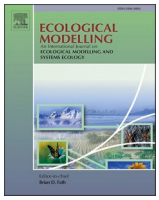
Ecological Modelling 483 (2023) 110454

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Effect of study area extent on the potential distribution of Species: A case study with models for *Raoiella indica* Hirst (Acari: Tenuipalpidae)

George Amaro [a], Elisangela Gomes Fidelis [b], Ricardo Siqueira da Silva [c,*], Cesar Augusto Marchioro [d]

[a] *Embrapa Roraima, Boa Vista, RR, Brazil*
[b] *Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF, Brazil*
[c] *Departamento de Agronomia, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, MG, Brazil*
[d] *Departamento de Agricultura, Biodiversidade e Florestas, Universidade Federal de Santa Catarina, Curitibanos, SC, Brazil*

ARTICLE INFO

Keywords:
Modeling process
Species distribution models
Ecological niche
Study area extent
Invasive species

ABSTRACT

Ecological niche models are used to quantify the relationships between known occurrence records of a given species and environmental variables at these locations. Maxent is among the most widely used algorithms for modeling species distribution and has demonstrated better performance compared to other methods. However, the extent of the study area is a critical issue in the development of presence-only species distribution models because it encompasses the region used to extract the background points employed to characterize the environments accessible to the species. Thus, this study evaluated the effect of the extension of the study area on the species distribution modeling with the Maxent algorithm and occurrence data from the invasive species *Raoiella indica* Hirst (Acari: Tenuipalpidae). The increase in the study area extent inflated most of the threshold-dependent and -independent metrics used to assess model performance. The selection of the study area also affected the predicted suitable areas for the species (its potential distribution). The analysis shows that models developed with smaller study areas resulted in model overfitting and an increase in false-negative predictions. The extent of the area used during model training has a strong influence on the model outputs, with significant consequences for predicting the potential distribution of invasive species and thus for the areas under risk of invasion.

1. Introduction

The interest in describing, understanding, and predicting the geographic and environmental distribution of species has existed for a long time (Wallace, 1860; Grinnell, 1917). Therefore, in recent decades several methods have been proposed for this purpose (Stockwell and Peterson, 2002; Phillips et al., 2006; Soberón et al., 2017; Valavi et al., 2021). In general, such models use associations between known species occurrence records and environmental variables to estimate potential geographic distributions (Elith, 2017). Since the late 1990s, two main terminologies have been used to refer to modeling methods that correlate known occurrence records and environmental conditions: "species distribution models" and "ecological niche models (ENMs)" (Peterson and Soberón, 2012). ENMs are extensively employed in numerous studies on ecology and evolution, including those focused on identifying suitable areas for invasive species (e.g. Zhu et al., 2012; Zhang et al., 2021, Marchioro and Krechemer, 2023). This aids in decision-making regarding the implementation of phytosanitary measures aimed at preventing new instances of biological invasion (Bradshaw et al., 2016).

When employing correlative models to estimate species distributions, three common approaches are used, which vary based on the characteristics of the occurrence data: (1) using presence-only data; (2) using presence-absence data when available; and (3) using presence data and a sample of background data (Sillero et al., 2021). Background data represents a sample from the study area that helps characterize the environmental conditions throughout the region under investigation, including the accessible sites where the species could potentially exist (Fernandez et al., 2022). In recent decades, methods that do not rely on absence data have become popular, with the maximum entropy (Maxent) algorithm, in particular, gaining prominence (Phillips et al., 2006).

Maxent is a machine learning algorithm that was developed specifically to estimate potential distribution of species for scenarios where

* Corresponding author.
E-mail address: ricardo.siqueira@ufvjm.edu.br (R.S. da Silva).

https://doi.org/10.1016/j.ecolmodel.2023.110454
Received 10 February 2023; Received in revised form 29 June 2023; Accepted 3 July 2023

only presence data are available (Phillips et al., 2006; Elith et al., 2011; Merow et al., 2013), and it shows good performance compared to other methods (Elith et al., 2006; Heikkinen et al., 2012; Venette, 2017; Feng et al., 2019; Jha et al., 2022). This algorithm estimates habitat suitability by contrasting environments where the species occurs with those sampled as backgrounds to determine which combinations of variables best predict the known distribution of the species. Therefore, background data establish the environmental domain of the study, while presence data should establish under what conditions a species is more likely to be present than average (Hijmans and Elith, 2021).

Studies have shown that the extent of the study area is critical for the development of presence-only niche models with pronounced effects on model performance and predictions, particularly for Maxent models (VanDerWal et al., 2009; Barbet-Massin et al., 2012; Anderson and Raza, 2010; Khosravi et al., 2016; Cooper and Soberón, 2018; Machado-Stredel et al., 2021). Some studies state that it is preferable to restrict background sampling based on an ecological reasoning or in a way that explains sampling bias, rather than sampling the entire background environment (Philips et al., 2009; Rodda et al., 2011). Conversely, others indicate that a smaller area closer to known occurrence points produces more accurate distributional predictions, whereas extensive study areas can have a negative impact on model performance and occasionally lead to inaccurate predictions (VanDerWal et al., 2009; Elith et al., 2010).

The size and methodology used to define the study area are particularly relevant for invasive species. Numerous approaches can be found in the literature for delineating the study area, including the creation of buffers (e.g. Fourcade et al., 2014), minimum convex polygons (e.g. Chetan et al., 2014), or the largest possible rectangle around the occurrence records (e.g. Jarnevich et al., 2017, Amaro et al., 2021). Alternatively, some studies use ecological classifications such as climate zones or biomes (e.g. Hill and Terblanche, 2014; Mota et al., 2022). These methods result in study areas of varying extents. Furthermore, depending on the purpose of the study, it can be essential to capture the full extent of a species' presence to ensure the representation of entire ecological niche. However, in contexts where occurrence records exist in invaded areas, the inclusion of these records remains a subject of debate. The common argument is that these occurrences provide a closer approximation of a species' fundamental niche and its potential for a future invasion, while records from the native area better reflect the realized niche (Elith, 2017). Previous studies have found that distribution models for invasive species benefit from incorporating data from the invaded area (Mau-Crimmins et al., 2006; Broennimann and Guisan, 2008; Beaumont et al., 2009; Sales et al., 2017), as it enables the identification of conditions under which species can establish and thrive. However, contrasting results have also been reported in the literature (Vaclavık and Meentemeyer 2012; Barbet-Massin et al., 2018). Depending on the approach employed to define the study area, the inclusion of invasive records, which are typically geographically distant from native records, can significantly expand the background extent.

The red palm mite, *Raoiella indica* Hirst (Acari: Tenuipalpidae), was used as a model organism to evaluate the effects of study areas with different extents generated with three methods on model performance and the resulting predictions. This is a phytophagous species that feeds on several palm species (family Arecaceae), including economically important monocotyledonous plants such as coconuts and bananas (Roda et al., 2012; Otero-Colina et al., 2016; Gondim et al., 2012). Predicting the potential distribution of invasive species like *R. indica*, which have the capacity to cause significant economic losses, is crucial for implementing phytosanitary measures aiming at preventing new events of biological invasion (Peterson, 2003; McGeoch et al., 2010; Jiménez-Valverde et al., 2011). In this context, understanding how the extent of the study area impacts Maxent outputs can contribute to the development of more accurate models, which, in turn, can assist in identifying regions with a higher risk of invasion.

