

RESEARCH ARTICLE

Evaluating species distribution model predictions through time against paleozoological records Ignacio A. Lazagabaster^{1,2}  | Chris D. Thomas³ | Juliet V. Spedding⁴ | Salima Ikram^{5,6} | Irene Solano-Regadera¹ | Steven Snape⁴ | Jakob Bro-Jørgensen¹¹Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool, UK²Centro Nacional de Investigación Sobre Evolución Humana CENIEH, Burgos, Spain³Leverhulme Centre for Anthropocene Biodiversity, Department of Biology, University of York, York, UK⁴Department of Archaeology, Classics and Egyptology, University of Liverpool, Liverpool, UK⁵Department of Sociology, Egyptology and Anthropology, The American University in Cairo, New Cairo, Egypt⁶Max Planck Institute for Meteorology, Hamburg, Germany

Correspondence

Ignacio A. Lazagabaster, Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool L69 3BX, UK.
Email: ignacio.aguilar.lazagabaster@gmail.com

Funding information

Leverhulme Trust, Grant/Award Number: RPG-2021-104

Abstract

Species distribution models (SDMs) are widely used to project how species distributions may vary over time, particularly in response climate change. Although the fit of such models to current distributions is regularly enumerated, SDMs are rarely tested across longer time spans to gauge their actual performance under environmental change. Here, we utilise paleozoological presence/absence records to independently assess the predictive accuracy of SDMs through time. To illustrate the approach, we focused on modelling the Holocene distribution of the hartebeest, *Alcelaphus buselaphus*, a widespread savannah-adapted African antelope. We applied various modelling algorithms to three occurrence datasets, including a point dataset from online repositories and two range maps representing current and 'natural' (i.e. hypothetical assuming no human impact) distributions. We compared conventional model evaluation metrics which assess fit to current distributions (i.e. True Skill Statistic, TSS_c , and Area Under the Curve, AUC_c) to analogous 'paleometrics' for past distributions (i.e. TSS_p , AUC_p , and in addition $Boyce_p$, $F2\text{-score}_p$ and $Sorensen_p$). Our findings reveal only a weak correlation between the ranking of conventional metrics and paleometrics, suggesting that the models most effectively capturing present-day distributions may not be the most reliable to hindcast historical distributions, and that the choice of input data and modelling algorithm both significantly influences environmental suitability predictions and SDM performance. We thus advocate assessment of model performance using paleometrics, particularly those capturing the correct prediction of presences, such as $F2\text{-score}_p$ or $Sorensen_p$, due to the potential unreliability of absence data in paleozoological records. By integrating archaeological and paleontological records into the assessment of alternative models' ability to project shifts in species distributions over time, we are likely to enhance our understanding of environmental constraints on species distributions.

KEYWORDS

accuracy, hartebeest, Holocene, model evaluation, past, SDM, sensitivity, true skill statistic

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Ecology and Evolution* published by John Wiley & Sons Ltd.

JEL CLASSIFICATION

Biodiversity ecology, Biogeography, Conservation ecology, Ecosystem ecology, Global change ecology, Paleocology, Spatial ecology, Zoology

1 | INTRODUCTION

Species distribution models (SDMs) have emerged as a valuable and versatile tool that is now widely applied in multiple research domains (Araújo et al., 2019; Austin, 2007; Elith & Leathwick, 2009; Franklin et al., 2013; Payne & Bro-Jørgensen, 2016a; Smith et al., 2013; Svenning et al., 2011; Varela et al., 2011). These models predict the suitable geographic area for a species based on the environmental conditions (e.g., climate, vegetation, topography, soil) of the localities where the species is known to occur. In historical biogeography, SDMs have proven highly useful for understanding the impact of past climatic and environmental changes on the distribution and diversification of fauna and flora (Lazagabaster et al., 2021; Lentini et al., 2018; Martin et al., 2023; Svenning et al., 2011). Furthermore, understanding species' distribution shifts in response to climate change may assist in developing optimal conservation and management plans, especially for species vulnerable to extinction and invasive taxa (Barnosky et al., 2017; Dietl et al., 2015; Jeschke & Strayer, 2008; Lima-Ribeiro et al., 2017; Payne & Bro-Jørgensen, 2016b; Svenning et al., 2011).

Despite their wide use, the reliability of SDM outputs depend heavily on the quality of occurrence data (Araújo et al., 2019; Barbet-Massin et al., 2012; Duputié et al., 2014; Fourcade, 2016; Lobo et al., 2010; Varela et al., 2011). Ideally, the distributional data should come from comprehensive sampling (Tessarolo et al., 2014) so that it captures the environmental conditions of all the habitats that the species occupies (Barbet-Massin et al., 2010; Varela et al., 2009). Online data repositories are the main sources of occurrence records but they often contain heterogeneous and biased datasets (Fourcade, 2016), for example, due to bias in human observations towards certain locations, such as national parks and sites of scientific studies, and differences in museum data digitalisation and sharing data policies (Beck et al., 2014).

Another critical aspect of SDM analysis is the choice of modelling algorithms, as these determine the relationship between the presence-absence dataset and the chosen environmental variables, thus forming the basis for generating habitat suitability predictions (Elith & Graham, 2009; Hallgren et al., 2019; Li & Wang, 2013; Pearce & Ferrier, 2000; Tsoar et al., 2007). Various algorithms are available, each with its advantages and disadvantages (Carlson, 2020; Elith & Leathwick, 2009; Hallgren et al., 2019; Li & Wang, 2013; Pearce & Ferrier, 2000), and it is critical to understand the potential and pitfalls of the various modelling procedures.

To assess the appropriateness of using alternative occurrence datasets and model algorithms, it is essential to evaluate the performance of SDMs in terms of their reliability and applicability across

different contexts. Multiple evaluation metrics comparing the predicted distributions with the actual species occurrences have been developed (Allouche et al., 2006; Fourcade et al., 2018; Wunderlich et al., 2019). These evaluation metrics are categorised into threshold-dependent and threshold-independent measures.

Threshold-dependent metrics require the conversion of continuous probability outputs of SDMs into binary presence-absence predictions based on a threshold value. Among threshold-dependent metrics, the True Skill Statistic (TSS; Allouche et al., 2006) is one of the most common in the literature. TSS balances the True Positive Rate (TPR), also known as Sensitivity or Recall, and the True Negative Rate (TNR), also known as Specificity. TPR measures the proportion of actual positives that are correctly identified as such by the model, thus focusing on the model's effectiveness in detecting positive outcomes. By contrast, TNR measures the proportion of actual negatives that are correctly identified as such by the model, thus focusing on the model's ability to detect negative outcomes. TSS is particularly valued for its independence of prevalence, which makes it robust in scenarios where the species' presence is rare or unevenly distributed (Allouche et al., 2006; but see Wunderlich et al., 2019). Additionally, Accuracy and Precision are often reported alongside TSS to give a comprehensive view of model performance. Accuracy represents the overall proportion of true results (both true positives and true negatives) among the total predictions made, providing a general measure of how well the model performs across all predictions. Meanwhile, Precision measures the proportion of detected presences that are true, highlighting the reliability of positive predictions made by the model (Liu et al., 2011).

Other notable threshold-dependent metrics include F-scores and the Sørensen similarity index. The F-scores combine TPR and Precision, providing a balanced measure of model fitness by considering both false positives and false negatives (Liu et al., 2011). This metric is especially useful in situations where failing to detect a presence (false negative) is as costly as incorrectly signalling a presence (false positive). The Sørensen index is a derivative of the Jaccard index, which is calculated by taking the number of sites where both the model and observations agree on the presence of the species and dividing this by the total number of sites where the species is predicted or observed to be present (Pottier et al., 2013). By focusing solely on the presence records, the Sørensen Index is particularly useful when knowing where a species is likely to be found is more critical than knowing where it is not.

Among threshold-independent metrics, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is the most used. AUC assesses the model's capability to distinguish between presence and absence locations over all potential threshold values. Despite its widespread use, AUC may be influenced by the spatial distribution of presence and absence data points

(Lobo et al., 2008). Another significant threshold-independent metric is the Boyce Index, which evaluates how effectively the model predicts presences as compared to random expectation. This metric is particularly valuable for analysing presence-only data (Jiménez & Soberón, 2020). While AUC focuses on the model's ability to differentiate presences from pseudoabsences, the Boyce Index highlights the model's capacity to identify presences in areas of higher habitat suitability (Radomski et al., 2022).

Evaluation metrics are typically used to assess the ability of SDMs to predict areas that are known to be suitable at present. However, SDMs are rarely tested across sufficient time spans and degrees of environmental change (Inman et al., 2018; Myers et al., 2015) to properly understand their hindcasting performance. When models are applied to past or future conditions—where environmental variables may deviate significantly from present-day norms—their effectiveness can vary (Inman et al., 2018; Varela et al., 2009; Werkowska et al., 2017). SDMs typically use input data from distributions reflecting the species realised niche at present, but the realised niche is influenced by the dynamic interplay of environmental conditions, biotic interactions, such as competition and predation, and dispersal constraints (Cranston et al., 2024). These factors collectively shape the realised niche within the broader fundamental niche and are inherently dynamic, potentially changing over time (Soberón, 2010). Establishing these responses is particularly challenging when projecting to past or future climates that have no present-day analogue (Varela et al., 2009). The assumption of niche conservatism—that a species' ecological niche remains stable over time—can further be challenged by adaptive change over evolutionary time (Losos, 2008; Pearman et al., 2014; Stigall et al., 2014; Stigall, 2012). Validating SDM predictions with truly independent data, such as fossil records, offers a solution to interpret the realism of the scenarios generated (Lazagabaster et al., 2021; Lentini et al., 2018; Lima-Ribeiro et al., 2017; Martin et al., 2023; Roberts & Hamann, 2012; Smith et al., 2013).

The main goal of this study is to evaluate the performance of alternative combinations of presence/absence data and modelling algorithms by using measures of model efficiency that compare predicted geographic distributions in the past against actual changes in species distributions over relatively long periods of time as estimated from archaeological evidence. To illustrate our approach, we investigated the historical biogeography of the hartebeest *Alcelaphus buselaphus*, a large widespread savanna-adapted African antelope. We calculated various evaluation metrics that compare model-predicted probabilities of occurrence with the hartebeest's modern distribution and with its distribution over the past 8000 years, as deduced from the paleozoological record from Holocene Africa. The combinations of input data and modelling algorithms that represent the hartebeest's paleozoological record most effectively are assumed to provide the most realistic historical reconstructions of suitable environments that can be generated with the available data. This allowed us to test whether the models that were best at predicting the hartebeest's modern distribution also outperformed other models in predicting its historical distribution.

2 | MATERIALS AND METHODS

2.1 | Test case scenario: the historical biogeography of the hartebeest

The hartebeest (143–177 kg, depending on subspecies) is an African antelope that lives in diverse environments in terms of primary production, habitat structure and seasonality; its habitats include sub-grasslands, bushlands, and woodland savannas (Gosling & Capellini, 2013; Kingdon, 2015). Today, it is confined to sub-Saharan Africa, but its historical distribution includes large parts of Mediterranean North Africa, the Levant, and the Saharo-Saharan region. Dates of extinction from most of the Sahara occur during the mid-Holocene drying trend following the African Humid Period (AHP, ~11,500–5000 ka), but the now extinct subspecies known as the bubal (*Alcelaphus buselaphus buselaphus*) continued to survive in Mediterranean areas from Morocco to Egypt until the early 20th century when it was extirpated by over-hunting (Loggers et al., 1992). Like many other large African mammals, hartebeests survive today predominantly in and around protected areas (Gosling & Capellini, 2013; Kingdon, 2015).

2.2 | Workflow for the species distribution modelling

We divided the modelling workflow into seven steps focused on: (1) the study area and environmental data, (2) occurrence data, (3) absence data, (4) variable selection, (5) modelling algorithms, (6) hindcasting projections, and (7) model evaluation (Figure 1).

2.2.1 | Study area

The study area was selected to encompass the known paleozoological record of the hartebeest, including Africa and the Levant (23°W to 63°E longitude and 44°S to 45°N latitude).

2.2.2 | Environmental data

Earth System Models (ESMs) often face challenges in simulating past climatic conditions, which is indeed the case for precipitation patterns in North Africa, where their outputs sometimes diverge from proxy-derived estimates, such as pollen-based reconstructions (Braconnot et al., 2012). In this study, we use climate data for the past 7850 years obtained from a transient simulation conducted with the comprehensive ESM MPI-ESM1.2, with a raster cell size of 1.8°×1.8° (~200 km²; Dallmeyer et al., 2020). Notably, this simulation demonstrates a stronger concordance with precipitation reconstructions derived from pollen records during the African Humid Period (AHP) when compared to simulations within the Paleoclimate Model Intercomparison Project (PMIP) framework.

