Theoretical and Applied Climatology (2025) 156:45
https://doi.org/10.1007/s00704-024-05289-8

RESEARCH

Mapping the potential distribution and invasion risk of Watermelon mosaic virus using MaxEnt ecological niche modeling

Kayo Heberth de Brito Reis[1] · Mayara Moledo Picanço[2] · Poliana Silvestre Pereira[1] · Hugo Daniel Dias de Souza[1] · Mônica Carvalho de Sá[3] · George Correa Amaro[4] · Ricardo Siqueira da Silva[3,5] · Marcelo Coutinho Picanço[2] · Renato Almeida Sarmento[1]

Received: 3 August 2024 / Accepted: 25 October 2024

Abstract

Watermelon mosaic virus (WMV) frequently infects crops in the Cucurbitaceae family, posing a significant challenge in their production. Managing viruses in crops remain a challenge, primarily due to the limited number of available strategies. The most effective strategy for controlling WMV is to prevent its introduction into regions currently free of the disease. To achieve this, it is necessary to map the locations where the WMV is present and identify areas at risk of invasion. This can be achieved through maximum entropy modeling (Maxent). This study aimed to map the countries with potential distribution for WMV and determine the environmental factors related to its ecological niche. The generated model was robust and reliable according to the 21 metrics used to evaluate it. The response curves of the selected variables revealed that the survival of WMV is directly linked to specific conditions of temperature, precipitation, and altitude, with the virus having a higher probability of survival in warm regions, at altitudes below 1000 m, and with good rainfall availability. The suitability map showed that 46.08% of the planet presents some probability of WMV survival, with the areas of highest probability located in countries in southern Europe, as well as in the United States, Brazil, Argentina, China, Turkey, and Iran. Additionally, the climate zoning map indicated that WMV occurs most frequently in areas classified as Cfa (humid subtropical), Csa (Mediterranean), Aw (tropical savanna), and BSk (cold semi-arid) according to the Köppen-Geiger classification.

✉ Kayo Heberth de Brito Reis
kayoheberthdebritoreis@gmail.com

Mayara Moledo Picanço
mayara.moledo@ufv.br

Poliana Silvestre Pereira
poliana_silvestre@yahoo.com.br

Hugo Daniel Dias de Souza
hugo.daniel@mail.uft.edu.br

Mônica Carvalho de Sá
monica.carvalho@ufvjm.edu.br

George Correa Amaro
george.amaro@embrapa.br

Ricardo Siqueira da Silva
ricardo.siqueira@ufvjm.edu.br

Marcelo Coutinho Picanço
picanco@ufv.br

Renato Almeida Sarmento
rsarmento@uft.edu.br

[1] Graduate Program in Plant Production, Federal University of Tocantins, Gurupi, Tocantins 77402-970, Brazil

[2] Department of Entomology, Federal University of Viçosa, Viçosa, Minas Gerais 36570-900, Brazil

[3] Department of Agronomy, Federal University of Vales of the Jequitinhonha and Mucuri, Diamantina, Minas Gerais 39100-000, Brazil

[4] Embrapa Roraima, Boa Vista, Roraima 69301-970, Brazil

[5] Department of Ecological Modelling, Helmholtz Centre for Environmental Research—UFZ Leipzig, Permoserstr. 15, 04318 Leipzig, Germany

Published online: 21 December 2024

1 Introduction

Cucurbits are a family of plants with a global production exceeding 200 million tons (FAO 2021). Despite their economic importance, cucurbits face severe phytosanitary problems in many countries, including the incidence of Watermelon mosaic virus (WMV), one of the most critical viruses significantly impacting these crops (Lecoq and Desbiez 2008). The economic impact caused by the pathogen depends on the timing of the infection. In melon crops, for example, production losses can reach 50% if the virus is introduced before flowering (Alonso-Prados et al. 1997).

WMV is a pathogen belonging to the genus *Potyvirus* and the family *Potyviridae* (Yakoubi et al. 2008). In addition to cucurbits, this virus infects plant species from other families (Ayazpour and Vahidian 2016). Its primary means of dissemination is through aphids in a non-persistent manner (Lecoq and Desbiez 2008). The most common symptoms of the infection include mosaic patterns on the leaves, plant atrophy, deformations of leaves and fruits, and pale coloration of the fruits, all of which reduce their commercial value (Moradi et al. 2012).

The lack of information about the ecology and distribution of WMV has contributed to its rapid spread, leading to significant annual productivity losses worldwide (Mumford et al. 2016). Once introduced to a new location, managing the virus becomes challenging (Jones 2006). Therefore, one of the most effective strategies to deal with WMV is to prevent its introduction in areas where it is not yet present. This can be achieved through legislative control (Rubio et al. 2020). Although it faces challenges in implementation, such as lack of infrastructure, insufficient international cooperation, and even interference from economic interests—especially in countries with limited resources—it remains the primary tool for controlling viral pests.

Legislative control aims to prevent the entry of exotic pests into areas free from their occurrence through actions by Phytosanitary Defense Bodies (Gallo et al. 2002). For these actions to be effective, it is essential to know where the pest is established and to map areas with potential for invasion (Waage and Mumford 2008). This mapping is fundamental because identifying regions at risk of new species invasion can prevent economic damage (Amaro et al. 2022). This measure aids in decision-making related to the implementation of phytosanitary measures that prevent new cases of biological invasion (Bradshaw et al. 2016). Determining the invasion potential of an organism not only supports legislative control but also provides information to guide genetic improvement programs in developing cultivars resistant to potential pest introductions.

Factors influencing the probability of biological invasion include both positive elements, such as ideal weather conditions, and detrimental ones, such as natural physical barriers or extreme weather conditions (Elith and Leathwick 2009; Elith and Franklin 2013). Understanding and interpreting these abiotic and biotic factors allows us to 'track' potential invasive species (Thorso et al. 2016) and map pest-free areas with potential for invasion.

Maxent is a popular tool for predicting an organism's distribution potential. It uses environmental variables to forecast areas with climatic suitability (Phillips and Elith 2010; Galdino et al. 2016) and performs well with small datasets (Pearce and Boyce 2006; Pearson et al. 2007). Maxent can also predict suitable locations under climate change scenarios (Warren and Seifert 2011) by combining ecological niche and species distribution models to predict the organism's occurrence points (Bentlage et al. 2013). These predictions are made through the creation of species distribution models (SDMs), which are essential tools that support and guide the development and implementation of environmental policies, phytosanitary measures, and management programs (Addison et al. 2013; Martin et al. 2020), including legislative control efforts.

Although the Watermelon mosaic virus is one of the most widespread viruses affecting cucurbits globally and poses challenges for control, there are limited studies in the literature that determine its global distribution potential or the environmental characteristics associated with its spread. Therefore, this study aimed to use Maxent to determine the global distribution potential of WMV and identify the environmental factors that facilitate its introduction and survival.

2 Material and methods

2.1 Occurrence data

The occurrence data for the *Watermelon mosaic virus* was obtained from the Global Biodiversity Information Facility (GBIF) database using the {rgbif} package, version 3.8.1 (Chamberlain et al. 2022), accessed on 09/05/2024. These data were also supplemented with information from published studies and field research conducted across various regions and climates.

Procedures were implemented for data cleaning (Hijmans and Elith 2013; Zizka et al. 2019; Ribeiro et al. 2022). These procedures included: a) retaining only records with a spatial resolution of ≤ 1 km for analysis; b) removing occurrence records within a radius of 10 km from the centers of capital cities and 5 km from the centers of countries, states, and municipalities; c) eliminating records with identical longitude and latitude, a 0.5-degree radius around the GBIF headquarters, duplicated coordinates, and zero values; and d) discarding records located over water or those not associated with all selected environmental variables.

Accounting for sampling bias in species distribution models, whether presence-only or presence-background models, is one of the greatest challenges in the modeling process. Regardless of the model type chosen, the use of biased data will obscure the true relationship between occurrences and predictor variables (Barber et al. 2022; Schartel and Cao 2024). An environmental filter (Castellanos et al. 2019; Velazco et al. 2022) was applied to reduce sampling bias. These environmental filters are sensitive to the bin size; thus, four bin sizes (3, 4, 5, and 7) were tested. For each bin size, spatial autocorrelation was calculated among the filtered records based on Moran's I and the number of filtered records. Then, the number of bins was selected based on the lower quartile of Moran's I, choosing the one with the highest number of records among those (Velazco et al. 2021). A regular multidimensional grid was subsequently created in the environmental space defined by the predictor variables. The cell size of this grid was determined by the number of selected bins to divide the range of variable values into class intervals (Varela et al. 2014; Castellanos et al. 2019). After that, a single occurrence was randomly selected within each cell of the grid.

The partitioning of occurrence data for model performance evaluation was done using spatial block cross-validation, as this method allows for better control of potential spatial autocorrelation between training and test data of the model, and it assesses transferability more adequately compared to other partitioning methods (Roberts et al. 2017; Valavi et al. 2019). Methods for partitioning geographically structured data are particularly useful for assessing the transferability of models to different regions or time periods (Roberts et al. 2017; Santini et al. 2021).

To select the best grid size (square blocks, similar to a checkerboard pattern), 30 grids were generated with resolutions ranging from 0.5 (~ 56 km) to 5 degrees (~ 557 km), in four partitions, with a minimum of five occurrences per partition and using 80% of the presences to test autocorrelation. The grid selected had: a) the lowest spatial autocorrelation, as measured by Moran's I; b) the maximum environmental similarity, considering Euclidean distance; and c) the minimal difference in the number of records between training and testing data, as indicated by the standard deviation (Velazco et al. 2019).

2.2 Environmental data

A set of 19 bioclimatic variables derived from temperature and precipitation data from the WorldClim database version 2.1 (Fick and Hijmans 2017) was used, with spatial resolutions of 30 s (~ 1 km at the equator) and 2.5 min (~ 5 km at the equator) for the years 1970–2000. These were obtained using the {geodata} package version 0.6–2 (Hijmans et al. 2023) to represent current climatic conditions, as they are capable of capturing annual variations and limiting factors known to influence the geographic distribution of species (O'Donnell and Ignizio 2012).

An elevation variable was added, with its primary source being the Shuttle Radar Topography Mission (SRTM). The data are available between −60º and 60º latitude, supplemented with GTOP30 data for higher latitudes (> 60º). Additionally, a variable related to the global harvested area (ha) of watermelon in the year 2020 was included from the CROPGRIDS dataset (Tang et al. 2024), where the values represent the average number of hectares harvested per grid cell area.

Collinearity is a problem for creating models, especially when there is interest in project them to a new geographic location or time, particularly if the correlation structure between the variables is not constant. As a result, not all predictor variables are used for fine-tuning the final model. Therefore, the selection of variables for the model during the modeling procedure was carried out through an iterative (data-driven) process based on adjustments and refinements of Maxent models. The resulting variables were evaluated and supplemented by considering their biological relevance.

2.3 Background selection and calibration area

The calibration area (CA) was considered equivalent to the M region of the BAM Diagram (Soberon and Peterson 2005; Phillips et al. 2009; Soberón 2010; Elith et al. 2011; Owens et al. 2013). The accessible area approach of the BAM framework was used, meaning the CA is the theoretical target to define the area accessible to the species. These areas depend on opportunities and constraints on the movement of species (M), including areas where the species might potentially be present (Soberón, 2010; Barve et al. 2011; Cooper and Soberón, 2018; Mendes et al. 2020).

The size of the calibration area affects the model's performance metrics. The models' discrimination ability (i.e., the ability to correctly distinguish between presence and absence localities), for example, usually increases with the size of the calibration area (Anderson and Raza 2010; Barbet-Massin et al. 2012; Amaro et al. 2023). This mainly happens because larger areas tend to include absences that are ecologically more distant from presences, making them easier to distinguish (Lobo et al. 2008; Vanderwal et al. 2009). The model's ability to predict the probability of occurrence decreases with the size of the calibration area, as larger areas tend to include regions far from the presence locations, which are irrelevant for inferring the interaction between the species and the environment (Acevedo et al. 2012). In certain situations, different calibration areas, considering the characteristics of the occurrences, may be useful to explore different dynamics of a phenomenon (Elith et al. 2011), meaning that other areas beyond those defined by the occurrences can be included. To achieve the goal of capturing the potential distribution, the location data used to develop the model should preferably be drawn from the widest possible geographical and environmental range (Jarnevich et al. 2015), as long as scientific criteria are used to define the extent and boundaries of the calibration area (Sillero et al. 2021).

Using the biogeographic entities method (Rojas-Soto et al. 2024), Köppen-Geiger climate zones were used to delimit the calibration area (Brunel et al. 2010; Webber et al. 2011; Hill and Terblanche 2014; Marchioro 2016; Hill et al. 2017; Datta et al. 2019) as biotic regions. These regions are climatic and geographical units that share the species' environmental and historical adaptations (Barve et al. 2011) to create a continuous polygon, considering areas that contain at least one occurrence. WMV is an invasive species, so considering that its dispersal is influenced by human activity (Pyšek et al. 2020), occurrences from both native and invaded areas were used (Broennimann and Guisan 2008; Beaumont et al. 2009; Zhang et al. 2020), mainly because the species may be undergoing a niche climate shift during the invasion process. This allows for a clear representation of the species' dispersal ability and history (Barve et al. 2011).

