Predicting Habitat Suitability for Asian Elephants in Non-analog Ecosystems with Bayesian Models

Ryoko Noda[a,*], Michael Francis Mechenich[b], Juha Saarinen[c], Aki Vehtari[d], Indrė Žliobaitė[b,c]

[a]*School of Science, Aalto University, P.O. Box 11000, FI-00076, Finland*
[b]*Department of Computer Science, University of Helsinki, P.O. Box 68, FI-00014, Finland*
[c]*Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, FI-00014, Finland*
[d]*Department of Computer Science, Aalto University, Tietotekniikan laitos, P.O.Box 15400, FI-00076, Finland*

Abstract

Rewilding is an ambitious approach to conservation aiming at restoring and protecting natural processes. As the world is rapidly transitioning into conditions that have not been observed before, we need to be able to extrapolate and predict how natural processes would act under new conditions. Species distribution models have a good potential to inform rewilding decisions by the predictive modeling of potential species presence under various habitat conditions. A critical requirement when utilizing these models is to be able to express the uncertainty in the environment or its predictions. This study demonstrates the use of Bayesian statistical models to address this challenge. As a case study, we explore Bayesian logistic regression and Bayesian generalized additive models in order to predict suitable habitats for Asian elephants (*Elephas maximus*) until the year 2070 under the worst case working scenario of climate change. In this comparative study predictions of habitat suitability are solely based on climatic conditions. The results of the two Bayesian models are compared to two benchmark models, maximum-likelihood estimated logistic regression and random forest. We analyze and discuss trade-offs, relative advantages and limitations of these modelling choices. The results of our analysis suggest that the Bayesian logistic regression fit on scaled spatial-CV features with adjusted, wider priors gives the most robust predictions in this setting, which tend to correspond with the distribution of woodland biomes broadly similar to those in the species' historical range.

Keywords: Bayesian Statistical Models, Species Distribution Modelling, Rewilding, Non-analog Ecosystems, *Elephas maximus*

1. Introduction

Rewilding is a conservation strategy which proposes ecosystem restoration through the reintroduction of extirpated megafaunal species [1, 2]. While relatively few such wild experiments [3] have been conducted, several projects have provided insight into the potential for megafauna to reestablish important top-down tropic cascades [4]. In one notable case, gray wolves reintroduced in Yellowstone National Park suppressed the Rocky Mountain elk population, allowing woody riparian plants to recover after seventy years of overbrowsing [5].

Naturally, such projects must be considered carefully from different angles, both ecological and cultural, especially when introducing a species outside its native geographic range. A first criterion to consider is the suitability of climatic conditions for the species, at present and under future climate change, which adds uncertainty to assessments of long-term survival in the introduced range.

Species distribution models (SDMs) [6], also known as ecological niche models or climatic envelope models, help manage this uncertainty. These are mathematical or statistical models that predict habitat suitability for a certain species given climatic inputs, such as temperature, precipitation, or vegetation. They provide a data-driven assessment of to what extent the species can thrive in its new range or the ways in which the range can shift over time [7, 8, 9].

SDMs are usually statistical, machine learning or process-based models that connect climatic features with species observations or the physiology of the

*Corresponding author, 13 Rue de la Commune, 1210 Saint-Josse-ten-Noode, Belgium. Email: ryokonoda.1811@gmail.com

species. These types of models are limiting because they do not model the random effects and fluctuations in the environment nor signal the uncertainty of predictions. Additionally, they require a great amount of data, distribution data for machine learning models and physiological data for process-based models. For both types of models, the data is difficult and costly to collect. This poses a bottleneck for building SDMs for practical use.

Bayesian models may offer a solution to these limitations. These are statistical models that predict with uncertainty by combining human knowledge with observed data. They utilize probability distributions to present all uncertainties, including predictions and parameters that determine the relationship between the data and the predictions. Furthermore, since they incorporate prior assumptions about the problem that they attempt to solve, they can potentially work more robustly with relatively small amounts of data. In this case, the predictions are initially closer to prior hypotheses based on human knowledge and become increasingly data driven as more data arrives. In addition to the possibility of using additional prior information, Bayesian inference is attractive because of the integration over the uncertainty presented by the posterior, which also has a regularizing effect on the predictions.

In order to explore how Bayesian models can be used to predict species distributions for rewilding projects, we conduct a methodological case study. We explore two types of models: Bayesian logistic regression models and Bayesian generalized additive models (Bayesian GAMs) [10]. These are iteratively built and evaluated using different modeling options: feature selection, feature scaling, weakly informative priors of different strengths, and for the Bayesian GAMs, different non-linearity settings. The results are then compared to the results from two baseline non-Bayesian models: maximum-likelihood estimated (MLE) logistic regression and random forest. These are also iteratively built on the same modeling choices, but only for the ones that make sense for non-Bayesian models (feature selection and feature scaling).

The iterative model exploration in this work roughly follows the Bayesian workflow described by Gelman et al. [11]. The Bayesian and non-Bayesian model pairs are chosen so that one represents a linear model (Bayesian logistic regression and MLE logistic regression) and one represents a non-linear model (Bayesian GAM and random forest).

The models are trained to predict species distributions to inform a potential rewilding project. In this study, we attempt to locate feasible areas for Asian elephants (*Elephas maximus*) that would remain climatically suitable in the year 2070 in the worst scenario of climate change, known as RCP 8.5 [12], which does not include any climate mitigation actions. The search area includes the Earth's entire terrestrial surface, excluding Antarctica and regions at elevations greater than 3,000 meters MSL, the species' upper elevation limit [7].

2. Background on *Elephas* ecology and range

As a megaherbivore species having the potential to reestablish top-down trophic cascades [4] in trophically downgraded ecosystems [13], the Asian elephant has been proposed as a candidate for conservation translocation in South America [14], North America [2], Europe [15], Australia [16], and the tropical Asia-Pacific region [17].

According to the current understanding, the genus *Elephas* originated in Africa during the Pliocene [18, 19]. While the identity of many of the formerly recognized "subspecies" of the "*Elephas recki*" - species complex from the Pliocene and Pleistocene of Africa have recently been re-identified as members of the genus *Palaeoloxodon* [19, 20], the earliest confirmed *Elephas* species in the most recent phylogenetic analyses [19] is *E. atavus* from the Early Pleistocene of East Africa, where it lived in savanna environments and had grass-dominated diet [21]. It is possible, however, that some Pliocene elephants from Africa do belong to the genus *Elephas*, such as *E. ekorensis* from the Early Pliocene of East Africa [18]. Also the Pliocene to earliest Pleistocene taxa "*Elephas recki brumpti*" and "*Elephas recki shungurensis*" from East Africa have been tentatively assigned to the genus *Elephas* in recent revisions [22, 23].

The genus *Elephas* went extinct in Africa during the Pleistocene (e.g., [18]), but was widespread across Southern Asia from the Pleistocene to the Holocene (e.g., [24]). The extant Asian elephant (*Elephas maximus*) had a much wider range from the Late Pleistocene to most of the Holocene than it has today. The historical range extended from the Southeast Asian archipelago to eastern central China in the north, Southern Asia South of the Tibetan Plateau, and parts of Western Asia through the narrow corridor of the fertile crescent all the way to its westernmost occurrence around the Gavur Lake in central Anatolia [25]. The current range of *E. maximus* is limited to parts of India, Sri Lanka and Southeast Asia, including Borneo and Sumatra, and it is heavily influenced by the loss of suitable habitat by land use, as well as other human influence (e.g. [24, 26]).

Today *E. maximus* occupies a variety of environments from tropical rainforests in Borneo to seasonally dry

grassy woodlands in Sri Lanka. It prefers ecotone habitats where the vegetation consists of a mixture of forest, low-growing woody plants, and grasses [27]. In terms of diet, *E. maximus* is a mixed-feeder including varying proportions of grass and browse in its diet, following both seasonal and regional variation in available vegetation. In dry forests and woodlands, the diet is usually browse-dominated during dry seasons, while grasses can form the majority of the diet during wet seasons [26]. In tropical rainforests the diet can consist almost entirely of browse and fruit [26]. It has been noted that unlike the African savanna elephant (*Loxodonta africana*), *E. maximus* does not seem to extensively transform woodlands into more open habitats [26].

3. The Study Area and Data

For predictive modeling we use present-day climatic features as input, and the estimated presence or absence of *E. maximus*, based on an expert-delineated range map, as binary labels. The training dataset for all models consists of hexagonal grid cells on the globe, where each cell is a learning instance. Instances where the species is assumed to be present were sampled from the estimated present-natural range, and instances where the species is assumed to be absent were sampled from the region surrounding the species' range: areas to which the species has access via dispersal, but which were not included in the estimated range.

