This is a “preproof” accepted article for Weed Science. This version may be subject to change in the production process, *and does not include access to supplementary material*.
DOI: 10.1017/wsc.2025.5

Short title: Modeling Weed Habitat in Rapeseed Fields

Habitat Suitability Modeling of Dominant Weed in Rapeseed (*Brassica napus*) Fields Using Machine Learning Techniques

Emran Dastres[1], Ghazal Shafiee Sarvestani[2], Mohsen Edalat[3], and Hamid Reza Pourghasemi[4]

[1]PhD, Department of Plant Production and Genetics, School of Agriculture, Shiraz County, Fars Province, Iran
[2]PhD, Department of Plant Production and Genetics, School of Agriculture, Shiraz County, Fars Province, Iran
[3]Associate Professor, (ORCID 0000-0002-9601-0769), Department of Plant Production and Genetics, School of Agriculture, Shiraz County, Fars Province, Iran
[4]Professor, Department of Soil Science, School of Agriculture, Shiraz County, Fars Province, Iran

Author for correspondence: Mohsen Edalat: Email: edalat@shirazu.ac.ir

Abstract

Weed infestations have been identified as a major cause of yield reductions in rapeseed (*Brassica napus* L.), a vital oil crop that has gained significant prominence in Iran, especially within Fars Province. Weed management using machine learning algorithms has become a crucial approach within the framework of precision agriculture for enhancing the efficacy and efficiency of weed control strategies. The evolution of habitat suitability models for weeds represents a significant advancement in agricultural technology, offering the capability to predict weed occurrence and proliferation accurately and reliably. This study focuses on the issue of dominant weed infestation in rapeseed cultivation, particularly emphasizing the prevalence and impact of wild oat (*Avena fatua* L.) as the dominant weed species in rapeseed farming in 2023. We collected data on 12 environmental variables related to topography, climate, and soil properties to develop habitat suitability models. Three "machine learning techniques", including "random forest (RF)", "support vector machine (SVM)", and "boosted regression tree (BRT)", were estimated based on the “receiver operating characteristic (ROC) and area under the curve (AUC)” to model the distribution of *A. fatua*. Model performance was quantified using the “ROC curve and AUC” metrics to identify the best predictive algorithm. The findings indicated that "Random Forest (RF), boosted regression tree (BRT), and support vector machine (SVM)" models exhibited accuracies of 99%, 97%, and 96% for the habitat suitability of *A. fatua*, respectively. The Boruta feature selection method identified the slope variable as significantly influential in wild oat habitat suitability modeling, followed by plan curvature, clay, temperature, and silt. This study serves as a case study that highlights the utility of machine learning for habitat suitability predictions when information on multiple environmental variables is available. This approach supports effective weed management strategies, potentially enhancing rapeseed productivity and mitigating the ecological impacts associated with weed infestation.

Keywords: Weed management, Habitat suitability, Precision agriculture, Machine learning, Ecological modeling

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Introduction

Rapeseed (*Brassica napus* L.) has gained global significance as a valuable oilseed crop that is widely cultivated because of its high-quality oil and protein-rich by-products (Neik et al. 2020). Rapeseed is the second major oilseed crop globally with the increasing world demand and production, followed by soybean oil (Tu et al. 2024). Its versatility as a source of edible oil, animal feed, and biofuel contributes to its pivotal role in food security and renewable energy sectors (Tileuberdi et al. 2022). Rapeseed exports have increased in recent decades, and by 2025 they are expected to expand by 40% (Tiwari et al. 2020).

Since 1996, rapeseed production in Iran has grown consistently in the international oilseed marketplace (Spörl et al. 2022). The increasing demand for sustainable agriculture highlights the necessity of efficient rapeseed cultivation practices, promoting its resilience to environmental stressors, and optimizing yield (Majidian et al. 2024). A notable challenge in rapeseed production is weed management, which can significantly reduce crop yield and quality by competing for resources such as nutrients, water, and sunlight (Hassan et al. 2023). This significant threat not only affects grain production and yield but may also compromise the quality of rapeseed oil, showing the urgent need for the agricultural sector to explore innovative practices and technologies to mitigate this challenge (Walia and Kumar 2020). A critical component in the formulation of effective management plans is a comprehensive understanding of weed flora and its geographical distribution. Such knowledge facilitates the application of herbicides and development of other appropriate management techniques (Krähmer et al. 2020; Nath et al. 2024).

Several weed species have been recognized for their significant effects on rapeseed yield and cultivation. The management and control of these weeds are crucial for maintaining the productivity and profitability of rapeseed crops (Asaduzzaman et al. 2020). *Avena fatua* (wild oats), belonging to the Gramineae family, is one of the most dominant weeds in rapeseed and is currently found in approximately 50 countries globally (Matsuhashi et al. 2021). Some studies have shown that wild oats can significantly reduce crop yield, highlighting their severe impact on agricultural productivity (Tang et al. 2024). Moreover, wild oats present a significant challenge because of their substantial resistance to herbicides, increasing control efforts, and causing an ongoing threat to rapeseed cultivation (Onkokesung et al. 2022).

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

GIS is one of the most effective and precise tools for producing weed distribution maps. These systems leverage advanced technologies to accurately identify areas infested by weeds, thereby facilitating targeted management approaches (Mohan and Giridhar 2022). Detailed species distribution and habitat suitability modeling enabled by geographic information system (GIS) technology play a critical role in environmental management by providing in-depth assessments of the interactions between different species and their environments.

In recent years, machine learning algorithms have emerged as powerful tools for modeling the habitat suitability of weeds based on environmental variables (Rather et al. 2020). By analyzing data on soil composition, climate conditions, and other ecological factors, machine learning models can predict the likelihood of weed proliferation in specific areas (Bi et al. 2024). These insights can aid in preemptive weed management strategies tailored to environmental conditions, thereby enhancing the precision of crop management practices in rapeseed farming (Akhter et al. 2020).

The integration of machine learning techniques (MLTs) into habitat suitability modeling (HSM) represents a cutting-edge approach that enhances the prediction and understanding of geographical distribution (Beery et al. 2021). By utilizing the power of algorithms and computational models, machine learning can analyze complex environmental and biological data to identify patterns and relationships that influence the presence or absence of species across different fields (Jeon et al. 2023). The use of habitat suitability as a measure for assessing the risk of weed infestation has increased globally (Hartl et al. 2024). HSM has been increasingly employed to identify areas that are potentially vulnerable to various weed species over extensive geographical areas (Qazi et al. 2023; Wang et al. 2023). Schartel et al. (2021) determined the habitat suitability of eight exotic species that were invasive in Baja California and assessed their distribution and invasion risk. Wan and Wang (2019) evaluated the compatibility of habitats for ten dangerous weed species and proposed a strategy for mitigating the risks posed by these weed species by modifying prevention and control methods.

Several studies have used MLTs to predict species distribution and habitat suitability modeling, such as support vector machine (SVM), random forest (RF), boosted regression trees (BRT), classification and regression trees (CARTs), generalized additive models (GAMs), and generalized linear models (GLMs) (Gholami et al. 2021; Mondal and Bhat 2021). RF is a group-learning method that uses multiple decision trees to improve prediction accuracy and is ideal for

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

assessing habitat suitability by evaluating diverse environmental variables (Renjana et al. 2024). Environmental research widely employs the support vector machine (SVM) framework, rooted in statistical learning theory. Although SVM demonstrates significant utility, its effectiveness in modeling habitats that favor the growth of specific plant species remains an area of ongoing investigation (Tazikeh et al. 2022). The BRT model combines the principles of boosting, a machine learning technique, with regression trees, and creates a powerful predictive model (Salditt et al. 2023). In predicting natural events and hazard backgrounds, models such as RF, SVM, and BRT have gained prominence because of their simplicity and efficacy (Berhane et al. 2021; Hasannejadasl et al. 2023; Hasan et al. 2024). However, the utilization of these models to assess habitat suitability for weed species in rapeseed fields remains relatively underexplored in scientific literature.

This study set forth two primary aims to address the key challenges in rapeseed farming within the Fars Province of Iran. First, it sought to identify and document the predominant weed species affecting rapeseed cultivation across the region, thereby contributing essential data to local agronomic research. Second, we implemented and compared three advanced modeling approaches, RF, SVM, and BRT, to predict habitat suitability for the identified dominant weed species. The assessment of influential environmental factors facilitated by the Boruta algorithm further enhances the model interpretability and ecological insight. Additionally, the selection of the optimal model based on the receiver operating characteristic (ROC) curve and area under the curve (AUC) maximizes predictive accuracy, pioneering the application of these machine learning techniques in weed habitat modeling. These aims collectively address a significant research gap, offering foundational knowledge that can improve precision in weed management strategies, reduce yield losses, and promote sustainable rapeseed production.

Materials and Methods

Study area

This investigation was performed in the southwestern region of the Fars Province, Iran, in 2023 (Fig. 1). The research region is situated between 27° 15′ 29″ and 30° 24′ 36″ N, and between 51° 29′ 32″ and 54° 28′ 49″ E. According to the FAO (2024), Iran has expanded its rapeseed cultivation significantly, reaching a total of around 200,000 hectares. This growth is part of the country's efforts to boost self-sufficiency in oilseed production, with regions like Fars Province

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

playing a crucial role. Geographical analyses based on topographic maps show that Fars Province encompasses both mountainous terrain and plains. The province is also distinguished by its climatic diversity, with the four seasons exerting distinct effects on regional flora. This variation in climate is largely attributed to the varied elevation, ranging from 182 to 3,183 meters above sea level. The Fars Province has an "average annual rainfall" of 315 mm and an "average annual temperature" of 15 °C (Kheiri et al. 2024). The average slope of the Fars Province is 7 °, which is particularly favorable for rapeseed cultivation.

Methodology

This research followed a five-stage methodology: (1) data collection, (2) preparation of influential factors, (3) habitat suitability modeling using three models: RF, SVM, and BRT, (4) evaluation of models and selection of the best model, and (5) variable importance analysis, as illustrated in Fig. 2.

Data collection and Sampling

In the present study, sampling was conducted through 114 rapeseed fields in 28 different counties of the Fars Province, based on the cultivation area of this crop. Some studies have demonstrated that the presence of weeds at the 6–8 leaf growth stage significantly reduces rapeseed yield (Bečka et al. 2021). Chao et al. (2023) stated that the critical period for weeds in autumn rapeseed growth can reduce plant performance by more than 10%; therefore, rapeseed should be maintained without weeds. Therefore, sampling was carried out during the winter season in 2023, when rapeseed is in the 6-8 growing leaf stage. Sampling was conducted using a 0.25-m² quadrat in the form of a W-shaped field based on the cultivation area (Fried et al. 2022) in each country (Table 1 and Fig. 3).

In addition to weed sampling, the geographic coordinates of each farm (latitude and longitude) were determined by using a GPS device. After collecting weeds from various rapeseed fields, they were accurately identified and counted based on genus and species. Soil samples from each point were transferred to the laboratory to determine the chemical and physical properties of the soil in each rapeseed field. Based on the equations presented (1–5), the "frequency %, uniformity %, mean field density (plant/m^2), and abundance index of different species" (Thomas 1985) were evaluated in Fars Province:

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

$$F_k = \frac{\sum Y_i}{n} \times 100 \quad (1)$$

where n is the "number of fields visited," Y_i is the "presence or absence of species k in field i," and F_k is the frequency of species k across all the quadrats. The following formula was used to obtain the uniformity index for species k (U_k).

$$U_k = \frac{\sum_i^n \sum_j^m Xij}{\sum_j^m m} \quad (2)$$

where Xij indicates the "presence or absence of species k in the i-th quadrat" and "j-th field, with n fields and m quadrats".

$$Dki = \frac{\sum_1^m Zj}{m} \times 4 \quad (3)$$

m is the number of thrown quadrats and "Zj is the number of plants in the quadrat." "Dki is the density (number of plants per meter) of the k species at field number i.

$$MDSK = \frac{\sum_1^n Dki}{n} \times 4 \quad (4)$$

Equation (4) states that "n is the number of fields visited", "Dki is the density (number of plants per meter) of k species on field number i", and MDSK is the mean density of species k.

$$Aik = F_k + U_k + MFDk \quad (5)$$

Finally, Equation (5) was used to determine the dominance index of the weeds. Using this equation, the "frequency (F_k), field uniformity (U_k), and mean density of species k (MDSK)" were combined to determine the predominant weed species.