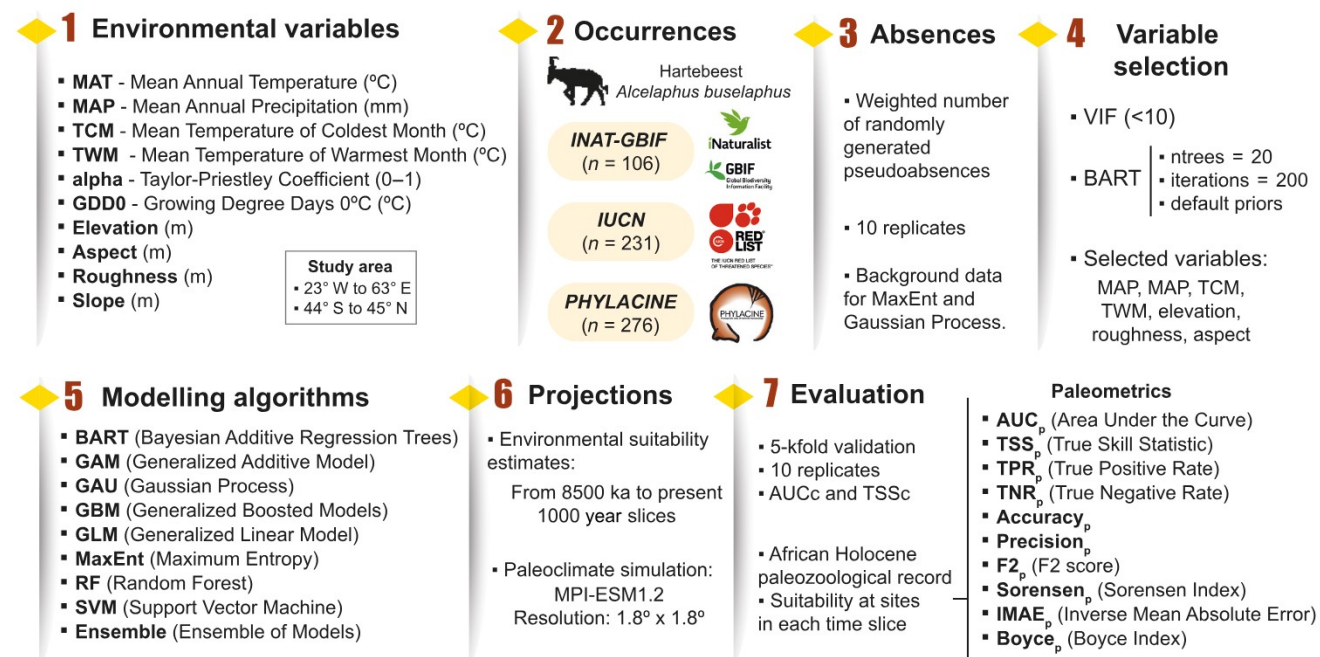


FIGURE 1 Main components of the analyses followed to model the Holocene geographic distribution of the hartebeest in this study.

The variables used include running means over 100 years of monthly mean precipitation ($\text{kg/m}^2\text{s}$) and mean 2 m air-temperature (°K), both derived from the MPI-ESM1.2 BIOME 1 model (Dallmeyer et al., 2020; Prentice et al., 1992). We also included the mean temperature of the coldest month (TCM, in°C), the mean temperature of the warmest month (TWM, in°C), the Taylor-Priestley coefficient ('alpha'), and the Growing Degree Days based on 0°C (GDD0). The Taylor-Priestley coefficient is an indicator of moisture availability ranging from 0 to 1 and is calculated as the ratio of actual evaporation to potential evaporation, where values under 0.18 are associated with deserts. The GDD is calculated as the difference between the mean daily air temperature (°C) and a threshold base temperature (0°C in this case).

We used the ice sheet extent data from the CHLSA-Trace21k paleo-orographic simulations (Karger et al., 2023) to mask the bioclimatic rasters and to calculate four topographic variables, i.e. elevation, aspect, roughness, and slope. The CHLSA-Trace21k dataset provides 30arcsec-resolution paleo-orographic reconstructions in 100-year time steps for the last 21,000 years, created by combining high-resolution information on glacial cover interpolated with a dynamic ice sheet model (ICE6G) (Karger et al., 2023).

All the topographic and bioclimatic layers were averaged into eight rasters. The raster 600 represents the years 200 to 1000 BP, the raster 1500 represents the years 1001 to 2000 BP, the raster 2500 represents the years 2001 to 3000 BP, the raster 3500 represents the years 3001 to 4000 BP, the raster 4500 represents the years 4001 to 5000 BP, the raster 5500 represents the years 5001 to 6000 BP, the raster 6500 represents the years 6001 to 7000 BP, and the raster 7500 represents the years 7001 to 7850 BP.

2.2.3 | Occurrence data

Data relating to the present-day distribution of hartebeest were obtained from the combination of two open-access online repositories of sightings and other records, i.e. the iNaturalist (INAT) and the Global Biodiversity Information Facility (GBIF). Visual examination of the INAT-GBIF occurrences revealed that many observations were heavily clustered (Figure 2 and Figure S1). To reduce biases due to data point clustering, we applied spatial thinning by adjusting the presence data to align with the climate raster's resolution (Velazco et al., 2022). Moreover, we also generated pseudo-presence data from distribution ranges produced by the International Union of Conservation of Nature (IUCN) Red List of Threatened Species (actual distribution assessed by IUCN experts) and the reconstructed natural ranges provided in the Atlas of Mammal Macroecology v.1.2.1 PHYLACINE dataset, which represent hypothetical distributions in the absence of present and historical human impacts (Faurby et al., 2018).

INAT-GBIF

This dataset was obtained from two online repositories of species observations. First, we downloaded research-quality-only occurrences (georeferenced observations verified by two or more experts and backed by photographs) with a coordinate uncertainty <10,000 m from the citizen science website iNaturalist at <https://www.inaturalist.org/>. The iNaturalist query with taxon name = 'Alcelaphus buselaphus' produced 1661 results. We used spatial thinning to reduce the sampling bias inherent to this type of observational datasets; specifically, we adjusted the presence data to the climate raster resolution allowing only one presence point per raster cell, resulting in a total of 92 occurrences. In addition, we obtained occurrence data also from

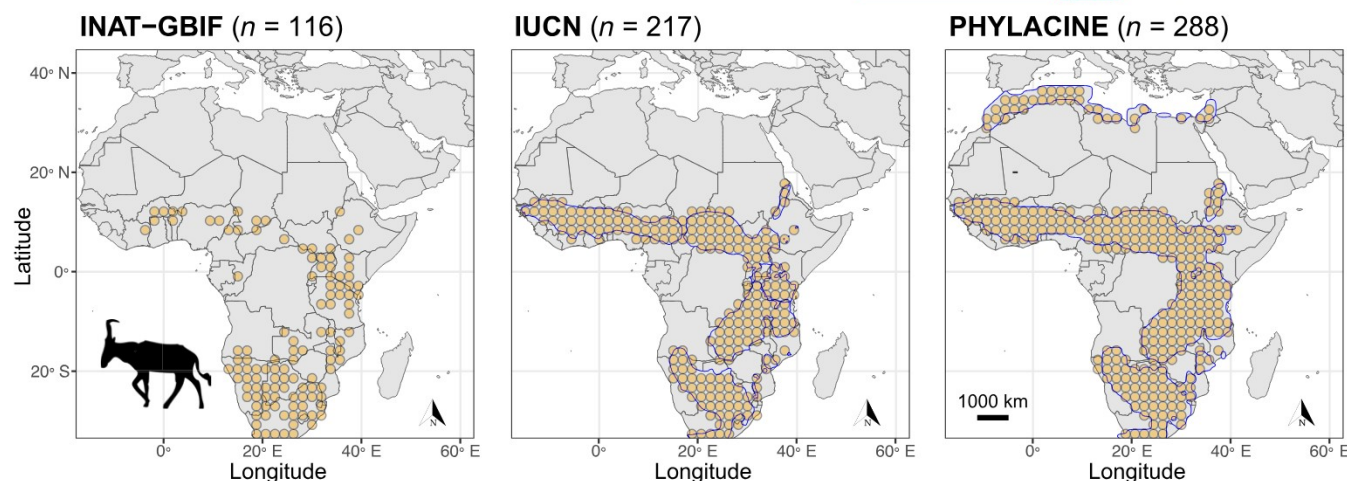


FIGURE 2 Map of the study area showing the three hartebeest's occurrence datasets used in the construction of species distribution models.

the Global Biodiversity Information Facility (GBIF) at <https://www.gbif.org/>. The query was filtered to select 'human observations' and 'fossil specimens' from the year 1800 to the present and with a coordinate uncertainty <10,000m. The data was cleaned to eliminate coordinate uncertainty, taxonomic problems, and data incongruency. We further selected only data points that were present within the hartebeest's current natural range (from the PHYLACINE dataset, see below) plus an extra 100-km buffer around this range. Of 812 initial results, the filtered and cleaned dataset included 734 records. After adjusting the presence data to the climate raster resolution, the GBIF dataset had 82 occurrences. We then combined the INAT and GBIF cleaned datasets, and after removing duplicates, the final dataset included 106 occurrences (Figure 2).

IUCN

This dataset represents the actual current distribution of the hartebeest as assessed by International Union of Conservation of Nature (IUCN) Red List of Threatened Species experts. Polygon shapefiles are available for download at <https://www.iucnredlist.org/>. All cells available at the raster resolution within the IUCN polygon were taken as occurrence (=pseudopresence) records ($n=231$, Figure 2).

PHYLACINE

This dataset represents the hypothetical present-day natural range for the hartebeest as provided by the Atlas of Mammal Macroecology v.1.2.1 PHYLACINE database (Faurby et al., 2018). These natural ranges estimate the current potential distribution of species in the absence of anthropogenic impacts such as extinctions or extirpations. Notably, the authors suggest that, without human influence, hartebeests would likely inhabit most of the northern African Mediterranean coast today. The PHYLACINE estimations are derived from methods that include fossil co-occurrence modelling and relatively recent historical sightings, so it is anticipated that these ranges would more accurately predict the fossil occurrences of the hartebeest than the IUCN or the INAT-GBIF, which only include present-day observations and ranges. The shapefile polygons

used were downloaded from https://megapast2future.github.io/PHYLACINE_1.2/. We included all cells within the PHYLACINE polygon at the dataset's raster resolution as occurrence records ($n=276$, Figure 2).

2.2.4 | Absence data

For each of three occurrence datasets, we generated 10 sets of randomly selected (pseudo)absence points (i.e. considering only raster cells with no occurrence points) and one set of background points (i.e. considering all raster cells within the entire study area, including cells with occurrence points). For the INAT-GBIF dataset, the pseudoabsence points were selected from cells outside 100-km buffers around each presence point. For the IUCN and PHYLACINE datasets, the absence points were taken outside of the respective polygons. The number of absences equalled the number of occurrences, following the recommendation of weighted sample sizes (Barbet-Massin et al., 2012). Background data are recommended for presence-only species distribution modelling algorithms such as MaxEnt (Phillips et al., 2009) and are sometimes used also for other methods, including tree classification and machine learning techniques (Valavi et al., 2022). We selected all raster cells in the study area as background points ($n=978$, Figure S2).

2.2.5 | Variable importance and selection

For each of the three occurrence datasets, we first removed variables with Variance Inflation Factor (VIF) >10 to avoid collinearity between covariates from affecting the results. We removed additional variables by applying the automated stepwise variable set reduction procedure in the Bayesian additive regression trees algorithm (BART), available in the *embarcadero* R library (Carlson, 2020). This procedure eliminates the least informative variable based on the root mean square error (RMSE) for a specified number of model iterations ($n=200$) and a set

number of trees ($n=20$; based on recommendation of Carlson, 2020, cfr. also Chipman et al., 2010). Normally, there is a strong agreement between the selected variables and those that ranked higher in variable importance (Carlson, 2020), and this was the case also in our analyses. Variable importance was calculated by counting the number of times a given variable is used by a tree split across the full posterior draw of trees after 200 iterations. This method weighs the relative importance of each variable across all decision trees and thus spreads variable contribution across all covariates.

The selected variables, in order of importance, were MAP, TCM, TWM, elevation, and roughness (Figure S3). The relationship between predictor variables and environmental suitability is shown by the partial dependence plots in Figures S4–S6. Partial dependence plots show the response curves of an individual variable in relation to environmental suitability estimations. In all cases, MAP follows approximately a normal distribution, with a peak of environmental suitability values between ~500 and ~1200 mm. The results also indicate that hartebeests prefer habitats with TCM above 10°C and TWM below 28°C. The primary importance of precipitation agrees well with precipitation being the major determinant of vegetation production in tropical areas (Chamaillé-Jammes & Fritz, 2009).