The target prediction area is all land areas on Earth under present-day and future conditions, excluding Antarctica; further, areas above the species' upper elevation limit (3 000 meters MSL[7]) were removed from both the training and target datasets. Excluding the present-day areas that we used for training, the target prediction area does not have label assignments. The training and target prediction areas are visualized in Figures 1a and 1b respectively.

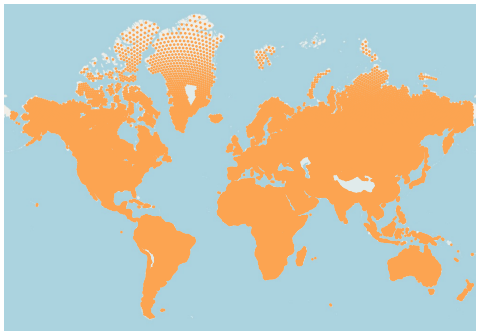
3.1. The Discrete Global Grid System

A discrete global grid system (DGGS) partitions the Earth's surface into areal grid cells, which serve as units of observation and analysis. It provides a systematic spatial framework, to which model training and target data may be mapped.

In ecology, rectangular grid systems based on latitude and longitude are most commonly used [28]. However, in developing our SDMs we utilized the ISEA3H DGGS [29], a hexagonal grid system. The cells of this system are equal-area and maximally compact, and as a result, have a statistical advantage over those of latitude and longitude-based systems [29].

We used climate and species distribution data from the Eco-ISEA3H spatial database [30, 31] at resolution nine, a high-resolution grid in which the distance between the centroids of the hexagonal cells is approximately 50 kilometers globally.

(a) The training area with the present-natural range from the PHYLACINE database (presences) [32, 33] in yellow and pseudo-absence areas (absences) in blue. We have not labeled areas above the species' upper elevation limit (3,000 meters MSL[7]), namely the Himalayas and Tibetan Plateau, as these present barriers to dispersal.



(b) The target prediction area in orange, with areas above 3,000 meters MSL excluded.

Figure 1: The training area and target prediction area visualized by grid cell centroids. Because a high-resolution grid was used, centroids are visible as discrete points only at high latitudes, where the map projection exaggerates inter-centroid distances.

3.2. Climate Datasets

The climate data used in this study comprises the 27 climate extremes indices defined by the Expert Team on Climate Change Detection and Indices (ETCCDI), and the 19 bioclimatic variables from WorldClim [34]. For both sets of climatic variables, we chose observational and simulation data averaged for the years 1950-2000 to represent the present-day climate, and simulation data averaged for the years 2061-2080 to represent the expected climate in approximately 2070.

We use present-day and future ETCCDI indices derived from results of the Community Climate System Model version 4 (CCSM4) [35], one of a number of Coupled Model Intercomparison Project Phase 5 (CMIP5) global climate models for which these indices were calculated [36, 37]. Present-day WorldClim bioclimatic variables were derived from monthly temperature and precipitation, interpolated from weather station observations using thin-plate smoothing splines [34]. Future bioclimatic variables were derived from CCSM4 results, downscaled and bias-corrected using the WorldClim present-day interpolations as the baseline climate.

For all future prediction data we selected the representative concentration pathway 8.5 (RCP 8.5) scenario, also known as the 'business as usual' scenario [38]. It assumes the absence of climate change policies, high energy demands, and slow technological change. Since it has the highest predicted greenhouse gas emissions, we chose it to represent the worst case scenario of global warming.

3.3. Species Distribution

The species distribution data used in this study is based on the 'present-natural' range of *E. maximus* from the PHYLACINE database [32, 33]. The present-natural range, mapped in yellow in Figure 1a, represents areas estimated to have been habitats for the Asian elephant under present-day climate, if not for human activity. The grid cells within this range are labeled as presences, or areas where Asian elephant populations are able to survive. In using the present-natural range for model training, we attempt to avoid the biasing effect of anthropogenic range contraction [39].

Because our models also required unsuitable habitat areas as examples for training, there was a need to add 'pseudo-absence' points to represent habitats that would be climatically unsuitable for our target species [40]. This area of pseudo-absence was defined by iteratively buffering the present-natural range, incorporating adjacent land area accessible by dispersal, but not included in the species' estimated range. As shown in Figure 1a, this pseudo-absence buffer surrounds the present-natural range, excluding areas of high topographic relief (above 3,000 meters MSL), which present barriers to movement.

The grid cells within this buffer are labeled as absences, or areas where Asian elephant populations cannot maintain viable populations. The pseudo-absence area was constructed so that the number of cells with the 'species absent' labels approximately equals the number of cells with the 'species present' labels. After adding the pseudo-absence area, the total grid cells with habitat suitability labels were 7,331, 3,765 of which were marked as the species being present.

4. Statistical Models

For comparative analysis of alternative computational modeling choices, we used linear and non-linear models in the Bayesian and non-Bayesian frameworks.

4.1. Linear models

4.1.1. Baseline model: MLE logistic regression

Logistic regression in general is a statistical model that is used to predict binary outcomes. If we denote the probability of an observation being positive (species present, in our case) to be $P(y = 1)$, and the input features as $x_1, x_2, \cdots, x_k$, it is modeling the logit of the probability with a linear hyperplane as

$$\log \frac{P(y=1)}{1-P(y=1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \quad (1)$$

where the coefficients and intercept $\beta_0, \beta_1, \cdots, \beta_k$ are the parameters of the model. The logit of the probability, $\log(P(y = 1)/(1 - P(y = 1)))$, is also called the 'log-odds', and it is an intermediate quantity for which the linear model is often justified.

The log-odds is transformed into a probability using the inverse-logit function

$$P(y=1) = \frac{1}{1-e^{-z}}, \quad (2)$$

where z denotes the log-odds. We used maximum-likelihood estimation to infer parameter estimates.

4.1.2. Bayesian logistic regression

Bayesian logistic regression uses the same model as MLE logistic regression (Section 4.1.1), but uses Bayesian inference for the parameters. The Bayesian inference produces posterior distributions for the parameters and posterior predictive distributions for the predictions.

4.2. Non-linear models

4.2.1. Baseline model: Random forest

Random forest [41] is a non-linear modeling technique. It utilizes an ensemble of classification or regression trees; each tree is grown from a bootstrap sample of the training dataset, and represents a series of sequential decisions, in which each node of the tree is a binary split made on a predictive feature (e.g., mean annual temperature above 25°C or not). Further, a random subset of

features is considered when finding the optimal split at each node (we use a value of sqrt(n), n being the number of potential predictors, for the size of this subset). When used for classification, the outputs from the component trees are put through a majority vote to create a single output. This is known to be a simple but powerful method of retaining the complex non-linearity of decision trees while avoiding overfitting.

The random forest models within this study were designed to match the implementation of our previous research [42]. We used the prediction of a random forest model with 100 decision trees for the final prediction outputs.

4.2.2. *Bayesian generalized additive model*

Bayesian generalized additive model (GAM) has a structure similar to Bayesian logistic regression, but uses a sum of non-linear (usually) univariate functions $s_k(\cdot)$ of the features instead of linear functions to model log-odds.

$$\log \frac{P(y=1)}{1-P(y=1)} = \beta_0 + \beta_1 s_1(x_1) + \beta_2 s_2(x_2) + \cdots + \beta_k s_k(x_k). \quad (3)$$

4.3. *Implementation*

Modeling was done with various packages within R [43]. For modeling MLE logistic regression we used the glm function of R's stat package. For random forest we used the randomForest package [44]. And for the Bayesian models we used the brms [45] package.

5. Analysis protocol

5.1. *Modeling options*

For each Bayesian model, we used four modeling options in order to find models that best fit the data: feature selection, feature scaling, priors, and basis dimension (Bayesian GAM only). For non-Bayesian models, we used feature selection and feature scaling (since the rest were not applicable).

5.1.1. *Feature selection*

SDM development requires two feature selection steps, namely (1) removing highly collinear features [46], and (2) identifying a minimum subset of available features with greatest ecological relevance and explanatory power [6]. First, collinear features were removed from the initial set of 46 climatic predictors via the *vifcor* algorithm [47]; this procedure identifies the most correlated pair of predictors, removes the predictor with the greater variance inflation factor (VIF), and iterates until all pairwise correlations fall below a specified threshold.

Following this, forward feature selection (FFS) was performed. FFS is a common stepwise selection procedure [48], which begins with an empty model, and iteratively selects the next feature from the initial set which most improves model performance. Candidate logistic regression and random forest models were evaluated via (1) random cross-validation (CV) and (2) spatial CV [49], in which the full training dataset was partitioned into spatial blocks like those shown in Figure 2.