Important factors

In general, for habitat suitability modeling, it is necessary to identify the factors that affect weed growth and development. For example, some studies have demonstrated that environmental factors including topography, soil chemical and physical properties, road development, temperature, and rainfall can affect weed distribution (Jehangir et al. 2024). Twelve layers were used as influencing factors, including "elevation, slope degree, slope aspect, plan curvature, distance from rivers, mean annual precipitation, mean annual temperature, pH, EC, and soil clay, silt, and sand percentages," which were considered to affect the growth and development of weed species. These 12 study layers were then converted to 30-meter resolution for future analyses (Kabiri et al. 2022) in ArcGIS version 10.8.1 (https://www.esri.com/en-us/arcgis/products/arcgis-desktop/overview). The annual mean rainfall and temperature data

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

were gathered from 29 meteorological organizations in the counties of the Fars Province. The data were then converted to a point map using ArcGIS version 10.8.1 software. The point map and study area were converted into temperature and rainfall maps using a 30-meter resolution by the IDW algorithm (Fig. 4 (A–B)).

In total, 189 soil samples were collected at a depth of 30 cm. A hydrometer was used to determine the physical characteristics of the soil, such as the amounts of "sand, silt, and clay" (Feng et al. 2024). A pH meter and a "conductivity meter" were used to test the pH and EC of the soil, respectively. "Sand, silt, clay, pH, and EC" layers were also converted into a raster map with 30-m resolution (Fig.4 (C-G)). A "digital elevation model (DEM)" of Fars Province was applied to assess "elevation, slope degree, slope aspect, and plan curvature" with a 30-m resolution (Fig. 4 (H–K)). Using topographic maps at a resolution of 1:25,000, a raster map of the distance from the rivers was created to assess the impact of the rivers on habitat suitability (Fig. 4 L).

Random forest (RF)

Random Forest (RF) is a supervised learning method developed by Breiman (2001) and consists of an ensemble of decision trees used for both classification and regression tasks. The RF model operates by constructing multiple trees during training and outputting the mode of the classes or mean prediction for classification and regression, respectively. This approach, which enhances model robustness and accuracy, is particularly effective for complex data, making it highly suitable for habitat suitability modeling (Talhami et al. 2024).

In this study, the key parameters for RF, such as n_estimators (number of trees in the forest), max_depth (maximum depth of each tree), and min_samples_split (minimum number of samples required to split a node), were optimized. We used grid search cross-validation to tune these parameters, with n_estimators ranging from 100 to 500, max_depth from 10 to 50, and min_samples_split set to identify the optimal values. The performance of the model was evaluated using accuracy, F1 score, and AUC/ROC metrics, providing a comprehensive assessment of model accuracy and threshold-specific performance. The RF model was implemented using the random forest package in R (https://cran.r-project.org/web/packages/randomForest/index.htm), which facilitates parameter tuning and cross-validation.

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Support vector machine (SVM)

The Support Vector Machine (SVM), introduced by Vapnik (1997), is a nonparametric statistical method that does not assume any particular distribution of the dataset. SVM is effective for high-dimensional data with a relatively small number of samples, making it suitable for species distribution modeling (Kumar et al. 2024).

For our SVM model, the key parameters included C (penalty parameter) and the kernel type (linear, polynomial, sinusoidal, or radial basis function). The C parameter was tuned from a range of 0.1 to 10 on a log scale to balance the margin and misclassification tolerance, while kernel selection was optimized based on model performance. Accuracy, F1 score, and AUC/ROC metrics were used for model evaluation, emphasizing precision and recall owing to potential class imbalance. The SVM model was implemented using the e1071 package in R (https://cran.r-project.org/web/packages/e1071/index.html), which provides comprehensive support for parameter optimization and evaluation.

Boosted regression trees (BRT)

A BRT is an ensemble method that combines the predictions of several weak classifiers into a stronger overall model (Alnahit et al. 2022). It uses the CART framework to iteratively add trees that correct errors made by previous ones, optimizing both the learningrate (learning speed) and n_estimators.

For this study, learning rate and estimators were optimized using a range of 0.01 to 0.1 for learning_rate and up to 500 trees for n_estimators. We also tuned max_depth to control tree complexity and prevent overfitting. The model was evaluated using accuracy, F1 score, and AUC/ROC metrics to provide a robust assessment of the predictive accuracy across thresholds.

We implemented BRT using the gbm package in R (https://cran.r-project.org/web/packages/gbm/index.html), which facilitates parameter tuning and cross-validation, including early stopping based on AUC/ROC performance.

Boruta algorithm

A critical component of this research involves evaluating the importance of variables in spatial modeling for habitat suitability to guide optimal management strategies (López-Torres et al. 2023). The Boruta algorithm was chosen for this purpose because it effectively identifies

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

influential variables by leveraging the random forest model's capacity for variable selection (Li et al., 2023). The algorithm operates by iteratively comparing the importance of actual features to shadow features, which are randomized duplicates, thus distinguishing truly important predictors from noise (Xiao et al. 2024).

For the implementation, we used the Boruta package in R (https://cran.r-project.org/web/packages/Boruta/index.html). The key parameters included maxRuns, set to 500 to ensure sufficient iterations for stable results, and doTrace, set to 2 for detailed output during the execution of the algorithm. The maxRuns parameter influences the stability and reliability of the variable importance ranking. Higher values provide more robust assessments by allowing more comparisons across iterations. Additionally, we used a p-value threshold of 0.05 to statistically identify significant variables.

The Boruta algorithm outputs three categories of variables: Confirmed, Tentative, and Rejected (Han et al. 2022). This categorization helps refine the selection process by confirming variables with a statistically meaningful impact on habitat suitability while excluding non-informative features (Wang et al. 2022). The results of the Boruta algorithm provide a clear ranked list of predictor variables crucial for understanding and managing habitat suitability patterns across different regions (Prasad et al. 2022). The variable importance derived from Boruta was instrumental in identifying which factors were most relevant in weed habitat suitability modeling, thereby guiding targeted management strategies.

Accuracy of models

In habitat suitability modeling, where the goal is to forecast the "presence or absence" of a species in various locations, ROC and AUC metrics are essential tools for assessing model performance (Jamali et al. 2024). For this purpose, 70% of the presence data of the dominant weed were used in the modeling process, while the remaining 30% of the data were utilized for validation and to evaluate the model's projected accuracy.

In this study, the 70:30 split between training and validation datasets was selected based on its established utility in predictive modeling and its practical alignment with the dataset size. This ratio is widely used in ecological and machine learning applications as a standard practice (Fielding and Bell 1997), balancing the competing requirements of sufficient data for model training and a reasonable subset for validation. The chosen split minimizes overfitting risk while allowing the evaluation of model performance on an independent dataset.

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Given the dataset size, this split is particularly well-suited to maximize the reliability of model parameter estimation and predictive accuracy. Despite the relatively modest dataset size, ecological modeling often operates with limited datasets due to challenges such as field collection constraints and environmental variability (Elith et al. 2006). While larger datasets are ideal, the 70:30 split effectively uses the available data to produce statistically sound results, consistent with studies in similar contexts (Hameed and Alamgir 2022).

The ROC curve and AUC are widely used metrics for assessing prediction models' accuracy. The ROC curve, a graphical representation, plots two parameters to show how well a classification model performs: the "True Positive Rate (TPR)" or sensitivity, and the "False Positive Rate (FPR)" or 1-specificity, across different threshold values (Muschelli 2020). The TPR, represented on the y-axis, indicates the proportion of real positives correctly identified by the model, while the FPR, shown on the x-axis, represents the proportion of real negatives that are incorrectly classified as positives (Carrington et al. 2022). A single aggregate performance metric across all potential classification thresholds is provided by the ROC curve and AUC (Verbakel et al. 2020; Saha et al. 2023). The AUC value ranges from 0 to 1 and is classified into four performance categories: "0.5-0.6 (poor), 0.6-0.7 (moderate), 0.7-0.8 (good), 0.8-0.9 (very good), and 0.9-1.0 (excellent)" (Table 2). In this study, the ROC-AUC was utilized to evaluate the RF, BRT, and SVM models using SPSS software version 26 (http://www.ibm.com).

Collinearity Test of Effective Factors

The Collinearity test of useful elements is a crucial technique of statistical analysis employed to diagnose the extent of “multi-collinearity” among independent variables within a regression model (Barman et al. 2024). To quantitatively assess “multi-collinearity,” the "variance inflation factor (VIF)" and tolerance indices were utilized. These metrics offer insights into the degree of linear association between an independent factor and the remaining independent variables in the model. A VIF value of 5 or 10 and above is generally regarded as demonstrating a problematic level of multi-collinearity, indicating an exaggerated variance in an "estimated regression coefficient" by a factor of 5 or 10 because of its linear relationship with other variables (Cheng et al. 2022). The percentage of volatility of an independent variable that cannot be accounted for by other independent variables is called the tolerance. Hence, a lower tolerance value indicates a higher overlap of explanatory information among variables, signifying a potential multi-

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

collinearity issue. Typically, a tolerance value of less than 0.20 or 0.10 is considered indicative of significant multi-collinearity (Negash and Alelgn 2022).

Results and Discussion

Determine the dominant weed

Frequency percentages of genera and species were used to assess the dominant weeds. The initial findings from the sampling process indicated that *A. fatua* emerged as the most prevalent weed species, signifying its significant presence and impact in the sampled areas. Notably, 32 dominant weed species were identified, with wild oats being the primary dominant species, with a frequency of 58.48% (Table 3). This indicated the critical importance of *A. fatua* in terms of their abundance and ecological influence on the studied environments.

Multi-collinearity test

Table 4. shows the collinearity between the factors affecting the species distribution modeling of *A. fatua* in the study area. Thus, based on the findings obtained, the tolerance coefficient is not less than 0.1 in any of the indices, and the variance inflation factor was not five or greater in any of the indices; therefore, there was no collinearity between the indices used. Otherwise, there will be multi-collinearity between the independent parameters and parameter estimates, and statistical significance standards will be targeted (Rovetta et al., 2023). This leads to a lack of acceptable accuracy for spatial analysis, especially in RF, BRT, and SVM modeling.

Machine-learning techniques (MLTs)

The final maps of the RF, SVM, and BRT models were divided into four classes to determine the suitability of the wild oat habitats (Fig. 5A).

RF Algorithm

According to the RF model, the low (66.56%), moderate (16.35%), high (11.71%), and very high (5.38%) classes had the largest relative areas (Table 5). In addition, the RF model map showed that the northern, northwestern, central, eastern, western, southeastern, and southwestern regions of the study area had the highest "habitat suitability" for *A. fatua* (Fig. 5A). However, some

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

centers had low "habitat suitability" for *A. fatua* (Fig. 5A). However, the northeast and parts of the center were not affected by this weed invasion.

BRT Algorithm

The habitat suitability map of *A. fatua* by BRT showed that the low (56.65%), moderate (26.93%), high (11.71%), and very high (4.55%) classes had the largest relative areas (Table 5). The situation of the counties regarding the suitability of the habitat of *A. fatua* based on the BRT model, was the same as that of the RF model (Fig. 5B). This demonstrated that these models had the same performance in terms of predicting the habitat suitability of this weed.

SVM Algorithm

The SVM model had different classification conditions such that the moderate (37.89%), low (37.89%), high (19.81%), and very high (6.74%) classes had the highest relative areas (Table 5). The suitability map of the SVM model showed that parts of the " northern, northwestern, and southern study areas had a greater habitat suitability for *A. fatua* (Fig. 5C). In this model, small portions of the research area ("west, northwest, southeast, east, and north") had the highest habitat suitability. According to the findings of the SVM model, it can be emphasized that the east had the highest habitat suitability for *A. fatua* (Fig. 5C). In addition, counties in the "southwest, southeast, and a large portion of the center of the research area" had low habitat suitability.

Evaluation of algorithms

In this study, the models were evaluated using the “ROC curve” and “AUC.” The most accurate models were the "RF, BRT, and SVM" models according to the ROC curve (Fig. 6). Also, the "area under the curve" confirmed the accuracy of the RF (0.99%), BRT (0.97), and SVM (0.96) models (Table 6). Huang et al. (2021) have reported that the areas under the curve are 0.5–0.6 (poor), 0.6–0.7 (moderate), 0.7–0.8 (good), 0.8–0.9 (very good), and 0.9–1 (excellent), respectively. Therefore, the RF, BRT, and SVM models were excellent in this study.