2.2.6 | Model implementation

We modelled potential geographic distribution of the hartebeest using eight modelling algorithms (Figure 1), i.e. BART, generalised linear model (GLM), generalised additive model (GAM), Gaussian process (GAU), support vector machine (SVM), maximum entropy (MaxEnt), random forest (RF), and generalised boosted regression (GBM). The main characteristics of each algorithm are described in Supporting Information S6. BART is implemented in the R library *embarcadero* (Carlson, 2020), with all other SDM modelling algorithms using the *flexsdm* library in R (Velazco et al., 2022). In addition, we constructed an ensemble model by combining all the models except BART (which is not implemented in the *flexsdm* library), using a TSS-based weighted average approach. The nine resulting modelling frameworks were applied to each combination of the three occurrence datasets with the 10 random pseudoabsence replicates.

To generate simple, easy-to-interpret models, GLM was set to work with polynomials of second order only and with no interactions, and correspondingly, the k value (number of knots) in GAM was set to two; this is based on the consideration that whereas n - or U -shaped quadratic relationships between explanatory variables and probability of occurrence often make biological sense, this is not the case for higher order polynomials. For both MaxEnt and GAU, the models were fitted with presences and background points but were validated against presences and pseudoabsences (see model evaluation section below). This procedure makes MaxEnt and Gaussian processes comparable to other presence-absence models (Velazco et al., 2022). The rest of the algorithms, which usually produce more complex models, were run with the default R *flexsdm* settings. Finally, BART models were run with the default parameters

in the library *embarcadero* (Carlson, 2020), i.e. 200 trees and 1000 back-fitting Markov chain Monte Carlo (MCMC) iterations and discarding 20% as burn-ins.

All combinations of occurrence-pseudoabsence dataset and modelling algorithm were used to project distributions for eight past time slices, totalling 2160 projections. The projections were then evaluated against the paleozoological record to compare their performance (Sections 2.3 and 2.4).

2.3 | The paleozoological dataset

We obtained paleozoological presence/absence data for the Holocene from the compilation of dated faunal remains from Africa in Phelps et al. (2020) (<https://doi.pangaea.de/10.1594/PANGAEA.904942>). Note that this dataset is limited to the African continent and does not include areas of the Middle East (including Palestine and Israel), where the hartebeest is known to have been present in the Holocene (Tsahar et al., 2009). The dataset includes earliest and latest possible dates derived from calibrated radiocarbon dates or known cultural or typological periods, and we included each occurrence record in all of the time slices with which the reported date interval overlapped either fully or partially. For this assignment process, we developed an R function *dates_to_time_slices()*, which allows the user to either select the full time interval, as in our analysis (option = 'interval'), or just the midpoint of each interval (option = 'midpoint') when assigning records to time slices.

In our analysis, the relatively low resolution of the environmental raster meant that certain sites, particularly coastal locations in eastern and southern Africa, were inadvertently omitted due to their placement outside of the information-containing pixels. To prevent such data loss and ensure comprehensive coverage, we also developed a custom R function *coordinates_to_nearest_valid()*, which systematically adjusts the coordinates of the occurrence data by aligning them with the nearest valid raster cells, thereby preserving as much data as possible.

2.4 | Model evaluation

2.4.1 | Fit to present-day distribution

Model performance was evaluated with AUC_c and TSS_c using a 5k-fold cross-validation approach with 10 replicates (Allouche et al., 2006; Márcia Barbosa et al., 2013). TSS was rescaled to the 0–1 range to facilitate comparison with other evaluation metrics (Barbosa, 2015). AUC and TSS values <0.4 are considered poor, 0.4–0.8 useful, and >0.8 excellent (Zhang et al., 2015). Independent thresholds were calculated for each model such that the sensitivity and specificity were maximised (Figures S7 and S8). These thresholds represent cut-off values with which continuous environmental suitability projections are converted to binary scores with discrete favourable or unfavourable areas.

2.4.2 | Fit to past distribution

We calculated both threshold-dependent and threshold-independent metrics to evaluate models fit to past distribution (Table 1). The calculation of threshold-dependent metrics is based on the construction of a confusion matrix (Table 2), where *a* is the number of fossil localities for which presence was correctly predicted by the model (true positives); *b* is the number of localities in which remains of the species were not found but the model predicted suitable habitats (false positives); *c* is the number of localities for which the species was found but the model predicted unsuitable habitats (false negatives); and *d* is the number of localities for which absence was correctly predicted by the model (true negatives) (Allouche et al., 2006). From this confusion matrix, we calculated the following metrics:

True Positive Rate (TPR_p), True Negative Rate (TNR_p), False Positive Rate (FPR_p), False Negative Rate (FNR_p), True Skill Statistic (TSS_p), Accuracy_p, Precision_p, F2 Score_p, and Sorensen_p.

The threshold independent measures included the Area Under the Curve (AUC_p) and the Boyce Index (Boyce_p). AUC_p was calculated through the *roc* and *auc* functions in the R library 'pROC' (Robin et al., 2011), whereas Boyce_p was obtained using the *evalContBoyce* function in the *enmSdmX* package (Hirzel et al., 2006; Smith, 2023). We also computed the Index of Model Absolute Error (IMAE_p) to quantify the absolute deviation of model predictions from actual observed outcomes. This metric calculates the absolute difference between the predicted probability and the actual observed value for each data point, sums these absolute errors and divides the sum by the total number of observations to achieve a normalised measure.

TABLE 1 Description of evaluation metrics included in the *paleo.eval* function and used in this study.

Metric	Formula	Description
Threshold-dependent		
True positive rate (TPR)	$\frac{a}{a+c}$	Also known as Sensitivity or Recall, measures the proportion of actual positives (presence of the taxon) that are correctly identified by the model
True negative rate (TNR)	$\frac{d}{b+d}$	Also known as Specificity, measures the proportion of actual negatives (absence of the taxon) that are correctly identified
False positive rate (FPR)	$\frac{b}{b+d}$	Measures the proportion of actual negatives that are incorrectly classified as positives
False negative rate (FNR)	$\frac{c}{a+c}$	Measures the proportion of actual positives that are incorrectly classified as negatives
True skill statistic (TSS)	$TPR + TNR - 1$	Quantifies the model's ability to avoid both false positives and false negatives
Accuracy	$\frac{a+d}{n}$	Quantifies the overall ability of the model to correctly classify both presences and absences
Precision	$\frac{a}{a+b}$	Indicates the proportion of positive identifications that were actually correct
Prevalence	$\frac{a+c}{a+b+c+d}$	Reflects the proportion of actual positives in the dataset
F2 score	$2 * \frac{\text{precision} * \text{TPR}}{\text{precision} + \text{TPR}}$	Is the harmonic mean of precision and recall, providing a balance between them
False discovery rate (FDR)	$\frac{b}{a+b}$	Measures the proportion of positive results that were false positives
Sorensen	$2 * \frac{a}{c + (2*a) + d}$	Similar to F1 Score, measures the similarity between predicted and observed presences
Jaccard	$\frac{a}{a+c+d}$	Another similarity measure between the predicted and observed data
Odds ratio (OR)	$\frac{a*d}{b*c}$	Assesses the odds of prediction success relative to prediction failure
Kappa	$\frac{P_o - P_e}{1 - P_e}$	Where P_o is the observed agreement, and P_e is the expected agreement under independence. Measures inter-rater agreement for qualitative items, adjusting for chance agreement
Threshold-independent		
Area under the curve (AUC)	$\frac{R - n_1 * (n_1 + 1) / 2}{n_1 * n_2}$	From the Receiver Operating Characteristic (ROC) analysis, where <i>R</i> is the sum of ranks of positives, <i>n</i> ₁ is the number of positives, and <i>n</i> ₂ is the number of negatives. It measures the model's discriminative ability, i.e., its capability to correctly classify observations into presence or absence categories across all possible threshold values
Boyce	$B = \rho(S, E)$	Simplified conceptual formula where ρ is Spearman's rank correlation coefficient, <i>S</i> are the mid-points of suitability bins, and <i>E</i> is the cumulative sum of observed presences up to each bin. It evaluates the model's predictive performance by comparing the observed presences against a background of predictions
Index of model absolute error (IMAE)	$\frac{\sum i = \text{real} - \text{pred} }{n}$	Where, $ i = \text{real} - \text{pred} $ is the absolute error for each data point and <i>n</i> is the total number of observations in the dataset. It quantifies the absolute deviation of predictions from reality, offering a normalised error rate across the dataset

TABLE 2 Confusion matrix used to calculate the thresholded paleometrics. After Allouche et al. (2006), but using paleozoological records as validation sites.

	Paleozoological record	
	Presence	Absence
SDM projections		
Presence	a (correctly classified presence)	b (incorrectly classified absence)
Absence	c (incorrectly classified presence)	d (correctly classified absence)

To facilitate the calculation of these metrics, which we refer to as 'paleometrics', we provide the R script with the function *paleo.eval* (Appendix S1, Table S1). This function requires a raster of model predictions for the time slices to be evaluated, the specification of a threshold value, and a dataset indicating latitude and longitude of paleozoological localities, species composition, and time range. We computed paleometrics using the *paleo.eval* function to compare environmental suitability projections obtained using different occurrence datasets and modelling algorithms for each of the eight time slices to assess how datasets and algorithms influenced the reliability of our results (Figures S10–S18).

All metrics were scaled to range from 0 to 1 to allow standardised comparisons across different models. The statistical comparisons were done utilising non-parametric Kruskal-Wallis and post-hoc Dunn tests with Bonferroni correction to ensure robust comparisons. All analyses were performed in R v.4.3.1. and RStudio v.2023.06.2-561.

2.4.3 | Sensitivity analysis

Not all organisms present in an ecosystem are preserved, and not all preserved remains are discovered (Behrensmeyer et al., 2000; Jablonski et al., 1986), therefore paleozoological records are inherently biased taxonomically, spatially, and temporally (Behrensmeyer et al., 2000; Inman et al., 2018; Lloyd et al., 2012). Because absence of a taxon at a specific site and time in the paleozoological record does not necessarily indicate its true absence from the area, false positives (parameter *b* in the confusion matrix) can skew results towards artificially low values of the TNRp and thereby reduce estimates of paleometrics that incorporate evaluation of absences, such as TSSp and AUCp. In the paleozoological dataset used in this study, there is a disproportionate number of localities with only few species (Figure S19) and we observed that hartebeest presence was recorded more times at sites where four species or more had been documented. Thus, we conducted a sensitivity analysis to examine the potential effect of sampling bias in the paleozoological dataset. Sites with the lowest recorded species richness were progressively eliminated and paleometrics recalculated until all sites with species richness of less than five were excluded.

3 | RESULTS

3.1 | Evaluating model performance for the present day

Projected environmental suitability estimates based on modern climate were generally similar for all the distribution models generated; high variation in certain geographical areas was mainly dependent on the occurrence dataset used and, to a lesser extent, on the algorithm applied (environmental suitability maps averaged for all BART models for each occurrence dataset shown in Figure 3). Notable differences among occurrence datasets were seen in large areas of southern and western Africa. Moreover, the INAT-GBIF and PHYLACINE projections showed potentially suitable environments in the Iberian Peninsula, while the IUCN map did not. Only the PHYLACINE projections showed certain areas of northern Africa as potentially suitable.

Comparing model performance across occurrence datasets, significant differences were found (evaluated using AUC and TSS; Kruskal-Wallis, $p < .001$; Figure 4 and Tables S2 and S6). The models generated with the IUCN dataset had the highest scores (AUC = mean of $0.96 \pm$ standard deviation of 0.01; TSS = 0.86 ± 0.03), followed by PHYLACINE (AUC = 0.94 ± 0.02 ; TSS = 0.80 ± 0.05) and INAT-GBIF (AUC = 0.88 ± 0.02 ; TSS = 0.70 ± 0.05). Comparing the modelling algorithms, the AUC scores were similar, but there were notable differences in TSS. Tree classifiers BART (TSS = 0.86 ± 0.05) and RF (TSS = 0.82 ± 0.07) slightly outperformed the others, while GBM showed significantly lower performance (TSS = 0.70 ± 0.08 ; Dunn: $p < .001$).

3.2 | Evaluating model performance for hindcasts

3.2.1 | Presence-absence datasets

Comparing the performance of models in predicting present-day distributions against their predictions of environmental suitability in the past, as assessed using the paleozoological record, revealed marked differences (for illustration, environmental suitability maps generated with MaxEnt and the PHYLACINE dataset is shown in Figure 5; model choice based on favourable evaluation metrics, see below). Most of the suitable areas in southern and eastern Africa remain broadly similar throughout the Holocene, and between 5 and 8ka, suitable habitats are also predicted in central tropical Africa, by the northern African Mediterranean coast, and on the Arabian Peninsula, although these areas contract over time.