Distribution models like Maxent, which are based on presence-background data, estimate the relative probability of presence by comparing occurrence locations with a background (an environmental context). This background consists of all locations in the calibration area—places where the species is present, as well as those without presence information (where its occurrence is unknown) (Phillips and Elith 2013; Halvorsen et al. 2015). The background sample should be selected in a way that reflects the environmental conditions of interest to contrast with the occurrences based on the spatial scale of the ecological questions at hand (Saupe et al. 2012). Although Maxent uses the default of 10,000 background points (Phillips and Dudík, 2008; Barbet-Massin et al. 2012), it is necessary for this number to be representative of the underlying environment. Therefore, it may not be suitable for representing very large areas with diverse environments (Renner and Warton 2013). There are studies where the number of points used ranged from 75,000 to 300,000 (El-Gabbas and Dormann 2018). Thus, considering the extent of the model's calibration area, 20,000 points were selected, randomly distributed across the calibration area and equally stratified to the presence points in each partition (Hirzel and Guisan 2002).

2.4 Species distribution models

All data processing procedures, model development, maps, and graphics were performed using R software, version 4.4.0 "Puppy Cup" (R Core Team 2023). The software was used within a fully automated framework, developed based on best practices and recommendations for species distribution modeling with Maxent (Sillero 2011; Merow et al. 2013; Jarnevich et al. 2015; Araújo et al. 2019; Low et al. 2021; Santini et al. 2021; Sillero and Barbosa 2021; Srivastava et al. 2021; Moudrý et al. 2024; Rojas-Soto et al. 2024). For spatial data analysis and transformation, the following packages were used: {terra} version 1.7–78 (Hijmans 2023) and {sf} version 1.0–16 (Pebesma 2018). Additionally, the following packages were employed: a) {ENMeval} 2.0.4 (Kass et al. 2021) for variable selection; b) {flexsdm} version 1.3.4 (Velazco et al. 2022) for all species distribution modeling procedures, using tools from {maxnet} version 0.1.4 (Phillips 2021); c) {pROC} version 1.18.5 (Robin et al. 2011) for ROC curve plots and estimates; d) {tmap} version 3.3–4 (Tennekes 2018) for plotting all resulting maps; e) {ggplot2} 3.5.1 (Wickham 2016) for visualizing various results.

The maximum entropy model (Maxent) was used through a non-homogeneous Poisson point process (Phillips et al. 2006; Renner and Warton 2013; Renner et al. 2015; Phillips 2017; Phillips et al. 2017). This model was adopted because it is one of the most widely used for modeling species distribution and has shown good performance compared to others (Elith et al. 2006; Elith et al. 2011; Heikkinen et al. 2012; Hijmans 2012; Venette 2017; Helmstetter et al. 2021; Valavi et al. 2022). An important feature of this model is that the original Maxent output can be interpreted as a model of relative species abundance and can be transformed into a probability of presence using a cloglog function (Peay et al. 2023).

Presence-background models like Maxent compare environmental conditions, represented by predictor variables (available in the calibration area and defined by background points), with the conditions used by the species, which are represented by its occurrences (Hirzel et al. 2002; Phillips et al. 2006; Phillips and Dudík, 2008). All background locations where the species has not been recorded are considered available but unused conditions. Thus, these models can distinguish between suitable and unsuitable habitats, not by the probability of species occurrence at a given location, but through a habitat suitability index, i.e., the quality of the habitat for the species' survival and persistence (Sillero 2011). This index, used to assess habitat quality, is specific to each modeling method (Acevedo et al. 2012). Therefore, a suitable habitat for the species does not necessarily guarantee its presence, while an unsuitable habitat does not always ensure its absence. For identifying a species' habitat—i.e., which locations meet the environmental requirements of the species within the study area—presence-background or presence-only methods are preferable (Sillero et al. 2021).

The two main parameters that must be adjusted in Maxent models are the regularization multiplier and the combinations of feature classes (Elith et al. 2011; Merow et al. 2013). The regularization multiplier (RM) determines the penalty associated with including variables or their transformations (features) in the model. Higher RM values impose a stronger penalty on model complexity, resulting in simpler (flatter) predictions. Features determine the potential shape of the marginal response curves. Therefore, a model that can only include linear classes, for example, will likely be simpler than a model that can include all possible features.

Features are transformations of the original predictor variables used to build the model. They can be linear (L), quadratic (Q), threshold (T), hinge (H), product (P), and categorical (Merow et al. 2013). Hinge features tend to make linear and threshold features redundant. One way to obtain a relatively smoother model, similar to a generalized additive model (GAM), is to use only hinge features (Elith et al. 2010; Elith et al. 2011). Excluding product features creates an additive model that is easier to interpret, although less capable of representing complex interactions (Elith et al. 2011). Therefore, the types of features to be used should consider the sample size (number of occurrences). Thus, we use features as follows: a) linear when there are fewer than 10 occurrences, b) linear and quadratic between 10 and 15 occurrences, c) linear, quadratic, and hinge when there are 15 to 80 occurrences, and d) all features when there are more than 80 recorded occurrences (Phillips and Dudík 2008; Elith et al. 2011; Merow et al. 2013). Thus, the first Maxent model (base model) was generated using RM = 1 and FC = LQHP (Maxent's default settings), along with a dataset containing the coordinates of presence and background points and the values of all predictor variables (30-s resolution) at those points. The objective of this initial model was to evaluate and select the most important variables for the final model by using five-fold cross-validation on random folds (fivefold cross-validation).

After developing the initial model, a data-driven variable selection was carried out starting from the base model, beginning with the variable with the highest contribution (permutation importance). If a variable was correlated with other variables, considering Spearman's rank coefficient >|0.7|, a Jackknife test was performed. Additionally, among the correlated variables, those that reduced model performance the least when removed, based on the True Skill Statistic (TSS) metric, were discarded. TSS is the sum of sensitivity (percentage of successfully predicted presence points) and specificity (percentage of successfully predicted background points) minus one (Allouche et al. 2006). This process was repeated until the remaining variables were no longer correlated at the fixed threshold (Vignali et al. 2020). After this, to optimize the model's parsimony, the maximum number of variables was removed while maintaining its performance. This procedure was evaluated based on a single-variable exclusion Jackknife test and according to the TSS metric, considering a threshold that kept only variables with a permutation importance greater than 2%. Reducing predictors is a beneficial procedure as it can limit overfitting, thereby resulting in a model with better generalization. This makes it possible to produce more accurate predictions for data not used during training (Vignali et al. 2020). Research results indicate that for projecting models under different conditions (such as other locations and under climate change), it is better to adopt a few variables with clear biological significance rather than many variables with uncertain effects on species distribution (Araújo and Guisan 2006; Austin and Van Niel 2011; Santini et al. 2021).

There are studies suggesting that different combinations of FC and RM should be tested to choose the best hyperparameter configurations specific to the species and datasets used in the modeling (Merow et al. 2013; Syfert et al. 2013; Radosavljevic and Anderson 2014; Moreno-Amat et al. 2015). Therefore, the final model was defined through fine-tuning of 152 models, consisting of eight feature classes (FC = "L", "H", "LQ", "QH", "LQH", "LQP", "QHP", "LQHP") and 19 values for the regularization multiplier (RM = 0.5 to 5, with increments of 0.25), considering only the previously selected variables. To identify the best combination of hyperparameters, the model with the highest TSS estimate was chosen, using a threshold that maximizes the sum of sensitivity and specificity – maxSSS (Liu et al. 2005; Liu et al. 2016), with values restricted to the training range (clamp) and clog-log output format, representing the estimated probability of occurrence between 0 and 1 (Phillips et al. 2017). The permutation importance of the variables for the model was estimated using the varImportance function from the {fitMaxnet} package (Wilson 2024).

To validate the performance of a model, it is necessary to use multiple evaluation metrics (Castellanos et al. 2019; Sofaer et al. 2019; Konowalik and Nosol 2021), as they may vary in their dependence on thresholds (Liu et al. 2009; Liu et al. 2013) and sensitivity to prevalence (Leroy et al. 2018). Therefore, various metrics were calculated to assess the classification, discrimination, and calibration ability of the final model, which was projected on a global scale. Probability maps of occurrence were generated considering local environmental conditions, with values ranging from 0 to 1 and a spatial resolution of 2.5 min. The projection result was divided into five fixed probability classes, classifying the likelihood of occurrence as: a) unsuitable (0–10%); b) marginal (10–20%); c) moderate (20–50%); d) optimal (50–80%); and e) high (80–100%). The area of each class was estimated and serves as the reference for assessing the effect of climate change.

For the application of this study in environmental management and phytosanitary policies, such as legislative control, presence/absence maps of WMV were created, which may be more useful when compared to the generated environmental suitability maps (Liu et al. 2013). However, with this discretization to produce binary maps, there is a possibility of losing information (Liu et al. 2016). Therefore, binary maps were also generated using the maxSSS threshold (Liu et al. 2005; Liu et al. 2016), resulting in two types of binary maps.

2.5 Map of the world's largest producers and climate zone map based on Köppen-Geiger

Data on the harvested area (in hectares) of the main cucurbits cultivated worldwide was collected from the FAO website. This data was used to map the most significant global producers based on harvested area, divided into four classes. The WMV occurrence points were then plotted on this map using ArcGIS software version 10.8 (Esri 2020). Furthermore, using R software, a climate zone map based on the Köppen-Geiger classification was generated from the filtered recorded occurrences. This map shows which climate types each WMV occurrence point falls into, along with a frequency histogram that highlights the most common climate types associated with the virus.

3 Results

3.1 Species distribution model performance and variable contribution

A total of 109 occurrence points of the Watermelon mosaic virus were collected from GBIF, and 277 points were obtained from field experiment articles. After the data "cleaning" process, 280 presence points were retained. Following this, these points were reduced to 231 after applying the environmental filter (4 bins, with Moran's I = 0.4017) (Fig. 1).

In Fig. 2, we have the covariates used, grouped according to hierarchical cluster analysis, where their correlation was measured using Spearman's rank correlation coefficient (ρ), based on their values at the occurrence coordinates.

The data-driven variable selection process, using Maxent with default parameters, resulted in seven variables: "Bio06," "Bio10," "Bio15," "Elev," "Bio08," "Bio02," "Bio19," "Bio18," and "Watermelon." In Table 1, we present the descriptive statistics of the selected variables and the others used, considering their values at the occurrence coordinates of Watermelon mosaic virus.

Based on the Köppen-Geiger climate zones, the calibration area of the model was defined, where occurrences of WMV were recorded. The total calibration area was estimated at 105,675,844 km^2, considering the vast dispersion of the pest.

The fine-tuning resulted in a Maxent model with FC = LQHP and RM = 0.5. The developed model demonstrated adequacy in discriminating between occurrences in the test dataset and background points, with the aim of identifying the potential distribution of the Watermelon mosaic virus. This model adequacy can be corroborated by some of the most commonly used metrics (Table 2), with the top five being: 1) TPR (Fielding and Bell 1997; Elith et al. 2006; Liu et al. 2013), which measures the model's sensitivity, allowing for the identification of 83% of the areas of actual occurrence of the species; 2) TNR (Hanley and McNeil 1982; Fawcett 2006), which evaluates specificity and allows for the identification of 73% of the areas where the species is not present; 3) TSS (Allouche et al. 2006), which assesses the