2. Material and methods

2.1. Occurrence records and environmental data

A total of 220 records (from native and invaded regions) of *R. indica* were obtained from published literature (Amaro et al. 2021) and were cleaned via CoordinateCleaner v.2.0–20 R package (Zizka et al., 2019) using the tests "zeros", "seas", "equal", "institutions", "duplicates", "centroids", "gbif", "validity", "capitals", resulting in 203 observations. Sampling bias resulting from heterogeneous geographical sampling (Moua et al., 2020) was reduced by applying an environmental filter (Varela et al., 2014) on the occurrences via the "occfilt_env" function in flexsdm R package v.1.3.0 (Velazco et al., 2022), using five classes. A final count of 150 observations was used for model calibration in this study (Fig. 1).

Nineteen bioclimatic variables derived from the WorldClim 2.1 database (Fick; Hijmans, 2017) at 2.5 min spatial resolution (~ 5 km at the equator) were used in the study. These variables were selected because they represent temperature and precipitation conditions that are known to constrain the distribution of different organisms (Slater and Michael, 2012). Furthermore, such bioclimatic variables were previously used in a modeling study performed with *R. indica* (Amaro et al., 2021). Correlated variables were removed from the analysis using the variance inflation factor (VIF; Marquaridt, 1970), performed with *usdm* package (Naimi et al., 2014). Variables with VIF $>$ 10 (Naimi et al., 2014) were excluded. A final set of 9 variables (Bio2, Bio3, Bio8, Bio9, Bio13, Bio14, Bio15, Bio18, and Bio19) were used in subsequent analysis.

2.2. Delimitation of the study area

Overall, six different study areas (calibration areas) were evaluated using three approaches: (1) buffers of different sizes (200, 400, 800, and 1200 km) around the occurrence records, (2) the Minimum Convex Polygon (MCP) approach, and (3) the enclosing rectangle method (ER). The calibration areas were delimited with the "calib_area" function in the flexsdm R package. MCP is the smallest polygon in which no internal angle exceeds 180° and which contains all occurrence records. On the other hand, ER method creates the largest possible rectangle encompassing the occurrence records. These methods were selected because they are commonly employed in different studies with the Maxent algorithm (Rodda et al., 2011; Chetan et al., 2014; Fourcade et al., 2014; Barga et al., 2018; Amaro et al., 2021, among others).

2.3. Modeling process

The models were developed with the Maxent algorithm using the flexsdm R package through a non-homogeneous Poisson process (Phillips et al., 2017). This machine-learning algorithm is widely used to predict the potential distribution of insect pests due to its robust statistical performance (Elith et al., 2010). Maxent has two main settings that affect model performance and transferability into geographical space: (1) features classes (FCs), and (2) regularization multiplier values (RM) (Merow et al., 2013; Sutton and Martin, 2022). A feature corresponds to a mathematical transformation of the different covariates used in the model to allow complex relationships to be modeled (Elith et al., 2010). The regularization multiplier is a parameter that adds new constraints; that is, it is a penalty imposed on the model. The main objective is to avoid excessive complexity and/or overfitting, controlling the intensity of the features used to build the model (Elith et al., 2010; Shcheglovitova and Anderson, 2013). In this context, the "tune_max" function of the flexsdm R package was used to generate 50 Maxent models with different combinations of FCs (linear -L, quadratic - Q, hinge -H, LQ and LQH) and RMs ranging from 0.5 to 5 (with an increment of 0.5). Product features were omitted because the (marginal) response curves for each predictor variable completely define the model

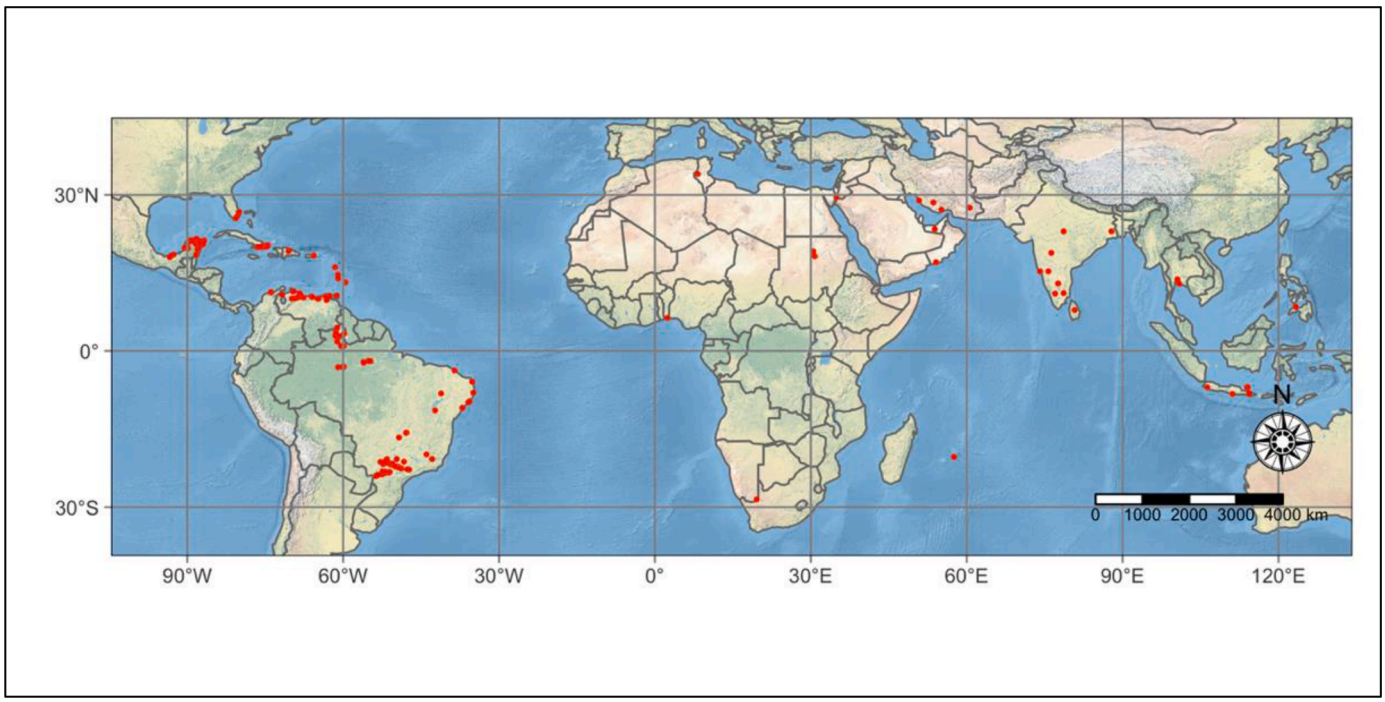


Fig. 1. Data on the presence (record of occurrences) of Raoiella indica Hirst (Acari; Tenuipalpidae) (red dots), as described in Amaro et al., 2021. Studies show that Africa or the Middle East is the most probable origin of the Raoiella (Dowing et al. 2012). Thus, registers in America are invasive.

and are easier to interpret than those that depend on the values of other variables. The threshold that maximizes the sum of sensitivity and specificity was used to obtain the binary maps of environmental suitability.