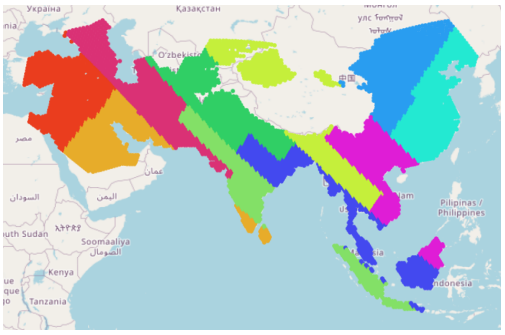


Figure 2: One spatial-CV fold pattern used in feature selection. This particular fold was also utilized when calculating numerical scores for the models.

Because model performance estimated via spatial CV is dependent on the spatial configuration of the CV folds, the procedure was repeated over 100 random blocking configurations; those features selected most frequently in the best-performing feature set were retained for the final model [42]. While this does not fully eliminate a possibility for information leakage from the ground truth labels, our sensitivity experiments reported later on in Section 5.2 show that the potential impact of such a leakage to the modelling accuracy would be negligible in this setting. Random and spatial CV selected different feature sets, and these are compared in our final modeling and analysis, under the names 'random-CV' and 'spatial-CV' features, respectively. Lists of features and short descriptions are presented in Table 1.

5.1.2. *Feature scaling*

We tested both raw and scaled versions of features when modeling. Thus there were four feature sets to fit per model type (raw random-CV features, scaled random-CV features, raw spatial-CV features, and scaled spatial-CV features).

For our dataset, feature scaling refers to preprocessing the input features so that they have range [0,1] but

retains the shape of its original distribution. This is done through min-max scaling

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}. \quad (4)$$

Maximums and minimums of the variables were computed over the full modelling dataset.

5.1.3. Selecting priors

We started by assigning weak priors to Bayesian models. For Bayesian logistic regression, we used the priorsense package [50] to assess prior sensitivity. priorsense is an R package that automates the diagnosis by using slightly perturbed versions of the prior and likelihood distributions.

For the GAM models, the fast method used by priorsense did not work well, and the prior sensitivity analysis was made simply by re-running the inference with different prior choices.

The list of initial and final priors used for Bayesian logistic regression and Bayesian GAM can be found in Table 2.

5.1.4. GAM Basis Dimension

The flexible functions we chose in Equation 3 for Bayesian GAM uses an eigendecomposed approximation of a thin-plate regression spline [51].

The approximation of the spline uses eigendecomposition to extract the first k eigenvectors of the original piecewise function. From these, the algorithm builds a reconstruction that preserves the essence of the original function while reducing its computational complexity. The number of eigenvectors k in this process is called 'the basis dimension' or 'the dimension of the basis', and reducing it indirectly affects the complexity of the approximated spline.

By default, in brms, the R package we used, the basis dimension is automatically adjusted. However, the user can set an upper bound on it through an argument of the smoothing function. We analysed the sensitivity of the results with regard to this argument. The initial and final values of k are shown with the priors in Table 2.

5.2. Evaluation methods

All Bayesian predictions within this study are made with 500 posterior draws. The point estimates are the medians of the predictions, and they are reported with their interquartile ranges (IQRs), or the width of the 50% credible intervals.

The models along with their design choices are evaluated quantitatively with numerical scores and calibration plots, and qualitatively through predictions mapped on geographic locations on QGIS [52]. For numerical scores, we present area under the curve (AUC) [53] and maximum attainable true skill statistic (TSS) [54].

For numerical scores, we examined both training scores and validation scores to diagnose signs of overfitting. The training scores were calculated from the models fit on all 7,331 grid cells. Their outputs for the area shown in Figure 1a were compared to their true labels to compute the values. The validation scores of the models were calculated through a 10-fold CV using one fold pattern of spatial-CV feature selection, shown in Figure 2. The scores are the average of the 10 validation folds.

Note that, because a supervised feature selection procedure was utilized, final model assessments not calculated via a nested approach (for example, one in which the CV-based FFS process is replicated within an outer CV or hold-out procedure) may overestimate model performance [55, 56]. However, nested CV experiments for *E. maximus* indicate that, given the size of the training dataset and predictive strength of climatic features, this bias is very slight: over 100 experimental iterations, AUC values were overestimated by a mean of less than 0.0002. Thus, to consistently estimate and compare the performance of Bayesian and ML modeling approaches, and to assess all feature sets via spatial CV (which accounts for the spatial autocorrelation present in environmental datasets [49]), post-selection, spatial CV-based AUC and TSS estimates are used in the following discussion.

The calibration plots, reported along with the visual presentations of model predictions, diagnose whether the outputs of a model are well calibrated probabilities. A calibration plot is created by first dividing the predictions into a user-defined number of bins according to their predicted values. Then the proportion of positive labels within each bin is calculated and plotted as dots or a regression line. If the proportion of positives in the bins are roughly equal to their predicted probabilities, the plot is lined up with the line $y = x$, indicating that the model is calibrated and outputs can be interpreted as probabilities.

Table 1: The list of features. We utilize two feature sets selected through different CV processes in a previous study [42], which we name the 'random-CV' feature set and the 'spatial-CV' feature set, respectively. The two feature sets share six common features.

Feature set	Feature Name	Source	Definition
random-CV	BIO03	WorldClim	Isothermality: Mean of monthly (maximum - minimum temperature) divided by the annual temperature range.
	TN10P	ETCCDI	Proportion of days when the minimum temperature is lower than the 10th percentile of historical data.
	GSL	ETCCDI	Growing season length.
	TNX	ETCCDI	Maximum daily minimum temperature.
spatial-CV	BIO08	WorldClim	Mean temperature of the wettest quarter.
	TXX	ETCCDI	Maximum daily maximum temperature.
	BIO02	WorldClim	Mean diurnal range: Mean of monthly difference between maximum and minimum temperature.
	TN90P	ETCCDI	Proportion of days when the minimum temperature is higher than the 90th percentile of historical data.
common for both sets	ID	ETCCDI	Icing days.
	BIO14	WorldClim	Precipitation of the driest month.
	BIO18	WorldClim	Precipitation of the warmest quarter.
	CWD	ETCCDI	Maximum length of wet spell.
	RX1DAY	ETCCDI	Monthly maximum one-day precipitation.
	WSDI	ETCCDI	Warm spell duration index: Count of days in year with at least six consecutive days when the maximum temperature is larger than the 90th percentile of historical data.

Table 2: Initial and adjusted prior and basis dimension (for GAM) settings for Bayesian models. Settings that are not changed from the initial set are denoted by '-'. The second parameter in the normal distributions indicate the standard deviation. The flat prior used for GAMs is an uniform distribution that gives a small amount of probability density to all values from minus infinity to positive infinity. The default basis dimension value, -1, allows the spline algorithm to find the optimal basis dimension in brms. When setting it to 1, it forces the splines to use the least possible basis dimension value. The scales of the priors for the coefficients and intercept for Bayesian GAM did not affect the prediction results once the non-linearity prior and basis dimension were restricted, hence we show both flat and normal priors as our adjusted priors in the final column.

Model type	Feature set	Scaling	Setting type		Initial	After adjustment
logistic regression	random-CV	yes	priors	coefficients	Normal(0,5)	Normal(0,10)
				intercept	Normal(0,10)	Normal(0,20)
		no	priors	coefficients	Normal(0,5)	Normal(0,10)
				intercept	Normal(0,10)	Normal(0,20)
	spatial-CV	yes	priors	coefficients	Normal(0,5)	Normal(0,10)
				intercept	Normal(0,10)	Normal(0,20)
		no	priors	coefficients	Normal(0,5)	-
				intercept	Normal(0,10)	-
GAM	all feature sets and scaling		priors	coefficients	Normal(0,5)	Normal(0,5) or flat
				non-linearity	Student-t(3,0,2.5)	Normal(0,1)
				intercept	Normal(0,10)	Normal(0,10) or flat
			argument	basis dimension (k)	-1 (default)	1 (least possible)

6. Results

In the main section, we show the results for the models trained on scaled features (for both random-CV and spatial-CV feature sets), with the adjusted prior and non-linearity settings for the Bayesian models. We present the events that occurred during the iterative modeling process for the Bayesian models in Appendix A. For a more casual and visual alternative to explore our results, we refer readers to our visual summary of experimental designs reported in an online supplement[1].

6.1. The linear models - Bayesian and MLE logistic regression

The numerical scores and visualizations for these models are presented in Figures 3 and B.11. The visualizations for MLE logistic regression are presented in Appendix B instead of in this section since they were almost identical to Bayesian logistic regression.