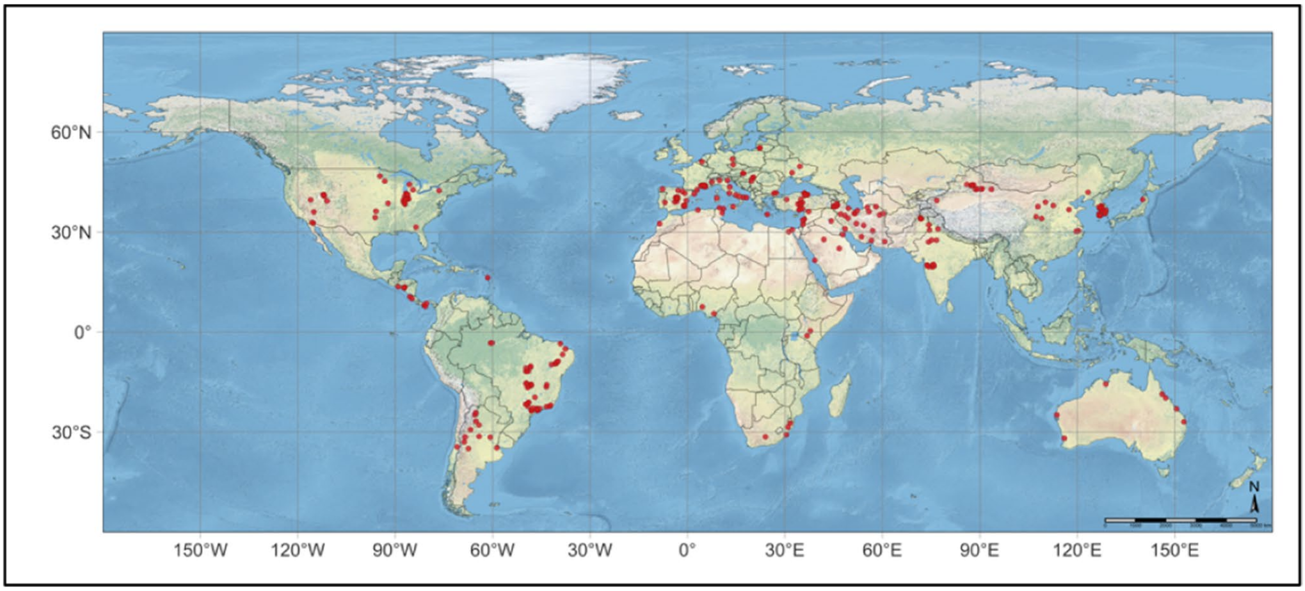


Fig. 1 Global distribution of Watermelon mosaic virus

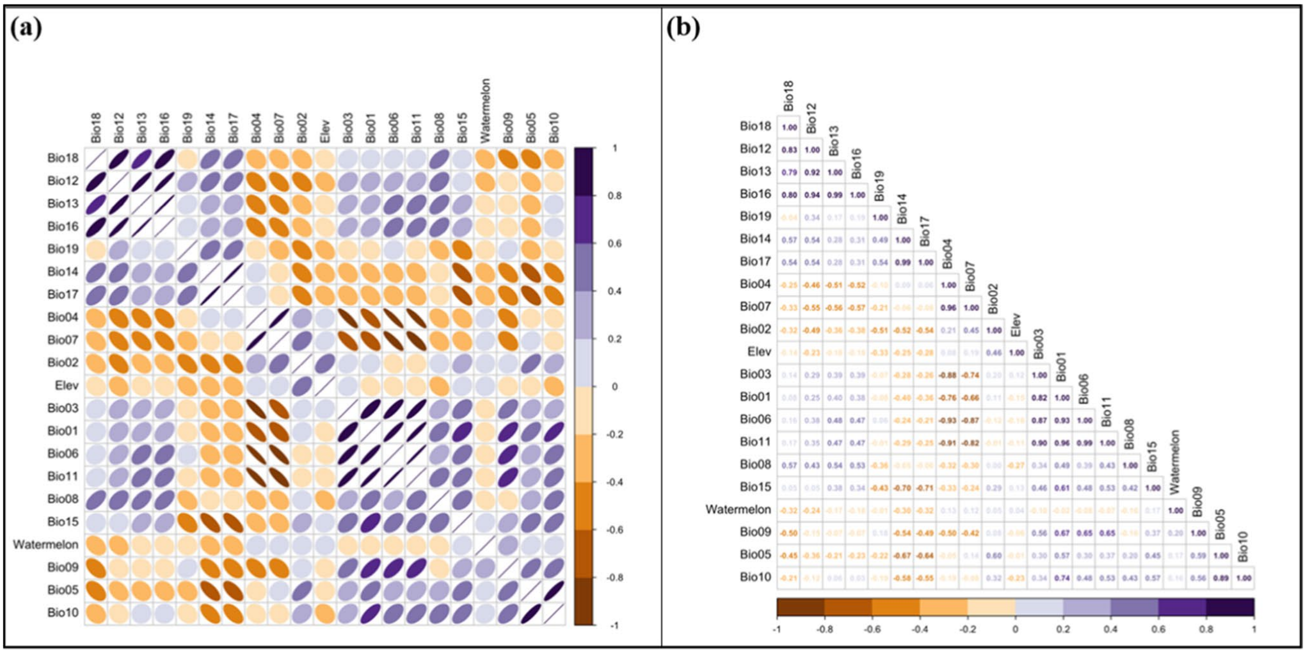


Fig. 2 Correlation between bioclimatic variables: **a** the light purple color with a rightward slope indicates a positive correlation, while the orange color with a leftward slope indicates a negative correlation. The intensity of the correlation coefficient increases as the shape changes from a circle (ρ = 0) to an ellipse (ρ = intermediate) and to a line (|ρ| = 1); correlated variables were grouped using Ward's method (the groups are more internally homogeneous and more heterogeneous among themselves) through hierarchical cluster analysis; **b** estimated values of the correlation coefficients between the variables, following the same color pattern

Table 1 Descriptive statistics of the covariates used in the models, considering their values at the occurrence coordinates of the Watermelon mosaic virus (variables in bold represent those selected for the Maxent model)

Variable	Variable Name	Minimum	Maximun	Median	Mean	SD
Bio01	Annual Mean Temperature	4.27	28.64	15.88	16.78	5.9
Bio02	**Mean Diurnal Range**	**6.73**	**17.52**	**11.47**	**11.46**	**2.1**
Bio03	Isothermality	21.46	84.6	38.04	43.19	15.46
Bio04	Temperature Seasonality	42.93	1,514.20	691.13	649.09	342.55
Bio05	Maximum Temperature of Warmest Month	22.2	46.3	31.2	31.67	4.25
Bio06	**Minimum Temperature of Coldest Month**	**−20.8**	**22.6**	**2.05**	**2.34**	**9.68**
Bio07	Temperature Annual Range	9.2	53.8	29.85	29.33	9.52
Bio08	**Mean Temperature of Wettest Quarter**	**−0.98**	**30.8**	**20.07**	**18.46**	**6.9**
Bio09	Mean Temperature of Driest Quarter	−12.25	36.78	20.35	15.97	11.28
Bio10	**Mean Temperature of Warmest Quarter**	**16.53**	**36.78**	**23.98**	**24.57**	**3.73**
Bio11	Mean Temperature of Coldest Quarter	−12.87	27.63	7.2	8.66	9.65
Bio12	Annual Precipitation	23	2,790.00	675.5	788.66	508.01
Bio13	Precipitation of Wettest Month	5	479	110	136.83	98.54
Bio14	Precipitation of Driest Month	0	96	9	16.95	17.51
Bio15	**Precipitation Seasonality**	**12.71**	**149.34**	**61.07**	**62.44**	**29.35**
Bio16	Precipitation of Wettest Quarter	13	1,282.00	280	358.68	255.77
Bio17	Precipitation of Driest Quarter	0	296	44	63.77	60.31
Bio18	**Precipitation of Warmest Quarter**	**0**	**1,076.00**	**164**	**223.62**	**206.81**
Bio19	**Precipitation of Coldest Quarter**	**2**	**978**	**118**	**144.1**	**155.62**
Elev	**Elevation (m)**	**−227**	**2,074.00**	**281**	**448.9**	**456.05**
Watermelon	**Watermelon (ha)**	**−0.06**	**37.24**	**0.83**	**2.67**	**5.41**

Table 2 Evaluation metrics of the final Maxent model of Watermelon mosaic virus

Metrics	Values
True Positive Rate, Sensitivity or Recall (TPR)	0.83311
True Negative Rate or Specificity (TNR)	0.7291
True Skill Statistic (TSS)	0.56221
Sorensen Index	0.07225
Jaccard Index	0.03761
F-measure on Presence-Background (FPB)	0.07522
Omission or False Negative Rate (OR)	0.16689
Boyce Index	0.88742
Area Under ROC Curve (AUC)	0.82754
Area Under Precision/Recall Curve (AUCPR)	0.06537
Inverse Mean Absolute Error (IMAE)	0.79156
False Positive Rate (FPR)	0.2709
Positive Predictive Value or Precision (PPV)	0.75462
Negative Predictive Value (NPV)	0.46671
Accuracy	0.7811
F1 Score	0.79193
Balanced Accuracy	0.7811
Matthews Correlation Coefficient (MCC)	0.56527
Minimum Training Presence (MTP)	0.01526
10th Percentile Training Presence (10TP)	0.12965
Symmetric Extremal Dependence Index (SEDI)	0.72265

discriminative capacity of the model (0.56221), taking into account true positives and true negatives; 4) Boyce Index (Boyce et al. 2002), the only metric specifically designed to evaluate presence-background models, which assesses the model's ability to discriminate between presence locations and background, indicating when the model tends to correctly predict suitable areas for the species (0.88742); and 5) AUC (Hanley and McNeil 1982; Fawcett 2006), which is the area under the ROC (Receiver Operating Characteristic) curve, a widely used statistic for characterizing the performance of SDMs (Yackulic et al. 2013), despite its limitations, as it indicates the model's ability to distinguish between presences and absences (or pseudo-absences, or background) in 83% of cases (Lobo et al. 2008; Hanczar et al. 2010).

The model also demonstrated satisfactory performance according to the omission rates, which represent the proportion of incorrectly predicted test locations when converted to a binary scale (0 or 1), indicated by MTP (0.01526) and 10TP (0.12965). The MTP defines the threshold value of the lowest predicted value by the model for a training locality. If a location in the test dataset produces a prediction above this threshold, it is identified as "presence" and assigned a value of 1 (Radosavljevic and Anderson 2014). The omission rate is the proportion of test locations with values below this threshold. The 10% threshold is similar, except that the threshold value is defined based on whatever omission occurs for the 10% of training locations with the lowest predicted values. Lower omission rates generally indicate high model performance. Omission rates greater than the expected theoretical values are potentially subject to overfitting.

To assess the reliability of the model, graphs were generated to evaluate the AUC and partial AUC (McClish 1989; Jiang et al. 1996), as well as a Presence Only Calibration graph (see Fig. 3) (Phillips and Elith 2010). These graphs are important because they visually complement the model evaluation through metrics. The AUC value was 0.826. The AUC provides an overall performance measure; however, the partial AUC with a 10% threshold allows for identifying a more specific region of the ROC curve for decision-making. Based on the pAUC, it can be stated that among the 10% of predictions with the highest probability of occurrence, 64% of the unsuitable areas were correctly identified by the model (specificity). Additionally, 64% of the locations where the species is actually present were correctly identified as suitable (sensitivity); therefore, the final model developed is reliable in identifying critical areas at this level.

3.2 Potential distribution under current climatic conditions

The most important bioclimatic variables for the model were selected based on their percentual permutation importance (Fig. 4). Of the seven variables used for the fine-tuning of the Maxent model for WMV, the five most important were: “Bio06” (minimum temperature of the coldest month), “Bio10” (mean temperature of the warmest quarter), “Bio15” (precipitation seasonality), “Elev” (elevation), and “Bio08” (mean temperature of the wettest quarter), respectively.

Figure 5 presents the marginal individual response curves (Partial Dependence Plots). These curves show the relationship between the probability of WMV occurrence and each of the covariates, where the response of each covariate is modeled for only one variable while the others are held constant at their mean (Friedman 2001). In addition to the curves, Fig. 5 also shows the frequency histograms and density curves of the variable values at the occurrences, for the five variables with the highest permutation importance for the model. When observing the response curves, it is possible to verify that there are no bimodal responses, as expected. Based on the evaluation of the response curves, we can infer that the Watermelon mosaic virus is an organism highly associated with warmer environments, with good rainfall availability, a well-defined seasonal pattern, and areas with an altitude below one thousand meters (Fig. 5).