When comparing paleometrics which focus on correctly predicting presences (Figure 4 and Tables S3 and S6), the PHYLACINE dataset (TPRp = 0.65 ± 0.15 ; Sorensen_p = 0.42 ± 0.07 ; F2p = 0.53 ± 0.08) was significantly better (Kruskal-Wallis: $p < .001$; Dunn: $p < .001$) than both IUCN (TPRp = 0.55 ± 0.18 ; Sorensen_p = 0.38 ± 0.05 ; F2p = 0.46 ± 0.09) and INAT-GBIF (TPRp = 0.49 ± 0.19 ; Sorensen_p = 0.37 ± 0.05 ; F2p = 0.42 ± 0.11). Conversely, when comparing the ability of models

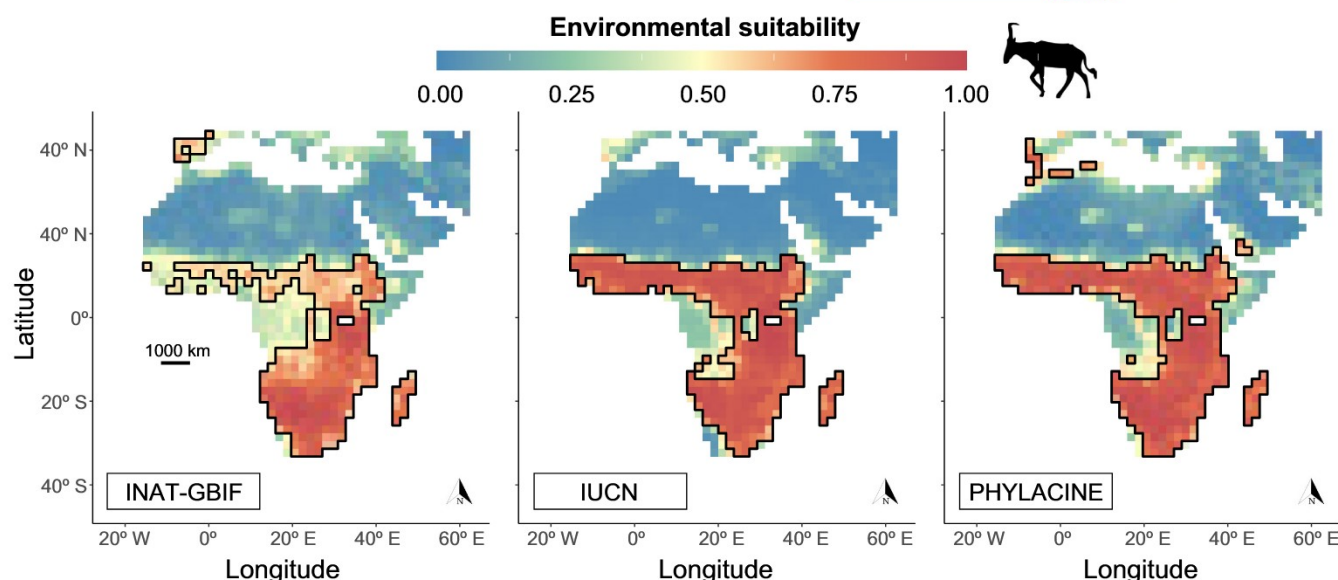


FIGURE 3 Projections of hartebeest's environmental suitability in the present day for the three occurrence datasets using BART. Environmental suitability scores vary from 0 (unsuitable) to 1 (ideal conditions). The areas delimited by the black line indicate suitable habitats above the model's optimal threshold.

generated with different datasets to predict absences (Figure 4), the INAT-GBIF dataset ($TNR_p = 0.63 \pm 0.13$) was significantly better (Kruskal-Wallis: $p < .001$; Dunn: $p < .001$) than both IUCN ($TNR_p = 0.57 \pm 0.12$) and the PHYLACINE dataset ($TNR_p = 0.50 \pm 0.11$).

When comparing metrics that balance the models' ability to correctly predict presences and absences (Figure 4 and Tables S4 and S6), TSS_p values were closely matched across all datasets, with PHYLACINE ($TSS_p = 0.57 \pm 0.05$) slightly outperforming both INAT-GBIF ($TSS_p = 0.56 \pm 0.05$) and IUCN ($TSS_p = 0.56 \pm 0.06$). For AUC_p , INAT-GBIF ($AUC_p = 0.61 \pm 0.07$) was significantly better (Kruskal-Wallis: $p < .01$; Dunn: $p = .007$) than IUCN ($AUC_p = 0.58 \pm 0.07$) but not significantly different from PHYLACINE ($AUC_p = 0.60 \pm 0.07$). For $Boyce_p$, INAT-GBIF ($Boyce_p = 0.58 \pm 0.16$) was significantly better (Kruskal-Wallis: $p < .01$; Dunn: $p = .007$) than PHYLACINE ($Boyce_p = 0.51 \pm 0.17$) and IUCN ($Boyce_p = 0.48 \pm 0.15$).

When comparing the overall error of the models in both binary ($Accuracy_p$ and $Precision_p$) and continuous ($IMAE_p$) form (Figure 4 and Tables S5 and S6), INAT-GBIF ($Accuracy_p = 0.58 \pm 0.07$; $IMAE_p = 0.57 \pm 0.04$) showed significantly less error (Kruskal-Wallis: $p < .01$; Dunn: $p = .007$) compared to PHYLACINE ($Accuracy_p = 0.53 \pm 0.07$; $IMAE_p = 0.52 \pm 0.05$), while IUCN showed intermediate values ($Accuracy_p = 0.55 \pm 0.06$; $IMAE_p = 0.55 \pm 0.05$). No significant differences were found in $Precision_p$ (INAT-GBIF = 0.33 ± 0.07 ; IUCN = 0.32 ± 0.07 ; PHYLACINE = 0.33 ± 0.09).

3.2.2 | Modelling algorithms

The performance of models when hindcasting distributions also depended on the modelling algorithm applied. When comparing

paleometrics which focus on correctly predicting presences (Figure 6 and Tables S2 and S7), the modelling algorithms with the highest values were MaxEnt ($TPR_p = 0.63 \pm 0.18$; $Sorensen_p = 0.42 \pm 0.06$; $F2_p = 0.51 \pm 0.09$) and GLM ($TPR_p = 0.63 \pm 0.15$; $Sorensen_p = 0.41 \pm 0.06$; $F2_p = 0.51 \pm 0.07$) and those with the lowest values were GBM ($TPR_p = 0.46 \pm 0.14$; $Sorensen_p = 0.35 \pm 0.06$; $F2_p = 0.40 \pm 0.08$), RF ($TPR_p = 0.50 \pm 0.21$; $Sorensen_p = 0.37 \pm 0.07$; $F2_p = 0.43 \pm 0.13$), and GAU ($TPR_p = 0.50 \pm 0.20$; $Sorensen_p = 0.35 \pm 0.05$; $F2_p = 0.41 \pm 0.10$). Most of these differences were significantly different (see Table S2). Conversely, when comparing the ability of models generated with different modelling algorithms to predict absences (Figure 6), the highest values were shown by RF ($TNR_p = 0.62 \pm 0.13$), GBM ($TNR_p = 0.60 \pm 0.13$), and GAU ($TNR_p = 0.59 \pm 0.14$), and the lowest by MaxEnt ($TNR_p = 0.52 \pm 0.13$), GLM ($TNR_p = 0.54 \pm 0.11$), and Ensemble ($TNR_p = 0.53 \pm 0.13$). These differences, however, were not significant (Kruskal-Wallis: $p > .05$).

When comparing metrics that balance the models' ability to correctly predict presences and absences (Figure 6 and Tables S3, S4, and S7), TSS_p values were significantly higher (Kruskal-Wallis: $p < .05$; Dunn: $p < .05$) in GLM, MaxEnt, GAM, Ensemble, and BART ($TSS_p = 0.57$ – 0.58), than in GAU and GBM ($TSS_p = 0.53$ – 0.54). AUC_p was significantly higher (Kruskal-Wallis: $p < .05$; Dunn: $p < .05$) in GLM and GAM ($AUC_p = 0.62 \pm 0.07$) than in GBM ($AUC_p = 0.56 \pm 0.07$), while $Boyce_p$ was significantly higher (Kruskal-Wallis: $p < .05$; Dunn: $p < .05$) in BART ($Boyce_p = 0.57 \pm 0.15$), GAM ($Boyce_p = 0.56 \pm 0.16$), and Ensemble ($Boyce_p = 0.55 \pm 0.16$), than in GBM ($Boyce_p = 0.46 \pm 0.16$).

When comparing the overall error of the models in both binary ($Accuracy_p$ and $Precision_p$) and continuous ($IMAE_p$) form, most modelling algorithms presented similar results and no significant differences were found (Figure 6 and Tables S6 and S7).

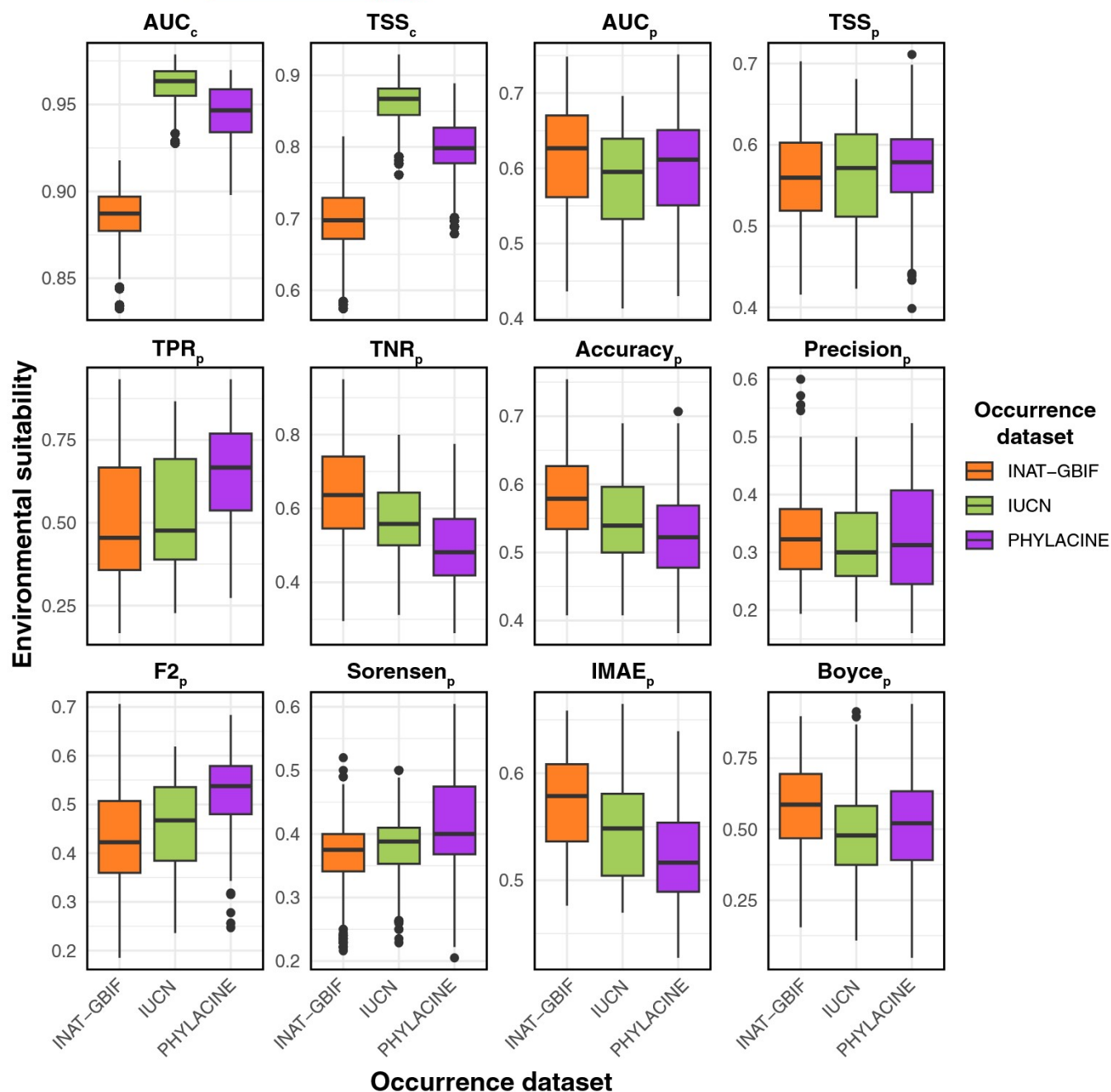


FIGURE 4 Comparison of AUC_c, TSS_c, and selected paleoevaluation metrics for the three occurrence datasets used in this study.