6.1.1. Bayesian logistic regression

The models trained on the two feature sets gave similar predictions for present-day climate conditions but differed greatly for the future conditions. The models fit on the random-CV features predicted a very generous future suitable habitat area for Asian elephants, including areas such as the Sahara desert and the southern parts of Alaska and the Nordic countries. The IQRs were only wider around the areas where the point predictions did not have clear presences or absences. The models fit on spatial-CV features had more conservative point predictions but with wider IQRs even in areas with strong presences or absences.

6.1.2. MLE logistic regression

MLE logistic regression gave results that are identical to Bayesian logistic regression save for slight details in the QGIS visualizations and calibration plots. In hindsight, this is unsurprising as the number of observations is large compared to the number of parameters and weak priors were used for Bayesian models. However, since this was an methodological research, we could not foresee that the model settings for the Bayesian logistic regression would be this similar to its non-Bayesian counterpart. We omit the results of MLE logistic regression from the main section of this article, but for full transparency, we present them in Figure B.11.

6.2. The non-linear models - Bayesian GAM and random forest

The numerical scores and visualizations for these models are presented in Figure 4.

6.2.1. Bayesian GAM

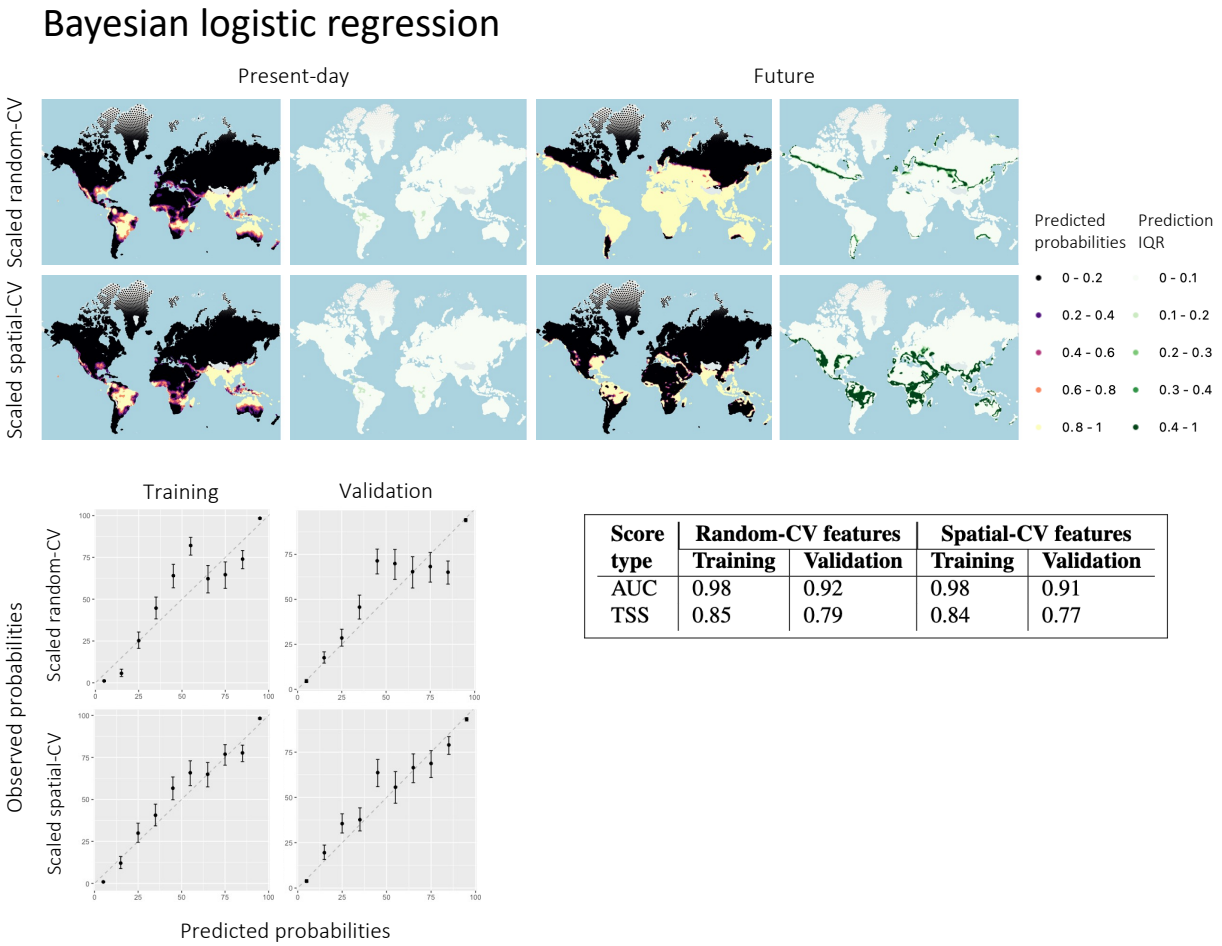
The final prior and basis dimension settings for Bayesian GAMs (in the column 'After adjustment' on Table 2) placed strong restrictions on the models' non-linearity. These we set in response to the overfit results we got from the initial settings, which are described in Appendix A.

The final settings restricted the non-linearity enough to make the models fit on random-CV features resemble the results from Bayesian and MLE logistic regression. The models fit on spatial-CV features, however, still gave very unrealistic predictions for the future, in which every area on Earth is a suitable habitat for *E. maximus*. It seemed that restricting the non-linearity of the model has only a limited effect that varied depending on the feature set used to fit the models.

6.2.2. Random Forest

When visualized on a world map, the outputs of the random forest models did not seem to indicate anything out of the ordinary. However, their calibration plots were distinctly different from the other models. The training calibrations implied that the models had found parameters that could almost completely separate the presences from the absences. This was an indication of severe overfitting, even though the numerical scores and the calibration plots of validation folds seemed to suggest that the symptom was mild. In fact, this gave us insight to what may have gone wrong with Bayesian GAM. It seemed that random forest and Bayesian GAM were too non-linear for the patterns in the data, and caused a problem called 'complete separation', described later in Section 7.2.

[1] `https://miro.com/app/board/uXjVOX_Zhf8=/?share_link_id=937959545296`



Score type	Random-CV features		Spatial-CV features	
	Training	Validation	Training	Validation
AUC	0.98	0.92	0.98	0.91
TSS	0.85	0.79	0.84	0.77

Figure 3: The prediction outputs, calibration plots, and numerical scores for the Bayesian logistic regression models with scaled features. The models were fit using the adjusted priors in Table 2. The results for MLE logistic regression are presented in Figure B.11 instead of this section since they were nearly identical to Bayesian logistic regression.