Figure 6 presents the global potential geographic distribution model of WMV under current climatic conditions. It is

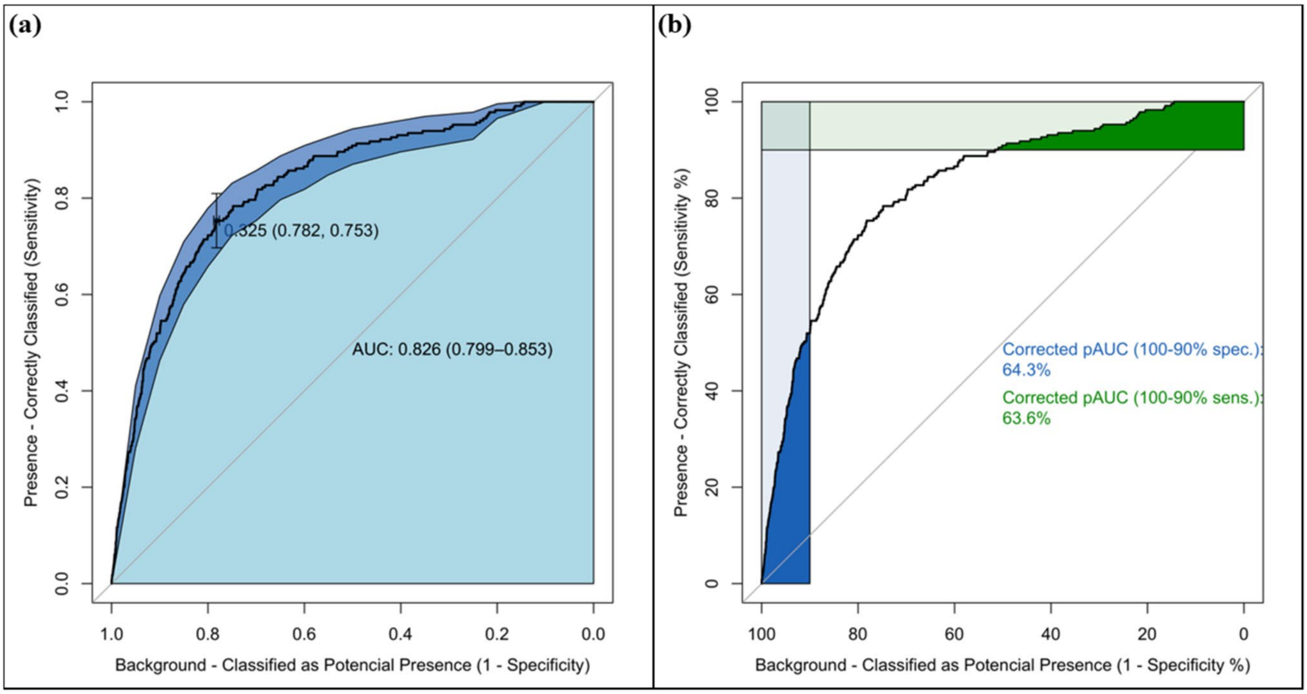
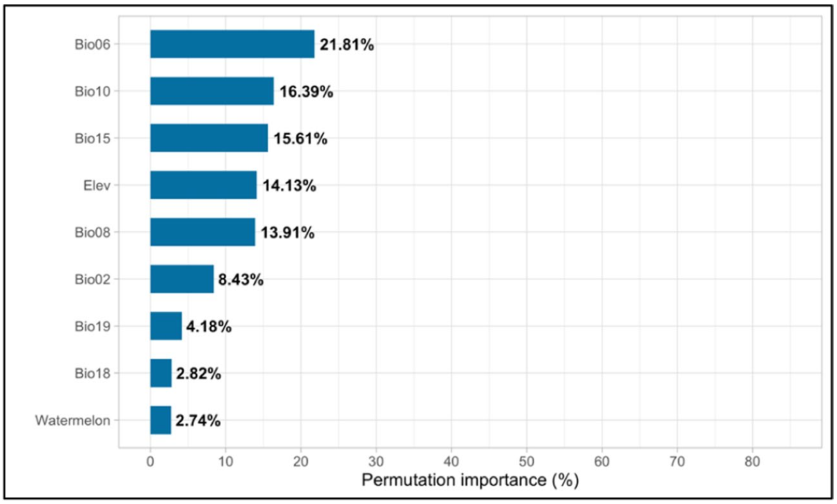


Fig. 3 Graphs showing the area under the ROC curve (AUC) (**a**) and the partial AUC at 10% (**b**) for the final Maxent model of Watermelon mosaic virus

Fig. 4 Permutation importance of variables in the final Maxent model for Watermelon mosaic virus



possible to identify that the highest probabilities of establishment, meaning the countries with the greatest potential distribution for WMV, are in southern European countries, as well as others such as the United States, Brazil, Argentina, China, Turkey, and Iran.

Figure 7 shows a map with the area (km^2) of each WMV survival class and its occurrence points, and Table 3 shows the percentage of the planet's territory occupied by each class. Thus, environments with a high probability of WMV occurrence cover 4,491,060 km^2, and environments with optimal probability represent 9,190,900 km^2, based on estimates with a spatial resolution of 2.5 min (Fig. 7). Additionally, Table 3 shows that 53.92% of the planet's territory is unsuitable for WMV survival. The remaining areas (46.08%) offer some probability of survival for the virus, with 16.85% classified as marginal probability, 19.23% as moderate, 6.72% as optimal, and 3.28% as high probability.

With the application of the Minimum Training Presence (MTP) value, it can be observed that there is a large area globally that shows minimal environmental suitability for the species' presence (marginal conditions). Using the 10th Percentile Training Presence (10TP), an indicator

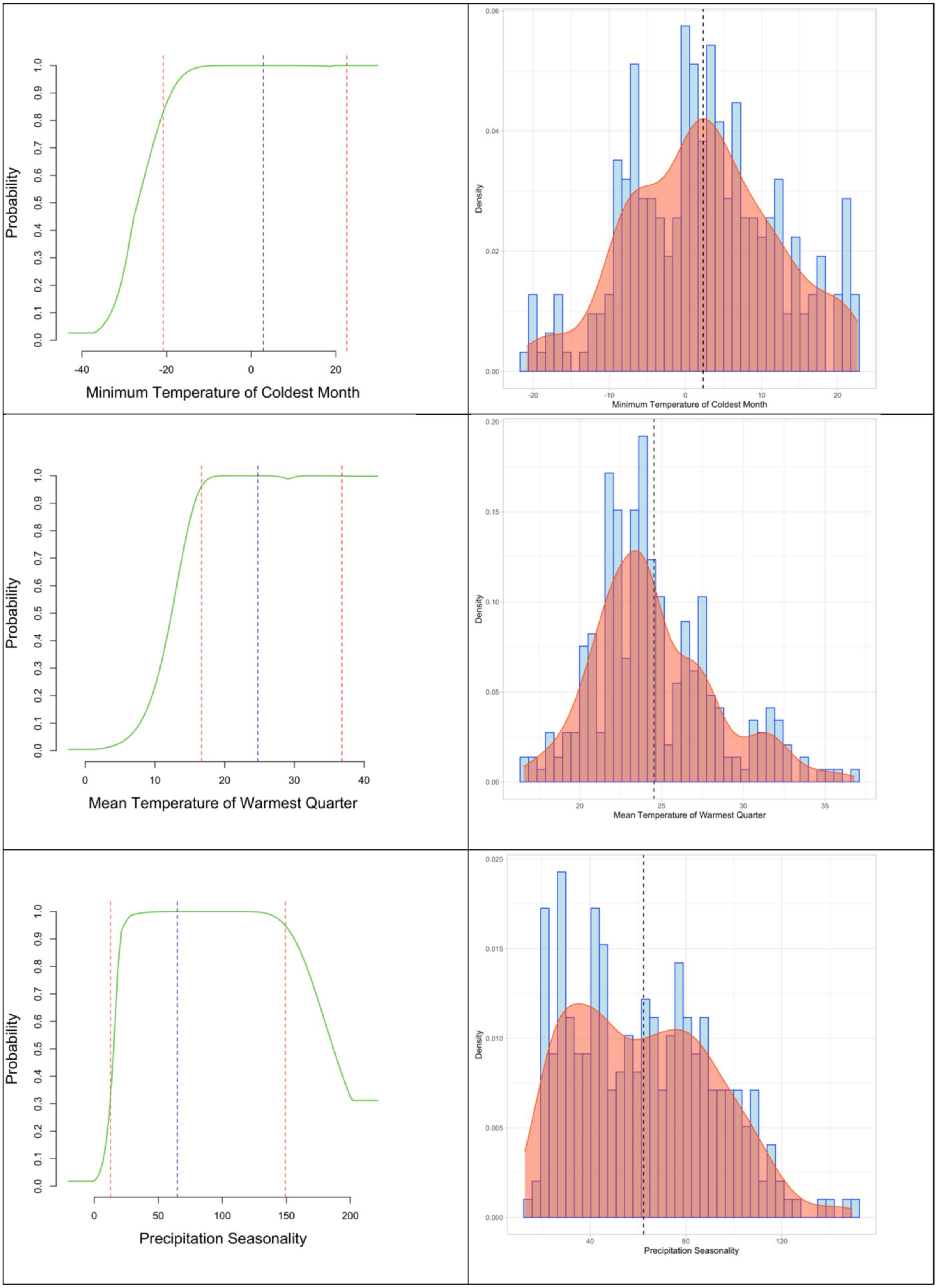


Fig. 5 Individual response curves (on the left) with the minimum, maximum (red dashed lines), and mean values (dark blue dashed line), along with frequency histograms (in light blue) and density curves (considering only the occurrences; on the right, in orange) and mean values (dashed lines), from the final Maxent model for Watermelon mosaic virus

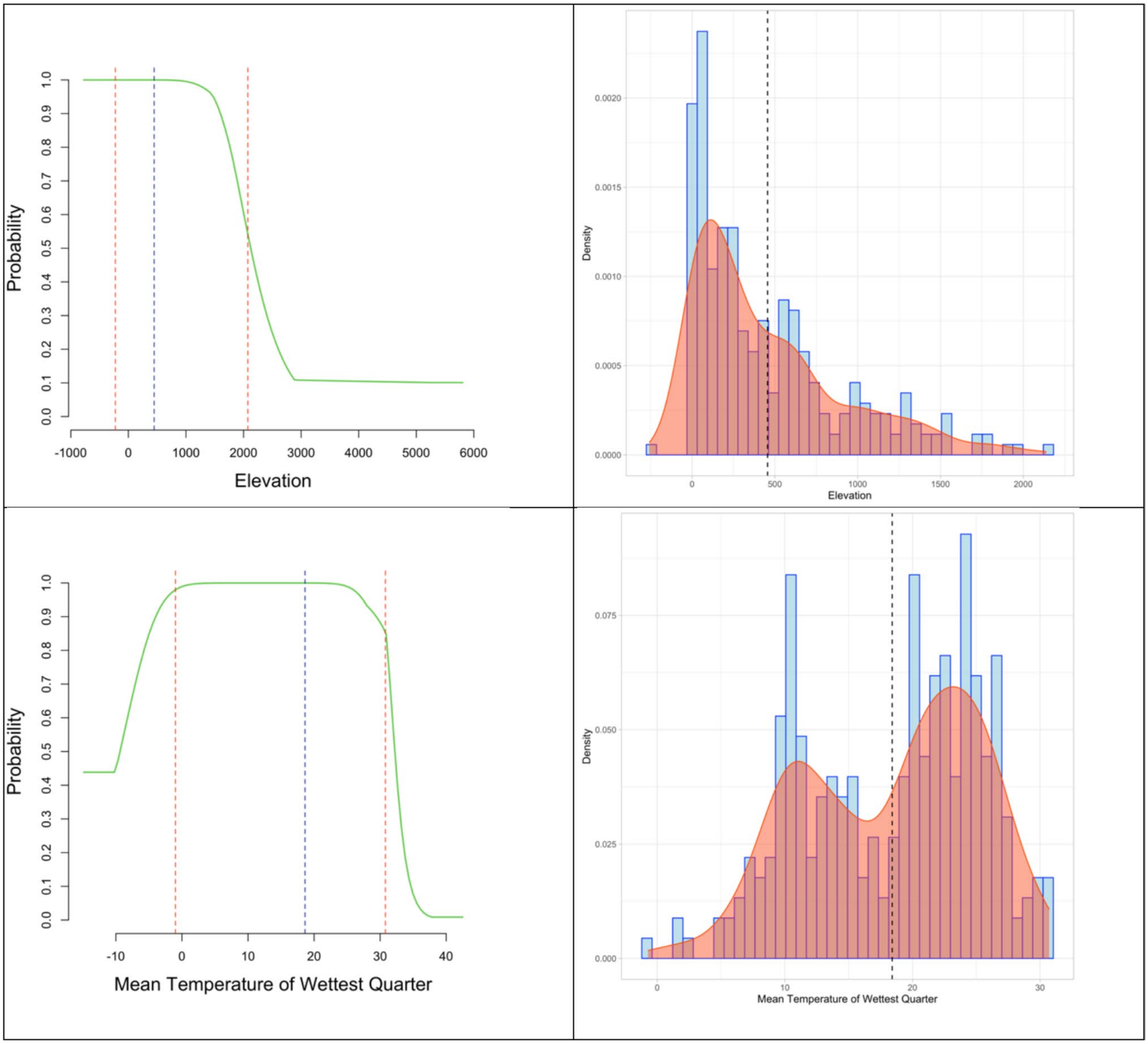


Fig. 5 (continued)

that considers only the top 90% most suitable areas, there is a high probability of WMV occurrence in the Brazilian Amazon region, Guyana, Suriname, French Guiana, Colombia, Venezuela, Panama, Costa Rica, Nicaragua, from North Africa to southern Iran, Sri Lanka, South Asia, the coastal region of Liberia, Ivory Coast, Nigeria, Cameroon, Equatorial Guinea, and a part of Central Africa (Fig. 8).

Figure 9 shows the global projection of the presence/absence of WMV, made by applying a threshold that maximizes the sum of sensitivity and specificity (estimated at 0.3747), identifying the areas with the highest probability of occurrence of the species, based on the conditions established by the selected climatic variables.