To control the potential spatial autocorrelation between model calibration and test data and improve its transferability, the spatial block cross-validation method was used for data partitioning (Roberts et al., 2017; Santini et al., 2021). Because calibration and evaluation localities are often close to each other, localities used to evaluate model performance are not truly independent of those used to calibrate it. Therefore, due to the spatial autocorrelation of the environment, they do not provide realistic tests of model performance, typically leading to

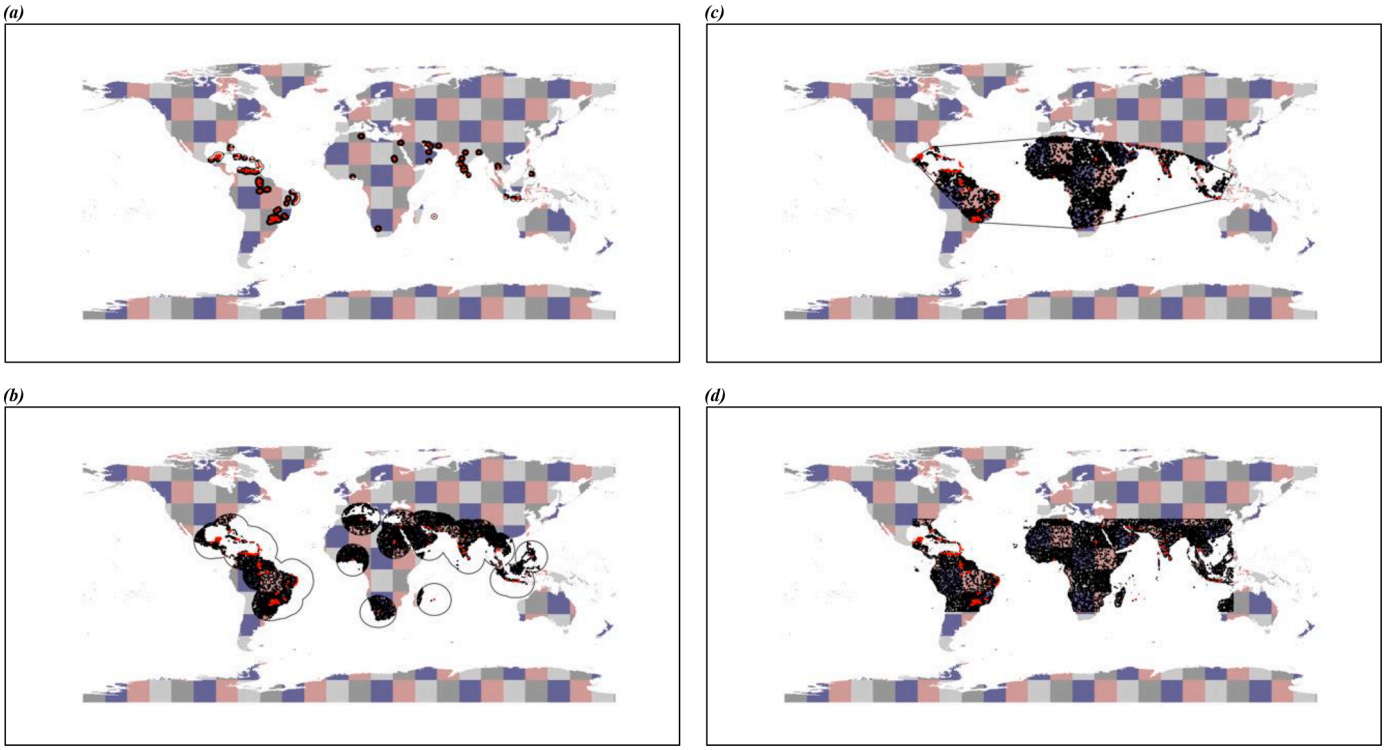


Fig. 2. Distribution of occurrences of Raoiella indica Hirst (Acari; Tenuipalpidae) (red dots), background samples (black dots), 200 km buffer (a), 1200 km buffer (b), minimal convex polygon (c) and rectangle wrapper (d) on the grid created by partitioning into blocks.

overestimates of performance (Radosavljevic and Anderson, 2014). The spatial block cross-validation allows the evaluation of the model using spatially independent data from the calibration data (Radosavljevic and Anderson, 2014). This was performed with the "part_sblock" function in the flexsdm R package, using four partitions and keeping the default values for the other options, resulting in 47, 63, 28, and 12 presence records for each of the created partitions. The samplings of background points in the calibration area were proportional to the occurrences for each partition using the random points method in the "sample_background" function within the flexsdm R package. Overall, 3133, 4200, 1867, and 800 background points were generated for each partition (10, 000 in total following Maxent default). Except for the calibration area, all other parameters were kept constant. The distribution of presence, background, buffer (800 km), and polygon (MCP) points can be viewed in Fig. 2, illustrating the differences in the methods used to delimit the study area.

The selection of optimal model settings was performed for each of the six background extents evaluated using the True Skill Statistics (TSS) (Allouche et al., 2006). The best model for each of the six background extents was then projected onto Brazil using the "sdm_predict" function. The area predicted as suitable was calculated based on the binary map generated with a threshold that maximizes the sum of sensitivity and specificity (MSS), using the "expand" function of the terra R package v. 1.5–21 (Hijmans, 2022).

2.4. Evaluation of model performance

Due to all the issues associated with distribution model evaluation and the lack of consensus on the best measure, following the recommendation of Konowalik and Nosol (2021), we chose to present multiple model metrics to enable comparisons and references: (1) Inverse Mean Absolute Error (IMA), (2) True Positive Rate (TPR), (3) True Negative Rate (TNR), (4) Sorensen, (5) Jaccard, (6) F Presence-Background (FPB), (7) Omission Rate at MSS (OR), (8) TSS, (9) Area Under the Curve (AUC), and (10) the continuous Boyce Index (BI). IMAE is calculated as 1 – Mean Absolute Error to be consistent with the other metrics, where the higher the value of a given performance metric, the higher the model accuracy (Velazco et al., 2022). TPR or sensitivity is the proportion of correctly predicted presences among all presences (Fielding and Bell, 1997). In contrast, TNR, or specificity, measures the proportion of correctly predicted absences among all absences (Fielding and Bell, 1997). Sorensen and Jaccard are the similarity indices of Sørensen (Sorensen, 1948; Leroy et al., 2018) and Jaccard (Jaccard, 1912; Leroy et al., 2018), respectively, widely used in ecology applications. FPB corresponds to the weighted harmonic mean between precision and sensitivity (Daskalaki et al., 2006), considering the presence and background points. The omission rate indicates the percentage of test sites that fall into areas not predicted to be suitable for the species (Phillips et al., 2006; Fielding and Bell, 1997). TSS is the mean prediction success rate for present and absent sites (Allouche et al., 2006). AUC of the Receiver Operating Characteristic Curve (ROC) (Fielding and Bell, 1997) expresses a threshold-independent metric and, despite several criticisms, continues to be used. Finally, the continuous Boyce index measures agreement between predicted gradients of habitat suitability and the distribution of retained test points (Hirzel et al., 2006). This metric requires only presence data and is independent of a threshold. A higher value indicates that the model's predictions are consistent with the observed distribution of presences in the test regions. All analyses were carried out in R statistical environment (R Core Team, 2022).