3.3 | Combined presence-absence datasets and modelling algorithms

The evaluation of the projections generated using different combinations of modelling algorithms and presence-absence datasets revealed illuminating patterns in predictive performance (Figure 7). When evaluating models for their ability to correctly predict presences in the paleozoological record (TPR_p, F2_p, Sorensen_p), the best performance was achieved by the PHYLACINE dataset in combination with MaxEnt, GLM, Ensemble, and GAM, all showing average values of TPR_p between 0.70 and 0.76, F2_p between 0.56 and 0.59 and Sorensen_p between 0.44 and 0.46. Other modelling algorithms

such as BART, SVM and RF, also performed well in combination with PHYLACINE, with average values of TPR_p between 0.62 and 0.67, F2_p between 0.51 and 0.53, and Sorensen_p between 0.42 and 0.43. GLM, MaxEnt, Ensemble, and GAM also performed relatively well when combined with IUCN and INAT-GBIF (TPR_p: 0.54–0.60; F2_p: 0.45–0.50; Sorensen_p: 0.38–0.41). However, GMB and GAU showed low values in these paleometrics across all datasets (TPR_p: 0.42–0.51; F2_p: 0.37–0.45; Sorensen_p: 0.33–0.37). Conversely, the best models at predicting absences in the paleozoological record were those that combined the INAT-GBIF dataset with RF, GBM, GAU, SVM, and BART, with TNR_p average values between 0.66 and 0.73. Most modelling algorithms combined performed similarly when combined

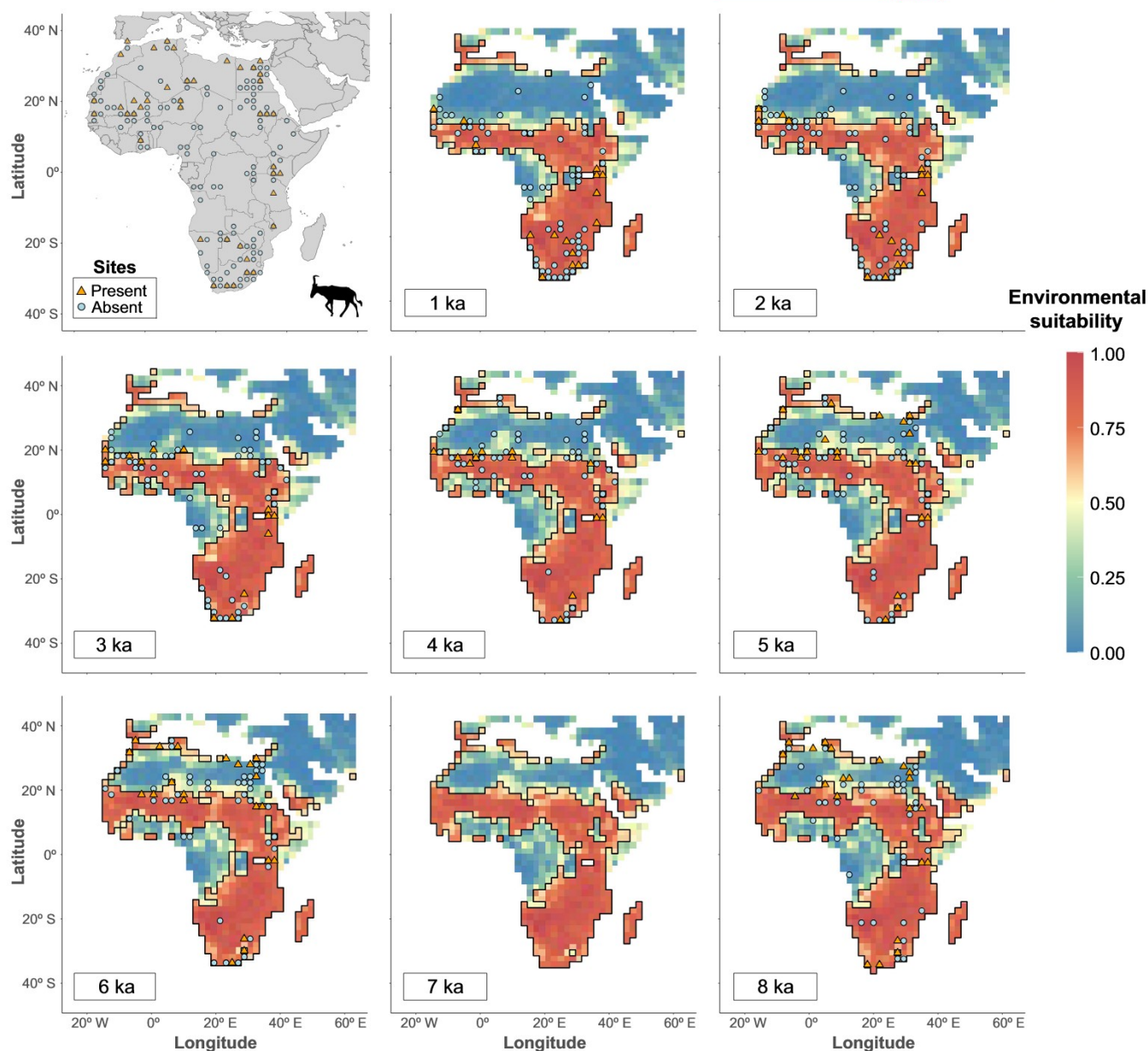


FIGURE 5 Projections of hartebeest's environmental suitability for each of the eight time slices using the PHYLACINE dataset and MaxEnt. Also shown is a map of the zooarchaeological localities from which presence and absence of hartebeest remains were used to calculate the paleometrics. Environmental suitability scores vary from 0 (unsuitable) to 1 (ideal conditions). The areas delimited by the black lines indicate suitable habitats above the model's optimal threshold. The orange triangles indicate the location of a site with identified hartebeest remains in that time slice and the blue circles indicate absence. Note that the northern expansion of suitable habitats into the Sahara is coincident with the appearance of zooarchaeological remains in that area.

with IUCN (TNR_p : 0.57–0.58), and all models tended to underperform in predicting absences when combined with PHYLACINE, but especially MaxEnt, GLM, SVM, and Ensemble (TNR_p : 0.43–0.48).

When comparing metrics that balance the correct prediction of presences and absences in the paleozoological record, the results show various differences depending on the paleometric considered. TSS_p average was higher in PHYLACINE datasets combined with Ensemble, GAM, GLM, and MaxEnt (TSS_p : 0.59–0.60) and lowest in IUCN combined with GBM and GAU (TSS_p : 0.52–0.54).

AUC_p showed the highest average values in RF, GLM, BART, and GAM when combined with PHYLACINE and BART, GAM, and GLM, when combined with INAT'GBIF (AUC_p : 0.62–0.64). The lowest average values of AUC_p were found in GBM when combined with PHYLACINE and IUCN (AUC_p : 0.54–0.55). The highest average values of $Boyce_p$ were found in BART and SVM when combined with INAT-GBIF ($Boyce_p$: 0.61–0.64), while the lowest values were shown by GBM and GAU when combined with IUCN or PHYLACINE ($Boyce_p$: 0.40–0.45).

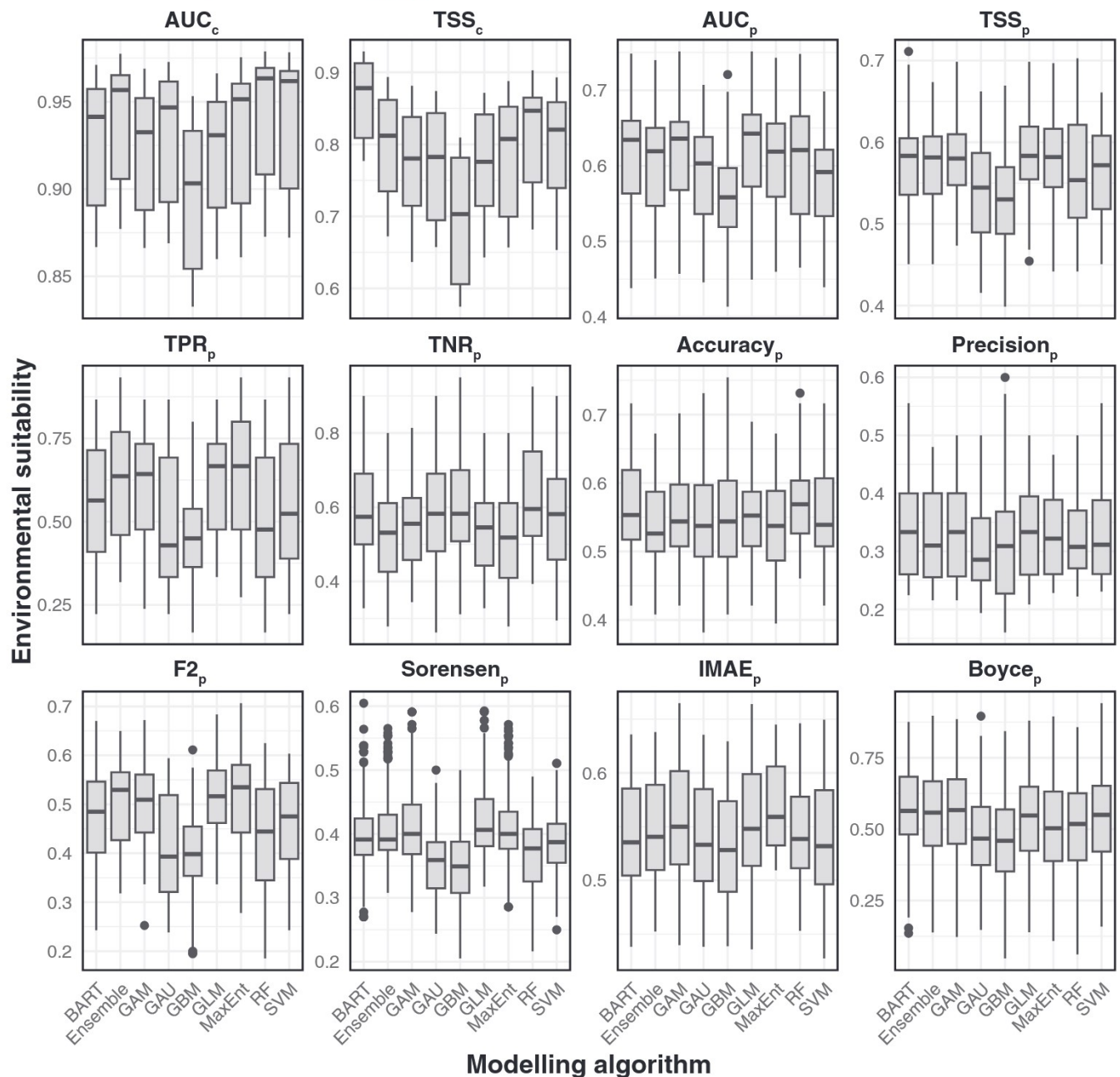


FIGURE 6 Comparison of AUC_c , TSS_c , and selected paleoevaluation metrics for the modelling algorithms used in this study.

When comparing the overall error of the models in both binary ($Accuracy_p$ and $Precision_p$) and continuous ($IMAE_p$), the results were also variable depending on the paleometrics considered. $Accuracy_p$ average was higher in INAT-GBIF when combined with BART, GBM, and RF ($Accuracy_p$: 0.60–0.62) and lowest in IUCN when combined with Ensemble, GAU, GBM, MaxEnt, and SVM ($Accuracy_p$: 0.51–0.52). $Precision_p$ was slightly higher in INAT-GBIF combined with BART and GBM, IUCN combined with GAM and GLM, and PHYLACINE combined with GAM, GBM, GLM ($Precision_p$: 0.34–0.35), while the lowest results were shown by PHYLACINE and IUCN when combined with GBM ($Precision_p$: 0.28–0.29). Finally, $IMAE_p$ show the highest average values in INAT-GBIF combined with GAM and GLM, and IUCN combined with MaxEnt ($IMAE_p$: 0.58–0.59),

while PHYLACINE combined with GAU, GBM, and SVM showed the lowest values ($IMAE_p$: 0.50–0.51).