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Variables importance

In this study, the relevance of these variables is evaluated through the application of the "Boruta algorithm." This method was used to determine the most influential factors in the analysis. The results of the “Boruta algorithm” demonstrated that the slope, plan curvature, clay, temperature, and silt factors had the greatest impact on the modeling of *A. fatua* habitat suitability (Table 7). Differences in the slope of the soil throughout the terrain may have affected the growth and expansion of *A. fatua*. This factor has a profound effect on vegetation dispersal patterns. One of the important effects of land slope is moisture absorption. For example, south-facing slopes subjected to higher solar irradiance typically exhibit reduced soil moisture levels, constraining plant growth.

Practical Implications and Conclusion

This study highlights the practical applications of machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Boosted Regression Tree (BRT), for modeling the habitat suitability of *A. fatua* in rapeseed fields. Each algorithm brings unique advantages to understanding weed distribution, which is crucial for devising sustainable and site-specific management strategies to mitigate the detrimental effects of *A. fatua* on crop productivity. By leveraging the strengths of these models, this research provides actionable insights that align with contemporary agricultural goals of improving efficiency while minimizing environmental impacts.

The RF model emerged as the most effective algorithm, achieving the highest accuracy (99%) in predicting habitat suitability. This model was instrumental in identifying key environmental predictors, such as slope, soil texture, and plan curvature, that significantly influence *A. fatua* distribution. Its embedded feature selection capabilities not only enhanced interpretability but also allowed for the refinement of management practices in heterogeneous agricultural landscapes. Studies by Kang et al. (2022) and Melash et al. (2023) further validate the efficacy of RF in handling complex ecological datasets with numerous interacting variables. Additionally, RF’s ensemble approach ensures model stability and robustness to outliers, making it particularly suitable for field-based ecological studies characterized by high variability in environmental conditions.

SVM also demonstrated its utility in analyzing high-dimensional datasets, with a classification accuracy of 96%. This algorithm excelled in differentiating between habitat suitability classes,

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

providing detailed ecological niche maps that are indispensable for spatially targeted weed management. The ability of SVM to handle complex interactions among environmental variables has been documented in recent works, including Akhtar et al. (2024) and O'Neill et al. (2023). These studies emphasize the importance of SVM in addressing challenges posed by diverse agro ecological conditions, where precision in habitat differentiation directly impacts the effectiveness of weed control measures.

The BRT model, with an accuracy of 97%, effectively captured nonlinear relationships between *A. fatua* occurrence and predictor variables. This capacity for addressing nonlinearity is particularly significant in weed science, where ecological interactions are rarely linear. The ensemble-based nature of BRT enhances its prediction precision, a feature corroborated by studies such as Montoya-Jiménez et al. (2022) and Kumari et al. (2024). By integrating BRT into habitat suitability modeling, this study adds to the growing body of evidence supporting its applicability in managing invasive species in agricultural systems.

Although the 70:30 training-validation split provides an efficient framework for ecological modeling, the dataset size remains a potential limitation of this study (Garcés et al. 2022). Smaller datasets inherently constrain the ability to capture rare patterns and subtle environmental interactions, which could impact model generalizability (Yu et al. 2024). However, this study operates within the boundaries of a case-study approach, where the primary goal is to explore and demonstrate a method's applicability rather than achieve universal generalizability. To address this limitation, the dataset size and split were carefully chosen to balance robustness in model training and reliable validation. Previous studies have demonstrated that even smaller datasets can yield valuable insights when the modeling methodology is rigorous (Wisz et al., 2008). Additionally, the model's performance metrics, assessed using cross-validation, support the inference that the chosen split is sufficient for the study's aims. Future research could address this limitation by expanding the dataset through additional sampling or leveraging synthetic data generation techniques to augment the dataset size. Nevertheless, for a case-study framework, this approach aligns well with established methodologies, and the results provide meaningful insights into the ecological processes under investigation.

The complementary strengths of RF, SVM, and BRT underscore their collective utility in ecological modeling. RF and BRT were particularly effective in assessing feature importance, while SVM provided the highest resolution in classification tasks. This integrated approach

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

offers a more comprehensive understanding of *A. fatua* habitat suitability and enables the creation of nuanced maps tailored to specific regional conditions. Such detailed mapping provides a critical basis for targeted interventions, ensuring that management resources are deployed efficiently and effectively in areas at high risk of weed invasion.

The practical implications of this study extend beyond theoretical modeling. By generating habitat suitability maps, this research equips agricultural practitioners with precise tools for implementing site-specific weed management strategies. This targeted approach not only minimizes herbicide usage but also supports environmentally conscious practices that align with the principles of sustainable agriculture. Topographic factors, such as slope and aspect, emerged as pivotal predictors, corroborating findings from Yang et al. (2023) and Vykydalová et al. (2024) that highlight the role of microclimatic conditions in shaping weed distribution. Similarly, the influence of soil texture and temperature on habitat suitability aligns with broader ecological studies, such as those by Dastres et al. (2023) and Yao (2023), emphasizing the adaptive strategies of *A. fatua* in diverse agro ecological contexts.

While the study showcases the effectiveness of RF, SVM, and BRT, it also acknowledges limitations inherent to these models. The accuracy of predictions is influenced by data quality and representativeness, as highlighted in recent works by Hasan et al. (2024) and Xu et al. (2024). Algorithmic biases, environmental variability, and scalability challenges further underscore the need for continuous refinement of these methods. For instance, temporal and spatial changes in environmental conditions may reduce the reliability of predictions over time, necessitating the development of more adaptive and scalable modeling frameworks. Future research should focus on addressing these limitations to enhance the robustness and generalizability of machine learning applications in weed science.

In conclusion, this research advances the field of weed science by demonstrating the potential of machine learning models to improve habitat suitability predictions for dominant weeds like *A. fatua*. By integrating ecological, agronomic, and computational insights, the study lays a foundation for the development of sustainable, data-driven weed management strategies. The findings not only highlight the efficacy of RF, BRT, and SVM in ecological modeling but also provide a roadmap for their broader application in addressing challenges associated with agricultural sustainability and biodiversity conservation.

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Funding statement.

his research was funded by Research Council of Shiraz University, Grant/Award Number: 98GCU1M75346

Competing interests.

No competing interests have been declared.

References

Ai Z, Zhang J, Liu H, Liang C, Xue S, Liu G (2020) Influence of slope aspect on the macro- and micronutrients in *Artemisia sacrorum* on the Loess Plateau in China. Environ Sci Pollut Res Int 27:20160–20172

Akhtar M, Tanveer M, Arshad M (2024) RoBoSS: A robust, bounded, sparse, and smooth loss function for supervised learning. IEEE Trans Pattern Anal Mach Intell https://doi.org/10.1109/TPAMI.2024.3465535

Akhter MJ, Jensen PK, Mathiassen SK, Melander B, Kudsk P (2020) Biology and management of *Vulpia myuros*—an emerging weed problem in no-till cropping systems in Europe. Plants 9:715 h

Aldayel M, Al-Nafjan A (2024) A comprehensive exploration of machine learning techniques for EEG-based anxiety detection. *PeerJ Comput Sci* 10. https://doi.org/10.7717/peerj-cs.1829

Alnahit AO, Mishra AK, Khan AA (2022) Stream water quality prediction using boosted regression tree and random forest models. *Stoch Environ Res Risk Assess* 36:2661–2680. https://doi.org/10.1007/s00477-021-02152-4

Asaduzzaman M, Pratley JE, Luckett D, Lemerle D, Wu H (2020) Weed management in canola (*Brassica napus* L): a review of current constraints and future strategies for Australia. *Arch Agron Soil Sci* 66:427–444. https://doi.org/10.1080/03650340.2019.1624726

Barman J, Biswas B, Rao KS (2024) A hybrid integration of analytical hierarchy process (AHP) and the multiobjective optimization on the basis of ratio analysis (MOORA) for landslide susceptibility zonation of Aizawl, India. *Nat Hazards* 1–26. https://doi.org/10.1007/s11069-024-06538-9

Bečka D, Bečková L, Kuchtová P, Cihlář P, Pazderů K, Mikšík V, Vašák J (2021) Growth and yield of winter oilseed rape under strip-tillage compared to conventional tillage. *Plant Soil Environ* 67:2. https://doi.org/10.17221/492/2020-PSE

Beery S, Cole E, Parker J, Perona P, Winner K (2021) Species distribution modeling for machine learning practitioners: a review. *In ACM SIGCAS Conf Comput Sustain Soc* 329–348. https://doi.org/10.1145/3460112.3471966

Berhane G, Kebede M, Alfarrah N (2021) Landslide susceptibility mapping and rock slope stability assessment using frequency ratio and kinematic analysis in the mountains of Mgulat area, Northern Ethiopia. Bull Eng Geol Environ 80:285–301. https://doi.org/10.1007/s10064-020-01905-9

Bi Z, Sun J, Xie Y, Gu Y, Zhang H, Zheng B, Ou R, Liu G, Li L, Peng X, Gao X, Wei N (2024) Machine learning-driven source identification and ecological risk prediction of heavy metal pollution in cultivated soils. J Hazard Mater 476:135109 . https://doi.org/10.1016/j.jhazmat.2024.135109

Breiman L (2001) Random forests. Mach Learn 45:5–32 . https://doi.org/10.1023/A:1010933404324

Carrington AM, Manuel DG, Fieguth PW, Ramsay T, Osmani V, Wernly B, Bennett C, Hawken S, Magwood O, Sheikh Y, McInnes M, Holzinger A (2022) Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Trans Pattern Anal Mach Intell 45:329–341 . https://doi.org/10.1109/TPAMI.2022.3145392

Chao WS, Anderson JV, Li X, Gesch RW, Berti MT, Horvath DP (2023) Overwintering camelina and canola/rapeseed show promise for improving integrated weed management approaches in the Upper Midwestern US. Plants 12:1329 . https://doi.org/10.3390/plants12061329

Cheng J, Sun J, Yao K, Xu M, Cao Y (2022) A variable selection method based on mutual information and variance inflation factor. Spectrochim Acta A Mol Biomol Spectrosc 268:120652 .https://doi.org/10.1016/j.saa.2021.120652

Christiansen, D. M., Römer, G., Dahlgren, J. P., Borg, M., Jones, O. R., Merinero, S., ... & Ehrlén, J. (2024). High-resolution data are necessary to understand the effects of climate on plant population dynamics of a forest herb. Ecology, 105(1), e4191. https://doi.org/10.1002/ecy.4191

da Costa Dias T, Silveira LF, Francisco MR (2024) Endemic and threatened birds as surrogates for identifying conservation priority areas and ecological corridors in the America's most endangered habitat. Sci Rep 14:21923 .https://doi.org/10.1038/s41598-024-72948-1

Dastres E, Jahangiri E, Edalat M, Zamani A, Amiri M, Pourghasemi HR (2023) Habitat suitability modeling of *Descurainia sophia* medicinal plant using three bivariate models. Environ Monit Assess 195:392 . https://doi.org/10.1007/s10661-023-10996-2

Drees TH, Shea K (2024) Elevated temperatures shift flower head height distributions and seed dispersal patterns in two invasive thistle species. Ecology 105. https://doi.org/10.1002/ecy.4201

Dubuc A, Collins GM, Coleman L, Waltham NJ, Rummer JL, Sheaves M (2021) Association between physiological performance and short temporal changes in habitat utilisation modulated by environmental factors. Mar Environ Res 170:105448. https://doi.org/10.1016/j.marenvres.2021.105448

Elith J, Graham H, Anderson R, Dudík M, Ferrier S, Guisan A, Zimmermann N (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129-151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Feng L, Khalil U, Aslam B, Ghaffar B, Tariq A, Jamil A, Farhan M, Aslam M, Soufan W (2024) Evaluation of soil texture classification from orthodox interpolation and machine learning techniques. Environ Res 246:118075. https://doi.org/10.1016/j.envres.2023.118075

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24(1):38-49 https://doi.org/10.1017/S0376892997000088