WMV occurrence points were plotted on a map showing the world's largest producers of cucurbits based on harvested area (Fig. 10). WMV occurs on every continent except Antarctica. Among the twenty-nine countries identified as the world's largest cucurbit producers, four (Russia, Kazakhstan, Cameroon, and Mexico) have no reported occurrences of the pathogen, which may be related to their ranking prominence. However, it is observed through the maps in Figs. 6, 7, 8, and 9 that these countries have areas with a probability of survival for WMV.

From the occurrence points of WMV, a climate zoning map was also created based on the Köppen-Geiger classification, showing where the virus occurs (Fig. 11). It is observed that WMV can survive in a wide range of climates, as it has

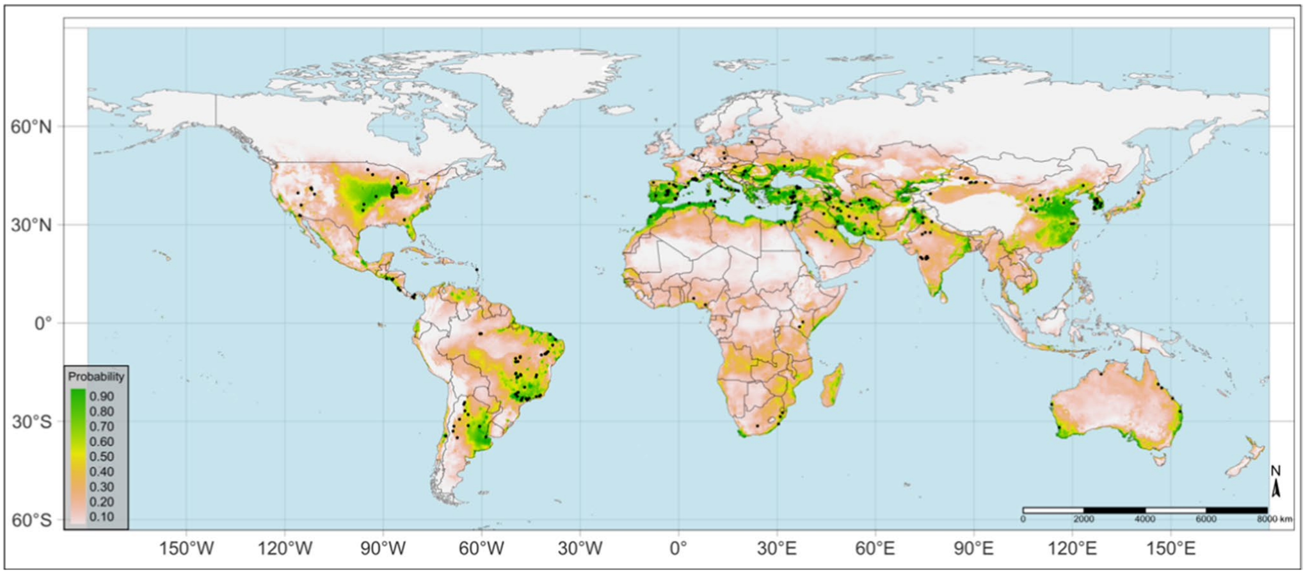


Fig. 6 Potential geographic distribution of Watermelon mosaic virus under current climatic conditions and confirmed occurrence points of the species

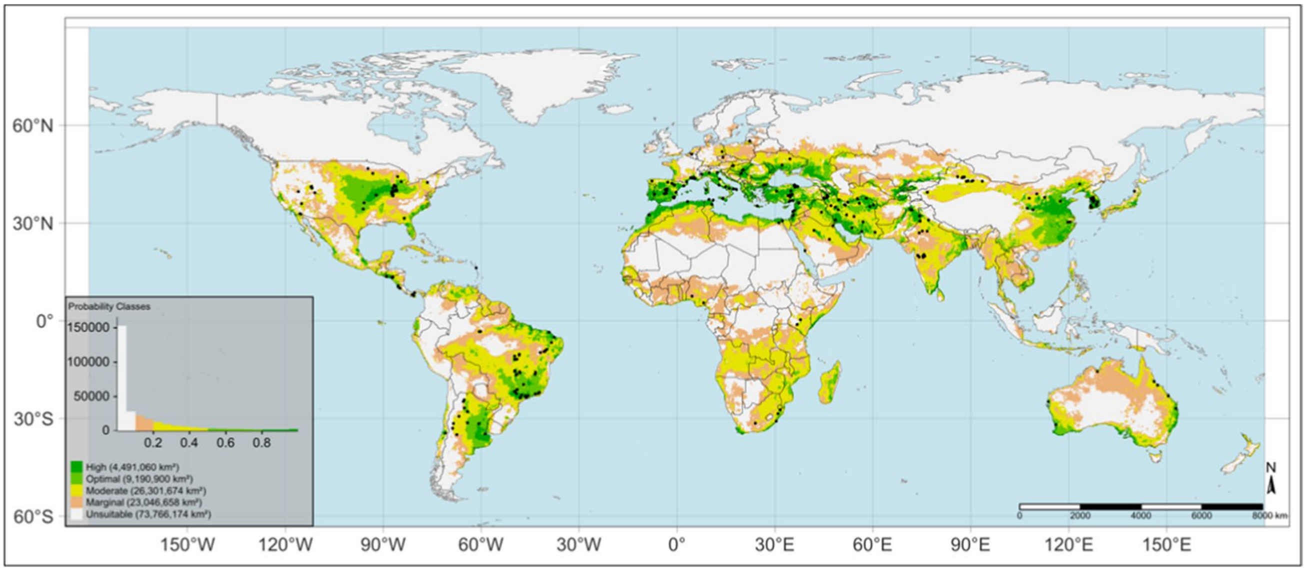


Fig. 7 Probability classes for the potential geographic distribution of Watermelon mosaic virus under current climatic conditions and area estimates (based on 2.5-min resolution) and confirmed occurrence points of the species

Table 3 Area of each WMV survival probability class in percentage

Probability class	Area (km^2)	Percentage (%)
High	4,491,060	3.28%
Optimal	9,190,900	6.72%
Moderate	26,301,674	19.23%
Marginal	23,046,658	16.85%
Unsuitable	73,766,174	53.92%
Total	136,796,466	100.00%

occurrence points across all types of climates in the Köppen-Geiger classification.

In addition to the climate zoning map, a histogram of the percentage of WMV occurrences in each type of climate was created. It is observed that WMV shows the highest frequencies of occurrence in climates classified as Cfa (subtropical climate), Csa (temperate climate with dry summer), Aw (tropical climate with dry winter), and BSk (cold semi-arid climate) (Kottek et al. 2006). These four climate

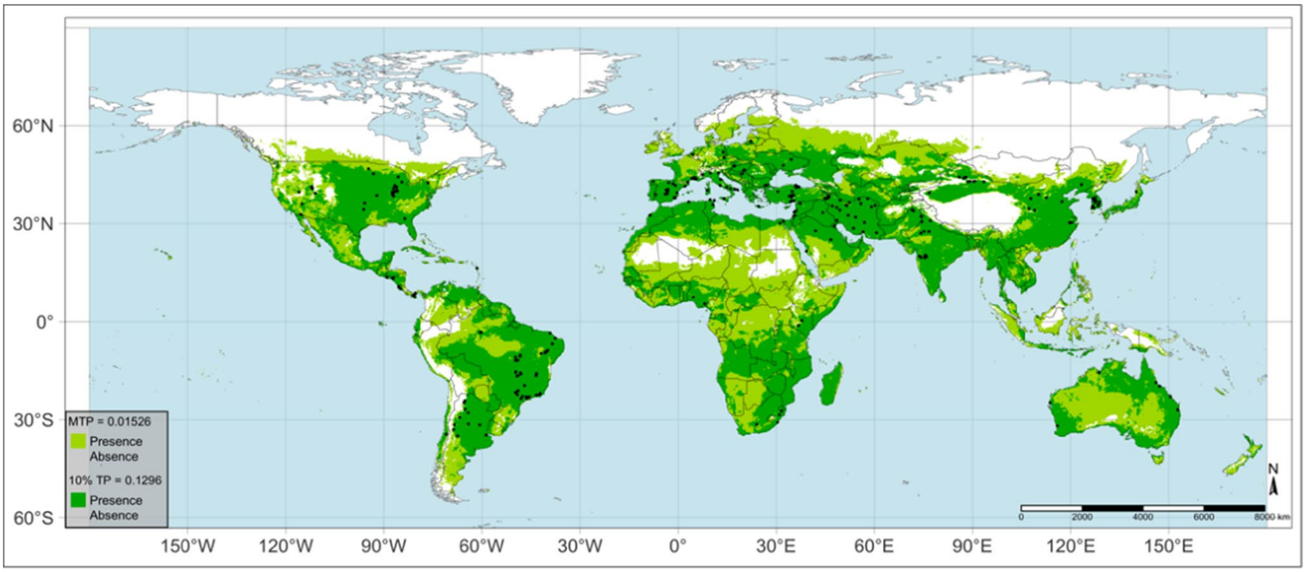


Fig. 8 Potential geographic distribution of Watermelon mosaic virus based on the application of the minimum training presence threshold (marginal probability of occurrence, MTP) and the threshold that identifies the most suitable areas (highest probability of occurrence, 10% TP)

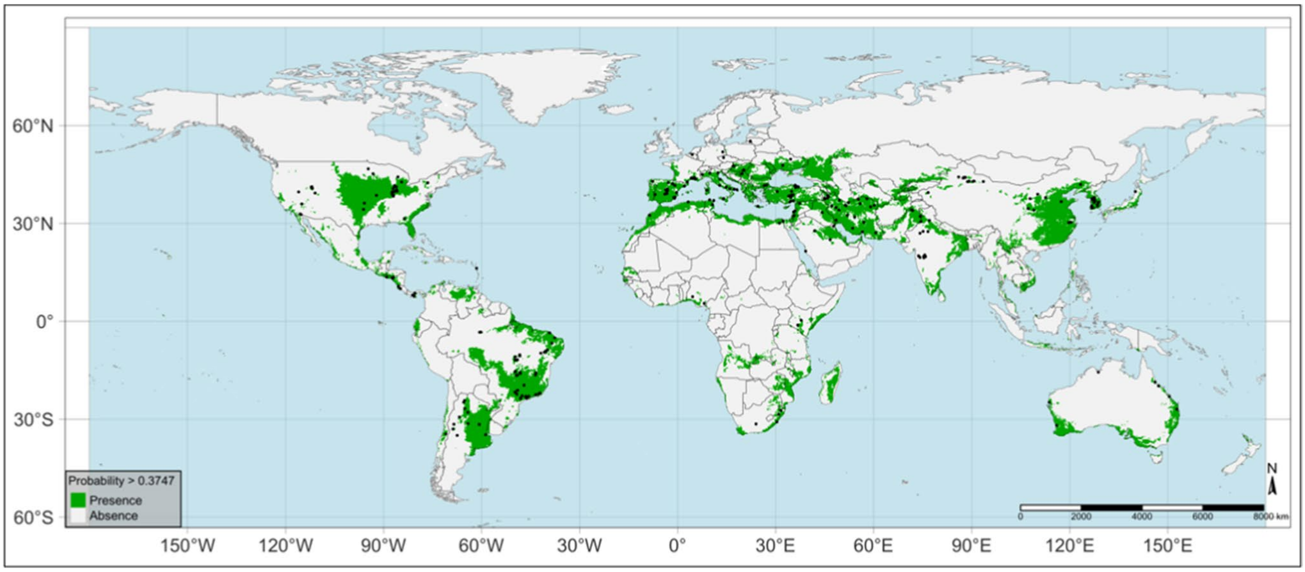


Fig. 9 Potential geographic distribution of Watermelon mosaic virus considering the application of the threshold that maximizes the sum of sensitivity and specificity (maxSSS = 0.3747) and confirmed occurrence points of the species

types account for 59.3% of all WMV occurrences globally (Fig. 12).