3.4 | Sensitivity analysis

The sensitivity analysis, where sites were removed incrementally depending on their species richness, showed that the differences between the presence-absence datasets relative to each other were maintained despite changes in their absolute average values (Figure 8). As sites were removed, a decreasing trend was observed in paleometrics that evaluate the ability of models to predict paleozoological absences (TNR_p) as well as those that balance the correct

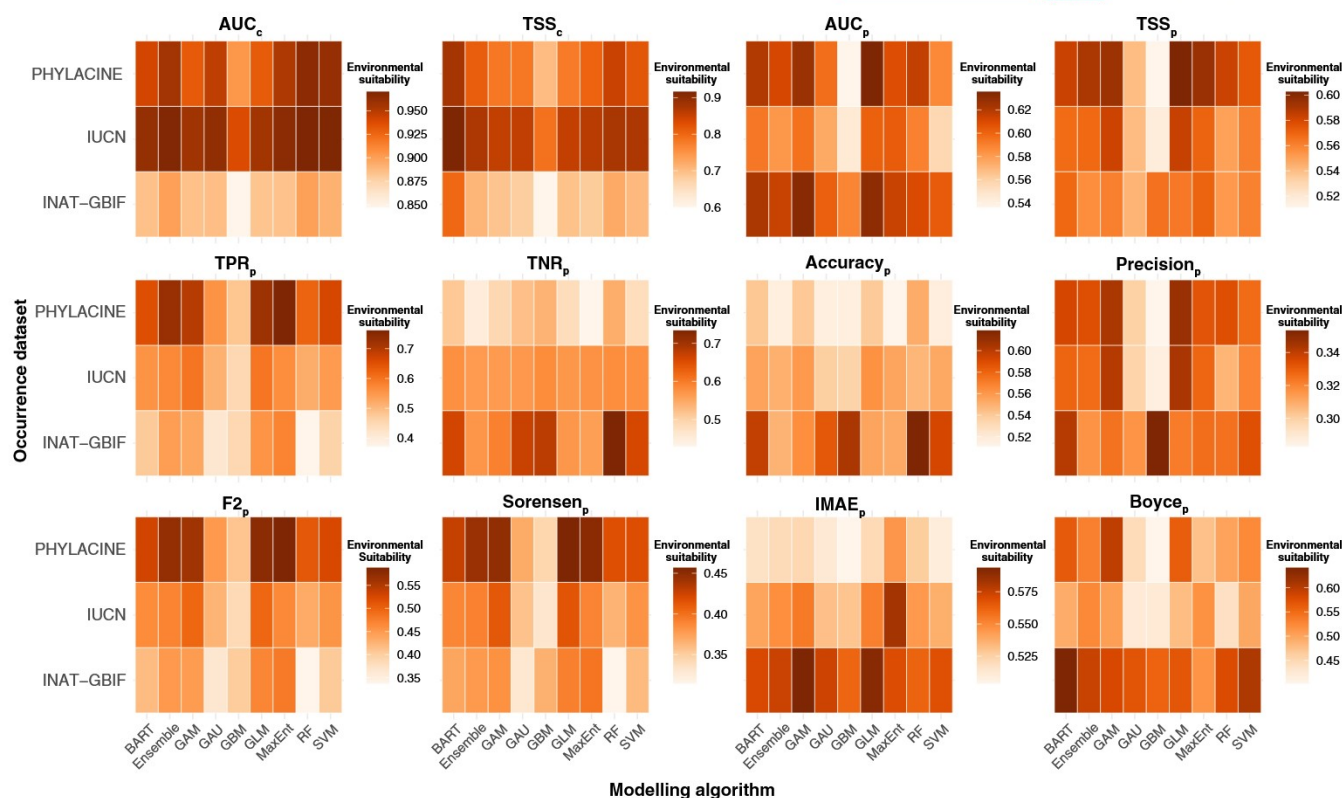


FIGURE 7 Heatmap comparing model performance across all occurrence datasets and modelling algorithms. Darker orange colours indicate higher values.

prediction of presences and absences (TSS_p , AUC_p , $Boyce_p$). By contrast, paleometrics that evaluate the ability of models to predict presences (TPR_p , $F2_p$, and $Sorensen_p$) tended to increase. Measures of error such as $Accuracy_p$ and $IMAE_p$ also show a decreasing pattern, while $Precision_p$ tended to increase.

4 | DISCUSSION

4.1 | Occurrence data, model algorithms and the reliability of SDMs

In this study, we have compared the outputs of hartebeest distribution models using three different sources of occurrence data, i.e. data extracted from online repositories and data from two alternative range maps. Each of them has advantages and disadvantages. Our results suggest that there is a trade-off between the correct classification of presences and the correct classification of absences in the paleozoological record. Models generated with PHYLOCINE were the best at predicting presences (high values of TPR_p , $F2_p$, and $Sorensen_p$) and the worst at predicting absences (low values of TNR_p). On the contrary, models generated with INAT-GBIF excelled at predicting absences but offered poor results at predicting presences. Finally, the models generated with IUCN ranges were the most balanced, as reflected in intermediate values across most metrics. This implies that PHYLOCINE models are likely better at capturing the extent of suitable habitats in the past although they could

be prone to overestimation. On the other hand, INAT-GBIF models offer a more conservative estimates and may fail to identify the full area that the hartebeest could have inhabited in the past.

Generally, modelling algorithms did not affect the SDM outputs as much as the occurrence dataset used but there were some notable differences that require explanation. GBM generated results that underperformed and were unrealistic judging from the projections of environmental suitability (Figure S10) and the lower values of TSS_c , AUC_c and paleometrics (Figure 4; Table S2). GBMs are a group of machine learning procedures that iteratively create new improved models based on previously generated ones (Natekin & Knoll, 2013). This allows the model to gradually improve its accuracy in each iteration but the approach can lead to very complex relationships between the environmental variables and the occurrence data (Elith et al., 2008), relationships that are not necessarily biologically meaningful and may generate spurious outcomes.

However, other relatively complex models, including BART and RF, demonstrated higher average AUC_c and TSS_c (Figure 4; Table S2). This is consistent with previous studies highlighting the tendency of tree-based methods to overfit SDMs (Cutler et al., 2007; Elith et al., 2008; Li & Wang, 2013). BART and RF also presented relatively high AUC_p and TSS_p , and these values were driven mainly by the correct prediction of absences in the paleozoological record (high values of TNR_p). This was especially noticeable when generating models combining these two classifiers with the INAT-GBIF dataset, which offered some of the best TNR_p averages but also the worst TPR_p values.

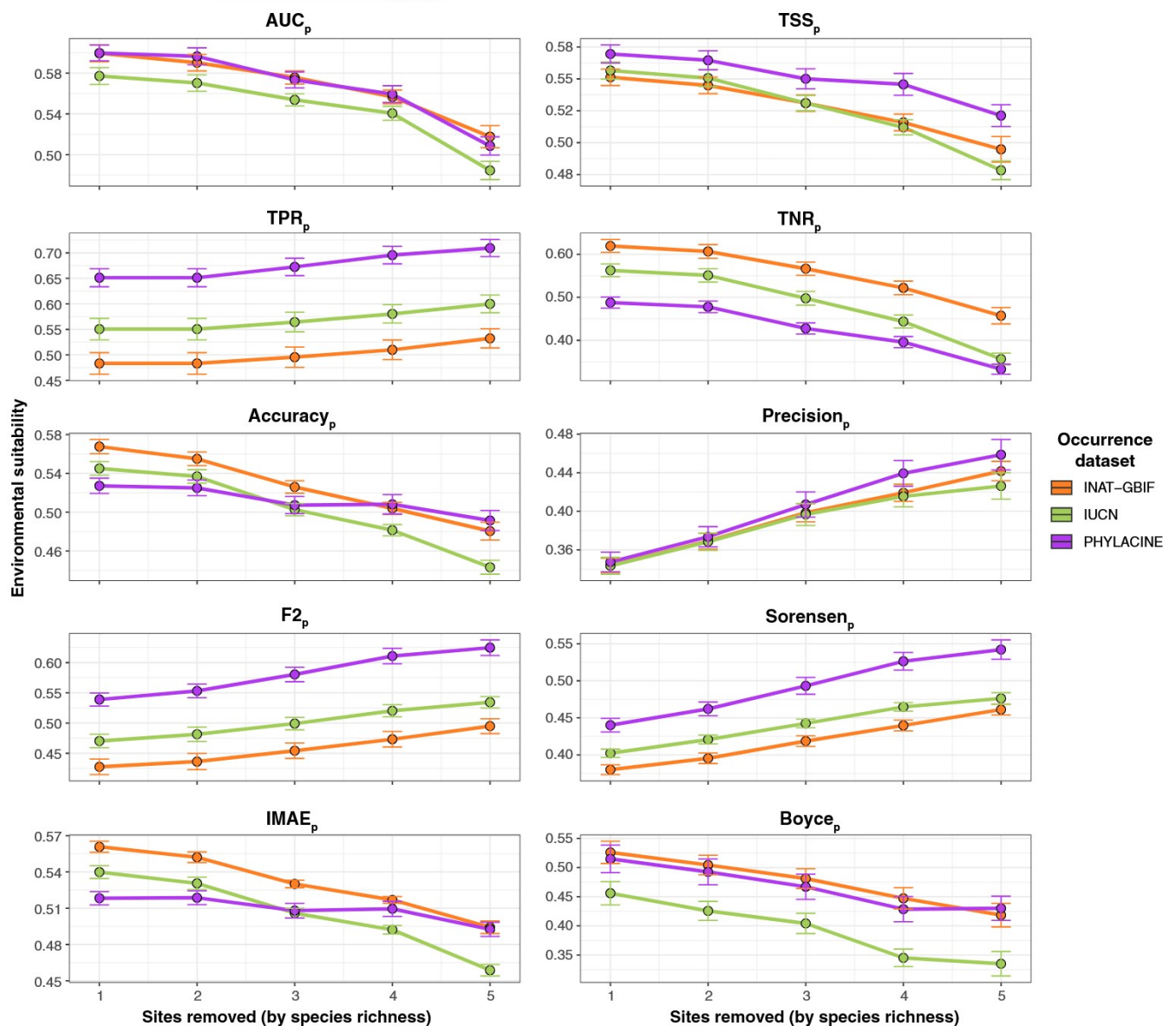


FIGURE 8 Sensitivity analysis showing the effect of paleozoological site removal on the paleometrics. The site removal is based on site species richness, whereby sites with lower species richness are progressively removed. This process continued until all sites with species richness of less than 10 are excluded.

Similarly, more complex models, such as those produced with GAU and SVM also showed low values of TPR_p and relatively high values of TNR_p, which again were accentuated when using the INAT-GBIF dataset (Figure 4 and Figure S13). The GAU algorithm consists of a machine learning procedure using maximum likelihood estimations (MLE) applied to Gaussian distributions characterised by a mean and variance. In the context of SDMs, the mean represents the optimal value of the environmental variable(s) for the species, and the variance represents the range of values over which the species can occur (Carl & Kühn, 2007; Golding & Purse, 2016; Lichstein et al., 2002). SVM is a machine learning algorithm that allow for complex non-linear relationships between variables, and it can often fit data better than more simple linear regression models (such as GLMs) and provide more accurate predictions of species

distributions (Li & Wang, 2013). But again, models working with complex functions such as GAU and SVM can result in unrealistic ecological relationships (Hallgren et al., 2019) and we show here that they are not the best at capturing the extent of past ranges of the hartebeest in the paleozoological record.

Quadratic GLMs and GAMs (limited to two knots), are relatively simple regression models commonly used in SDMs due to their ease of interpretation and implementation. These two modelling algorithms, together with MaxEnt and the Ensemble models, when combined with the PHYLACINE dataset, produced the highest TPR_p values (Figure 10 and Figure S17). These four models produced also high values of AUC_p and TSS_p across all datasets. MaxEnt works by fitting a probability distribution that maximises entropy subject to constraints imposed by the environmental data and the known

presences. To estimate the probability distribution, this algorithm first constructs a set of feature functions including both relatively simple functions such as linear or quadratic terms, as well as more

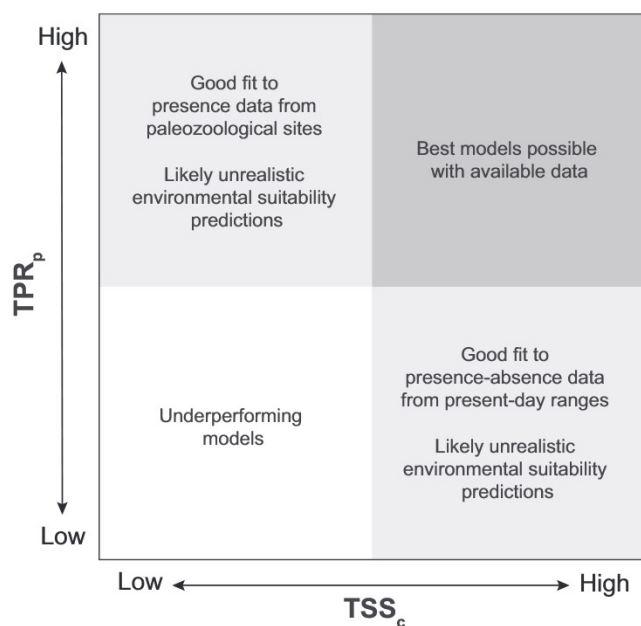


FIGURE 9 Modelling performance space comprised of TSS_c and TPR_p . Models with high TSS_c but low TPR_p represent a good fit to present-day distributions but are likely to generate unrealistic environmental suitability projections. Models with low TSS_c but high TPR_p represent a good fit to past distributions but are likely to generate unrealistic environmental suitability projections. Models with low TSS_c and low TPR_p are underperforming models, while the best models are those with high TSS_c and high TPR_p values.

complex functions allowing for interactions between variables (Phillips et al., 2009). This versatility explains why MaxEnt is one of the most used approaches in SDM and likely the reason why it performed so well in this study. The models generated with MaxEnt and the PHYLACINE occurrence dataset had a TSS_c over 0.8, and they were thus good at predicting both the modern and historical distribution of the hartebeest in our study (Figure 6).