Fraccica A, Romero E, Fourcaud T (2024) Effects of vegetation growth on soil microstructure and hydro-mechanical behaviour. Géotechnique 1–41. https://doi.org/10.1680/jgeot.23.00163

Fried G, Le Corre V, Rakotoson T, Buchmann J, Germain T, Gounon R, Chauvel B (2022) Impact of new management practices on arable and field margin plant communities in sunflower, with an emphasis on the abundance of *Ambrosia artemisiifolia* (Asteraceae). Weed Res 62:134–148. https://doi.org/10.1111/wre.12522

Garcés, P., Baumeister, S., Mason, L., Chatham, C. H., Holiga, S., Dukart, J., Jones, E. J. H., Banaschewski, T., Baron-Cohen, S., Bölte, S., Buitelaar, J. K., Durston, S., Oranje, B., Persico, A. M., Beckmann, C. F., Bougeron, T., Dell'Acqua, F., Ecker, C., Moessnang, C., Charman, T., … EU-AIMS LEAP group authorship (2022). Resting state EEG power spectrum and functional connectivity in autism: a cross-sectional analysis. Molecular autism, 13(1), 22. https://doi.org/10.1186/s13229-022-00500-x

Gholami H, Mohamadifar A, Rahimi S, Kaskaoutis DG, Collins AL (2021) Predicting land susceptibility to atmospheric dust emissions in central Iran by combining integrated data mining and a regional climate model. Atmos Pollut Res 12:172–187. https://doi.org/10.1016/j.apr.2021.03.005

Gxasheka M, Gajana CS, Dlamini P (2023) The role of topographic and soil factors on woody plant encroachment in mountainous rangelands: A mini literature review. Heliyon 9. https://doi.org/10.1016/j.heliyon.2023.e20615

Hameed MAB, Alamgir Z (2022) Improving mortality prediction in acute pancreatitis by machine learning and data augmentation. Comput Biol Med 150:106077.https://doi.org/10.1016/j.compbiomed.2022.106077

Han X, Wang L, Wang Y, Yang J, Wan X, Liang T, Rinklebe J (2022) Mechanisms and influencing factors of yttrium sorption on paddy soil: Experiments and modeling. Chemosphere 307:135688. https://doi.org/10.1016/j.chemosphere.2022.135688

Hartl T, Srivastava V, Prager S, Wist T (2024) Evaluating climate change scenarios on global pea aphid habitat suitability using species distribution models. Clim Chang Ecol 7:100084. https://doi.org/10.1016/j.ecochg.2024.100084

Hasan MM, Roy SK, Talha MD, Ferdous MT, Nasher NMR (2024) Predictive landslide susceptibility modeling in the southeastern hilly region of Bangladesh: application of machine learning algorithms in Khagrachari district. Environ Sci Pollut Res 1–18. https://doi.org/10.1007/s11356-024-34949-5

Hasannejadasl H, Osong B, Bermejo I, van der Poel H, Vanneste B, van Roermund J, Aben K, Zhang Z, Kiemeney L, Van Oort I, Verwey R, Hochstenbach L, Bloemen E, Dekker A, Fijten RRR (2023) A comparison of machine learning models for predicting urinary incontinence in men with localized prostate cancer. Front Oncol 13:1168219. https://doi.org/10.3389/fonc.2023.1168219

Hassan MS, Naz N, Ali H, Ali B, Akram M, Iqbal R, Ajmal S, Ali B, Ercisli S, Golokhvast KS, Hassan Z (2023) Ultra-responses of *Asphodelus tenuifolius* L. (Wild Onion) and *Convolvulus arvensis* L. (Field Bindweed) against shoot extract of *Trianthema portulacastrum* L. (Horse Purslane). Plants (Basel) 12(3):458 https://doi.org/10.3390/plants12030458

He B, Zhao Y, Liu S, Ahmad S, Mao W (2023) Mapping seagrass habitats of potential suitability using a hybrid machine learning model. Front Ecol Evol 11:1116083. https://doi.org/10.3389/fevo.2023.1116083

Hu H, Bao W, Huang L, Li F (2024) Shifting patterns in fine root distribution of four xerophytic species across soil structural gradients and years of growth. Ecol Evol 14: e10889. https://doi.org/10.1002/ece3.10889

Huang W, Liu H, Zhang Y, Mi R, Tong C, Xiao W, Shuai B (2021) Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM. Appl Soft Comput 109:107541. https://doi.org/10.1016/j.asoc.2021.107541

Jamali F, Amininasab SM, Taleshi H, Madadi H (2024) Ensemble forecasting of Persian leopard (Panthera pardus saxicolor) distribution and habitat suitability in south-western Iran. Wildl Res 51(3). https://doi.org/10.1071/WR23010

Jehangir S, Khan SM, Ahmad Z, Ejaz U, Ain QU, Lho LH, et al. (2024) Distribution of the Cannabis sativa L. in the Western Himalayas: A tale of the ecological factors behind its continuous invasiveness. Glob Ecol Conserv 49. https://doi.org/10.1016/j.gecco.2023.e02779

Jeon J, Lee S, Oh C (2023) Age-specific risk factors for the prediction of obesity using a machine learning approach. Front Public Health 10:998782. https://doi.org/10.3389/fpubh.2022.998782

Kabiri S, Allen M, Okuonzia JT, Akello B, Ssabaganzi R, Mubiru D (2022) Detecting wetland encroachment and urban agriculture land classification in Uganda using hyper-temporal remote sensing. AAS Open Res 3:18. https://doi.org/10.12688/aasopenres.13040.2

Kamal A, Mian I, Akbar W, Rahim H, Irfan M, Ali S, Zaman W (2023) Effects of soil depth and altitude on soil texture and soil quality index. Appl Ecol Environ Res 21(5). DOI: http://dx.doi.org/10.15666/aeer/2105_41354154

Kang W, Kim G, Park Y (2022) Habitat suitability and connectivity modeling predict genetic population structure and priority control areas for invasive nutria (*Myocastor coypus*) in a temperate river basin. PLoS One 17(12). https://doi.org/10.1371/journal.pone.0279082

Kheiri M, Kambouzia J, Rahimi-Moghaddam S, Moghaddam SM, Vasa L, Azadi H (2024) Effects of agro-climatic indices on wheat yield in arid, semi-arid, and sub-humid regions of Iran. Reg Environ Change 24(1):10. https://doi.org/10.1007/s10113-023-02173-5

Kim MK, Rouphael C, McMichael J, Welch N, Dasarathy S (2024) Challenges in and opportunities for electronic health record-based data analysis and interpretation. Gut Liver 18(2):201–208. https://doi.org/10.5009/gnl230272

Krähmer H, Andreasen C, Economou-Antonaka G, Holec J, Kalivas D, Kolářová M, Novák R, Panozzo S, Pinke G, Salonen J, Sattin M (2020) Weed surveys and weed mapping in Europe: State of the art and future tasks. Crop Prot 129:105010. https://doi.org/10.1016/j.cropro.2019.105010

Krishnan A, Singh A, Tamma K (2020) Visual signal evolution along complementary color axes in four bird lineages. Biol Open 9(9). https://doi.org/10.1242/bio.052316

Kumar A, Sinha S, Saurav S, Chauhan VB (2024) Prediction of unconfined compressive strength of cement–fly ash stabilized soil using support vector machines. Asian J Civ Eng 25(2):1149–1161. https://doi.org/10.1007/s42107-023-00833-9

Kumari G, Kotiyal PB, Singh H, Kumar M, Kumar N, Malik A, Singh S (2024) Predicting future climate change effects on biotic communities: A species distribution modeling approach. In: Forests and Climate Change: Biological Perspectives on Impact, Adaptation, and Mitigation Strategies. Springer Nature Singapore, pp 137–168. https://doi.org/10.1007/978-981-97-3905-9_7

Li Q, Wei Y, Zhang T, Che F, Yao S, Wang C, Shi D, Tang H, Song B (2023) Predictive models and early postoperative recurrence evaluation for hepatocellular carcinoma based on gadoxetic acid-enhanced MR imaging. Insights Imaging 14(1):4. https://doi.org/10.1186/s13244-022-01359-5

López-Torres JF, Sánchez-García JY, Núñez-Ríos JE, López-Hernández C (2023) Prioritizing factors for effective strategy implementation in small and medium-size organizations. Eur Bus Rev 35(5):694–712. https://doi.org/10.1108/EBR-11-2022-0230

Majidian P, Ghorbani HR, Farajpour M (2024) Achieving agricultural sustainability through soybean production in Iran: Potential and challenges. Heliyon 10(4). https://doi.org/10.1016/j.heliyon.2024.e26389

Matsuhashi S, Asai M, Fukasawa K (2021) Estimations and projections of Avena fatua dynamics under multiple management scenarios in crop fields using simplified longitudinal monitoring. PLoS One 16(1). https://doi.org/10.1371/journal.pone.0245217

Melash AA, Bogale AA, Migbaru AT, Chakilu GG, Percze A, Ábrahám ÉB, Mengistu DK (2023) Indigenous agricultural knowledge: A neglected human-based resource for sustainable crop protection and production. Heliyon 9(1). https://doi.org/10.1016/j.heliyon.2023.e12978

Mohan S, Giridhar MVSS (2022) A brief review of recent developments in the integration of deep learning with GIS. Geomatics Environ Eng 16(2):21–38.

Mondal R, Bhat A (2021) Comparison of regression-based and machine learning techniques to explain alpha diversity of fish communities in streams of central and eastern India. Ecol Indic 129:107922. https://doi.org/10.1016/j.ecolind.2021.107922

Monteiro A, Santos S (2022) Sustainable approach to weed management: The role of precision weed management. Agronomy 12(1):118. https://doi.org/10.3390/agronomy12010118

Montoya-Jiménez JC, Valdez-Lazalde JR, Ángeles-Perez G, De Los Santos-Posadas HM, Cruz-Cárdenas G (2022) Predictive capacity of nine algorithms and an ensemble model to determine the geographic distribution of tree species. iForest-Biogeosciences and Forestry 15(5):363. https://doi.org/10.3832/ifor4084-015

Muschelli III J (2020) ROC and AUC with a binary predictor: a potentially misleading metric. J Classification 37(3):696–708. https://doi.org/10.1007/s00357-019-09345-1

Nath CP, Singh RG, Choudhary VK, Datta D, Nandan R, Singh SS (2024) Challenges and alternatives of herbicide-based weed management. Agronomy 14(1):126. https://doi.org/10.3390/agronomy14010126

Negash BT, Alelgn Y (2022) Proper partograph utilization among skilled birth attendants in Hawassa city public health facilities, Sidama region, Ethiopia, in 2021. BMC Womens Health 22(1):539. https://doi.org/10.1186/s12905-022-02117-x

Neik TX, Amas J, Barbetti M, Edwards D, Batley J (2020) Understanding host-pathogen interactions in Brassica napus in the omics era. Plants (Basel, Switzerland) 9(10):1336. https://doi.org/10.3390/plants9101336

O'Neill H, Khalid Y, Spink G, Thorpe P (2023) A one-class support vector machine for detecting valve stiction. Digital Chemical Engineering 8:100116. https://doi.org/10.1016/j.dche.2023.100116

Onkokesung N, Brazier-Hicks M, Tetard-Jones C, Bentham A, Edwards R (2022) Molecular diagnostics for real-time determination of herbicide resistance in wild grasses. J Biotechnol 358:64–66. https://doi.org/10.1016/j.jbiotec.2022.09.004

Peters U (2022) Algorithmic political bias in artificial intelligence systems. Philos Technol 35(2):25. https://doi.org/10.1007/s13347-022-00512-8

Prasad P, Loveson VJ, Das B, Kotha M (2022) Novel ensemble machine learning models in flood susceptibility mapping. Geocarto Int 37(16):4571–4593. https://doi.org/10.1080/10106049.2021.1892209

Qazi AW, Saqib Z, Zaman-ul-Haq M, Gardezi SMH, Khan AM, Khan I, ... & Ahmed I (2023) Modelling impacts of climate change on habitat suitability of three endemic plant species in Pakistan. Pol J Environ Stud 32(4). https://doi.org/10.15244/pjoes/161876

Radočaj D, Jurišić M (2022) GIS-based cropland suitability prediction using machine learning: A novel approach to sustainable agricultural production. Agronomy 12(9):2210. https://doi.org/10.3390/agronomy12092210