3. Results

The results show a pronounced influence of the method used to delimit the calibration area on model selection, performance, and prediction (Table 1). Both RM values and FCs varied according to the calibration area. These parameters varied even when the same method was used to define the calibration area, as evidenced by the contrast outcomes when comparing results from the buffer method with varying radius sizes. For most calibration areas, the selected models consist of only hinge feature classes (three models). However, two models incorporated LQ feature classes. Regarding RM values, two selected models had RMs values of 0.5 and 1.0 (Table 1).

The area predicted as suitable varied widely among the methods and extent of the calibration area. For example, a difference of 91% in the predicted suitable areas was observed between a study area delimited with a buffer of 200 km and one delimited with 1200 km. Conversely, such a difference was only 2% between MCP and ER (Table 1). The differences shown in Figs. 3 and 4 clearly demonstrate that the method used to delimit the study area clearly affects the predicted suitable geographic regions for the species. For instance, contrary to the buffer and ER models, MCP models predicted suitable areas for *R. indica* in southern Brazil (Fig. 4, Fig. 5e). Similar divergences were also observed in midwestern Brazil, where the model with a 200 km buffer predicted unsuitable conditions for *R. indica*, contrary to the other models (Fig. 4, Fig. 5). In fact, numerous presence records, particularly in the states of Minas Gerais, São Paulo, and Paraná, were found outside the region considered suitable for the species (Fig. 5a). This indicates that the model exhibits identifiable errors of omission. Furthermore, the evaluation of the probability frequency histograms (Fig. 4) reveals that as the study area expands, the potential geographic distribution becomes more localized and less inclusive overall. Consequently, records in areas with lower probabilities of occurrence are not classified as suitable, particularly considering that the models generated with buffers of 200 and 400 km yielded higher threshold values (Fig. 5).

Our findings also show a clear influence of the calibration area on most metrics used to evaluate model performance (Table 2, Fig. 3). In general, except for TPR and OR, the increase in the calibration area inflated most of the metrics used. A different pattern was observed

Table 1
Results of models generated by the flexsdm R package for different buffer sizes (200, 400, 800, 1200 km), minimum convex polygon (MCP) and enveloping rectangle (ER), considering the extension entire study area.

Results	200 km	400 km	800 km	1.200 km	MCP	ER
RM	2.50	3.50	0.50	0.50	1.00	1.00
Features	Q	H	LQ	LQ	H	H
Threshold [A]	0.69	0.69	0.51	0.48	0.50	0.53
Area (km^2) [B]	5,492,933	12,954,210	28,696,074	45,292,591	48,048,774	63,768,341
Suitable area (km^2) [C]	1,827,308	2,515,338	3,449,437	3,487,684	2,928,716	2,852,826
% suitable area [D]	33.27	19.42	12.02	7.70	6.10	4.47
Background% [E]	14.79	6.23	2.79	1.75	1.71	12.67

[A] maximum (sensitivity + specificity).
[B] Extent of the area used to calibrate the model (study area).
[C] Area of potential geographic distribution, calculated based on the binary map (presence/absence).
[D] Percentage of potential geographic distribution area relative to the total extent of the study area.
[E] Percentage of the study area occupied by background points.

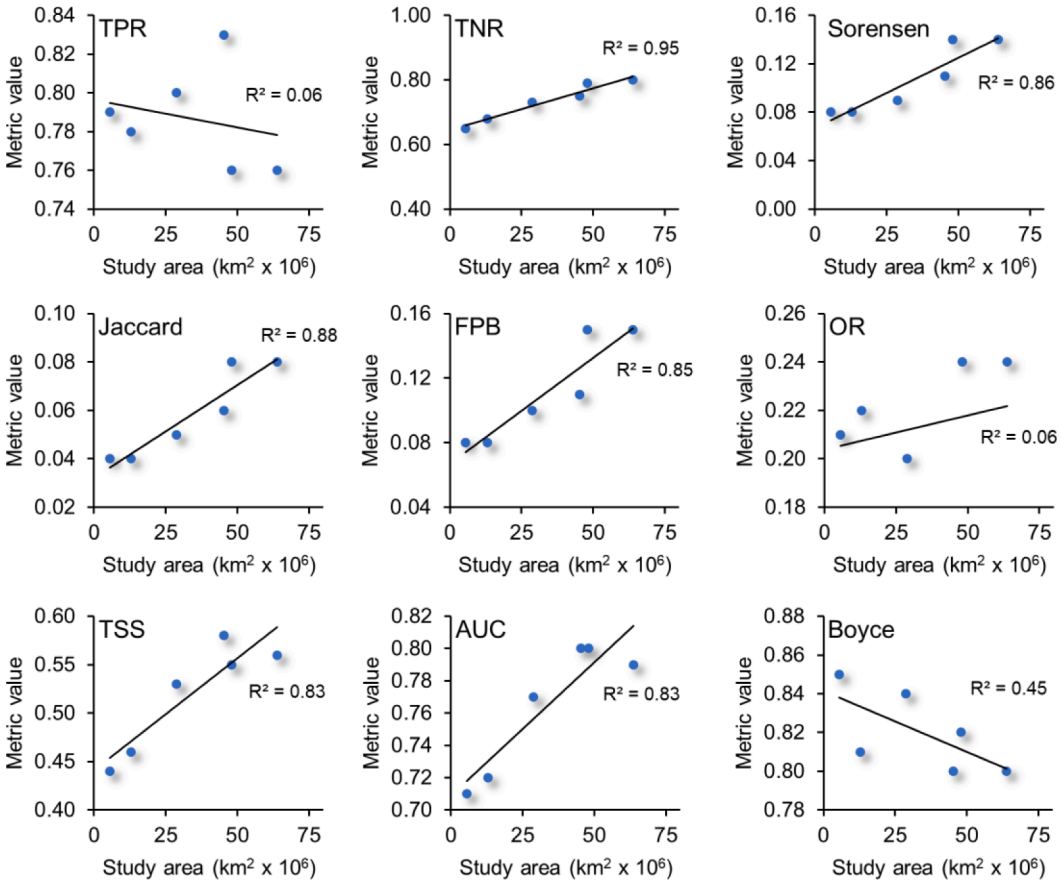


Fig. 3. Relationship between the extension of the study area (X-axis) and the metric values (Y-axis), including a simple linear regression (line) and their respective correlation coefficients (R^2), to illustrate trends related to area increase in each metric. The metric are true positive rate (TPR), true negative rate (TNR), Sørensen, Jaccard, F presence-background (FPB), omission rate (OR), true skill statistics (TSS), area ander the curve (AUC), and the continuous Boyce index (Boyce).

between areas delimited by buffer and polygon regarding the metrics TPR and OR. Both TPR and OP decreased as the buffer size increased. By contrast, the MCP and ER models presented very similar metrics (Table 2) and, although the total area of the extension used for the ER model was 33% larger than the extension area of the MCP model, its result, in terms of the predicted potential distribution area of *R. indica* (Fig. 5), was only 3% lower. Interestingly, regardless the observed variation in the metrics used to evaluate model performance, in general, all models exhibited good statistical performance.