4 Discussion and conclusions

In this study, a Maxent model was developed to identify potentially suitable areas for the Watermelon mosaic virus globally. The developed model demonstrated good discrimination capability for suitable areas for the species, as indicated by all the threshold-independent and threshold-dependent performance metrics used. However, it is evident that potential geographic distribution models predict better for specialist species, which have restricted geographic distribution, as well as specialist species with strict ecological requirements (i.e., limited ecological niche), than for generalist species (broader ecological niches) with large distribution areas (Stockwell and Peterson 2002; Seoane et al. 2005;

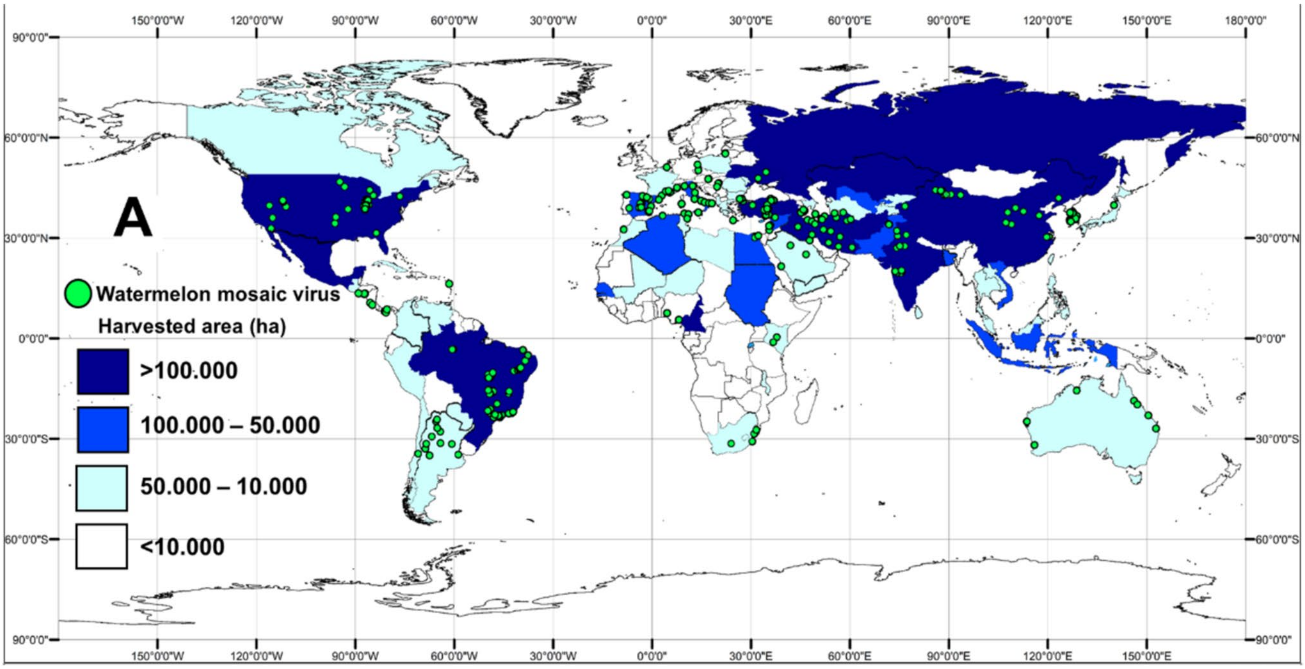


Fig. 10 Largest global producers of cucurbits by harvested area and points of occurrence of Watermelon mosaic virus plotted

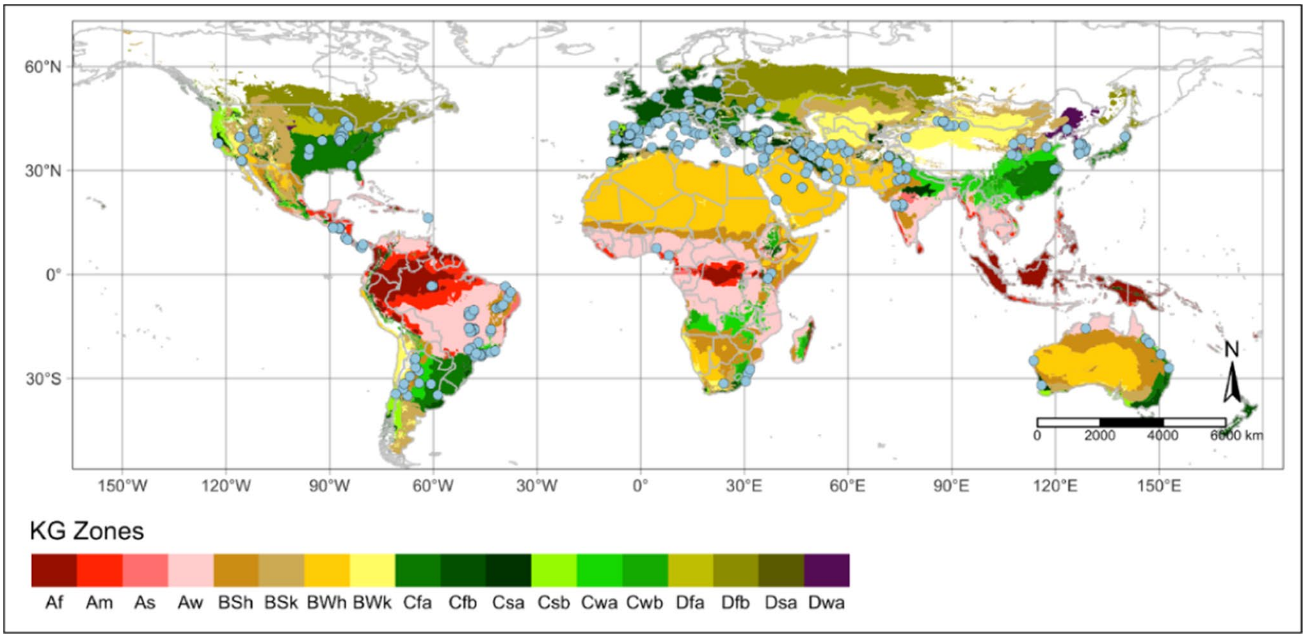


Fig. 11 Map of climate zones according to the Köppen-Geiger classification with WMV occurrence points plotted

Hernandez et al. 2006; Tsoar et al. 2007; Tessarolo et al. 2014; Proosdij et al. 2016).

Although a fully updated methodology was used in this study, it is acknowledged that the model projections may have limitations due to uncertainties related to the nature of invasive species, particularly niche shifts and their ability to adapt to climate change. There is growing evidence indicating the occurrence of niche shifts during biological invasions (Hill et al. 2017; Zhou et al. 2023). Therefore, these factors may lead to an underestimation of species distribution if only native occurrences are considered when developing the model. In the generated model,

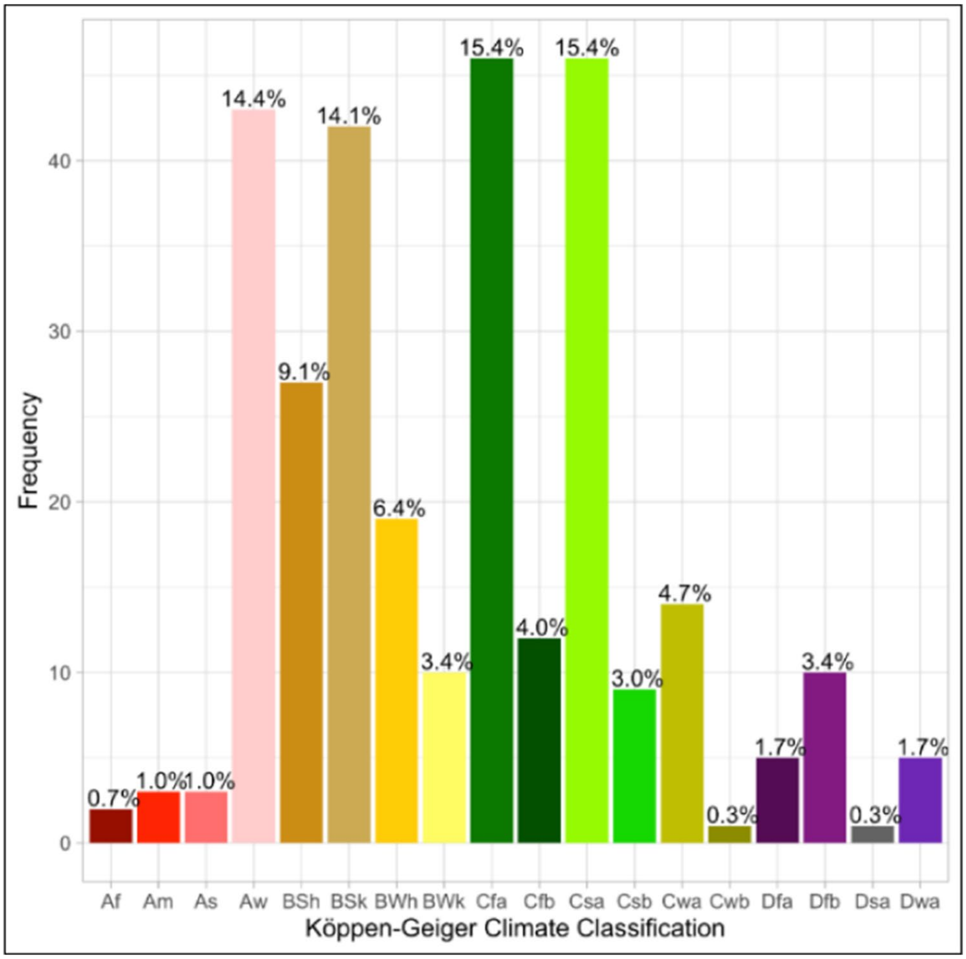


Fig. 12 Histogram frequency of Watermelon mosaic virus occurrence about each type of climate of the Köppen-Geiger classification

this issue was addressed by using occurrence records from both native and invaded areas. Additionally, only altitude and climatic variables were included in the model, along with the most recent information on the harvested area of the crop most affected by the Watermelon mosaic virus. Future studies should consider incorporating non-climatic factors, such as biotic interactions, dispersal capacity, introduction likelihood, and host plant presence, provided that their dynamics are also considered for projections under future climatic conditions if that is the study's objective.

The generated model was simple, containing only 5 variables, but with good quality, according to the AUC value (Hand and Anagnostopoulos 2013) and the other metrics used to evaluate its reliability (Tables 1 and 2). Temperature and precipitation are factors that condition the presence of an organism and are essential for understanding the distribution dynamics of an invasive species (Aidoo et al. 2022). Regarding Watermelon mosaic virus, the same dynamics are observed, where, of the five final variables, three are related to temperature (Bio06, Bio10, and Bio08), representing 52.11% of the model's variability, and one is related to precipitation (Bio15), representing 15.61% of the variability for the virus's probability of survival.

WMV is a pathogen with a non-systematic geographic distribution, meaning its presence is conditioned by specific climatic conditions (Kamberoglu et al. 2015). Regarding temperature, the response curves of the bioclimatic variables used in the final model indicate that the probability of virus survival is higher in regions with warmer temperatures for at least part of the year. This environmental characteristic can be observed in temperate climate regions, which is the climate type where WMV commonly occurs (Desbiez and Lecoq 2004). According to the response curves for Bios06, 08, and 10, there are conditions that considerably reduce the probability of virus survival, including a minimum temperature of the coldest month close to or below −20ºC, an average temperature of the coldest quarter below 20ºC, and an average temperature of the driest quarter below 0ºC. Thus, it is observed that this species survives poorly in regions of extreme cold or with extreme seasonal temperature variations. This indicates that the results obtained were consistent in biological terms, as it is known that seasonal temperature variations are an important factor associated with the severity of epidemics caused by WMV (Alonso-Prados et al. 2003). Moreover, another element that demonstrates the consistency of the response curves' results is the maps showing WMV survival probability, which indicate that regions in the extreme north of the planet (extreme cold regions) have few recorded occurrences.

The response curve for bioclimatic variable 15 provides information about WMV's survival capacity in relation to local precipitation conditions. It is observed that locations with good volume and low seasonal variation in precipitation present the best conditions for the virus to survive. Another important point is that this bioclimatic variable contributed more than 15% of the model's variability, highlighting the significance of precipitation as a factor in the virus's survival and in determining its ecological niche. This result is consistent because, according to Schoeneweiss (1978), water availability significantly affects plants' susceptibility to diseases. Furthermore, it is important to emphasize the significance of the data obtained, as there is limited information in the literature regarding the impact of precipitation on the virus's survival capacity and its invasion risk in new regions.

The last of the 5 variables used in the final model was the variable "Elev," which refers to the elevation of a location and the associated probability of WMV survival. This variable contributed 14.13% of the model's variability and indicated that locations with an altitude of less than one thousand meters offer the highest probability for the pathogen to survive, whereas the probability of WMV survival decreases drastically above two thousand meters in altitude.

After analyzing the variables associated with WMV survival, it is observed that this organism is highly associated with warmer environments for at least part of the year, with good availability of rainfall in a well-defined seasonal pattern, and in areas with an altitude of less than one thousand meters. This information can assist cucurbit producers in planning the cultivation of their crops, including defining the most critical periods concerning the probability of virus occurrence, thus establishing effective monitoring and pest management measures. Furthermore, the results we obtained can support the work of phytosanitary defense agencies in different countries by predicting which areas are at the highest risk of WMV invasion and identifying the environmental factors that favor this phenomenon. This way, these agencies can implement legislative measures to prevent the introduction of the pathogen in disease-free locations.

WMV is a globally dispersed virus that spreads on a continental scale through contaminated plant material and locally through disease vectors (Gagarinova et al. 2008). This broad dispersion becomes evident when analyzing the map in Fig. 1, which shows the points of disease occurrence. From this analysis, it is observed that, although WMV originated in northern China (Ben-Mansour et al. 2023), this pathogen can adapt and survive in diverse climates (see Fig. 11).