Rather TA, Kumar S, Khan JA (2020) Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm. Sci Rep 10(1):11473. https://doi.org/10.1038/s41598-020-68167-z

Renjana E, Firdiana ER, Angio MH, Ningrum LW, Lailaty IQ, Rahadiantoro A, Martiansyah I, Zulkarnaen R, Rahayu A, Raharjo PD, Abywijaya IK, Usmadi D, Risna RA, Cropper WP Jr, Yudaputra A (2024) Spatial habitat suitability prediction of essential oil wild plants on Indonesia's degraded lands. PeerJ 12. https://doi.org/10.7717/peerj.17210

Richardson E, Trevizani R, Greenbaum JA, Carter H, Nielsen M, Peters B (2024) The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns (New York, N.Y.) 5:100994. https://doi.org/10.1016/j.patter.2024.100994

Rovetta A (2023) A framework to avoid significance fallacy. Cureus 15:6. 10.7759/cureus.40242

Saha S, Bera B, Shit PK, Bhattacharjee S, Sengupta N (2023) Prediction of forest fire susceptibility applying machine and deep learning algorithms for conservation priorities of forest resources. Remote Sens Appl Soc Environ 29:100917. https://doi.org/10.1016/j.rsase.2022.100917

Salditt M, Humberg S, Nestler S (2023) Gradient Tree Boosting for Hierarchical Data. Multiv Behav Res 58:911–937. https://doi.org/10.1080/00273171.2022.2146638

Sampaio F, Batista MM, Marchioro CA (2024) Temperature-dependent reproduction of Spodoptera eridania: developing an oviposition model for a novel invasive species. Pest Manag Sci 80:1118–1125. https://doi.org/10.1002/ps.7842

Schartel TE, Cooper ML, May A, Daugherty MP (2021) Quantifying Planococcus ficus (Hemiptera: Pseudococcidae) invasion in Northern California vineyards to inform management strategy. Environ Entomol 50:138–148. https://doi.org/10.1093/ee/nvaa141

Serajian R, Sun JQ, Cobian-Iñiguez J, Ehsani R (2024) Predictive neural network modeling for almond harvest dust control. Sensors 24:2136. https://doi.org/10.3390/s24072136

Singh M, Kukal MS, Irmak S, Jhala AJ (2022) Water use characteristics of weeds: A global review, best practices, and future directions. Front Plant Sci 12:794090. https://doi.org/10.3389/fpls.2021.794090

Somerville GJ, Sønderskov M, Mathiassen SK, Metcalfe H (2020) Spatial modelling of within-field weed populations: A review. Agronomy 10:1044. https://doi.org/10.3390/agronomy10071044

Spörl J, Speer K, Jira W (2022) Simultaneous mass spectrometric detection of proteins of ten oilseed species in meat products. Foods 11:2155. https://doi.org/10.3390/foods11142155

Talhami M, Wakjira T, Alomar T, Fouladi S, Fezouni F, Ebead U, et al. (2024) Single and ensemble explainable machine learning-based prediction of membrane flux in the reverse osmosis process. J Water Process Eng 57:104633. https://doi.org/10.1016/j.jwpe.2023.104633

Tang W, Li Z, Guo H, Chen B, Wang T, Miao F, et al. (2024) Annual weeds suppression and oat forage yield responses to crop density management in an oat-cultivated grassland: A case study in Eastern China. Agronomy 14:583. https://doi.org/10.3390/agronomy14030583

Tazikeh S, Davoudi A, Shafiei A, Parsaei H, Atabaev TS, Ivakhnenko OP (2022) A comparison between the perturbed-chain statistical associating fluid theory equation of state and machine learning modeling approaches in asphaltene onset pressure and bubble point pressure prediction during gas injection. ACS Omega 7:30113–30124. https://doi.org/10.1021/acsomega.2c03192

Thomas AG (1985) Weed survey system used in Saskatchewan for cereal and oilseed crops. Weed Sci 33:34–43. https://doi.org/10.1017/S0043174500083892

Tileuberdi N, Turgumbayeva A, Yeskaliyeva B, Sarsenova L, Issayeva R (2022) Extraction, isolation of bioactive compounds and therapeutic potential of rapeseed (*Brassica napus* L.). Molecules 27:8824. https://doi.org/10.3390/molecules27248824

Tiwari AK, Nasreen S, Shahbaz M, Hammoudeh S (2020) Time-frequency causality and connectedness between international prices of energy, food, industry, agriculture and metals. Energy Econ 85:104529. https://doi.org/10.1016/j.eneco.2019.104529

Tu M, Wang R, Guo W, Xu S, Zhu Y, Dong J, Yao X, Jiang L (2024) A CRISPR/Cas9-induced male-sterile line facilitating easy hybrid production in polyploid rapeseed (*Brassica napus*). Horticult Res 11 https://doi.org/10.1093/hr/uhae139

Vapnik VN (1997) The support vector method. In: International Conference on Artificial Neural Networks, Springer, Berlin, Heidelberg, pp 261–271. https://doi.org/10.1007/BFb0020166

Verbakel JY, Steyerberg EW, Uno H, De Cock B, Wynants L, Collins GS, Van Calster B (2020) ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. J Clin Epidemiol 126:207–216. https://doi.org/10.1016/j.jclinepi.2020.01.028

Vykydalová L, Barroso PM, Děkanovský I, Neoralová M, Lumbantobing YR, Winkler J (2024) Interactions between weeds, pathogen symptoms and winter rapeseed stand structure. Agronomy 14:2273. https://doi.org/10.3390/agronomy14102273

Walia S, Kumar R (2023) Wild marigold (*Tagetes minuta* L.) biomass and essential oil composition modulated by weed management techniques. Ind Crops Prod 161:113183. https://doi.org/10.1016/j.indcrop.2020.113183

Wan JZ, Wang CJ (2019) Contribution of environmental factors toward distribution of ten most dangerous weed species globally. Appl Ecol Environ Res 17:14835–14846 .https://doi.org/10.1016/j.gecco.2020.e01142

Wang X, Liu X, Wang L, Yang J, Wan X, Liang T (2022) A holistic assessment of spatiotemporal variation, driving factors, and risks influencing river water quality in the northeastern Qinghai-Tibet Plateau. Sci Total Environ 851:157942. https://doi.org/10.1016/j.scitotenv.2022.157942

Wang ZW, Yin J, Wang X, Chen Y, Mao ZK, Lin F, Wang XG (2023) Habitat suitability evaluation of invasive plant species Datura stramonium in Liaoning Province: Based on Biomod2 combination model. J Appl Ecol 34:1272–1280. https://doi.org/10.13287/j.1001-9332.202305.017

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, NCEAS Predicting Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. Divers Distrib 14(5):763-773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Xiao X, Ma H, Gan G, Li Q, Zhang B, Xia S (2024) Robust k-Means-Type Clustering for Noisy Data. IEEE Trans Neural Netw Learn Syst PP:1–1. https://doi.org/10.1109/TNNLS.2024.3392211

Xu P, Liang S, Hahn A, Zhao VT, Lo WT, Haller BC, Sobkowiak B, Chitwood MH, Colijn C, Cohen T, Rhee KY, Messer PW, Wells MT, Clark AG, Kim J (2024) e3SIM: epidemiological-ecological-evolutionary simulation framework for genomic epidemiology. bioRxiv. https://doi.org/10.1101/2024.06.29.601123

Yang X, Zhang X, Zhang P, Bidegain G, Dong J, Hu C, Li M, Zhang Z, Guo H (2023) Ensemble habitat suitability modeling for predicting optimal sites for eelgrass (*Zostera marina*) in

the tidal lagoon ecosystem: Implications for restoration and conservation. J Environ Manag 330:117108. https://doi.org/10.1016/j.jenvman.2022.117108

Yao W, Nan F, Li Y, Li Y, Liang P, Zhao C (2023) Effects of different afforestation years on soil properties and quality. For 14:329. https://doi.org/10.3390/f14020329

Yu K, Sun L, Chen J, Reynolds M, Chaudhary T, Batmanghelich K (2024) DrasCLR: A self-supervised framework of learning disease-related and anatomy-specific representation for 3D lung CT images. Med Image Anal 92:103062.https://doi.org/10.1016/j.media.2023.103062

Zhang L, Du H, Yang Z, Song T, Zeng F, Peng W, Huang G (2022) Topography and soil properties determine biomass and productivity indirectly via community structural and species diversity in karst forest, Southwest China. Sustain 14:7644. https://doi.org/10.3390/su14137644

Zhou T, Geng Y, Ji C, Xu X, Wang H, Pan J, Bumberger J, Haase D, Lausch A (2021) Prediction of soil organic carbon and the C ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. Sci Total Environ 755:142661. https://doi.org/10.1016/j.scitotenv.2020.142661

Table 1. Reviewing the fields of any county in Fars province

Area under rapeseed cultivation (ha) in each county	**Number of fields measured**
< 500	2
500 to 1,000	3
1,000 to 5,000	4
5,000 to 1,0000	6
10,000 to 15,000	8
15,000 to 30,000	11
30,000 to 60,000	15
> 60,000	one field added to 15 for each 10,000 ha

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Table 2. The receiver operating characteristic (ROC) Curve Classification (Richardson et al., 2024)

Poor	Moderate	Good	Very good	Excellent
0.5 – 0.6	0.6 – 0.7	0.7 – 0.8	0.8 – 0.9	0.9 - 1

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Table 3. Frequency (%) of weeds in rapeseed fields

Number	Weed	Family	Frequency (%)
1	*Avena fatua* L.	Poaceae	58.48
2	*Sinapis arvensis* L.	Brassicaceae	45.68
3	*Malva neglecta* Wallr.	Malvaceae	36.62
4	*Chenopodium album* L.	Chenopodiaceae	34.46
5	*Convolvulus arvensis* L.	Convolvulaceae	31.64
6	*Sinapis arvensis* L.	Brassicaceae	29.61
7	*Centaurea depressa* M. Bieb.	Asteraceae	25.41
8	*Cerastium perfoliatum* L.	Caryophillaceae	24.46
9	*Triticum aestivum* L.	Gramineae	23.83
10	*Daucus carota* L.	Umbelferae	23.75
11	*Capsella bursa-pastoris* Medik.	Brassicaceae	23.81
12	*Descurainia sophia* (L.) Webb ex Prantl.	Brassicaceae	23.28
13	*Carthamus glaucus* M. Bieb.	Compositae	23.19
14	*Spergula arvensis* L.	Caryophyllaceae	23.01
15	*Trifolium pratense* L.	Leguminosae	22.70
16	*Tragopogoh collinus*	Asteraceae	22.67
17	*Amaranthus retroflexus L.*	Amaranthaceae	22.57
18	*Cynodon dactylon* (L.) Pers.	Poaceae	22.33
19	*Cardaria draba* (L.) Desv.	Brassicaceae	22.32
20	*Hordeum spontaneum* K. Koch	poaceae	19.76
21	*Sonchus oleraceus* L.	Compositae	19.55
22	*Eruca sativa* Lam.	Brassicaceae	16.53
23	*Echinochloa crus-galli* (L.) P. Beauv.	poaceae	15.50
24	*Portulaca oleraceae* L.	portulacaceae	15.44
25	*Eleusine indica* (L.) Gaertn.	Poaceae	14.24
26	*Suaeda aegyptiaca* (Hasselq.) Zohary	Amaranthaceae	14.23
27	*Eruca sativa* Lam.	Brassicaceae	14.19
28	*Centaurea depressa* M. Bieb.	Asteraceae	13.12
29	*Galium aparine* L.	Labiaceae	10.99
30	*Carduus nutans* L.	Compositae	10.96
31	*Tribulus terrestris* L.	Zigophalaceae	10.77
32	*Calendula arvensis* L.	Compositae	10.77

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Table 4. Variance inflation factor

	Collinearity statistics	
Factors	**Tolerance**	**Variance inflation factor (VIF)**
Slope aspect	0.259	1.796
Clay (%)	0.278	2.023
Elevation/DEM (m)	0.632	1.528
Electrical conductivity (EC) (ds/m)	0.566	3.000
pH	0.323	1.747
Plan curvature (100/m)	0.199	1.842
Mean annual rainfall (mm)	0.625	1.533
Distance from rivers (m)	0.627	1.531
Distance from roads (m)	0.289	2.002
Sand (%)	0.748	3.000
Silt (%)	0.600	1.549
Mean annual temperature (°C)	0.657	1.512
Slope degree	0.829	2.068

