist the reduction of rainfall caused by climate change. Keshtkar and Voigt (2016) showed that the three species (i.e. *Achillea millefolium*, *Festuca rupicola*, and *Centaurea jacea*) under the influence of the worst climate change scenario (i.e. RCP8.5) not only will not lose their land, but actually will increase their ranges. In addition, these plants are less sensitive to climate changes since because they still have the chance opportunity to migrate

to higher elevations and are thus less likely to face extinction. The absence of these plants in flat areas can be due to the fact that they have not had the opportunity to establish in these areas because they mainly are urban and agricultural land. In the event of a change in local and regional policies, and as the reduction of agricultural land and conversion to grassland, we will likely see the presence of these plants in such areas.

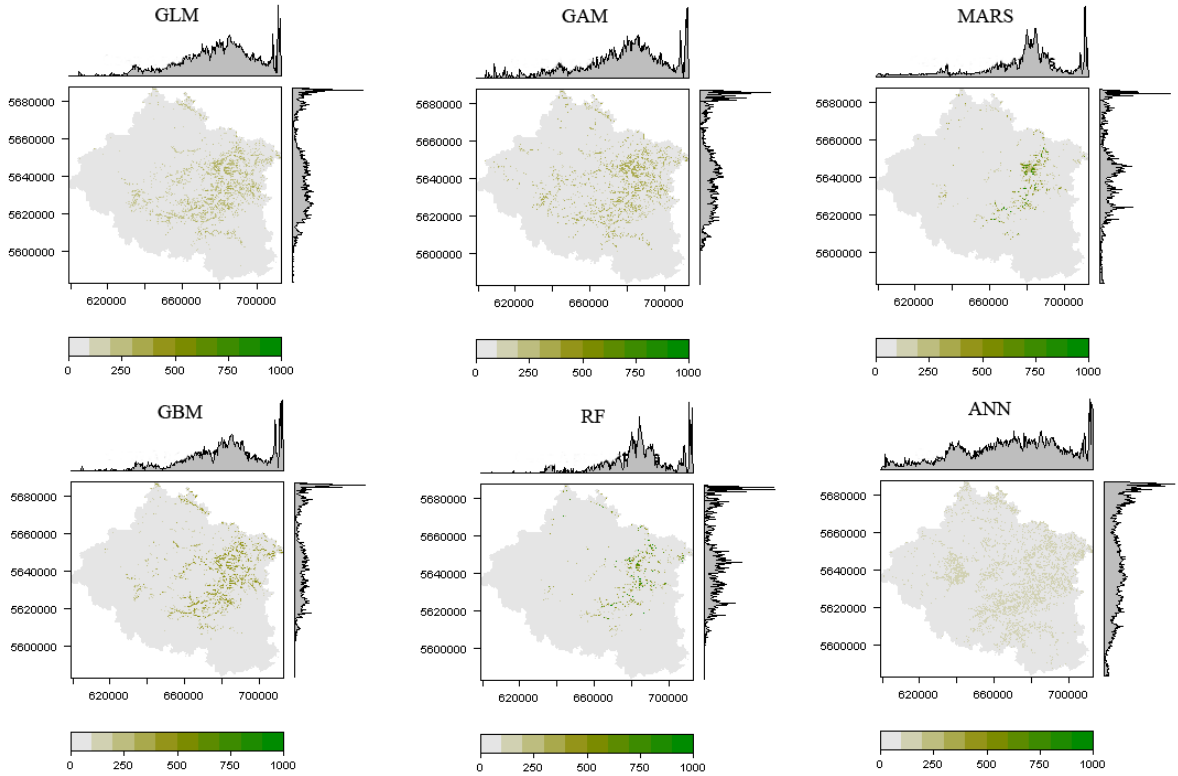

Figure 6. Predicted environmental suitability maps for C. jacea using six predictor algorithms. See Fig. 5 for more details

Although most models used in this study represented the predictions well, there are some uncertainties in predictions of plant species. First, all the models are sensitive to the qualities and quantities of the predictor and response variables. Although most of the factors used in this study demonstrated good predictive abilities for the projections of the species suitable habitats, these species were likely to be affected by other factors, the impacts of which were neglected in this study. However, to reduce uncertainties in this research, a combination of environmental and climate variables was utilized to display a better performance compared with those using only climate variables (Barbet-Massin et al., 2012).