A climatic suitability map for WMV, divided into classes, was created based on the generated Maxent model. It is observed that regions located on or near the Equator exhibit marginal probability or unsuitability for the virus's survival. This may be related to the fact that, in these low-latitude regions, there are minimal variations in the Sun's position throughout the year (Vieira et al. 2016). Consequently, many locations experience high temperatures associated with water scarcity. As seen through the response curve of Bio15, WMV has a higher probability of surviving in areas with greater water availability. Furthermore, although the virus is associated with warmer temperatures, these regions located on or near the Equator may experience temperatures that exceed the maximum threshold for optimal conditions for WMV.

Analyzing the map in Fig. 6, it can be observed that areas above the Tropic of Cancer, such as Central Europe, the United States, and eastern China, concentrate most occurrences of the virus. These regions also account for the percentage of areas with a high probability of WMV survival. However, there are occurrences located in areas with low survival probability for the pathogen, which is not uncommon when studying species with a wide geographic distribution. This phenomenon has several possible explanations. For instance, some occurrence points in the GBIF database may have been collected from cucurbit cultivation in greenhouses, which have significantly different environmental conditions compared to the standard conditions where the virus is typically found. Another possible explanation is that we are experiencing a scenario of constant climate change, which directly influences the development, survival, and dispersal of pests (Sentis and Desneux 2019), along with the fact that viruses commonly evolve and adapt quickly to new climatic conditions (Jones 2009; Elena et al. 2014). Lastly, a potential explanation for the mentioned occurrences may be that there is evidence that WMV originated through an ancestral recombination phenomenon (Desbiez and Lecoq 2004), possibly triggered by climate change, which may also have occurred in these marginal probability regions.

The area occupied by each probability class in km^2 was determined in Fig. 7, and the percentage of each class was described in Table 3. Based on these results, we find that, despite WMV being a widely distributed pathogen, about 53.92% of the planet's total area does not offer conditions for the pathogen to survive, and 16.85% presents only marginal conditions. However, it is known that we are living in a scenario of climate change due to greenhouse gas emissions increasing global temperatures (Kharin et al. 2013), which may lead to more areas becoming suitable for WMV and the pathogen being disseminated. These change scenarios tend to spread various pathogens to new regions and hosts (Baker et al. 2000; Etterson and Shaw 2001).

Through the map of the world's largest producers of the crop (Fig. 10) and the presence and absence maps (Figs. 8 and 9), it is observed that WMV poses a threat to the production system of major global cucurbit producers such as Russia, Kazakhstan, Mexico, and Cameroon. Regarding Russia, based on the results obtained, it is noted that it is a country with climatic conditions adverse to the introduction and survival of the pathogen, due to extreme temperatures throughout the year. However, we know that mutation or recombination events can allow this organism to adapt to new locations. In the case of the other countries, the risk of WMV introduction is higher, as they share borders with countries where the virus is already present and have a climate that may offer conditions for its survival. To prevent a possible introduction of WMV, all countries should strengthen the monitoring of their borders to avoid the entry of contaminated plant material, as early detection and detailed planning of control activities can help mitigate the impact of an invasive organism (Silva et al. 2014).

The occurrence points of WMV were plotted on a map containing the Köppen-Geiger climate zoning, and a histogram of the frequency of virus occurrences in each climate type was generated. It is observed that, despite WMV having ideal conditions for its survival as determined, it was found in all types of climates in the zone. However, climates BSk, Aw, Cfa, and Csa presented the highest percentages of occurrence frequency (59.3%). These climate types have in common elevated temperatures, especially in the summer, and good water availability (Kottek et al. 2006). This shows that the results presented by the response curves of the variables used in the model are consistent, as they exhibit environmental characteristics similar to those found in the climate types where WMV is most frequently encountered. Furthermore, although less frequent, there are occurrences of WMV in all types of climates classified by Köppen-Geiger, which may explain the global presence of the pathogen and the difficulty in determining its ecological niche.

This study was able to determine the environmental conditions associated with the best probabilities of WMV survival and the regions at risk of its introduction. Determining the ecological niche of a pest is an important tool for preventing its entry into countries threatened by its introduction. However, due to the ability of viruses to adapt and survive even in adverse climatic conditions, precautionary measures are necessary even in countries that present marginal survival conditions. Moreover, climate change may alter agroclimatic zones and the geographical distribution of pathogens and hosts (Yáñez-López et al. 2012). Thus, regions currently not included in the ecological niche of the species may become part of it.

Most pest invasions are related to failures in risk assessment and surveillance systems; therefore, continuous data collection on pest occurrences is necessary (Silva et al. 2014). In this sense, this work supports the risk analysis of phytosanitary agencies through the information provided, such as the ecological niche of the pest, locations of occurrence points, the ideal climatic conditions associated with its survival, and the main climate types where it is found.

Despite the challenges mentioned for its proper implementation, legislative control remains the most efficient measure to combat the virus, as it is an essential means to prevent the entry of diseases (Kassem et al. 2020; Ahmad 2021). Another possible management measure that can be adopted based on our results regarding the risk of species introduction is the development of resistant cultivars. Due to certain characteristics of a pest, if it invades a neighboring country, its entry into surrounding countries is almost inevitable, making preparation for the invasion the best strategy (Silva et al. 2014). In the case of viruses, the most efficient management measure after their entry into a region is the exploitation of genetic resistance in cultivars (Silveira et al. 2009). Therefore, our study is also useful for preventive genetic improvement programs, as breeders in countries threatened by the invasion of WMV can obtain resistant cultivars even before their arrival.

Through this work, it was possible to map the occurrence points of WMV and identify the countries already affected by the virus. Additionally, we determined its ecological niche and, consequently, the countries threatened by its introduction. The final model generated was simple, consisting of only five bioclimatic variables, but it was predictive and reliable according to various metrics. Bioclimatic variables 06, 10, and 15 (related to temperature and precipitation) accounted for over 53% of the model's variability, making them the most important. It was found that the warmest regions with good rainfall availability are ideal for the survival of WMV. Despite WMV having a higher probability of surviving in specific environmental conditions, pest occurrences were noted even in adverse climates. Thus, with the results obtained, this study aims to support the work of phytosanitary defense agencies in developing legislative control measures and improvement programs, specifically in developing resistant cultivars against pests, particularly in countries threatened by their introduction.

Acknowledgements The authors thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, for financial support (project: PROCAD-AMAZÔNIA: 88887.200564/2018-00), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (project: 458946/2014-1, 306011/2022-0, 405888/2023-6, 442689/2023-3 and 201237/2024-5), the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Brazil, and the Empresa Brasileira de Pesquisa Agropecuária (Embrapa), Brazil (project: 10.20.03.056.00.00).

Authors contribution Kayo H.: Conceptualization, Investigation, Data curation, Methodology, Formal analysis, Validation, Writing—original

draft, Writing—review & editing, Visualization. Mayara M.: Investigation, Methodology, Formal analysis, Validation, Visualization. Poliana S.: Investigation, Methodology, Formal analysis, Writing—review & editing, Visualization. Hugo D.: Investigation, Data curation, Formal analysis, Validation, Visualization. Mônica C.: Investigation, Methodology, Formal analysis, Validation, Visualization. George C.: Investigation, Methodology, Formal analysis, Validation, Visualization. Ricardo S.: Conceptualization, Data curation, Methodology, Formal analysis, Validation, Supervision, Writing—review & editing, Visualization. Marcelo C.: Conceptualization, Methodology, Supervision, Funding acquisition, Writing—review & editing, Visualization. Renato A.: Conceptualization, Supervision, Project administration, Funding acquisition, Writing—review & editing, Visualization.

Funding Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil—project: PROCAD-AMAZÔNIA: 88887.200564/2018-00.

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil—project: 458946/2014-1, 306011/2022-0, 405888/2023-6, 442689/2023-3 and 201237/2024-5.

Data availability Data will be made available on request.

Declarations

Competing interest The authors declare no competing interests.

References

Acevedo P et al (2012) Delimiting the geographical background in species distribution modelling. Journ Biog 39(8):1383–1390

Addison PF et al (2013) Practical solutions for making models indispensable in conservation decision-making. Divers Dist 19(5–6):490–502

Ahmad I (2021) Integrated pest management of *Rhynchophorus ferrugineus* olivier: an efficient approach to reduce infestation in date palm trees. Pak J Zool 54:927

Aidoo OF et al (2022) Climate-induced range shifts of invasive species (*Diaphorina citri* Kuwayama). Pest Manag Sci 78(6):2534–2549

Allouche O et al (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43(6):1223–1232

Alonso-Prados JL et al (2003) Epidemics of aphid-transmitted viruses in melon crops in Spain. Eur J Plant Pathol 109:129–138

Alonso-Prados JL et al (1997) Impact of cucumber mosaic virus and watermelon mosaic virus 2 infection on melon production in Central Spain. J Plant Pathol 79:131–134

Amaro G et al (2022) Risk analysis of the spread of the quarantine pest mite *Schizotetranychus hindustanicus* in Brazil. Exp Ap Acar 88(3–4):263–275

Amaro G et al (2023) Effect of study area extent on the potential distribution of species: a case study with models for *Raoiella indica* Hirst (Acari: Tenuipalpidae). Ecol Mod 483:110454

Anderson RP, Raza A (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. J Biogeogr 37(7):1378–1393

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. Jour Biog 33(10):1677–1688

Araújo MB et al (2019) Standards for distribution models in biodiversity assessments. Sci Adv 5(1):eaat4858

Austin MP, Van Niel KP (2011) Improving species distribution models for climate change studies: variable selection and scale. J Biogeogr 38(1):1–8.3

Ayazpour K, Vahidian M (2016) Study of Watermelon mosaic virus in cucurbit fields of Jahrom Area, Iran. Veget 29:4

Baker RHA et al (2000) The role of climatic mapping in predicting the potential geographical distribution of non-indigenous pests under current and future climates. Agric Ecosyst Environ 82:57–71

Barber RA et al (2022) Target-group backgrounds prove effective at correcting sampling bias in Maxent models. Diver Dist 28(1):128–141

Barbet-Massin M et al (2012) Selecting pseudo-absences for species distribution models: how, where and how many? Meth Ecol Evol 3(2):327–338

Barve N et al (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol Mod 222(11):1810–1819

Beaumont LJ et al (2009) Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. Divers Distrib 15(3):409–420

Ben-Mansour K et al (2023) Watermelon mosaic virus in the Czech Republic, its recent and historical origins. Plant Pathol 72:1528–1538

Bentlage B et al (2013) Plumbing the depths: extending ecological niche modelling and species distribution modelling in three dimensions. Glob Ecol Biogeogr 22(8):952–961

Boyce MS et al (2002) Evaluating resource selection functions. Ecol Mod 157(2-3):281-300

Bradshaw CJA et al (2016) Massive yet grossly underestimated global costs of invasive insects. Nat Commun 7:12986

Broennimann O, Guisan A (2008) Predicting current and future biological invasions: both native and invaded ranges matter. Biol Lett 4(5):585–589

Brunel S et al (2010) The EPPO prioritization process for invasive alien plants. EPPO Bul 40(3):407–422

Castellanos AA et al (2019) Environmental filtering improves ecological niche models across multiple scales. Meth Ecol Ev 10(4):481–492

Chamberlain S et al (2022) rgbif: Interface to the global biodiversity information facility API. R package version 3.7.3. https://CRAN.R-project.org/package=rgbif

Cooper JC, Soberón J (2018) Creating individual accessible area hypotheses improves stacked species distribution model performance. Glob Ecol Biogeogr 27(1):156–165

Datta A et al (2019) Niche expansion of the invasive plant species *Ageratina adenophora* despite evolutionary constraints. J Biogeogr 46(7):1306–1315

Desbiez C, Lecoq H (2004) The nucleotide sequence of Watermelon mosaic virus (WMV, Potyvirus) reveals interspecific recombination between two related potyviruses in the 5′ part of the genome. Arch Virol 149:1619–1632

Elena SF et al (2014) Evolution and emergence of plant viruses. Adv Virus Res 88:161–191

El-Gabbas A, Dormann CF (2018) Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography 41(7):1161–1172

Elith J et al (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129–151

Elith J et al (2010) The art of modelling range-shifting species: the art of modelling range-shifting species. Meth Ecol Evol 1:330–342

Elith J et al (2011) A statistical explanation of MaxEnt for ecologists. Divers Distrib 17(1):43–57

Elith J, Franklin J (2013) Species distribution modeling. In: Levin SA (ed) Encyc. of biod. Academic Press, Waltham, MA, pp 692–705

Elith J, Leathwick JR (2009). Species distribution models: ecological explanation and prediction across space and time. Ann Rev Ecol Evol Syst 40(1):677–97

Esri. (2020) ArcGIS Desktop: Release 10.8. Environmental Systems Research Institute, Redlands, CA

Etterson JR, Shaw RG (2001) Constraint to adaptive evolution in response to global warming. Science 294:151–154. https://doi.org/10.1126/science.1063656