4. Discussion

In this study, we evaluated the effects of the extension of the study area on both model performance and predicted suitability for an economically important agricultural pest. The findings clearly demonstrate that the choice of method used for defining the calibration area directly influences model performance and the resulting suitability maps. Thus, it is crucial to carefully consider this aspect when developing maxent models, especially for invasive species.

The projections of presence-background correlative models can be biased if the study area encompasses regions that are suitable for the species but remain unoccupied due to various factors, including: (1) limited dispersal or transient populations (Anderson and Raza, 2010; Elith et al., 2011); (2) biotic interactions (Anderson 2017; Jarnevich et al., 2015); and/or (3) environmental changes resulting from human activity (Anderson and Raza, 2010; Jiménez-Valverde et al., 2011). Therefore, ideally, the selection of the study area should consider the assumption that none of these three factors account for the species occupying an environmentally unsuitable subset. Consequently, despite the satisfactory evaluation metrics demonstrated by the MCP and ER models, their use could be deemed inappropriate for *R. indica*, given the extensive regions of Africa where no occurrences are known. However, it is not clear whether field surveys were conducted in this region and failed to record the presence of *R. indica* or if such surveys were not carried out.

The size of the buffers used to define the calibration area had a marked influence on model predictions. Predictions with very small areas such as the ones obtained with the 200 km buffer may indicate model overfitting and result in false-negative predictions. This is often the case when area sampling is insufficient to capture the full range of environmental conditions occupied by a species. On the other hand, extensive areas can lead to the identification of locations that are not occupied by the species, resulting in false-positive predictions. This may be attributed to areas that lie beyond the species' actual distribution or indicate dispersal limitations due to barriers or biotic interactions, such as competition and predation (Raxworthy et al., 2007).

The extent of the area used during the modeling process has a great influence on model performance and, if it is very limited, the importance of some factors in delimiting the distribution can be underestimated (Barve et al., 2011). Increasing the extent often includes absences that are more environmentally distant from the presence records and, consequently, some metrics like AUC and TSS tend to increase (Lobo et al., 2008). This was observed when the evaluation metrics of buffer models with different radius size were compared, as seen in Table 1. The threshold-dependent metric AUC, for example, increased from 0.71 in models with buffers of 200 km to 0.80 in models with calibration areas generated with buffers of 1200 km. By contrast, other metrics like CBI and OR did not follow this pattern.

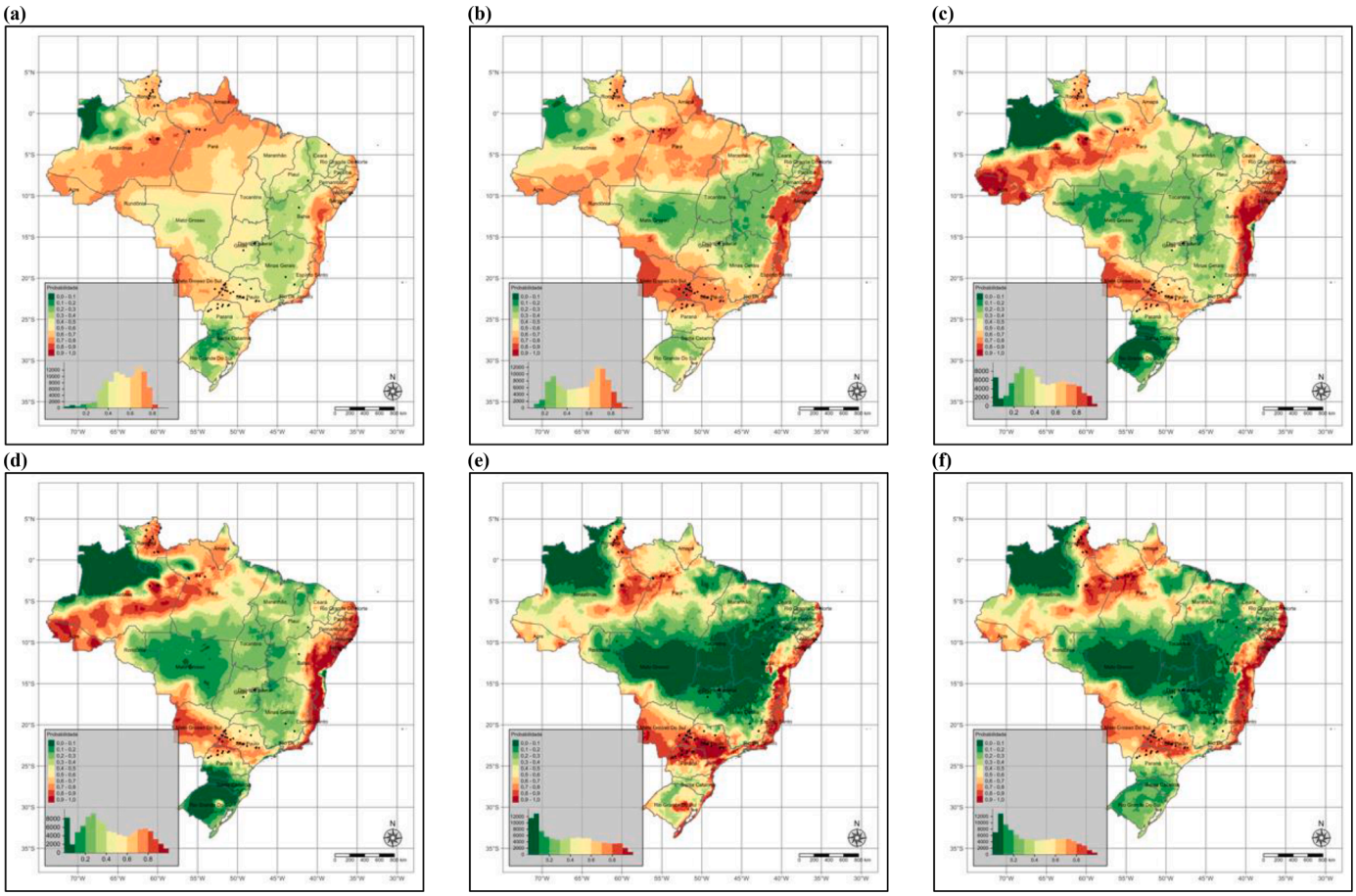


Fig. 4. Maps of the potential geographic distribution of Raoiella indica in Brazil, showing the relative probability of occurrence in intervals of 0.1 for buffers of 200 (a), 400 (b), 800 (c) and 1200 km (d), MCP (e), and enclosing rectangle (f).