7. Discussion

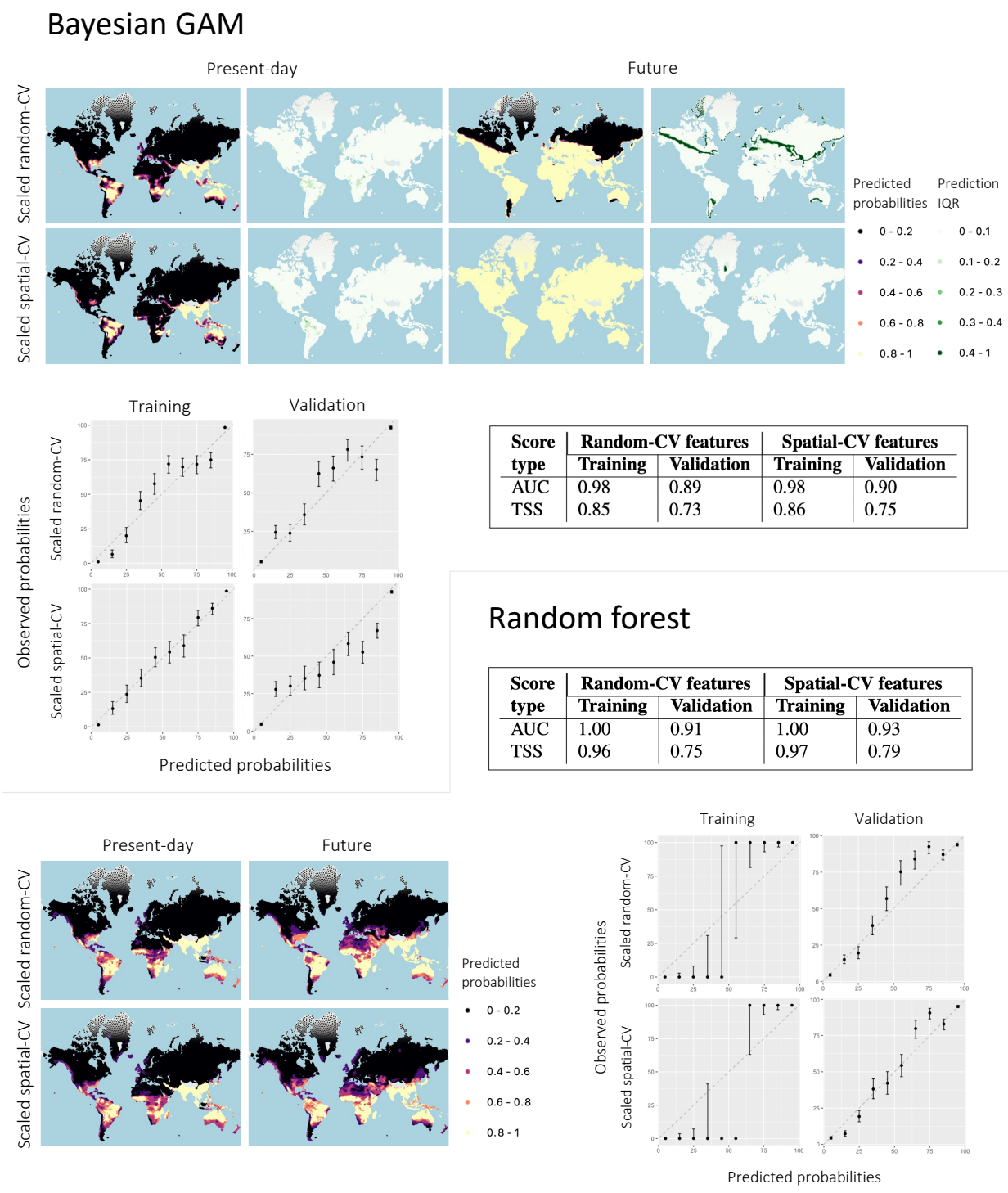
7.1. Modeling Decisions and Outcomes

This research iteratively assessed four model types, MLE logistic regression, random forest, Bayesian logistic regression, and Bayesian GAM with combinations of different modeling options.

MLE and Bayesian logistic regression had the simplest and stiffest structure, hence they did not show signs of overfitting. Rather, the calibration plots in Figures 3 and B.11 suggested that models fit on random-CV features underfit the true pattern, since the predictions undershoot the actual observed frequencies around the 50% bin and then overshoot them around the 80% bin.

Feature selection had an effect on the future predictions of both Bayesian and MLE logistic regression. Models fit on random-CV features gave a very optimistic outlook, while the models fit on spatial-CV features had a more conservative view. Also, though not presented in the main part of this article, we noticed that feature scaling affected Bayesian logistic regression by changing the relative relationship of the data against the priors. This is discussed more in Appendix A.

For random forest, a complex non-linear model, the calibration plots imply that the models were overfit to the extent where they could almost completely separate the presences from the absences. The other evaluation metrics, the numerical scores and the calibration plots for the validation folds, showed only slight symptoms of this problem.



Score type	Random-CV features		Spatial-CV features	
	Training	Validation	Training	Validation
AUC	0.98	0.89	0.98	0.90
TSS	0.85	0.73	0.86	0.75

Score type	Random-CV features		Spatial-CV features	
	Training	Validation	Training	Validation
AUC	1.00	0.91	1.00	0.93
TSS	0.96	0.75	0.97	0.79

Figure 4: The prediction outputs, calibration plots, and numerical scores for the Bayesian GAM models and random forest models with scaled features. The Bayesian models were fit using the adjusted priors and basis dimensions in Table 2.

Although not as drastic as the other model types, feature selection did have a small effect on the random forest predictions. This can be observed in South America, Africa, and the Oceania region in Figure 4. Also, like Bayesian and MLE logistic regression, feature scaling did not have an influence on the random forest models.

Bayesian GAM initially showed signs of extreme overfitting, which are discussed in Appendix A. Attempts to restrict the non-linearity made the outputs closer to Bayesian logistic regression, but only for random-CV features. Aside from feature selection, the effects of modeling decisions were either unclear or unpredictable for Bayesian GAM. Feature scaling did not have a large effect, though it did slightly alter the predictions for some prior and basis dimension combinations. And as we observed for models fit on spatial-CV features, restricting the non-linearity priors and basis dimension did not necessarily make all predictions closer to Bayesian logistic regression models. The priors for the coefficients and intercept did have an effect for some prior and basis combinations (not shown within this report), but once the non-linearity was restricted, did not influence the model. This is the reason why the adjusted settings for Bayesian GAM in Table 2 include both normal and flat priors - the scale of the priors for those parameters did not make a difference.

7.2. Analysis of anomalies in model predictions

By visualizing the conditional effects of individual features with brms's `conditional_effects` function, we discovered three patterns that most likely caused the models to misbehave. At least one pattern seemed to apply for all models, but the more unreliable models seemed to have frequent occurrences of these patterns for multiple features. We illustrate all patterns using examples for two features in the Bayesian GAM model fit on raw random-CV features with initial settings.

The first pattern is the high non-linearity and uncertainty in the conditional effects, shown in Figure 5a.

The second pattern can be observed in Figures 5b, 6, and the left part of Figure 5a. There is a challenge that there are regions in the feature space where there is severe class imbalance (only or almost only observed 'habitat unsuitable'). This causes the data to be weakly informative in a large part of the interesting sections of the feature space, and thus learning non-linearities is very difficult.

A further challenge is visualized in Figure 7, where the feature values of present-day and future Greenland lie outside of the observed data range. Extrapolation far away from the observed data with flexible non-linear models or simple latent linear models is likely to produce unrealistic predictions.

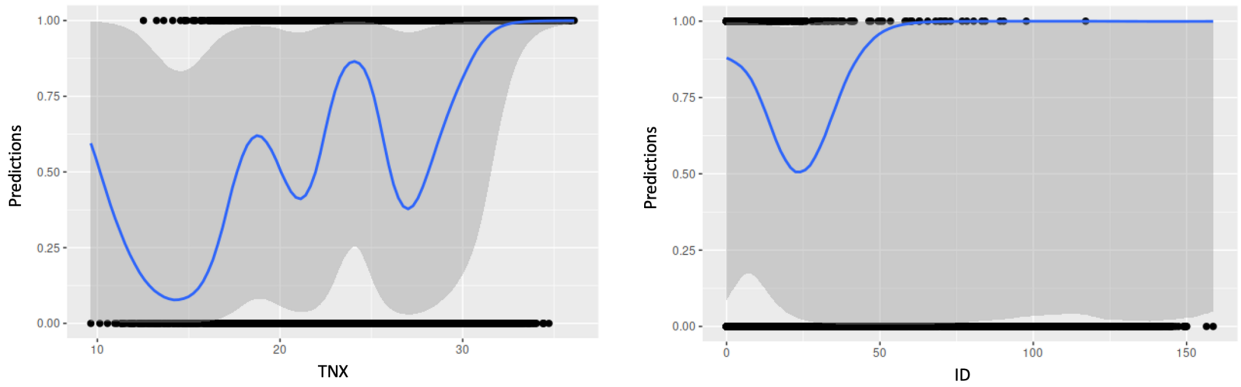
The first and second pattern and the random forest calibrations in Figure 4 may indicate a problem called 'complete separation', where the model finds a parameter combination that can perfectly separate the presences and the absences. These symptoms of complete separation appear to explain the conditional effects of TNX following a seemingly nonexistent pattern in Figure 5a, the model ignoring the effects of ID in Figure 5b, and additionally, the unusually large and variable posteriors for the Bayesian GAM models we present in Figure A.10.

Unlike the other two patterns, the third pattern seems primarily a consequence of extrapolation, though it may have been partially caused by the second pattern (icing days not reflected in the predictions). Since the models can only fit trends within the training data's range, they cannot guarantee their performance when predicting novel ranges. Therefore, the two features in Figure 7 show unreliable predictions for Greenland.

These plots are only one or two-dimensional snapshots, so they cannot express the full effects of the ten features used to create the models. However, they give us insight to what could have led to unexpected prediction results and how to avoid them. Supposing that they were indeed caused by complete separation, the first and second pattern of misbehavior can be prevented by using tighter informative priors that can restrict the model from forming unreliable posteriors. Also, since some of our variables seem to have a monotonic or unimodal effect on habitat suitability, placing a monotonicity constraint where applicable may be effective in suppressing the non-linear effects. Additionally, the third pattern might be mitigated by enhancing the training data with more examples of absences from climates that are physiologically not acceptable for Asian elephants, or by adding ranges of other proboscidean species as presences.

7.3. The most reliable model

Our assessment of the model performance also included qualitative evaluation of predicted habitats in relation to the known ecology of Elephantidae today as well as their ranges during the last 2.5 million years. In the context of this external knowledge, the most reliable model appeared to be the Bayesian logistic regression model fit on scaled spatial-CV features with adjusted priors. The outputs of this model are shown in Figure 3, in the bottom row of the world map visualizations. This model had fairly good validation scores with the most calibrated predictions. Additionally, the scaled features



(a) The conditional effect of TNX. The esimated effect is highly non-linear with high uncertainty, and it does not model well the data with small TNX values.