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Table 5. Habitat suitability classes areas for all applied models

Models	Classes	Relative area (%)
Random forest (RF)	Low	66.56
	Moderate	16.35
	High	11.71
	Very high	5.38
Boosted regression tree (BRT)	Low	56.65
	Moderate	26.93
	High	11.87
	Very high	4.55
Support vector machine (SVM)	Low	35.57
	Moderate	37.89
	High	19.81
	Very high	6.74

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Table 6. Area under the curve

Test Result Variable(s)	Area	Standard Error	Asymptotic Significant	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Boosted regression tree (BRT)	0.97	0.01	.00	0.96	0.99
Random forest (RF)	0.99	0.01	.00	0.98	1.000
Support vector machine (SVM)	0.96	0.01	.00	0.94	0.98

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

Table 7. Examining the Significance of Variables using the Boruta Algorithm

	Mean Importance	Mean Importance	Median Importance	Maximum Importance	Decision
Elevation/DEM*	-0.77	-0.93	-2.52	1.27	Confiremed
Aspect	13.28	13.20	11.79	14.60	Confiremed
Clay percent	19.61	19.69	18.42	20.93	Confiremed
EC*	10.87	10.73	10.23	11.98	Confiremed
pH	9.02	9.25	9.14	9.87	Confiremed
Plan curvature	20.77	20.60	19.99	21.73	Confiremed
Annual mean Rainfall	10.07	10.61	8.42	11.34	Confiremed
Distance from River	11.12	11.17	10.19	12.56	Confiremed
Sand percent	7.27	7.37	5.79	8.90	Confiremed
Silt percent	14.41	13.97	12.39	16.52	Confiremed
Slope degree	30.93	30.98	29.42	32.58	Confiremed
Annual mean Temperature	14.75	14.89	13.50	16.25	Confiremed

*DEM (Digital elevation model); *EC (Electrical conductivity)

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

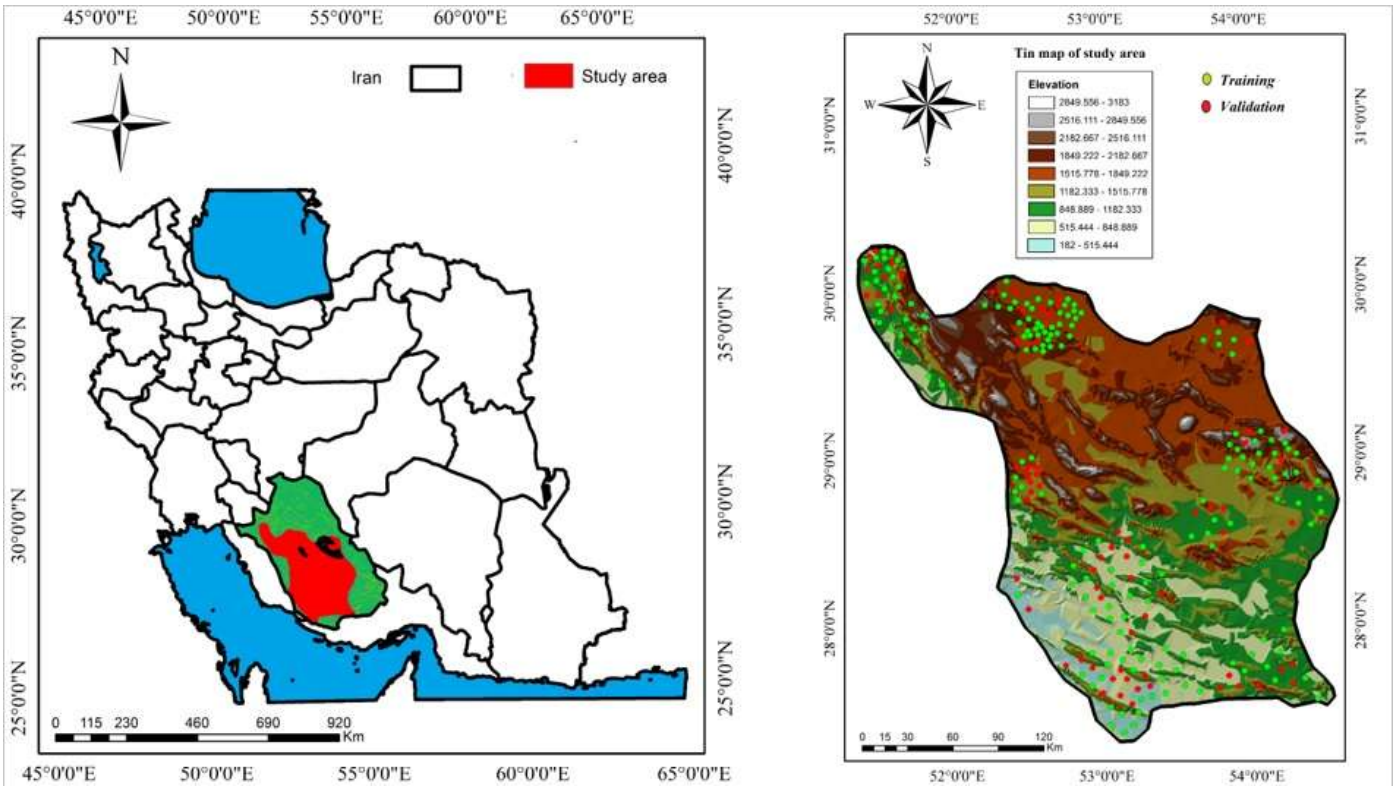


Figure 1. The research region is situated in Iran's Fars Province

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

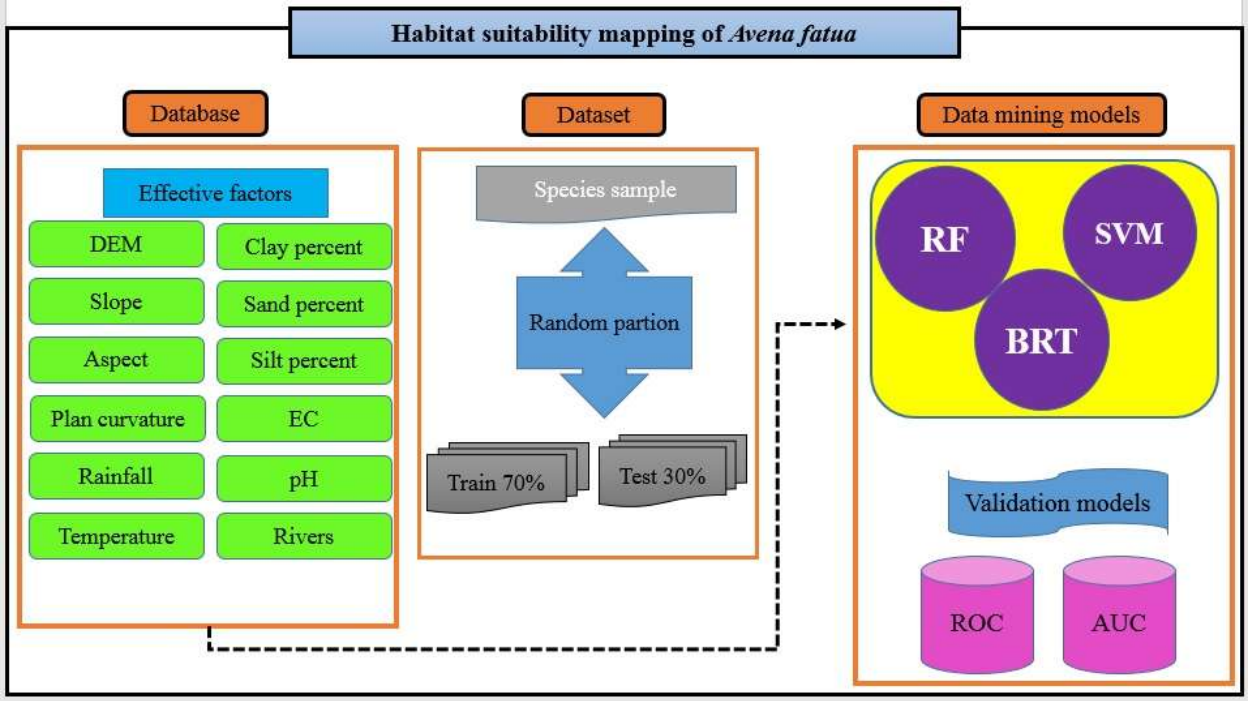
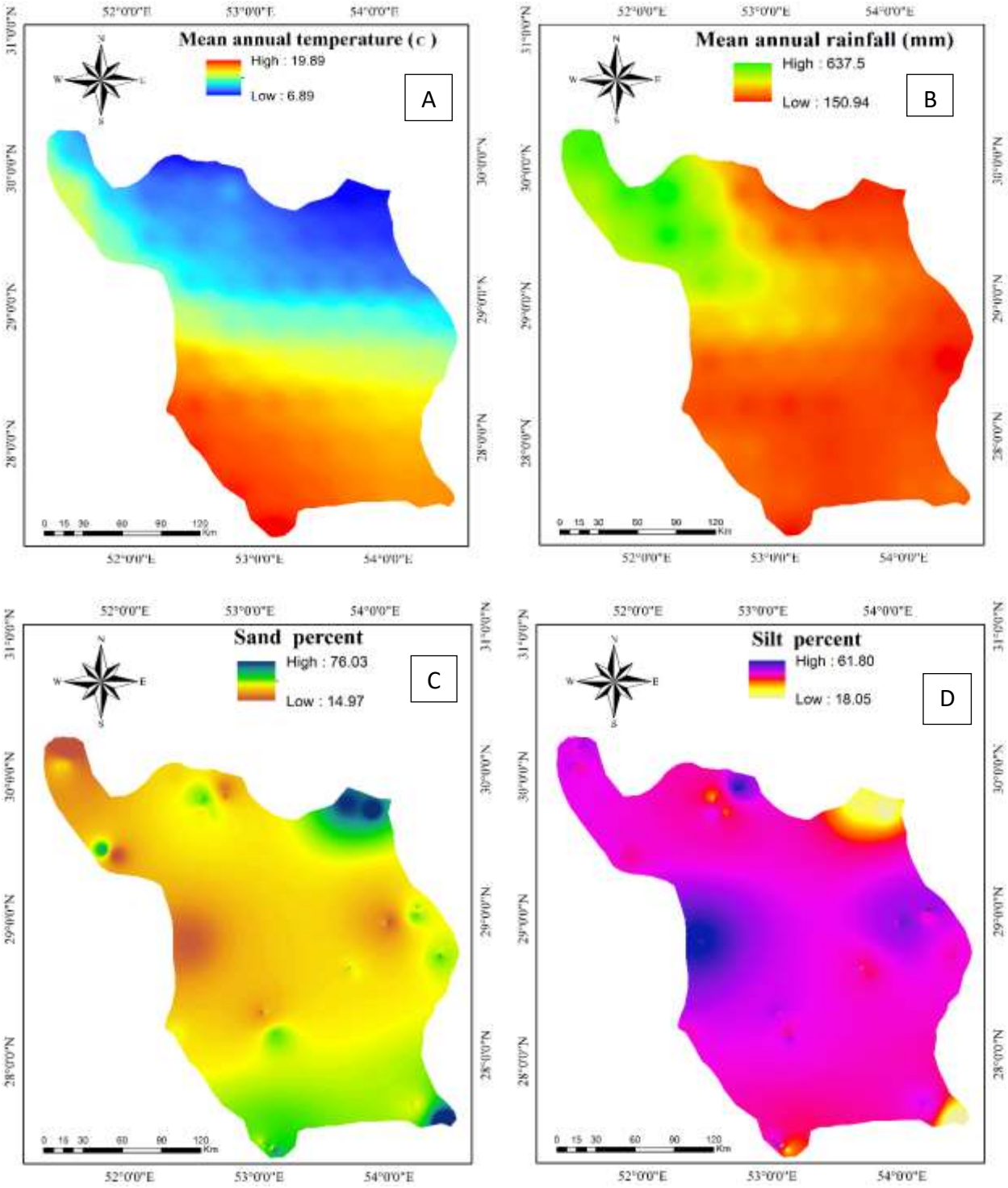