FAO (2021) Food and Agriculture Organization of the United Nations. Prod. Ind. Disponível em: https://www.fao.org/faostat/en/#data/QCL. Accessed 27 July 2023

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37(12):4302–4315

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

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. An Statist 29:1189–1232

Gagarinova AG et al (2008) Recombination analysis of Soybean mosaic virus sequences reveals evidence of RNA recombination between distinct pathotypes. Virol J 5:1–8

Galdino TVDS et al (2016) Mapping global potential risk of mango sudden decline disease caused by *Ceratocystis fimbriata*. PLoS ONE 11(7):e0159450

Gallo D et al (2002) Entomologia agrícola. Piracicaba: FEALQ. 920p

Halvorsen R et al (2015) Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt. Ecography 38(2):172–183

Hanczar B et al (2010) Small-sample precision of ROC-related estimates. Bioinformatics 26(6):822-830

Hand DJ, Anagnostopoulos C (2013) When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognit Lett 34(5):492–495

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36

Heikkinen RK et al (2012) Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35(3):276–288

Helmstetter NA et al (2021) Balancing transferability and complexity of species distribution models for rare species conservation. Divers Distrib 27(1):95–108

Hernandez PA et al (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29(5):773–785

Hirzel A et al (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecol 83(7):2027–2036

Hijmans RJ (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93(3):679–688

Hijmans RJ, Elith J (2013) Species distribution modeling with R. R Foundation for Statistical Computing. Available at: https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf

Hijmans RJ (2023) terra: Spatial Data Analysis. R package version 1.7–78. https://rspatial.github.io/terra/, https://rspatial.org/

Hijmans RJ et al (2023) geodata: Download geographic data. R package version 0.6-1. https://CRAN.R-project.org/package=geodata

Hill MP, Terblanche JS (2014) Niche overlap of congeneric invaders supports a single-species hypothesis and provides insight into future invasion risk: implications for global management of the *Bactrocera dorsalis* complex. PLoS ONE 9(2):e90121

Hill MP et al (2017) A global assessment of climatic niche shifts and human influence in insect invasions. Glob Ecol Biogeogr 26(6):679–689

Hirzel A, Guisan A (2002) Which is the optimal sampling strategy for habitat suitability modelling. Ecol Model 157(2–3):331–341

Jarnevich CS et al (2015) Caveats for correlative species distribution modeling. Ecol Inf 29:6–15

Jiang Y et al (1996) A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 201(3):745–750

Jones RAC (2006) Control of plant virus diseases. Adv Virus Res 67:205–244

Jones RAC (2009) Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus Res 141:113–130

Kamberoglu MA et al (2015) Characterization of an emerging isolate of Watermelon mosaic virus in Turkey. Int J Agric Biol 17:211–215

Kass JM et al (2021) ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. Methods Ecol Evol 12(9):1602–1608

Kassem HS et al (2020) Sustainable management of the red palm weevil: The nexus between farmers' adoption of integrated pest management and their knowledge of symptoms. Sustainability 12(22):9647

Kharin V et al (2013) Changes in temperature and precipitation extremes in the CMIP5 ensemble. Clim Chang 119(2):345–357

Konowalik K, Nosol A (2021) Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. Sci Rep 11(1):1482

Kottek M et al (2006) World map of the Köppen-Geiger climate classification updated. Meteorol Z 15(3):259–263

Lecoq H, Desbiez C (2008) Watermelon Mosaic virus and Zucchini yellow mosaic virus. In: Encyc. of Virol. Elsevier, Amsterdam, The Netherlands, pp 433–440

Leroy B et al (2018) Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. J Biogeogr 45(9):1994–2002

Liu C et al (2005) Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28(3):385–393

Liu C et al (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. J Biogeogr 40(4):778–789

Liu C et al (2016) On the selection of thresholds for predicting species occurrence with presence-only data. Ecol Evol 6(1):337–348

Liu C et al (2009) Measuring the accuracy of species distribution models: a review. In: Proc. 18th Worl. IMACs/MODSIM Cong. Cairns, Australia, vol 4241, p 4247

Lobo JM et al (2008) AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17(2):145–151

Low BW et al (2021) Predictor complexity and feature selection affect Maxent model transferability: Evidence from global freshwater invasive species. Diver Dist 27(3):497–511

Marchioro CA (2016) Global potential distribution of *Bactrocera carambolae* and the risks for fruit production in Brazil. PLoS ONE 11(11):e0166142

Martin GD et al (2020) Climate modelling suggests a review of the legal status of Brazilian pepper *Schinus terebinthifolia* in South Africa is required. S Afr J Bot 132:95–102

McClish DK (1989) Analyzing a portion of the ROC curve. Med Dec Mak 9(3):190–195

Mendes P et al (2020) Dealing with overprediction in species distribution models: How adding distance constraints can improve model accuracy. Ecol Model 431:109180

Merow C et al (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36(10):1058–1069

Moradi Z et al (2012) Nucleotide sequencing and symptomology of two new isolates of Watermelon mosaic virus from razavi and northern khorasan provinces. J Plant Prot 25:407–416

Moreno-Amat E et al (2015) Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. Ecol Model 312:308–317

Moudrý V et al (2024) Optimizing occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter. Ecography 2024:e07294

Mumford RA et al (2016) The role and challenges of new diagnostic technology in plant biosecurity. Food Secur 8:103–109

O'Donnell MS, Ignizio DA (2012) Bioclimatic predictors for supporting ecological applications in the conterminous United States. Data Series 691. US Geolog. Survey: Rest., VA. Available at: https://pubs.usgs.gov/ds/691/

Owens HL et al (2013) Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. Ecol Model 263:10–18

Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. J Appl Ecol 43:405–412

Pearson RG et al (2007) Predicting species distribution from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J Biogeogr 34:102–117

Peay WS et al (2023) A maximum entropy approach to defining geographic bounds on growth and yield model usage. Front For Glob Change 6:1215713

Pebesma EJ (2018) Simple features for R: standardized support for spatial vector data. R J 10(1):439

Phillips SJ, Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31(2):161–175

Phillips SJ, Elith J (2010) POC plots: calibrating species distribution models with presence-only data. Ecology 91:2476–2484

Phillips SJ, Elith J (2013) On estimating probability of presence from use–availability or presence–background data. Ecology 94(6):1409–1419

Phillips SJ et al (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190(3–4):231–259

Phillips SJ et al (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19(1):181–197

Phillips SJ et al (2017) Opening the black box: An open-source release of Maxent. Ecography 40(7):887–893

Phillips SJ (2017) A Brief Tutorial on Maxent. Available from url: http://biodiversityinformatics.amnh.org/open_source/maxent/. Accessed 01 Oct 2024

Phillips S (2021) maxnet: fitting 'maxent' species distribution models with 'glmnet'. R package version 0.1.4, https://github.com/mrmaxent/maxnet. Accessed 1 Oct 2024

Proosdij ASV et al (2016) Minimum required number of specimen records to develop accurate species distribution models. Ecography 39(6):542–552

Pyšek P et al (2020) Scientists' warning on invasive alien species. Biol Rev 95(6):1511–1534

R Core Team, R. (2023) R: A language and environment for statistical computing. Vienna, Austria: R found. for stat. comp., 171–203

Radosavljevic A, Anderson RP (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. J Biogeogr 41(4):629–643

Renner IW, Warton DI (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69(1):274–281

Renner IW et al (2015) Point process models for presence-only analysis. Methods Ecol Evol 6(4):366–379

Ribeiro BR et al (2022) bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. Methods Ecol Evol 13(7):1421–1428

Roberts DR et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929

Robin X et al (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:1–8

Rojas-Soto O et al (2024) Calibration areas in ecological niche and species distribution modelling: Unravelling approaches and concepts. J Biogeogr 51:1416–1428

Rubio L et al (2020) Detection of plant viruses and disease management: Relevance of genetic diversity and evolution. Front Plant Sci 11:1092

Santini L et al (2021) Assessing the reliability of species distribution projections in climate change research. Diver Dist 27(6):1035–1050

Saupe EE et al (2012) Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. Ecol Mod 237:11–22

Schartel TE, Cao Y (2024) Background selection complexity influences Maxent predictive performance in freshwater systems. Ecol Mod 488:110592

Schoeneweiss DF (1978) Water stress as a predisposing factor in plant disease. Wat Def Plant Grow 5:61–90

Sentis A, Desneux N (2019) Editorial overview: global change: integrating ecological and evolutionary consequences across time and space. Curr Opin Ins Sci 35:3–6

Seoane J et al (2005) Species-specific traits associated to prediction errors in bird habitat suitability modelling. Ecol Model 185(2–4):299–308

Sillero N (2011) What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecol Model 222(8):1343–1346

Sillero N, Barbosa AM (2021) Common mistakes in ecological niche models. Int J Geogr Inf Sci 35(2):213–226

Sillero N et al (2021) Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. Ecol Model 456:109671

Silva ML et al (2014) The role of natural and human-mediated pathways for invasive agricultural pests: a historical analysis of cases from Brazil. Agric Sci 5:634–646

Silveira LM et al (2009) Serological survey of virus in cucurbit species in the Lower Middle São Francisco River Basin, Brazil. Trop Plant Pat 34(2):123–126

Soberón JM (2010) Niche and area of distribution modeling: a population ecology perspective. Ecography 33(1):159–167

Soberon J, Peterson AT (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. Biod Infor 2:1–10

Sofaer HR et al (2019) Development and delivery of species distribution models to inform decision-making. BioScience 69(7):544–557

Srivastava V et al (2021) Oh the places they'll go: improving species distribution modelling for invasive forest pests in an uncertain world. Biol Inv 23:297–349

Stockwell DR, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148(1):1–13

Syfert MM et al (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PLoS ONE 8(2):e55158

Tang FH et al (2024) CROPGRIDS: a global geo-referenced dataset of 173 crops. Sci Data 11(1):413

Tennekes M (2018) tmap: Thematic Maps in R. J Stat Soft 84:1–39

Tessarolo G et al (2014) Uncertainty associated with survey design in species distribution models. Div Distrib 20(11):1258–1269

Thorson JT et al (2016) Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. Glob Ecol Biogeogr 25(9):1144–1158

Tsoar A et al (2007) A comparative evaluation of presence-only methods for modelling species distribution. Diver Distrib 13(4):397–405

Valavi R et al (2019) (2019) blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol 10:225–232

Valavi R et al (2022) Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol Monogr 92(1):e01486

Vanderwal J et al (2009) Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. T Amer Nat 174(2):282–291

Varela S et al (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37(11):1084–1091

Velazco SJE et al (2019) A dark scenario for Cerrado plant species: Effects of future climate, land use and protected areas ineffectiveness. Diver Distrib 25(4):660–673

Velazco SJE et al (2021) On opportunities and threats to conserve the phylogenetic diversity of Neotropical palms. Diver Distrib 27(3):512–523

Velazco SJE et al (2022) flexsdm: An r package for supporting a comprehensive and flexible species distribution modelling workflow. Methods Ecol Evol 13(8):1661–1669

Venette RC (2017) Climate analyses to assess risks from invasive forest insects: simple matching to advanced models. Curr For Rep 3:255–268

Vieira RG et al (2016) Comparative performance analysis between static solar panels and single-axis tracking system on a hot climate region near to the equator. Renew Sustain Energ Rev 64:672–681

Vignali S et al (2020) SDMtune: An R package to tune and evaluate species distribution models. Ecol Evol 10(20):11488–11506

Waage JK, Mumford JD (2008) Agricultural biosecurity. Philos Trans R Soc B Biol Sci 363(1492):863–876

Warren DL, Seifert SN (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecol Appl 21(2):335–342

Webber BL et al (2011) Modelling horses for novel climate courses: insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models. Diver Distrib 17(5):978–1000

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4

Wilson PD (2024) fitMaxnet: fit MaxEnt niche models using Maxnet. R Package Version 0.4.7. https://github.com/peterbat1/fitMaxnet

Yackulic CB et al (2013) Presence-only modelling using MAXENT: when can we trust the inferences? Methods Ecol Evol 4(3):236–243

Yakoubi S et al (2008) Algerian watermelon mosaic virus (AWMV): a new potyvirus species in the PRSV cluster. Virus Genes 37:103–109

Yáñez-López R et al (2012) The effect of climate change on plant diseases. Afr J Biotechnol 11(10):2417–2428

Zhang Z et al (2020) To invade or not to invade? Exploring the niche-based processes underlying the failure of a biological invasion using the invasive Chinese mitten crab. Sci Total Environ 728:138815

Zhou Y et al (2023) Niche shifts and range expansions after invasions of two major pests: the Asian longhorned beetle and the citrus long horned beetle. Pest Manag Sci 79(9):3149–3158

Zizka A et al (2019) Coordinate Cleaner: Standardized cleaning of occurrence records from biological collection databases. Methods Ecol Evol 10(5):744–751

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.