The metrics employed to assess model performance indicate that all models showed satisfactory results, regardless of the method used to delimit the study area. For instance, despite substantial variations in the suitability maps generated by models using study areas defined with buffers of 200 km and1200 km, both models consistently achieved high evaluation metrics. This is evident from the obtained AUC and Boyce index values, which were consistently above 0.70 and 0.80, respectively. This result reinforces the importance of using an ecological criterion in the definition of study area (Andersen et al., 2022; Elith et al., 2011; Marchetto et al., 2023; Phillips and Dudík, 2008; Kramer-Schadt et al., 2013). Ideally, dispersal and habitat processes should be considered to adequately delimit the extension area for adequately modeling species dispersion. However, this information is not always available, and, in the special case of invasive species, the dispersion characteristics of the species do not depend only on its capacity but mainly on other agents that act as vectors. Some studies employ climate zones or biomes with one or more occurrence records to delimit the extent of the study area (Hill and Terblanche, 2014; Hill et al., 2017). In this context, the use of an "exploratory modeling" process, as done in this work, can help to understand the details underlying the current and potential species dispersion, which can be decisive for the adequate choice of calibration parameters of the models, especially considering the economic risks involved, resulting from the presence of invasive species.

Considering that *R. indica* is an invasive species classified as a quarantine pest present in Brazil, the objective is to identify areas where it represents a more significant risk of invasion. Together with assessments related to the economic importance of host plants and dispersion routes, maps of potential distribution can be used to guide public policies for phytosanitary control. Therefore, the cost of false negatives would potentially be greater than the cost of false positives for those locations where host plants are economically relevant.

5. Conclusions

In summary, this study assessed the effects of six study area extents generated with three different methods on model performance and the resulting environmental suitability maps. Our findings clearly demonstrate that both the extent and the method used to delimit the study area affect model performance and environmental suitability. Although all models showed good performance regardless of the extent area used, pronounced differences were observed in the resulting suitability maps. Such differences were also observed in models with study areas created with buffers of different sizes. This is particularly relevant for invasive species, such as the case of *R. indica*, when models are used to identify areas with a higher risk of invasion. In such circumstances, errors in the identification of areas at risk of pest establishment may result in significant economic and environmental losses.

CRediT authorship contribution statement

George Amaro: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Elisangela Gomes Fidelis:** . **Ricardo Siqueira da Silva:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Cesar Augusto Marchioro:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

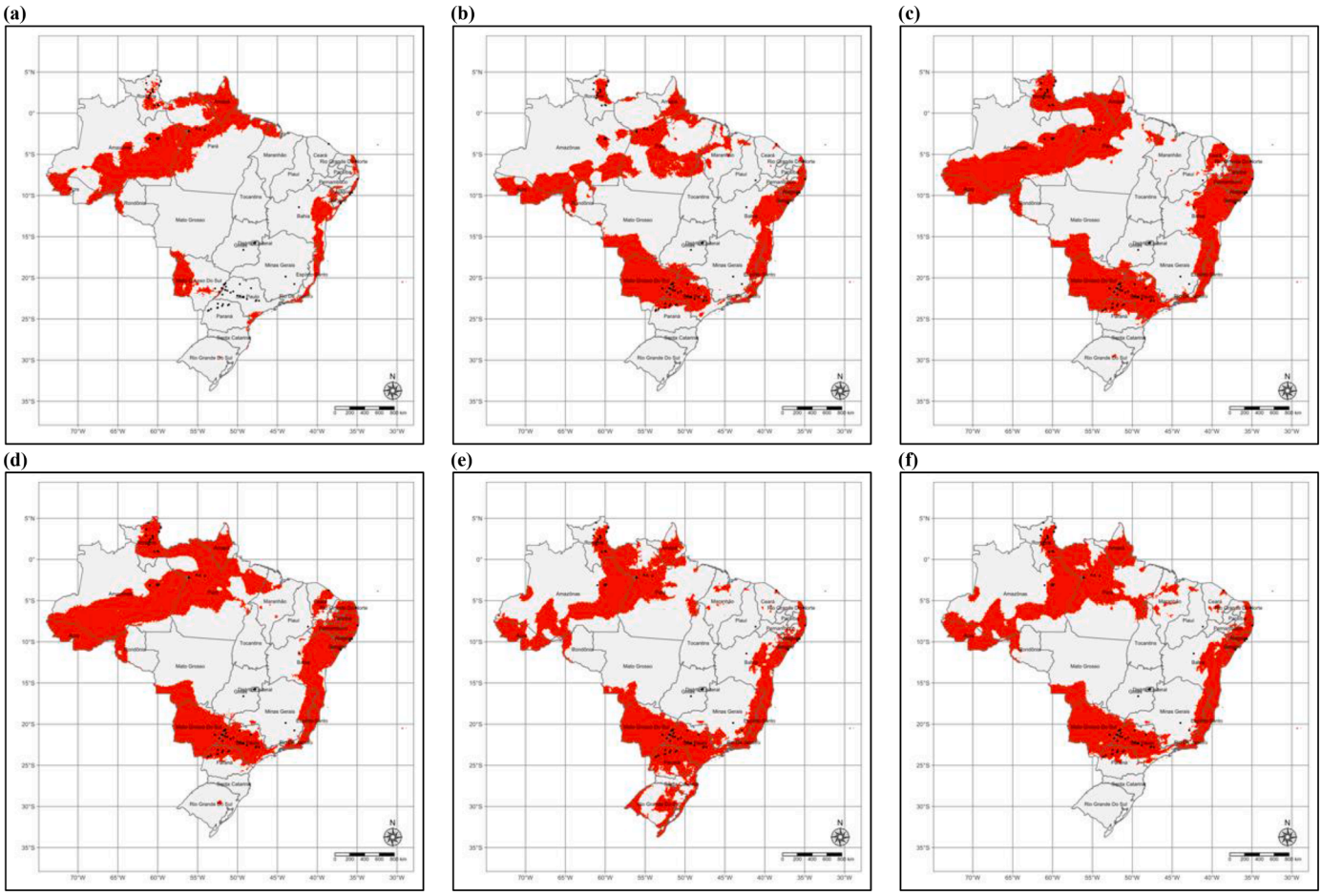


Fig. 5. Maps of presence (red) / absence of Raoiella indica in Brazil created using the threshold that maximizes the sum of sensitivity and specificity for buffers of 200 (a), 400 (b), 800 (c) and 1200 km (d), MCP (e) and enclosing rectangle (f).

Table 2
Threshold-dependent and -independent metrics (average values and standard deviations) used to evaluate the models developed using six different study area extensions derived from buffer of different sizes (200, 400, 800, 1200 km), minimum convex polygon (MCP) and enveloping rectangle (ER).