Figure 2. An *Avena fatua* habitat suitability mapping flowchart (*DEM (Digital elevation model); *EC (Electrical conductivity)

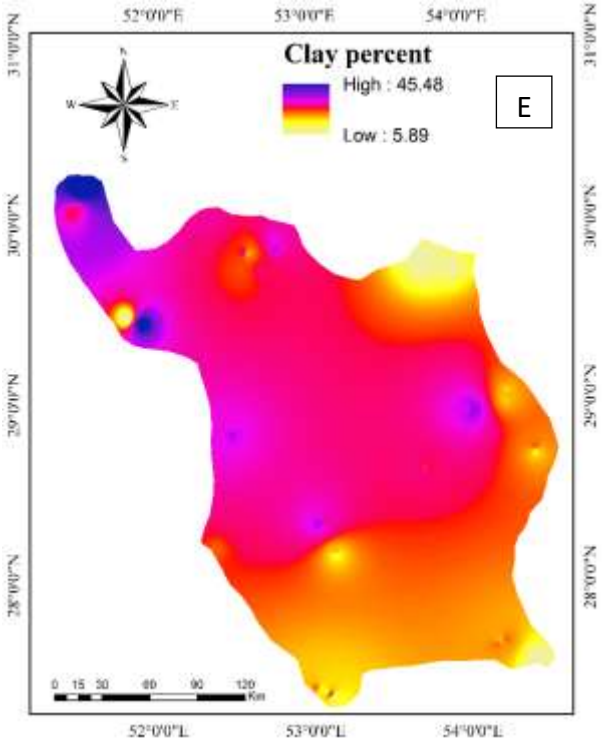
https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

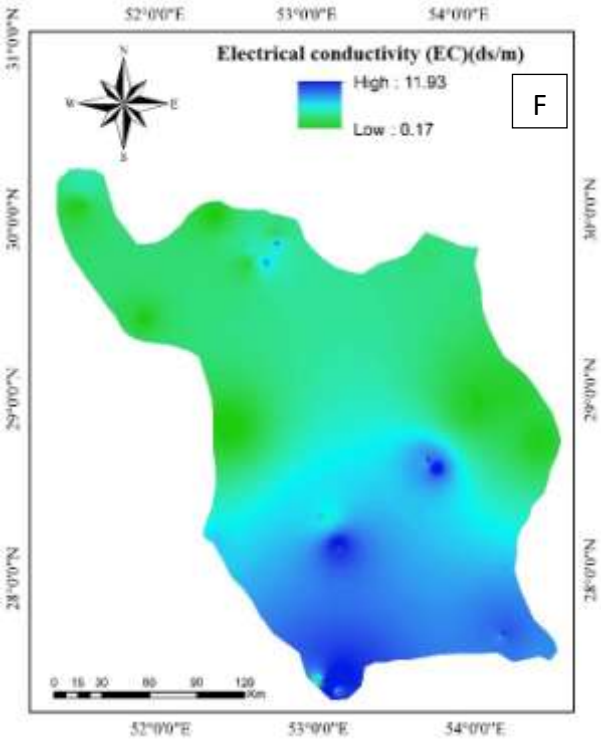
Figure 3. Spatial distribution of rapeseed, and weed samp

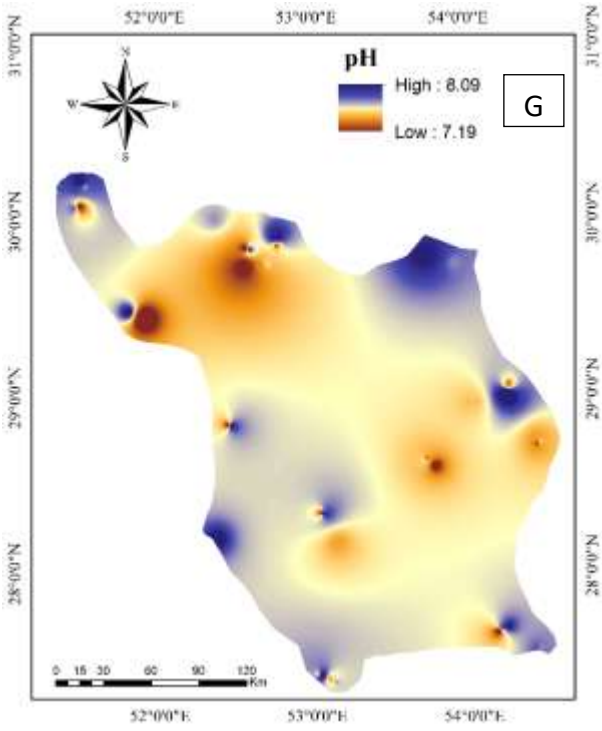
https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

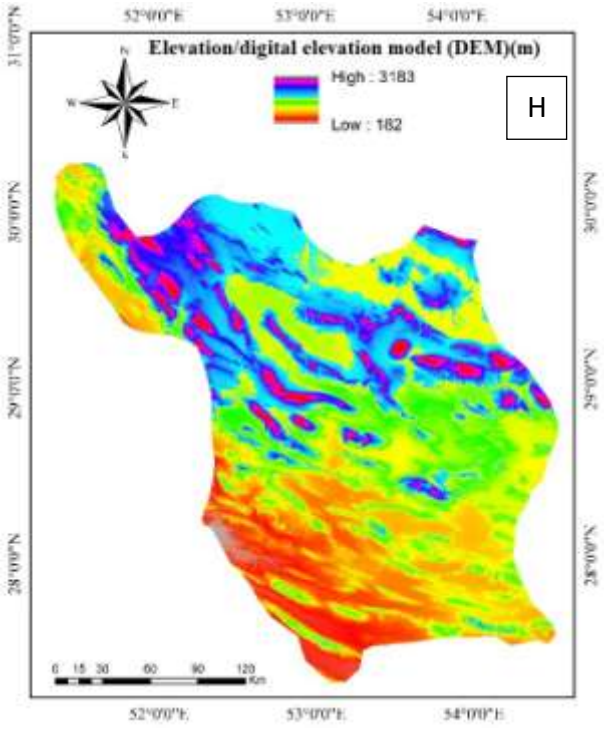


https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press









https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

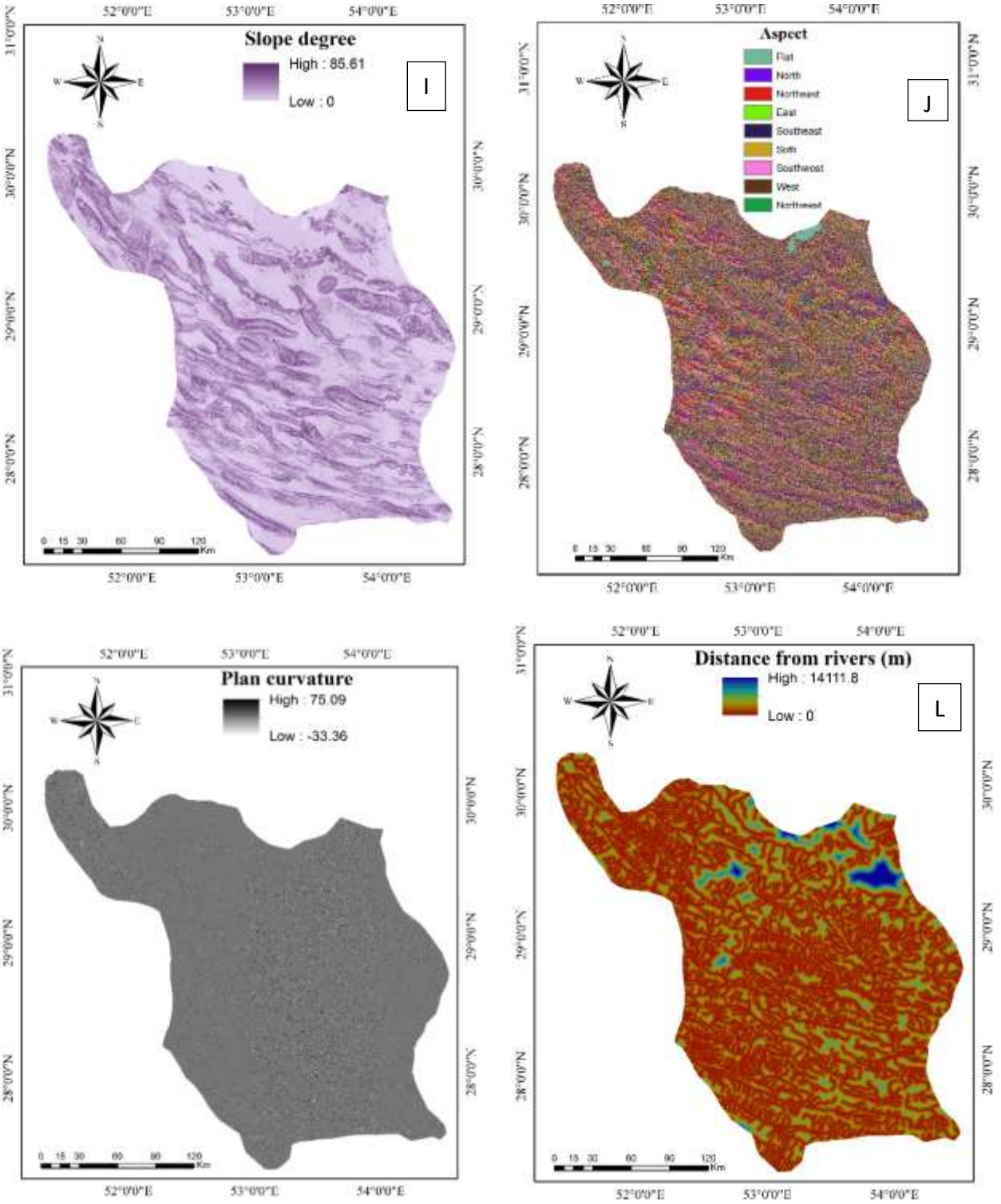
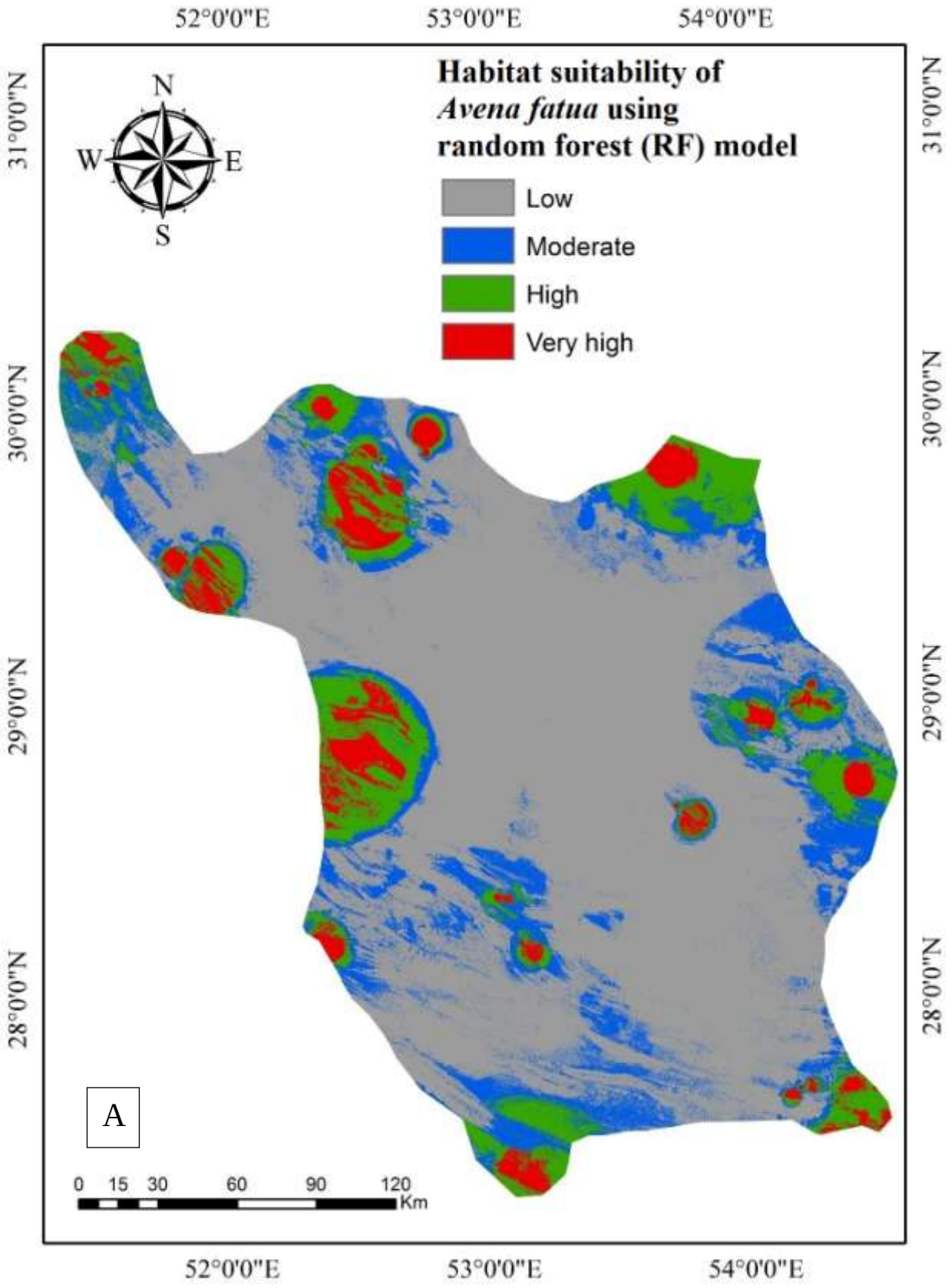
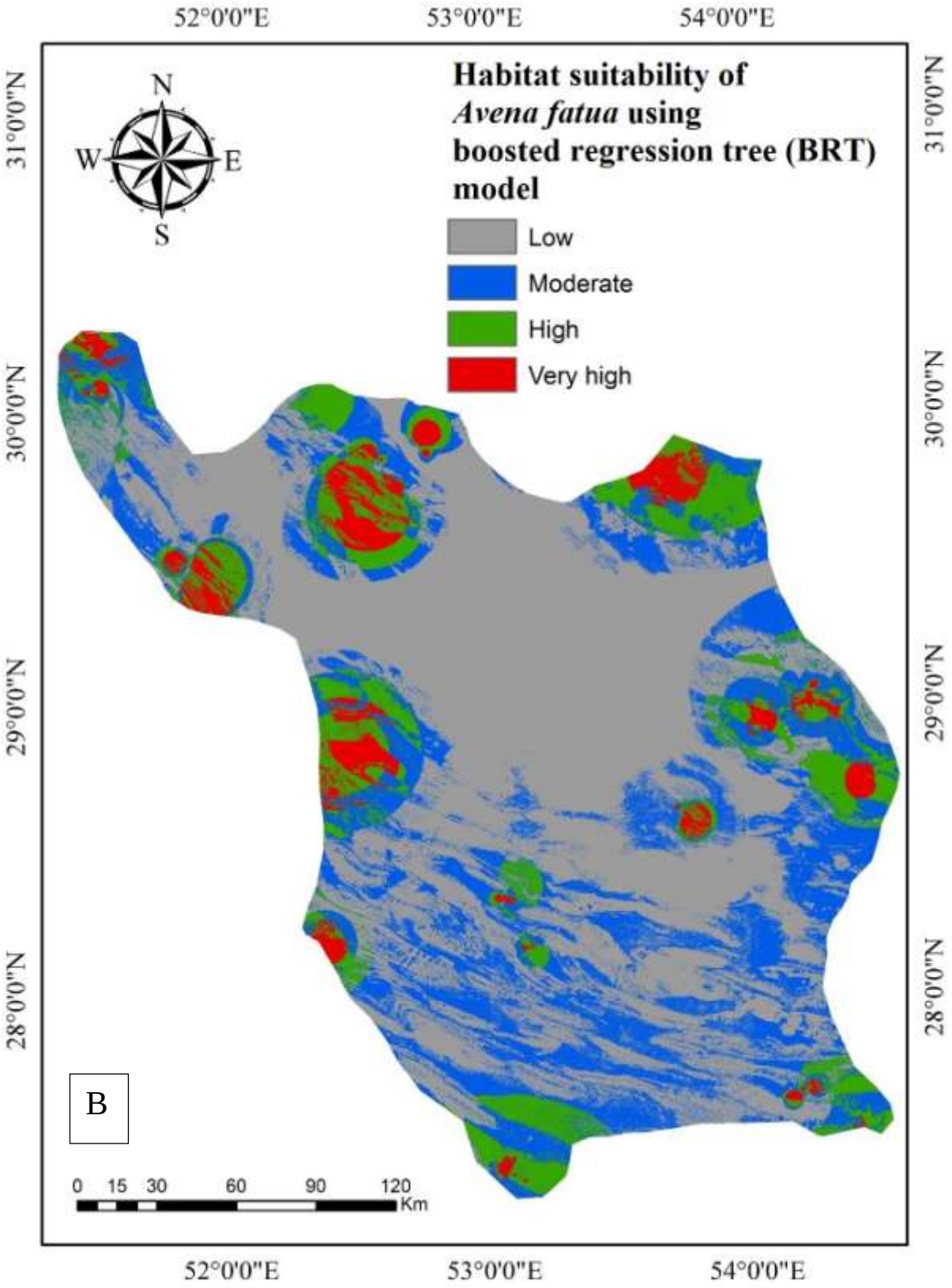