4.2 | Integrating the zooarchaeological record with species distribution models

Our results suggest that regular evaluation metrics, such as TSS_c and AUC_c , fitted to present-day distributions may not always be ideal to select the best-performing models, and that integrating the paleozoological record in the evaluation process by using paleometrics can significantly improve projections. However, we show that alternative paleometrics vary in their assessment of model performance. The zooarchaeological and paleontological records are usually strongly biased (Behrensmeyer et al., 2000; Lloyd et al., 2012), such that the probability of false negatives (=false absences) is likely to be much higher than that of false positives (=false presences). So which evaluation metrics should we choose and how do we select the best models?

TSS_p balances the prediction accuracy of both presences (TPR_p) and absences (TNR_p). Models with high TSS_c but low TSS_p should represent a good fit to the current presence-absence data, but these are likely to produce less realistic results when making model projections through time (Figure 9). Conversely, those models with high TSS_p but low TSS_c , predict the paleozoological record better, but

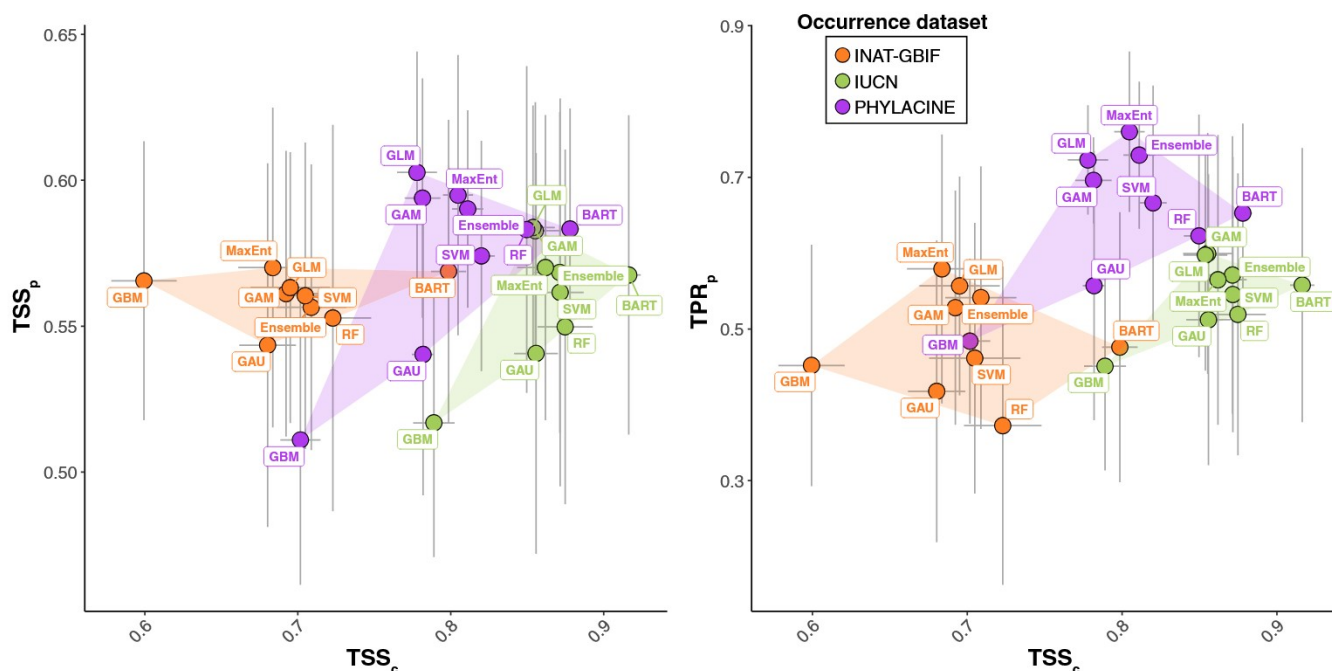


FIGURE 10 Biplots comparing TSS_c and the paleometrics TSS_p and TPR_p across all models, grouped and coloured by the various occurrence datasets and labelled by the different modelling algorithms used in this study. The circles represent the average across 10 replicates and the lines represent the range (minimum and maximum).

they may overestimate the range of suitable habitats as TSS_p can be significantly influenced by the uncertainty surrounding predictions of localities with unverified absences (TNR_p). Thus, good models are those that combine both a high TSS and high TSS_p , while models with low TSS and low TSS_p should be considered inadequate. For hartebeest, the PHYLACINE dataset combined with MaxEnt, GLM, GAM, or the Ensemble model offers the best results when taking into account both the present and past distribution fit (Figure 10). Paleometrics that emphasise the accurate prediction of presences may generally be more reliable indicators of model performance. These include not only TPR_p (how well presences are predicted) but also $F2_p$ and $Sorensen_p$. Unfortunately, high TPR_p values, especially when accompanied by low TNR_p values, can also result from excessive overprediction, indicating that the models might not be performing well.

As mentioned in the Introduction, when SDMs have a low capability to hindcast distributions, it may not necessarily reflect data or model limitations but could also result from shifts in the realised niche of the target species over time (Losos, 2008; Soberón, 2007; Stigall, 2012, 2014). The fundamental niche describes the environmental conditions under which a species can survive, without consideration of biotic interactions, and it may remain relatively broad and static over time. In contrast, the realised niche describes the actual environmental and biotic interactions where the species exists, and it is more dynamic due to ecological processes and altered biotic relationships (Soberón, 2010). This distinction is pivotal for SDMs, which often assume niche conservatism and do not account for changes in biotic interactions or environmental conditions that affect the realised niche (Pearman et al., 2014; Stigall et al., 2014). The issue is especially pertinent when predicting responses to novel climates lacking present-day analogues as combinations of biotic and/or abiotic conditions that have not previously been experienced are not incorporated in the model's baseline climate and ecological settings (Soberón, 2010; Varela et al., 2009). Recent discussions by Holt et al. (2020) and Norberg et al. (2019) underscores the importance of integrating niche dynamics and external ecological pressures in SDMs where possible to enhance their accuracy and ecological relevance. However, the frequent unavailability of necessary data often limits the inclusion of these factors, as was the case in our study.

5 | CONCLUSION

Our analysis unveiled significant variations in model performance of environmental suitability projections, both in the present and in the past, depending on the occurrence dataset used and the modelling algorithm applied. To select the most reliable modelling framework and interpret the outputs, we demonstrate the power of evaluating predicted distributions against the archaeological and fossil records of the species over extended periods of time. Although some modelling algorithms performed better than others in our study, we do

not suggest that any particular model is universally the best. Thus, it is advisable to test various alternative models and choose those that perform the best to address the specific research question with the data available. Specifically, high values of traditional metrics like AUC_c and TSS_c fitted to present-day distributions may be insufficient to guarantee that the habitat suitability projections represent realistic scenarios. We suggest adopting a combination of input data and modelling algorithms that is balanced to predict both past and present occurrences, assuming that absence data is less reliable than presence data in paleozoological records. Thus, we recommend using paleometrics such as TPR_p , $F2_p$, or $Sorensen_p$, but also ensure that values for more balanced metrics such as AUC_c , TSS_c , AUC_p , TSS_p , and $Boyce_p$ are reasonably high. Especially, when a model is to be used to make projections of changes in distribution in response to changing climates, we strongly advocate that its predictive power is tested through time and, given the potential for slow ecosystem responses and distributional lags, that the archaeological and paleontological records over an appropriately long time interval is used to evaluate the model performance.

AUTHOR CONTRIBUTIONS

Ignacio A. Lazagabaster: Conceptualization (lead); formal analysis (lead); investigation (lead); methodology (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). **Chris D. Thomas:** Conceptualization (lead); formal analysis (supporting); funding acquisition (supporting); investigation (equal); methodology (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Juliet V. Spedding:** Conceptualization (supporting); resources (supporting); writing – review and editing (equal). **Salima Ikram:** Conceptualization (supporting); funding acquisition (supporting); methodology (supporting); validation (supporting); writing – review and editing (equal). **Irene Solano-Regadera:** Methodology (supporting); writing – review and editing (supporting). **Steven Snape:** Conceptualization (supporting); funding acquisition (supporting); writing – review and editing (equal). **Jakob Bro-Jørgensen:** Conceptualization (lead); formal analysis (supporting); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); supervision (lead); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGEMENTS

We are grateful to BL Payne for helpful suggestions and discussion. We thank Anne Dallmeyer and Peter Hopcroft for sending paleoclimatic data and their helpful suggestions.

FUNDING INFORMATION

This work is part of the Biodiversity in Ancient Egypt during Societal Transitions (BEAST) project, funded by a Leverhulme Trust Research Project Grant with No. RPG-2021-104 awarded to JB-J.

CONFLICT OF INTEREST STATEMENT

No conflict of interest is reported.

OPEN RESEARCH BADGES



This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://osf.io/wck7n/?view_only=64874fd144754af28b6e4684f568bf44.

DATA AVAILABILITY STATEMENT

All data and code used in this work are available as Supporting Information in the online version of this manuscript and at Open Science Framework: https://osf.io/wck7n/?view_only=64874fd144754af28b6e4684f568bf44.

ORCID

Ignacio A. Lazagabaster <https://orcid.org/0000-0001-9149-7371>

REFERENCES

- Allouche, O., Tsao, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *The Journal of Applied Ecology*, 43(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5(1), eaat4858. <https://doi.org/10.1126/sciadv.aat4858>
- Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1), 1–19. <https://doi.org/10.1016/j.ecolmodel.2006.07.005>
- Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, 3(2), 327–338. <https://doi.org/10.1111/j.2041-210x.2011.00172.x>
- Barbet-Massin, M., Thuiller, W., & Jiguet, F. (2010). How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography*, 33(5), 878–886.
- Barbosa, A. M. (2015). Re-scaling of model evaluation measures to allow direct comparison of their values. *Journal of Brief Ideas*, 10, 1. <https://doi.org/10.5281/zenodo.15487>
- Barnosky, A. D., Hadly, E. A., Gonzalez, P., Head, J., Polly, P. D., Lawing, A. M., Eronen, J. T., Ackerly, D. D., Alex, K., Biber, E., Blois, J., Brashares, J., Ceballos, G., Davis, E., Dietl, G. P., Dirzo, R., Doremus, H., Fortelius, M., Greene, H. W., ... Zhang, Z. (2017). Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science*, 355(6325), eaah4787. <https://doi.org/10.1126/science.aah4787>
- Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19, 10–15. <https://doi.org/10.1016/j.ecoinf.2013.11.002>
- Behrensmeyer, A. K., Kidwell, S. M., & Gastaldo, R. A. (2000). Taphonomy and paleobiology. *Paleobiology*, 26(S4), 103–147. <https://doi.org/10.1017/S0094837300026907>
- Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., & Zhao, Y. (2012). Evaluation of climate models using palaeoclimatic data. *Nature Climate Change*, 2(6), 417–424. <https://doi.org/10.1038/nclim.ate1456>
- Carl, G., & Kühn, I. (2007). Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modelling*, 207(2), 159–170. <https://doi.org/10.1016/j.ecolmodel.2007.04.024>
- Carlson, C. J. (2020). Embarcadero: Species distribution modelling with Bayesian additive regression trees in R. *Methods in Ecology and Evolution*, 11(7), 850–858. <https://doi.org/10.1111/2041-210x.13389>
- Chamaillé-Jammes, S., & Fritz, H. (2009). Precipitation–NDVI relationships in eastern and southern African savannas vary along a precipitation gradient. *International Journal of Remote Sensing*, 30(13), 3409–3422.
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266–298.
- Cranston, A., Cooper, N. and Bro-Jørgensen, J., 2024. Using joint species distribution modelling to identify climatic and non-climatic drivers of Afrotropical ungulate distributions. *Ecography*, e07209. <https://doi.org/10.1111/ecog.07209>
- Cutler, D. R., Edwards, T. C., Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792. <https://doi.org/10.1890/07-0539.1>
- Dallmeyer, A., Claussen, M., Lorenz, S. J., & Shanahan, T. (2020). The end of the African humid period as seen by a transient comprehensive earth system model simulation of the last 8000 years. *Climate of the Past*, 16(1), 117–140. <https://doi.org/10.5194/cp-16-117-2020>
- Dietl, G. P., Kidwell, S. M., Brenner, M., Burney, D. A., Flessa, K. W., Jackson, S. T., & Koch, P. L. (2015). Conservation Paleobiology: Leveraging knowledge of the past to inform conservation and restoration. *Annual Review of Earth and Planetary Sciences*, 43(1), 79–103. <https://doi.org/10.1146/annurev-earth-040610-133349>
- Duputié, A., Zimmermann, N. E., & Chuine, I. (2014). Where are the wild things? Why we need better data on species distribution. *Global Ecology and Biogeography: A Journal of Macroecology*, 23(4), 457–467. <https://doi.org/10.1111/geb.12118>
- Elith, J., & Graham, C. H. (2009). Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), 66–77. <http://www.jstor.org/stable/30244651>
- Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 677–697. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of Animal Ecology*, 77(4), 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Faurby, S., Davis, M., Pedersen, R. Ø., Schowaneck, S. D., Antonelli, A., & Svenning, J.-C. (2018). PHYLACINE 1.2: The phylogenetic atlas of mammal macroecology. *Ecology*, 99(11), 2626. <https://doi.org/10.1002/ecy.2443>
- Fourcade, Y. (2016). Comparing species distributions modelled from occurrence data and from expert-based range maps. Implication for predicting range shifts with climate change. *Ecological Informatics*, 36, 8–14. <https://doi.org/10.1016/j.ecoinf.2016.09.002>
- Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography: A Journal of Macroecology*, 27(2), 245–256. <https://doi.org/10.1111/geb.12684>
- Franklin, J., Davis, F. W., Ikegami, M., Syphard, A. D., Flint, L. E., Flint, A. L., & Hannah, L. (2013). Modeling plant species distributions under future climates: How fine scale do climate projections need to be? *Global Change Biology*, 19(2), 473–483. <https://doi.org/10.1111/gcb.12051>
- Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology*