Metrics	200 km	400 km	800 km	1.200 km	MCP	ER
IMAE (sd)	0.08	0.04	0.03	0.03	0.03	0.01
TPR (mean)	0.79	0.78	0.80	0.83	0.76	0.76
TPR (sd)	0.18	0.12	0.11	0.08	0.05	0.02
TNR (mean)	0.65	0.68	0.73	0.75	0.79	0.80
TNR (sd)	0.26	0.18	0.15	0.16	0.14	0.14
Sorensen (mean)	0.08	0.08	0.09	0.11	0.14	0.14
Sorensen (sd)	0.03	0.02	0.03	0.04	0.10	0.09
Jaccard (mean)	0.04	0.04	0.05	0.06	0.08	0.08
Jaccard (sd)	0.02	0.01	0.02	0.02	0.06	0.05
FPB (mean)	0.08	0.08	0.10	0.11	0.15	0.15
FPB (sd)	0.03	0.03	0.04	0.05	0.12	0.10
OR (mean)	0.21	0.22	0.20	0.17	0.24	0.24
OR (sd)	0.18	0.12	0.11	0.08	0.05	0.02
TSS (mean)	0.44	0.46	0.53	0.58	0.55	0.56
TSS (sd)	0.15	0.11	0.13	0.14	0.14	0.15
AUC (mean)	0.71	0.72	0.77	0.80	0.80	0.79
AUC (sd)	0.09	0.07	0.08	0.07	0.05	0.06
Boyce (mean)	0.85	0.81	0.84	0.80	0.82	0.80
Boyce (sd)	0.11	0.27	0.17	0.24	0.22	0.20

Declaration of Competing Interest

The authors declare that they have no conflicts of interest. This manuscript has not been published previously, also it is not under consideration for publication elsewhere. All authors approve this manuscript, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically.

Data availability

Data will be made available on request.

Acknowledgments

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. This research was supported by Empresa Brasileira de Pesquisa Agropecuária -Embrapa (Project number: 10.20.03.056.00.00), Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Grant Number: 307852/2021–0), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES -Code finance 001 and Fundação de Amparo à Pesquisa do Estado de Minas Gerais-FAPEMIG.

References

Andersen, D., Litvinchuk, S.N., Jang, H.J., Jiang, J., Koo, K.S., Maslova, I., Kim, D., Jang, Y., Borzée, A., 2022. Incorporation of latitude-adjusted bioclimatic variables increases accuracy in species distribution models. Ecol. Modell. 469, 109986.

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

Amaro, G., et al., 2021. Current and potential geographic distribution of red palm mite (*Raoiella indica* Hirst) in Brazil. Ecol. Informa. 65, 101396.