Figure 4. Important layers, including: "mean annual temperature (A)", "mean annual precipitation (B)", " sand percent (C)", "silt percent (D)", "clay percent (E)", "electrical conductivity (EC) (F)", "pH (G)", "elevation/digital elevation model (H)", "slope degree (I)", "slope aspect (J)", "plan curvature (K)" and "distance from rivers (L)".

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press



https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press



https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

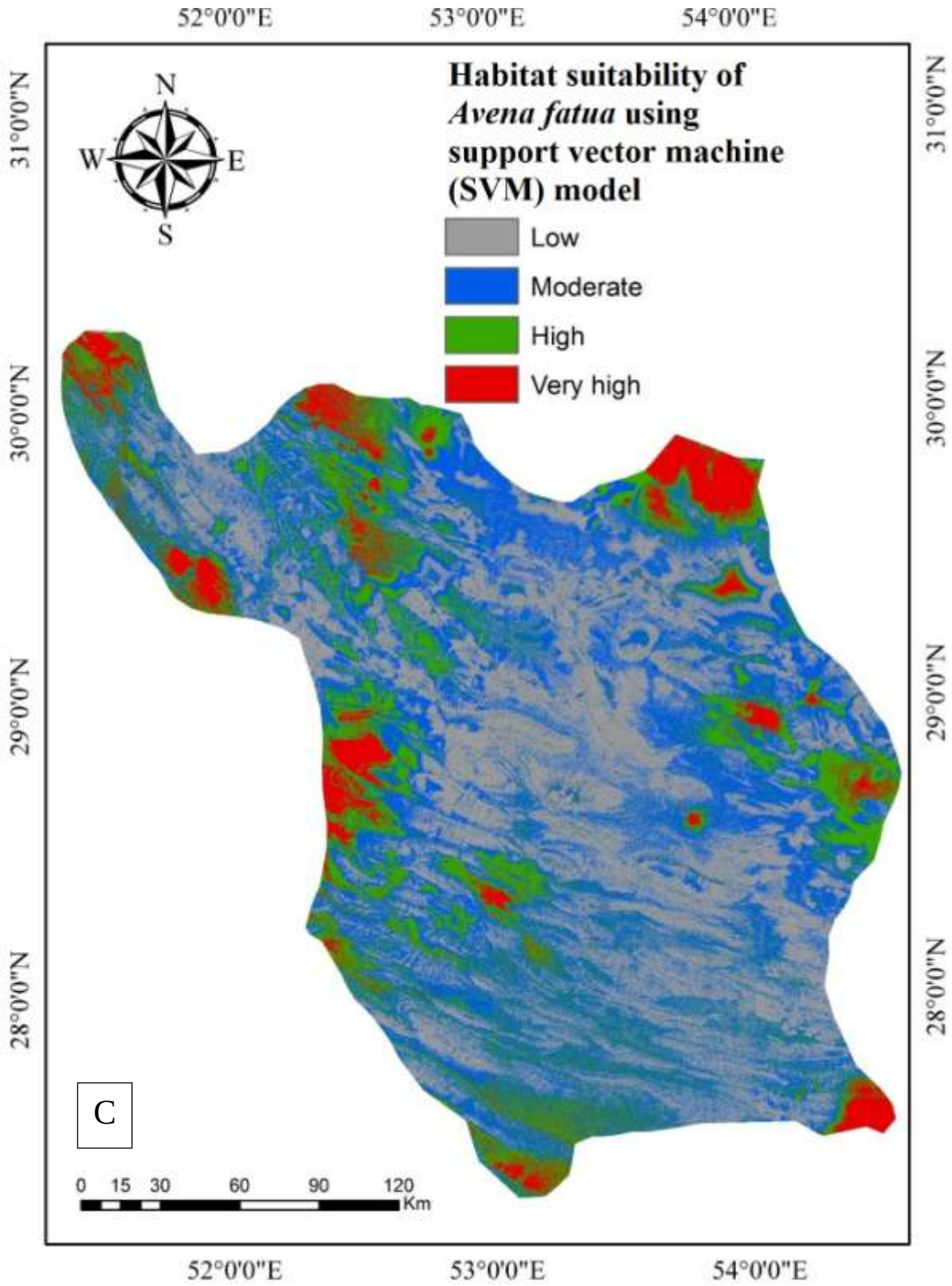


Figure 5. Habitat suitability maps of *Avena fatua* based on random forest (RF)(A), boosted regresion trees (BRT)(B), support vetor machine (SVM) (C)

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press

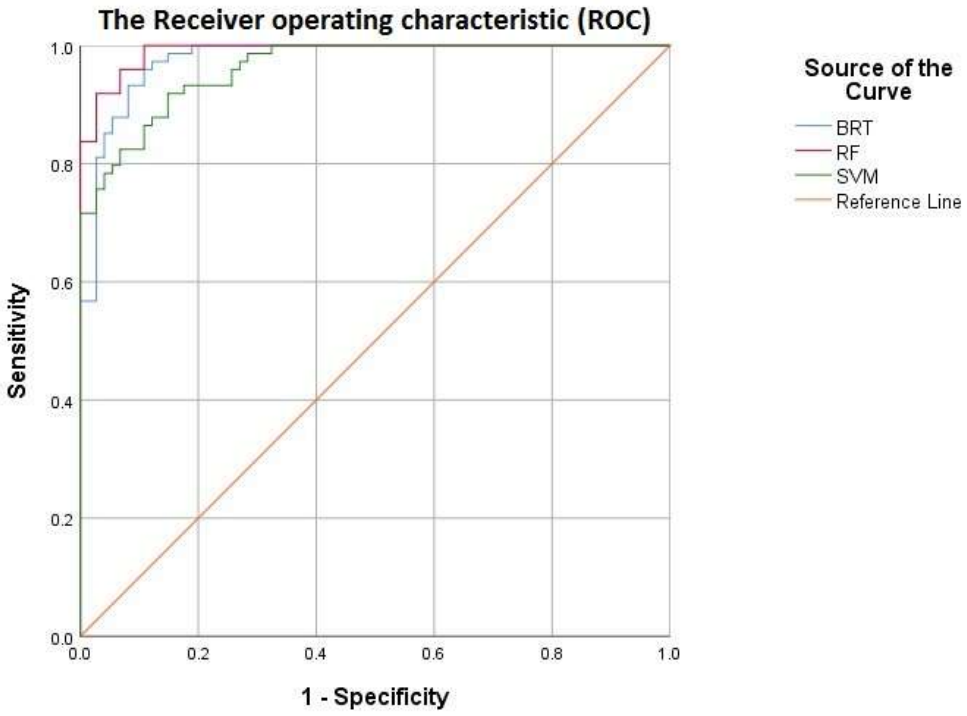


Figure 6. The Receiver operating characteristic (ROC) curve for evaluating algorithms

https://doi.org/10.1017/wsc.2025.5 Published online by Cambridge University Press