(b) The conditional effect of ID. The estimated effect has very high uncertainty, and does not follow the trend the ID values indicate.

Figure 5: The conditional effects of features TNX (maximum daily minimum temperature) and ID (icing days) on habitat suitability for the Bayesian GAM model fit on raw random-CV features with initial priors and basis dimension settings. The other features are held constant at their means. The blue line is the mean conditional effect, the shaded area is the 95% credible interval, and the black dots plotted on the top and bottom are training data points.

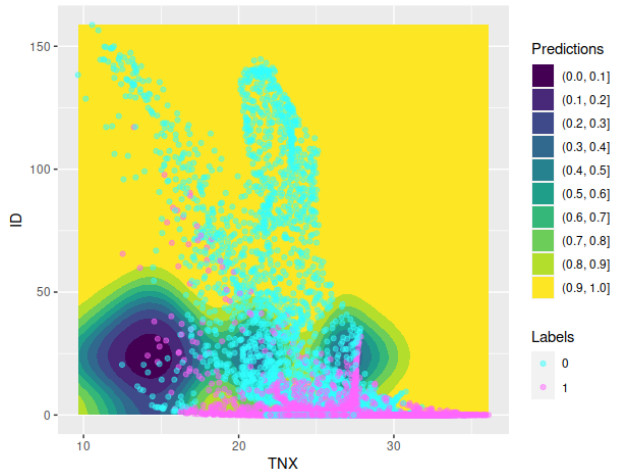


Figure 6: The interaction effect of TNX and ID on habitat suitability from the same model as Figures 5a and 5b. The other features are held constant at their mean values. The overlaid dots are the data points colored by their labels (0: habitat unsuitable, 1: habitat suitable).

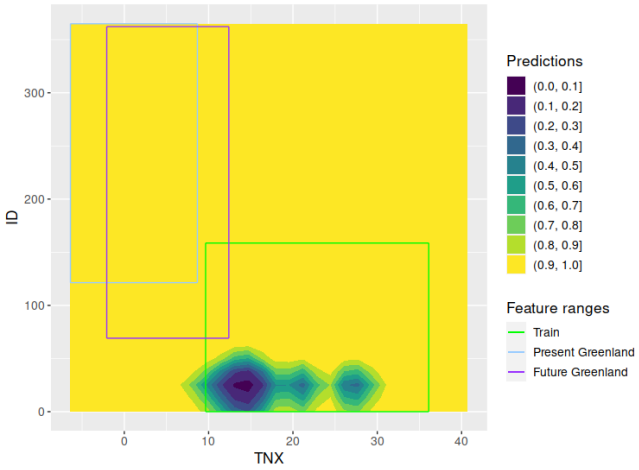


Figure 7: The same interaction effect as Figure 6 visualized on a range wider than that of the training data. The hollow rectangles overlaid on the plot are ranges of the data points that belong to the training data, present-day Greenland, and future Greenland.

and the simple model structure allow straightforward interpretations of posteriors. Above all, this model had the most convincing future predictions, as we discuss in Section 7.4.

Since the Bayesian logistic regression had the best predictions, one could argue that MLE logistic regression could have the same predictions while requiring less effort, in terms of planning and setting appropriate priors, to fit. However, though not demonstrated within this study, Bayesian models can incorporate knowledge through priors, which are powerful when predicting for areas with climates that are not observed within the range of the training data.

7.4. Ecological interpretations of the results

The goal of our case study was to model the suitable climatic range of the Asian elephant, *Elephas maximus*, today and in the near future to guide possibilities to introduce this species outside its current, heav-

ily reduced range, including areas beyond its historical range, for trophic rewilding purposes. Introducing *Elephas maximus* to environments such as the cerrado in South America [14], Australian woodlands [16] and even European temperate forests [15], has been suggested as a replacement of ecologically equivalent Pleistocene megafauna species in order to restore top-down trophic cascades in terrestrial ecosystems [4]. Such plans should, however, be made with caution, accounting for further factors beyond the suitable climatic or habitat range of the species, as discussed below.

The historical range of *Elephas maximus* from the Late Pleistocene until recent demonstrates that this species ranged from tropical to subtropical or warm-temperate climates and avoided arid climatic conditions. Current habitats of *E. maximus* range from rainforests to seasonally dry woodlands in India, Sri Lanka, Southern China, the Malayan Peninsula, Borneo and Sumatra. Despite dental adaptations that suggest the original adaptation of the genus *Elephas* to grazing in open grassland savannas during the Plio-Pleistocene in Africa, such as high hypsodonty and lamellar count (e.g. [18, 21, 22, 23]), the extant species *E. maximus* occupies mostly dry-season woodland environments and has browsing to mixed-feeding diet [26]. The currently suitable climatic range of *E. maximus* predicted from the scaled spatial-CV Bayesian logistic regression model, which we concluded as the 'most relisable model' in Section 7.3, corresponds with the distribution of biomes that are broadly similar to the woodland and forest habitats of this species in Asia. The predicted currently suitable geographic areas outside the historical range are located in Northern South America, South-Eastern Africa, Madagascar (except the arid southernmost part), most of the Southeast Asian archipelago, Northern Australia, and parts of Western Africa. In East Africa, the predicted suitable range corresponds with the distribution of miombo-woodlands and other types of tropical woodland, typically with grassy undergrowth. In South America the predicted suitable area overlaps with large parts of the Amazon and Atlantic rainforests, but most importantly it covers the entire cerrado woodland or savanna biome in central Brazil. In Australia, the predicted range covers the northern part of the continent, where the environments are predominantly tropical to subtropical woodlands and savannas, with some more open and dry vegetation and moist broadleaf forest. Marginally suitable areas according to this model include the Mediterranean coast, the fertile crescent in Western Asia, and parts of Central Africa and Southern (especially South-Eastern) North America.

The future prediction of suitable range for *E. maximus* based on the scaled spatial-CV Bayesian logistic regresssion model expands into covering some further areas, the most notable of which are the Northern and Eastern Mediterranean coast (except the Iberian Peninsula), the fertile crescent and most of Eastern North America, which is currently characterized by temperate to subtropical mixed forest environments. The predicted suitable area in southern Alaska in the future seems unrealistic. The natural range of *Elephas maximus* has never extended to the American continents, but until the end-Pleistocene mass extinction, other species of proboscideans were widespread there, including the mixed-feeding to grazing *Mammuthus columbi* in North America and the similarly mixed-feeding last gomphothere genera *Cuvieronius* and *Notiomastodon* in South America. Similarly, Southern Europe is outside of the historical range of *E. maximus*, but during the Pleistocene, the large, mixed feeding elephant *Palaeoloxodon antiquus* was present there and has been considered by some [15] broadly equivalent in terms of ecology. On the other hand, the fertile crescent represents the westernmost extent of the historical range of *E. maximus*. The extinction of *E. maximus* from this area during the Holocene has been attributed to periods of increased aridity, while suitable conditions for the species occurred during more humid climatic events [25]. It is thus plausible that with increased humidity the fertile crescent could support a population of *E. maximus* in the future, unless other factors such as habitat loss and other human influence prevent it.

7.5. Potential implications for conservation

It is important to note that these predictions are entirely based on climatic variables and they do not take into account other ecologically important factors. Such factors could include competition with other herbivorous mammal species, unpredictable unsuitability of resources (e.g., due to properties of plant communities, such as plant defense mechanisms) outside the natural range of *E. maximus*, and challenges due to land use by humans.

As the largest existing megaherbivores, elephants are not likely to be suppressed by competition with other large herbivorous mammals, but they may interact with and affect other species in ways that may be hard to predict. Studies of interactions between cattle, zebras and African savanna elephants (*Loxodonta africana*) have indicated that while there is strong competition between cattle and zebras over grass resources, and both species also interact with elephants, the presence of elephants can in fact mitigate such competitive relationships by

affecting the distribution of plant resources in the environment [57]. Such observations support the positive role of elephants in restoring trophic cascades in ecosystems, but they have not been extensively studied for *E. maximus*. In Africa, potentially competitive interaction of *E. maximus* with the African elephant species, especially the savanna elephant (*L. africana*), might be expected, and this would have to be taken into account in any rewilding effort involving *E. maximus* there. Both *E. maximus* and *L. africana* are mixed-feeders, and their dietary composition varies following resource availability both seasonally and in different environments [21, 26, 58]. However, of these species, *L. africana* occupies a wider range of habitats ranging from forests to deserts, and it can have a stronger effect on vegetation structure depending on other factors such as rainfall, although negative effects of elephants on their environment have been exaggerated [59]. *E. maximus* has not been shown to significantly affect vegetation structure in its habitats, except in artificially high population densities [26, 60]. The outcome of potential interaction between *E. maximus* and *L. africana* is difficult to predict, and introducing *E. maximus* to areas occupied by *L. africana* seems potentially risky for both species. Fossil record shows that in the past, sympatric occurrence of several proboscidean species has not been uncommon, but in such cases, the species usually show clearly different adaptations and evidence of niche partitioning, especially in diet [61].