Second, the validity of adequate information on species used to run niche models is contingent upon the potential biases in the availability of adequate information on the presence or absence of species. Several studies have shown that if absence data is collected along with presence data, niche models would be strengthened and the results could be closer to reality (Wiens et al., 2009). In the current study, the true absence data of the species was not available. Collecting species data during the growing season and performing replications within several successive years, like what was done in this study, can significantly enhance the quality of the observed data (Wiens et al., 2009).

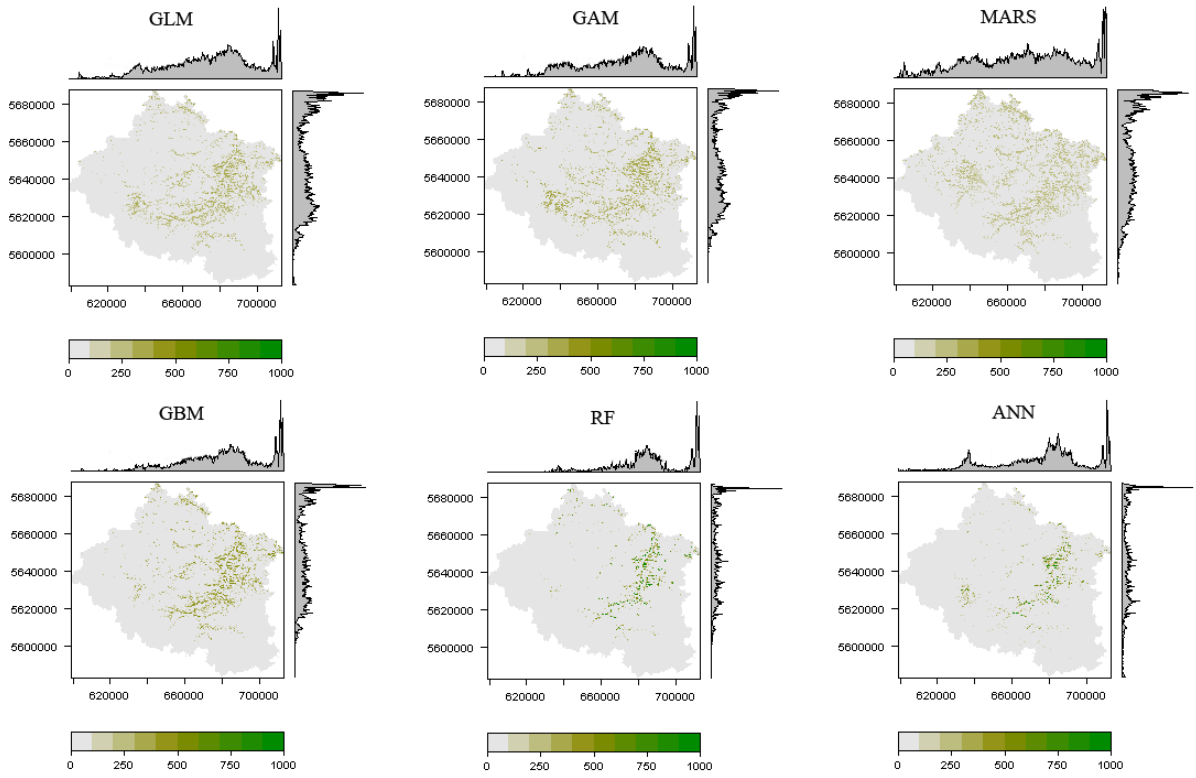

Figure 7. Predicted environmental suitability maps for A. millefolium using six predictor algorithms. See Fig. 5 for more details

All in all, the accuracy of the results obtained from the GAM model show that, at least at a regional level, informative suitable habitat maps for species can be produced. These can provide key information about the environmental tolerance of the studied plants that can be used to protect susceptible habitats, such as the semi-natural grasslands in Germany, from future invasion and the effect of climate change.

References

Akaike H. 1998. A New Look at the Statistical Model Identification. In E. Parzen, K. Tanabe, and G. Kitagawa (Eds.), Selected Papers of Hirotugu Akaike (pp. 215-222): Springer New York.

Akhter S, Mcdonald MA, van Breugel P, Sohel S, Kjær ED, Mariott R. 2017. Habitat distribution modelling to identify areas of high conservation value under climate change for Mangifera sylvatica Roxb. of Bangladesh. Land Use policy, 60, 223–232.

Al-Qaddi N, Vessella F, Stephan J, Al-Eisawi D, Schirone B. 2016. Current and future suitability areas of kermes oak (Quercus coccifera L.) in the Levant under climate change. Regional Environmental Change, 17, 143-156.

Allouche O, Tsoar A, Kadmon, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology, 43, 1223-1232.