Anderson, R.P., 2017. When and how should biotic interactions be considered in models of species niches and distributions? J. Biogeogr. 44 (1), 8–17.
Anderson, R.P., 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, 1378–1393.
Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Meth. Ecol. Evol. 3, 327–338.
Barbet-Massin, M., Rome, Q., Villemant, C., Courchamp, F., 2018. Can species distribution models really predict the expansion of invasive species? PLoS ONE 13, e0193085.
Barga, C.S., Dilts, T.E., Leger, E.A, 2018. Contrasting climate niches among co-occurring subdominant forbs of the sagebrush steppe. Divers. Distrib. 23, 1291–1307.
Barve, N., et al., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol. Model. 222, 1810–1819.
Beaumont, L.J., Gallagher, R.V., Thuiller, W., Downey, P.O., Leishman, M.R., Hughes, L., 2009. Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. Divers. Distrib. 15, 409–420.
Bradshaw, C.J.A., Leroy, B., Bellard, C., 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, 585–589.
Chetan, N., Praveen, K.K., Vasudeva, G.K., 2014. Delineating ecological boundaries of hanuman langur species complex in Peninsular India using MaxEnt modeling approach. PLoS ONE 9, e87804.
Cooper, J.C., Soberón, J., 2018. Creating individual accessible area hypotheses improves stacked species distribution model performance. Glob. Ecol. Biogeogr. 27, 156–165.
Daskalaki, S., Kopanas, I., Avouris, N., 2006. Evaluation of classifiers for an uneven class distribution problem. App. Artif. Intell. 20, 381–417.
Dowling, A.P.G., Ochoa, R., Beard, J.J., Welbourn, W.C., Ueckermann, E.A., 2012. Phylogenetic investigation of the genus *Raoiella* (Prostigmata: tenuipalpidae): diversity, distribution, and world invasions. Exp. Appl. Acarol. 57, 257–269.
Elith, J., 2017. Predicting distributions of invasive species. In: Robinson, A.P., Walshe, T., Burgman, M.A., Nunn, M. (Eds.), Invasive species: Risk Assessment and Management. Cambridge University Press, Cambridge, UK, pp. 93–129.
Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A.R., Hijmans, J., Huettmann, F.J., Leathwick, R., Lehmann, A., et al., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.
Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342.
Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57.
Feng, X., Park, D.S., Walker, C., Peterson, A.T., Papes, M., 2019. A checklist for maximizing reproducibility of ecological niche models. Nat. Ecol. Evol. 3, 1382–1395.
Fernandez, M., Sillero, N., Yesson, C., 2022. To be or not to be: the role of absences in niche modelling for highly mobile species in dynamic marine environments. Ecol. Model. 471, 110040.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.
Fourcade, Y., Engler, J.O., Rödder, D., Secondi, J., 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. PLoS ONE 9, e97122.
Gondim JR., M.G.C., Castro, T.M.M.G., Marsaro, A.L., Navia, D., Melo, J.W.S., Demite, P. R., Moraes, G.J., 2012. Can the red palm mite threaten the Amazon vegetation? Syst. Biodivers. 10, 527–535.
Grinnell, J., 1917. The niche-relationships of the California thrasher. Auk 34, 427–433.
Heikkinen, R.K., Marmion, M., Luoto, M., 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35, 276–288.
Hijmans, R.J., Elith, J. 2021. Species distribution models. Disponível em https://rspatial.org/raster/sdm/raster_SDM.pdf. Acessado em: 12 jul. 2022.
Hill, M.P., Terblanche, J.S., 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, e90121.
Hijmans, R.J.. Terra: spatial data analysis. https://CRAN.R-project.org/package=terra. Download.
Hill, M.P., Gallardo, B., Terblanche, J.S., 2017. A global assessment of climatic niche shifts and human influence in insect invasions. Glob. Ecol. Biogeogr. 26, 679–689.
Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A, 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199, 142–152.
Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. N. Phytol 11, 37–50.
Jarnevich, C.S., Stohlgren, T.J., Kumar, S., Morisette, J.T., Holcombe, T.R., 2015. Caveats for correlative species distribution modeling. Ecol. Inform. 29, 6–15.
Jarnevich, C.S., Talbert, M., Morisette, J., Aldridge, C., Brown, C.S., Kumar, S., Mannier, D., Talbert, C., Holcombe, T., 2017. Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: an example with background selection. Ecol. Model. 363, 48–56.
Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P., Lobo, J.M., 2011. Use of niche models in invasive species risk assessments. Biol. Invasions. 13, 2785–2797.
Jha, A., Praveen, J., Nameer, P.O., 2022. Contrasting occupancy models with presence-only models: does accounting for detection lead to better predictions? Ecol. Model. 472, 110105.
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, 1482. https://doi.org/10.1038/s41598-020-80062-1.
Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Diver. Distrib. 19 (11), 1366–1379.
Khosravi, R., Hemani, M.R., Malekian, M., Flint, A., Flint, L., 2016. Maxent modeling for predicting potential distribution of goitered gazelle in central Iran: the effect of extent and grain size on performance of the model. Turk. J. Zool. 40, 574–585.
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, 1994–2002.
Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151.
Machado-Stredel, F., Cobos, M.E., Peterson, A.T., 2021. A simulation-based method for selecting calibration areas for ecological niche models and species distribution models. Front. Biogeogr. 13.4, e48814.
Marchetto, E., Da Re, D., Tordoni, E., Bazzichetto, M., Zannini, P., Celebrin, S., Chieffallo, L., Malavasi, M., Rocchini, D., 2023. Testing the effect of sample prevalence and sampling methods on probability-and favourability-based SDMs. Ecol Modell 477, 110248.
Marchioro, C.A., Krechemer, F.S., 2023. Climatic niche shift and distribution of *Melanagromyza sojae* under current and future climate scenarios: does this species pose a risk to soybean production? Entomol. Exp. Appl. 171, 461–474.
Marquaridt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12, 591–612.
Mau-Crimmins, T.M., Schussman, H.R., Geiger, E.L., 2006. Can the invaded range of a species be predicted sufficiently using only native-range data? Lehmann lovegrass (*Eragrostis lehmanniana*) in the southwestern United States. Ecol. Modell. 193, 736–746.
McGeoch, M.A., Butchart, S.H.M., Spear, D., Marais, E., Kleynhans, E.J., Symes, A., Chanson, J., Hoffmann, M., 2010. Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. Divers. Distrib. 16, 95–108.
Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36, 1058–1069.
Mota, J.S., Barbosa, L.R., Marchioro, C.A., 2022. Suitable areas for invasive insect pests in Brazil and the potential impacts for eucalyptus forestry. Pest. Manag. Sci. 78, 2596–2606.
Moua, Y., Roux, E., Seyler, F., Briolant, S., 2020. Correcting the effect of sampling bias in species distribution modeling - A new method in the case of a low number of presence data. Ecol. Inform. 57, 101086.
Naimi, B., et al., 2014. Where is positional uncertainty a problem for species distribution modelling? Ecography 37 (2), 191–203, 2014.
Otero-Colina, G., González-Gómez, R., Martínez-Bolaños, L., Otero-Prevost, L.G., López-Buenfil, J.A., Escobedo-Graciamedrano, R.M., 2016. Infestation of *Raoiella indica* Hirst (Trombidiformes: tenuipalpidae) on host plants of high socio-economic importance for tropical America. Neotrop. Entomol. 45, 300–309.
Peterson, A.T., 2003. Predicting the geography of species invasion via ecological niche modelling. Q. Rev. Biol. 78, 419–433.
Peterson, A.T., Soberón, J., 2012. Species distribution modeling and ecological niche modeling: getting the concepts right. Nat. Conservação 10, 102–107.
Phillips, S.J., Anderson, R.P., Dudík, M., Shapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography 40, 1–7.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.
Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.
Phillips, S.J., Dudík, M., Elith, J., et al., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.
Core Team, R., 2022. R: A Language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Radosavlevic, A., Anderson, R.P., 2014. Making better MAXENT models of species distributions: complexity, overfitting and evaluation. J. Biogeogr. 41, 629–643.
Raxworthy, C.J., et al., 2007. Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. Sys. Biol. 56, 907–923.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Warton, D.I., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.
Roda, A., Nachman, G., Hosein, F., Rodrigues, J.C., Peña, J.E., 2012. Spatial distributions of the red palm mite, *Raoiella indica* (Acari: tenuipalpidae) on coconut and their implications for development of efficient sampling plans. Exp. Appl. Acarol. 57, 291–308.
Rodda, G.H., Jarnevich, C.S., Reed, R.M., 2011. Challenges in identifying sites climatically matched to the native ranges of animal invaders. PLoS ONE 6, 14670.
Sales, L.P., Ribeiro, B.R., Hayward, M.W., Paglia, A., Passamani, M., Loyola, R., 2017. Niche conservatism and the invasive potential of the wild boar. J. Anim. Ecol. 86, 1214–1223.
Santini, L., et al., 2021. Assessing the reliability of species distribution projections in climate change research. Divers. Distrib. 27, 1035–1050.

Shcheglovitova, M., Anderson, R.P., 2013. Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. Ecol. Model. 269, 9–17.
Sillero, N., Barbosa, M., 2021. Common mistakes in ecological niche models. Int. J. Geogr. Inf. Sci. 35, 213–226.
Slater, H., Michael, E., 2012. Predicting the current and future potential distributions of lymphatic filariasis in Africa using maximum entropy ecological niche modelling. PLoS ONE 7, e32202.
Soberón, J., Osorio-oliveira, L., Peterson, T., 2017. Conceptual differences between ecological niche modeling and species distribution modeling. Rev. Mex. Biodivers. 88, 437–441.
Sorensen, T.A., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol. Skar. 5, 1–34.
Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Modell. 148, 1–13.
Sutton, G.F., Martin, G.D., 2022. Testing MaxEnt model performance in a novel geographic region using an intentionally introduced insect. Ecol. Model. 473, 110139.
Vaclavik, T., Meentemeyer, R., 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Divers. Distrib. 18, 73–83.
Valavi, R., Elith, J., Lahoz-Monfort, J.J., Fuillera-Arroita, G., 2021. Modelling species presence-only data with random forests. Ecography 44, 1731–1742.
Vanderwal, J., Sshoo, L.P., Graham, C., William, S.E., 2009. Selecting pseudo- absence data for presence-only distribution modeling: how far should you stray from what you know? Ecol. Model. 220, 589–594.
Varela, S., Anderson, R.P., García-Valdes, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37, 1084–1091.
Velazco, S.J.E., Rose, M.B., Andrade, A.F.A., Minoli, I., Franklin, J., 2022. flexsdm: an R package for supporting a comprehensive and flexible species distribution modelling workflow. Methods Ecol. Evol. 13, 1661–1669.
Venette, R.C., 2017. Climate analyses to assess risks from invasive forest insects: simple matching to advanced models. Curr. For. Reports 3, 255–268.
Wallace, A.R., 1860. On the zoological geography of the Malay Archipelago. P. Linn. London 4, 172–184.
Zhang, Y., Tang, J., Ren, G., Zhao, K., Wang, X., 2021. Global potential distribution prediction of *Xanthium italicum* based on Maxent model. Sci. Rep. 11, 16545.
Zhu, G., Bu, W., Liu, G., 2012. Potential geographic distribution of brown marmorated stink bug invasion (*Halyomorpha halys*). PLoSONE 7, e31246.
Zizka, A., et al., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 7.