E. maximus shows notable plasticity in its diet and ability to change dietary composition according to locally available resources, which suggests that it would probably be able to survive in suitable woodland environments outside its original range, although interactions between elephants and plant communities could to some extent be difficult to predict due to unknown factors. Perhaps the largest challenges for introducing *E. maximus* beyond its current range come from human interference. Even if suitable areas for *E. maximus* can be predicted to occur in the future for example in Eastern US and Southern Europe, introducing this species would be practically impossible across most of those areas because of heavy land use and other human effects. Direct conflicts between *E. maximus* and humans occur across the natural range of the species, especially due to agricultural damages caused by the elephants [26]. Poaching is a further risk for the conservation of *E. maximus*, and it should be taken into account in rewilding efforts. Due to such factors, introducing *E. maximus* to new ecosystems for rewilding purposes would have to begin experimentally in relatively small and monitored rewilding areas.

8. Conclusion

This research presented the results for two Bayesian models compared to two baseline non-Bayesian models for a hypothetical rewilding case. While exploring different models, we tested different modeling options that may affect the prediction outcome. Each modeling decision had an effect on at least one model, though the magnitude of their effects were influenced by the type of model and the modeling options.

The more complex Bayesian model, Bayesian GAM, showed symptoms of misbehavior that is likely caused by almost complete separation. The same problem was also observed for random forest. Bayesian and MLE logistic regression, being simpler, did not exhibit this problem but did give unlikely predictions depending on the features selected for modeling. All models had at least some issues when extrapolating or predicting far outside the range of the training data. Our analysis of anomalies showed that those deviations were largely related to non-analog or unique conditions present in the modelling dataset. This resonates with the widely acknowledged challenge of modelling potentially to non-analog ecosystems [62] from limited observational data.

The purpose of fitting the models was to identify a model that provides convincing predictions from the climatic perspective to inform about potential rewilding sites for Asian elephants that are likely to remain suitable up until the year 2070. After quantitative evaluation and visual inspection of the outputs, we concluded that the Bayesian logistic regression fit on scaled spatial-CV features with adjusted, wider priors is the most reliable model given the data, our iterative modeling process, and historical habitats of proboscidean species (outputs shown in Figure 3). The present-day and future outputs indicated that large candidate areas for rewilding include the north half of South America, coastal regions of east and west Africa, and the northern coastline of Australia. It also suggests that reintroducing Asian elephants to areas in India and southeastern China may be feasible as well. Additionally, the predictions implied that some areas that were formerly suitable habitats may become climatically unsuitable in the RCP 8.5 scenario.

The modelling results suggest suitable areas for *Elephas maximus* across its historical range as well as climatically suitable areas outside its historical range, which tend to correspond with the distribution of woodland biomes broadly similar to those in the species' present and historical range. However, it is important to note that such predictions of habitat suitability are purely based on climatic variables. Other ecologically

important factors and practical restrictions (for example due to interactions with other species and perhaps most importantly due to loss of suitable habitats as well as other human interference) should be carefully considered before any rewilding attempts are made.

Also, though we experimented with Bayesian models, we did not incorporate expert opinions as informative priors, which could have prevented complete separation and refined the predictions. Furthermore, many of the effects we observed in Section 7.2 can be assumed to have either monotonic or unimodal functional shape. Such models are more difficult to implement and thus were not considered in this work, but may be a possible area to explore. Therefore, enhancing the dataset, exploring informative priors, and restraining nonlinear effects with monotonicity constraints would be good topics to address in future research.

Additionally, we would like to add that even though we concluded that Bayesian logistic regression is the most suitable model for this project, this will not always be the case for other settings. Had we used different climatic variables as input or added additional variables, we might have chosen one of the non-linear models for use in prediction. In a methodological study like this, one should evaluate all models available with different configurations to choose the final, best-suited model.

9. Data and code availability

The data and the codes created in the research are available at `https://github.com/RyokoNod/sdm-asian-elephants`. Additionally, for informally exploring all outputs of models, an interactive mind map is available at `https://miro.com/app/board/uXjVOX_Zhf8=/?share_link_id=937959545296`.

10. Author contributions

Ryoko Noda wrote the code, designed the model exploration, and created the original draft. Michael Mechenich analyzed the prediction results from an early stage with Ryoko and performed writing, reviewing, and editing. He also curated and produced the data sets used in this research. Juha Saarinen performed writing and reviewed the article for possible conflicts with historical proboscidean ranges. Aki Vehtari reviewed the article for flaws in statistical theory and edited wording. Indrė Žliobaitė performed writing and editing and supervised the creation of this study. All the authors analyzed and interpreted the final results.

11. Funding

Research leading to these results was partially funded by the Research Council of Finland (grant no. 314803 to IZ). JS was funded by the Research Council of Finland during this work (grant nr. 340775, NEPA - Non-analogue ecosystems in the past).

12. Declaration of Generative AI and AI-assisted technologies in the writing process

The authors did not use any form of generative AI during the preparation of this work. The writing and editing process was done by human hand.

Appendix A. The Bayesian iterative modeling process

Appendix A.1. Bayesian logistic regression

While we were examining models fit on different features (the random-CV feature set and the spatial-CV feature set, raw and scaled), we noticed that scaling had a small effect on the models trained on spatial-CV features. As shown in Figure A.8, the model fit on the scaled features predicted slightly pessimistic future habitat suitability compared to the model fit on raw features when using the initial prior set.

This effect of scaling implied that the initial priors might be overpowering the features, so we used the priorsense package to test and adjust the priors to less restrictive distributions. After adjustment, scaling spatial-CV features still had a small effect on the prediction outcome, but the predictions from the model fit on scaled features was not as pessimistic.

Appendix A.2. Bayesian GAM

With our initial settings, every result from the Bayesian GAM models showed symptoms of severe overfitting. Some examples of these prediction outputs are presented in Figure A.9.

Examining the posteriors (example in Figure A.10), we observed that the non-linearity of the functions in Equation 3 was too excessive, so we explored two methods to restrict it: selecting a tighter non-linearity prior and restricting the basis dimension. However, this did not work as expected. Adjusting each non-linearity setting individually only resulted in more unconvincing predictions. Adjusting both settings at the same time could restrict the model enough so that it resembled outputs from Bayesian logistic regression, but only for the random-CV feature set. These final results are shown in the top half of Figure 4.

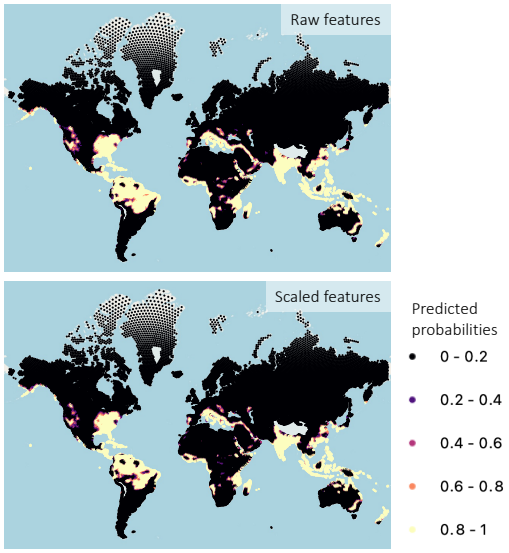


Figure A.8: The future predictions for Bayesian logistic regression models trained on raw and scaled spatial-CV features when using the initial prior settings in Table 2. The model fit on scaled features has a slightly pessimistic prediction output, especially visible in South America.

Appendix B. Results for MLE logistic regression

We have omitted these results from the main section since they were nearly identical to Bayesian logistic regression, but we present it here in the Appendix as Figure B.11 for transparency.

References

[1] J. Donlan, Re-wilding north america, Nature 436 (7053) (2005) 913–914. doi:10.1038/436913a.

[2] C. Josh Donlan, J. Berger, C. E. Bock, J. H. Bock, D. A. Burney, J. A. Estes, D. Foreman, P. S. Martin, G. W. Roemer, F. A. Smith, et al., Pleistocene rewilding: An optimistic agenda for twenty-first century conservation, The American Naturalist 168 (5) (2006) 660–681. doi:10.1086/508027.

[3] J. Lorimer, C. Driessen, Wild experiments at the Oostvaardersplassen: Rethinking environmentalism in the Anthropocene, Transactions of the Institute of British Geographers 39 (2) (2014) 169–332. doi:10.1111/tran.12030.