- and Evolution/British Ecological Society, 7, 598–608. <https://doi.org/10.1111/2041-210X.12523>
- Gosling, L. M., & Capellini, I. (2013). *Alcelaphus buselaphus* Hartbeest. In J. S. Kingdon & M. Hoffmann (Eds.), *The mammals of Africa* (pp. 511–526). Bloomsbury Publishing.
- Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., & Mackey, B. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, 408, 108719. <https://doi.org/10.1016/j.ecolmodel.2019.108719>
- Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2), 142–152.
- Holt, R. D. (2020). Some thoughts about the challenge of inferring ecological interactions from spatial data. *Biodiversity Informatics*, 15(1), 61–66.
- Inman, R., Franklin, J., Esque, T., & Nussear, K. (2018). Spatial sampling bias in the Neotoma paleoecological archives affects species paleo-distribution models. *Quaternary Science Reviews*, 198, 115–125. <https://doi.org/10.1016/j.quascirev.2018.08.015>
- Jablonski, D., Gould, S. J., & Raup, D. M. (1986, 1985). The nature of the fossil record: A biological perspective. In *Patterns and processes in the history of life: Report of the Dahlem workshop on patterns and processes in the history of life Berlin* (pp. 7–22). Springer.
- Jeschke, J. M., & Strayer, D. L. (2008). Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences*, 1134, 1–24. <https://doi.org/10.1196/annals.1439.002>
- Jiménez, L., & Soberón, J. (2020). Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. *Methods in Ecology and Evolution*, 11(12), 1571–1586. <https://doi.org/10.1111/2041-210X.13479>
- Karger, D. N., Nobis, M. P., Normand, S., Graham, C. H., & Zimmermann, N. E. (2023). CHELSA-TraCE21k-high-resolution (1 km) downscaled transient temperature and precipitation data since the Last Glacial Maximum. *Climate of the Past*, 19(2), 439–456.
- Kingdon, J. (2015). *The Kingdon field guide to African mammals* (2nd ed.). Bloomsbury Publishing.
- Lazagabaster, I. A., Rovelli, V., Fabre, P.-H., Porat, R., Ullman, M., Davidovich, U., Lavi, T., Ganor, A., Klein, E., Weiss, K., Nuriel, P., Meiri, M., & Marom, N. (2021). Rare crested rat subfossils unveil Afro-Eurasian ecological corridors synchronous with early human dispersals. *Proceedings of the National Academy of Sciences of the United States of America*, 118(31), e2105719118. <https://doi.org/10.1073/pnas.2105719118>
- Lentini, P. E., Stirnemann, I. A., Stojanovic, D., Worthy, T. H., & Stein, J. A. (2018). Using fossil records to inform reintroduction of the kakapo as a refugee species. *Biological Conservation*, 217, 157–165. <https://doi.org/10.1016/j.biocon.2017.10.027>
- Li, X., & Wang, Y. (2013). Applying various algorithms for species distribution modelling. *Integrative Zoology*, 8(2), 124–135. <https://doi.org/10.1111/1749-4877.12000>
- Lichstein, J. W., Simons, T. R., & Shriver, S. A. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, 72, 445–463. [https://doi.org/10.1890/0012-9615\(2002\)072\[0445:SAAAM\]2.0.CO;2](https://doi.org/10.1890/0012-9615(2002)072[0445:SAAAM]2.0.CO;2)
- Lima-Ribeiro, M. S., Moreno, A. K. M., Terribile, L. C., Caten, C. T., Loyola, R., Rangel, T. F., & Diniz-Filho, J. A. F. (2017). Fossil record improves biodiversity risk assessment under future climate change scenarios. *Diversity and Distributions*, 23(8), 922–933. <https://doi.org/10.1111/ddi.12575>
- Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, 34(2), 232–243. <https://doi.org/10.1111/j.1466-8238.2012.00790.x>
- Lloyd, G. T., Young, J. R., & Smith, A. B. (2012). Taxonomic structure of the fossil record is shaped by sampling bias. *Systematic Biology*, 61(1), 80–89. <https://doi.org/10.1093/sysbio/syr076>
- Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1), 103–114. <https://doi.org/10.1111/j.1600-0587.2009.06039.x>
- Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2), 145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Loggers, C. O., Thévenot, M., & Aulagnier, S. (1992). Status and distribution of Moroccan wild ungulates. *Biological Conservation*, 59(1), 9–18. [https://doi.org/10.1016/0006-3207\(92\)90708-U](https://doi.org/10.1016/0006-3207(92)90708-U)
- Losos, J. B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters*, 11(10), 995–1003. <https://doi.org/10.1111/j.1461-0248.2008.01229.x>
- Márcia Barbosa, A., Real, R., Muñoz, A.-R., & Brown, J. A. (2013). New measures for assessing model equilibrium and prediction mismatch in species distribution models. *Diversity and Distributions*, 19(10), 1333–1338. <https://doi.org/10.1111/ddi.12100>
- Martin, J. M., Short, R. A., Plumb, G. E., Markewicz, L., Van Vuren, D. H., Wehus-Tow, B., Otárola-Castillo, E., & Hill, M. E., Jr. (2023). Integrated evidence-based extent of occurrence for north American bison (*Bison bison*) since 1500 CE and before. *Ecology*, 104(1), e3864. <https://doi.org/10.1002/ecy.3864>
- Myers, C. E., Stigall, A. L., & Lieberman, B. S. (2015). PaleoENM: Applying ecological niche modeling to the fossil record. *Paleobiology*, 41(2), 226–244. <https://doi.org/10.1017/pab.2014.19>
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurobotics*, 7, 21. <https://doi.org/10.3389/fnbot.2013.00021>
- Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., & Foster, S. D. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3), e01370.
- Payne, B. L., & Bro-Jørgensen, J. (2016a). Disproportionate climate-induced range loss forecast for the most threatened African antelopes. *Current Biology*, 26(9), 1200–1205. <https://doi.org/10.1016/j.cub.2016.02.067>
- Payne, B. L., & Bro-Jørgensen, J. (2016b). A framework for prioritizing conservation translocations to mimic natural ecological processes under climate change: A case study with African antelopes. *Biological Conservation*, 201, 230–236. <https://doi.org/10.1016/j.biocon.2016.07.018>
- Pearce, J., & Ferrier, S. (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, 128(2), 127–147. [https://doi.org/10.1016/S0304-3800\(99\)00227-6](https://doi.org/10.1016/S0304-3800(99)00227-6)
- Pearman, P. B., Lavergne, S., Roquet, C., Wüest, R., Zimmermann, N. E., & Thuiller, W. (2014). Phylogenetic patterns of climatic, habitat and trophic niches in a European avian assemblage. *Global Ecology and Biogeography*, 23(4), 414–424.
- Phelps, L. N., Broennimann, O., Manning, K., Timpson, A., Jousse, H., Mariethoz, G., Fordham, D. A., Shanahan, T. M., Davis, B. A., & Guisan, A. (2020). Reconstructing the climatic niche breadth of land use for animal production during the African Holocene. *Global Ecology and Biogeography*, 29(1), 127–147.
- Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications: A Publication of the Ecological Society of America*, 19(1), 181–197. <https://doi.org/10.1890/07-2153.1>

- Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C. F., Vittoz, P., & Guisan, A. (2013). The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, 22(1), 52–63. <https://doi.org/10.1111/j.1466-8238.2012.00790.x>
- Prentice, I. C., Cramer, W., Harrison, S. P., Leemans, R., Monserud, R. A., & Solomon, A. M. (1992). Special paper: A global Biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*, 19(2), 117–134. <https://doi.org/10.2307/2845499>
- Radomski, T., Beamer, D., Babineau, A., Wilson, C., Pechmann, J., & Kozak, K. H. (2022). Finding what you don't know: Testing SDM methods for poorly known species. *Diversity and Distributions*, 28(9), 1769–1780. <https://doi.org/10.1111/ddi.13536>
- Roberts, D. R., & Hamann, A. (2012). Method selection for species distribution modelling: Are temporally or spatially independent evaluations necessary? *Ecography*, 35(9), 792–802. <https://doi.org/10.1111/j.1600-0587.2011.07147.x>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 1–8.
- Smith, S. E., Gregory, R. D., Anderson, B. J., & Thomas, C. D. (2013). The past, present and potential future distributions of cold-adapted bird species. *Diversity and Distributions*, 19(3), 352–362. <https://doi.org/10.1111/ddi.12025>
- Smith, A. B., Murphy, S. J., Henderson, D., & Erickson, K. D. (2023). Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecology and Biogeography*, 32(3), 342–355.
- Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, 10(12), 1115–1123.
- Soberón, J. M. (2010). Niche and area of distribution modeling: A population ecology perspective. *Ecography*, 33(1), 159–167. <https://doi.org/10.1111/j.1600-0587.2009.06074.x>
- Stigall, A. L. (2012). Using ecological niche modelling to evaluate niche stability in deep time. *Journal of Biogeography*, 39(4), 772–781. <https://doi.org/10.1111/j.1365-2699.2011.02651.x>
- Stigall, A. L. (2014). When and how do species achieve niche stability over long time scales? *Ecography*, 37(11), 1123–1132.
- Svenning, J.-C., Fløjgaard, C., Marske, K. A., Nógues-Bravo, D., & Normand, S. (2011). Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews*, 30(21–22), 2930–2947. <https://doi.org/10.1016/j.quascirev.2011.06.012>
- Tessarolo, G., Rangel, T. F., Araújo, M. B., & Hortal, J. (2014). Uncertainty associated with survey design in species distribution models. *Diversity and Distributions*, 20(11), 1258–1269. <https://doi.org/10.1111/ddi.12236>
- Tsahar, E., Izhaki, I., Lev-Yadun, S., & Bar-Oz, G. (2009). Distribution and extinction of ungulates during the Holocene of the southern Levant. *PLoS One*, 4(4), e5316. <https://doi.org/10.1371/journal.pone.0005316>
- Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, 13(4), 397–405. <https://doi.org/10.1111/j.1472-4642.2007.00346.x>
- Valavi, R., Guillera-Aroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, 92(1), e01486. <https://doi.org/10.1002/ecm.1486>
- Varela, S., Lobo, J. M., & Hortal, J. (2011). Using species distribution models in paleobiogeography: A matter of data, predictors and concepts. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 310(3), 451–463. <https://doi.org/10.1016/j.palaeo.2011.07.021>
- Varela, S., Rodríguez, J., & Lobo, J. M. (2009). Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. *Journal of Biogeography*, 36(9), 1645–1655. <https://doi.org/10.1111/j.1365-2699.2009.02125.x>
- Velazco, S. J. E., Rose, M. B., de Andrade, A. F. A., Minoli, I., & Franklin, J. (2022). Flexsdm: An R package for supporting a comprehensive and flexible species distribution modelling workflow. *Methods in Ecology and Evolution/British Ecological Society*, 13(8), 1661–1669. <https://doi.org/10.1111/2041-210x.13874>
- Werkowska, W., Márquez, A. L., Real, R., & Acevedo, P. (2017). A practical overview of transferability in species distribution modeling. *Environmental Review*, 25(1), 127–133. <https://doi.org/10.1139/er-2016-0045>
- Wunderlich, R. F., Lin, Y.-P., Anthony, J., & Petway, J. R. (2019). Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. *Nature Conservation*, 35, 97–116. <https://doi.org/10.3897/natureconservation.35.33918>
- Zhang, L., Liu, S., Sun, P., Wang, T., Wang, G., Zhang, X., & Wang, L. (2015). Consensus forecasting of species distributions: The effects of niche model performance and niche properties. *PLoS One*, 10(3), e0120056. <https://doi.org/10.1371/journal.pone.0120056>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lazagabaster, I. A., Thomas, C. D., Spedding, J. V., Ikram, S., Solano-Regadera, I., Snape, S., & Bro-Jørgensen, J. (2024). Evaluating species distribution model predictions through time against paleozoological records. *Ecology and Evolution*, 14, e70288. <https://doi.org/10.1002/ece3.70288>