Araujo MB, Guisan A. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography, 33, 1677-1688.

Austin MP. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches, Ecological Modelling, 200, 1–19.

Austin MP, Van Niel KP. 2011. Improving species distribution models for climate change studies: variable selection and scale. J Biogeogr, 38, 1–8.

Balazy R, Kamińska A, Ciesielski M, Socha J, Pierzchalski M. 2019. Modeling the Effect of Environmental and Topographic Variables Affecting the Height Increment of Norway Spruce Stands in Mountainous Conditions with the Use of LiDAR Data. Remote Sensing, 11 (20), 2407.

Barbet-Massin M, Thuiller W, Jiguet F. 2012. The fate of European breeding birds under climate, land-

use and dispersal scenarios. Global Change Biology, 18, 881-890.
Bateman BL, Murphy HT, Reside AE, Mokany K, VanDerWal J. 2013. Appropriateness of full-, partial- and no-dispersal scenarios in climate change impact modelling. Diversity and Distributions, 19, 1224-1234.
Breiman L. 2001. Random Forests. Machine Learning, 45, 5-32.
Bucklin DN, Basille M, Benscoter AM, Brandt LA, Mazzotti FJ, Romañach SS, Speroterra C, Watling J.I. 2015. Comparing species distribution models constructed with different subsets of environmental predictors. Diversity and Distributions, 21, 23-35.
Cianci D, Hartemink N, Ibanez-Justicia A. 2015. Modelling the potential spatial distribution of mosquito species using three different techniques. International Journal of Health Geographics, 14, 10-20.
Dirnböck T, Essl F, Rabitsch W. 2011. Disproportional risk for habitat loss of high-altitude endemic species under climate change. Global Change Biology, 17, 990-996.
Duan RY, Kong XQ, Huang MY, Fan WY, Wang ZG. 2014. The predictive performance and stability of six species distribution models. PLoSOne, 9, e112764.
Elith J, Leathwick J. 2009. species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology Evolution and Systematics, 40(1), 677-697.
Engler R, Randin CF, Vittoz P, Czáka T, Beniston M, Zimmermann NE, Guisan A. 2009. Predicting future distributions of mountain plants under climate change: does dispersal capacity matter? Ecography, 32, 34-45.
Eskildsen A, Roux PC, Heikkinen RK, Høye TT, Kissling WD, Pöyry J, Wisz MS, Luoto M. 2013. Testing species distribution models across space and time: high latitude butterflies and recent warming. Global Ecology and Biogeography, 22(12), 1293–1303.
Friedman J. 1991. Multivariate adaptive regression splines. Annals of Statistics, 19, 1–141.
Franklin J. 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge: Cambridge University Press.
Guisan A, Zimmermann NE, Elith J, Graham CH, Phillips S, Peterson AT. 2007. What Matters for Predicting the Occurrences of Trees: Techniques, Data, or Species' Characteristics? Ecological Monographs, 77, 615-630.
Guisan A. 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435.
Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29-36.
Hastie TJ, Tibshirani RJ. 1990. Generalized Additive Models. New York: Chapman and Hall.
Heikkinen RK, Marmion M, Luoto M. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography, 35, 276-288.
Holcombe TR, Stohlgren TJ, Jarnevich CS. 2010. From points to forecasts: predicting invasive species habitat suitability in the near term. Diversity, 2(5), 738–767.
Huberty CJ. 1994. Applied Discriminant Analysis. New York: Wiley.
Ibáñez I, Katz DS, Peltier D, Wolf SM, Connor Barrie BT. 2014. Assessing the integrated effects of landscape fragmentation on plants and plant communities: the challenge of multiprocess–multiresponse dynamics. Journal of Ecology, 102(4), 882-895.
Isabelle B, Damien G, Wilfried T. 2014. FATE-HD: a spatially and temporally explicit integrated model for predicting vegetation structure and diversity at regional scale. Global Change Biology, 20, 2368-2378.
Keshtkar H, Voigt W. 2016. Potential impacts of climate and landscape fragmentation changes on plant distributions: coupling multi-temporal satellite imagery with GIS-based cellular automata model. Ecological Informatics, 32, 145–155.
Kissling WD, Field R, Korntheuer H, Heyder U, Böhning-Gaese K. 2010. Woody plants and the prediction of climate-change impacts on bird diversity. Philosophical Transactions of the Royal Society B: Biological Sciences, 365, 2035-2045.
Kosanic A, Anderson K, Harrison S, Turkington T, Bennie J. 2018. Changes in the geographical distribution of plant species and climatic variables on the West Cornwall peninsula (South West UK). PloS One, 13(2), e0191021.
Kumar P. 2012. Assessment of impact of climate change on Rhododendrons in Sikkim Himalayas using Maxent modelling: limitations and challenges. Biodiversity and Conservation, 21, 1251-1266.
Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T. 2005. Using multivariate adaptive regression

splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biology, 50, 2034-2052.
Lehmann A, Overton J McC, Austin MP. 2002. Regression models for spatial prediction: their role for biodiversity and conservation. Biodiversity and Conservation, 11, 2085–2092.
Liu C, Berry PM, Dawson TP, Pearson RG. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography, 28, 385-393.
McCullagh P, Nelder JA. 1989. Generalized Linear Models, Chapman and Hall, London.
Naghipour AA, Teimoori Asl S, Ashrafzadeh MR, Haidarian M. 2021. Predicting the Potential Distribution of Crataegus azarolus L. under Climate Change in Central Zagros, Iran. Journal of Wildlife and Biodiversity, 5(4), 28-43.
Naimi B, Araújo MB. 2016. sdm: A reproducible and extensible R platform for species distribution modelling. Ecography, 39(4), 368–375.
Norberg A. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs, 89, 1–24.
Oppel S, Meirinho A, Ramírez I, Gardner B, O'Connell AF, Miller PI, Louzao M. 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation, 156, 94-104.
Parviainen M, Luoto M, Ryttäri T, Heikkinen RK. 2008. Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. Journal of Biogeography, 35, 1888-1905.
Pearson RG, Dawson TP, Berry PM, Harrison PA. 2002. SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Species. Ecological Modelling, 154, 289-300.
Phillips SJ, Dudik M, Elith J, Graham C, Lehmann A, Leathwick J. 2009. Sample selection bias and presence-only models of species distributions. Ecological Applications, 19, 181–197.
R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
Rafiee G, Jafari R, Matinkhah SH, Tarkesh isfahani M, karimzadeh HR, jafari Z. 2020. Predicting the Potential Habitat Distribution of Crataegus Pontica C. Koch, Using a Combined Modeling Approach in Lorestan Province. Ijae, 9(2), 45-59.
Rajpoot R, Adhikari D, Verma S, Saikia P, Kumar A, Grant KR. 2020. Climate models predict a divergent future for the medicinal tree Boswellia serrata Roxb. In India. Global Ecology and Conservation, 23, e01040.
Ridgeway G. 1999. "The state of boosting," Computing Science and Statistics. 31, 172-181.
Ripley BD. 1996. Pattern recognition and neural networks. Cambridge: Cambridge University Press.
Robertson MP, Peter CI, Villet MH, Ripley BS. 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. Ecological Modelling, 164, 153-167.
Sartz Richard S. 1972. Effect of topography on microclimate in southwestern Wisconsin. Research Paper NC-74. St. Paul, MN: U.S. Dept. of Agriculture, Forest Service, North Central Forest Experiment Station.
Shrestha N. 2020. Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8, 39-42.
Thuiller W, Georges D, Engler R, Araujo MB. 2013. biomod2: Ensemble platform for species distribution modelling. Ecography, 32, 369-373.
Thuiller W, Georges D, Engler R, Breiner F, Georges MD, Thuiller CW. 2016. Package 'biomod2'. https://cran.r-project.org/package= biomod2.
Xu Y, Huang Y, Zhao H, Yang M, Zhuang Y, Ye X. 2021. Modelling the Effects of Climate Change on the Distribution of Endangered Cypripedium japonicum in China. Forests. 12(4), 429.
Wiens JA, Stralberg D, Jongsomjit D, Howell CA, Snyder MA. 2009. Niches, models, and climate change: Assessing the assumptions and uncertainties. Proceedings of the National Academy of Sciences of the United States of America, 106, 19729-19736.
Wilson JP, Gallant JC. 2000. Terrain Analysis: Principles and Applications. Wiley.
Zimmermann NE, Edwards TC, Graham CH, Pearman PB, Svenning JC. 2010. New trends in species distribution modelling. Ecography, 33, 985-989.
Zimmermann NE, Kienast F. 1999. Predictive mapping of alpine grasslands in Switzerland: Species

versus community approach. Journal of Vegetation Science, 10, 469-482.
Zomer RJ, Xu J, Wang M, Trabucco A, Li Z. 2015. Projected impact of climate change on the effectiveness of the existing protected area network for biodiversity conservation within Yunnan Province, China. Biological Conservation, 184, 335-345.