[4] J.-C. Svenning, P. B. M. Pedersen, C. J. Donlan, R. Ejrns, S. Faurby, M. Galetti, D. M. Hansen, B. Sandel, C. J. Sandom, J. W. Terborgh, F. W. M. Vera, Science for a wilder Anthropocene: Synthesis and future directions for trophic rewilding research, Proceedings of the National Academy of Sciences 113 (4) (2016) 898–906. doi:10.1073/pnas.1502556112.

[5] R. L. Beschta, W. J. Ripple, Riparian vegetation recovery in yellowstone: The first two decades after wolf reintroduction, Biological Conservation 198 (2016) 93–103. doi:10.1016/j.biocon.2016.03.031.

[6] J. Elith, J. R. Leathwick, Species distribution models: Ecological explanation and prediction across space and time, Annual Review of Ecology, Evolution, and Systematics 40 (2009) 677–697. doi:10.1146/annurev.ecolsys.110308.120159.

[7] S. Jarvie, J.-C. Svenning, Using species distribution modelling to determine opportunities for trophic rewilding under future scenarios of climate change, Phil. Trans. R. Soc. B 373 (2018) 20170446.

[8] M. M. Barlow, C. N. Johnson, M. C. McDowell, M. W. Fielding, R. J. Amin, R. Brewster, Species distribution models for conservation: Identifying translocation sites for eastern quolls under climate change, Global Ecology and Conservation 29 (2021) e01735.

[9] A. C. Eyre, N. J. Briscoe, D. K. P. Harley, L. F. Lumsden, L. B. McComb, P. E. Lentini, Using species distribution models and decision tools to direct surveys and identify potential translocation sites for a critically endangered species, Diversity and Distributions 28 (2022) 700–711.

[10] A. Catalina, P.-C. Bürkner, A. Vehtari, Projection predictive inference for generalized linear and additive multilevel models, in: G. Camps-Valls, F. J. R. Ruiz, I. Valera (Eds.), Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, Vol. 151 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 4446–4461.
URL `https://proceedings.mlr.press/v151/catalina22a.html`

[11] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Brkner, M. Modrk, Bayesian workflow (2020). doi:10.48550/ARXIV.2011.01808.
URL `https://arxiv.org/abs/2011.01808`

[12] The Intergovernmental Panel on Climate Change. [link].
URL `https://www.ipcc.ch/`

[13] J. A. Estes, J. Terborgh, J. S. Brashares, M. E. Power, J. Berger, W. J. Bond, S. R. Carpenter, T. E. Essington, R. D. Holt, J. B. C. Jackson, R. J. Marquis, L. Oksanen, T. Oksanen, R. T. Paine, E. K. Pikitch, W. J. Ripple, S. A. Sandin, M. Scheffer, T. W. Schoener, J. B. Shurin, A. R. E. Sinclair, M. E. Soul, R. Virtanen, D. A. Wardle, Trophic downgrading of planet Earth, Science 333 (6040) (2011) 301–306. doi:10.1126/science.1205106.

[14] M. Galetti, Parks of the Pleistocene: Recreating the Cerrado and the Pantanal with megafauna, Natureza & Conservação 2 (1) (2004) 95–101.

[15] J.-C. Svenning, "Pleistocene re-wilding" merits serious consideration also outside North America, IBS Newsletter 5 (3) (2007) 3–10.

[16] D. Bowman, Bring elephants to Australia?, Nature 482 (7383) (2012) 30. doi:10.1038/482030a.

[17] J. Louys, R. T. Corlett, G. J. Price, S. Hawkins, P. J. Piper, Rewilding the tropics, and other conservation translocations strategies in the tropical Asia-Pacific region, Ecology and Evolution 4 (22) (2014) 4380–4398. doi:10.1002/ece3.1287.

[18] W. J. Sanders, E. Gheerbrant, J. M. Harris, H. Saegusa, C. Delmer, Proboscidea, University of California Press, 2010, pp. 124–146.

[19] H. Zhang, Evolution and systematics of the elephantidae (mammalia, proboscidea) from the late miocene to recent, Ph.D. thesis (07 2020).

[20] A. Larramendi, H. Zhang, M. R. Palombo, M. P. Ferretti, The evolution of palaeoloxodon skull structure: Disentangling phylogenetic, sexually dimorphic, ontogenetic, and allometric morphological signals, Quaternary Science Reviews 229 (2020) 106090. doi:10.1016/j.quascirev.2019.106090.

[21] T. E. Cerling, J. M. Harris, M. G. Leakey, Browsing and grazing in elephants: The isotope record of modern and fossil proboscideans, Oecologia 120 (3) (1999) 364–374.

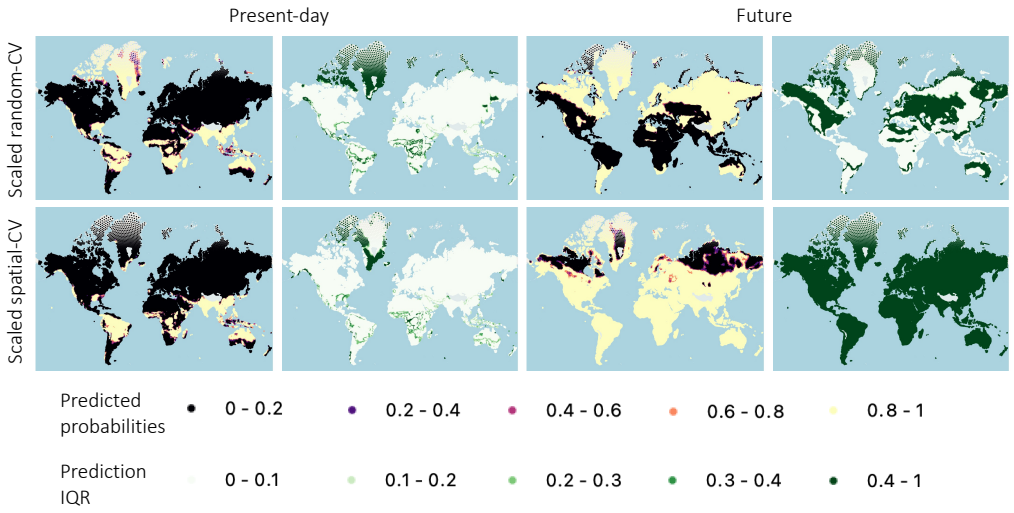


Figure A.9: Some examples of the overfit predictions from the Bayesian GAM models using the initial prior and basis dimension settings.

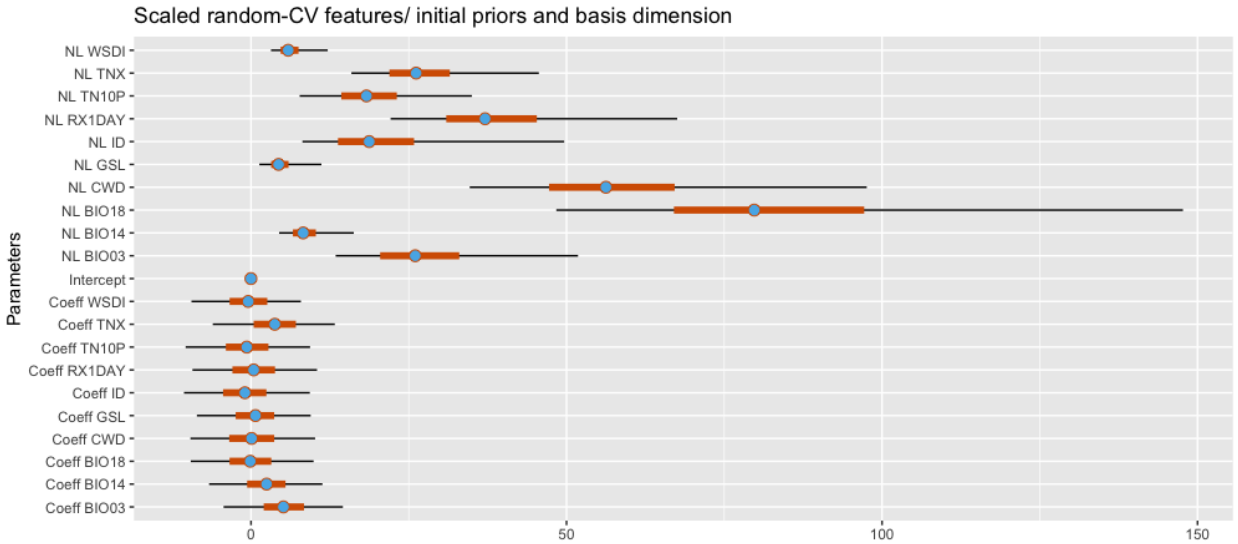
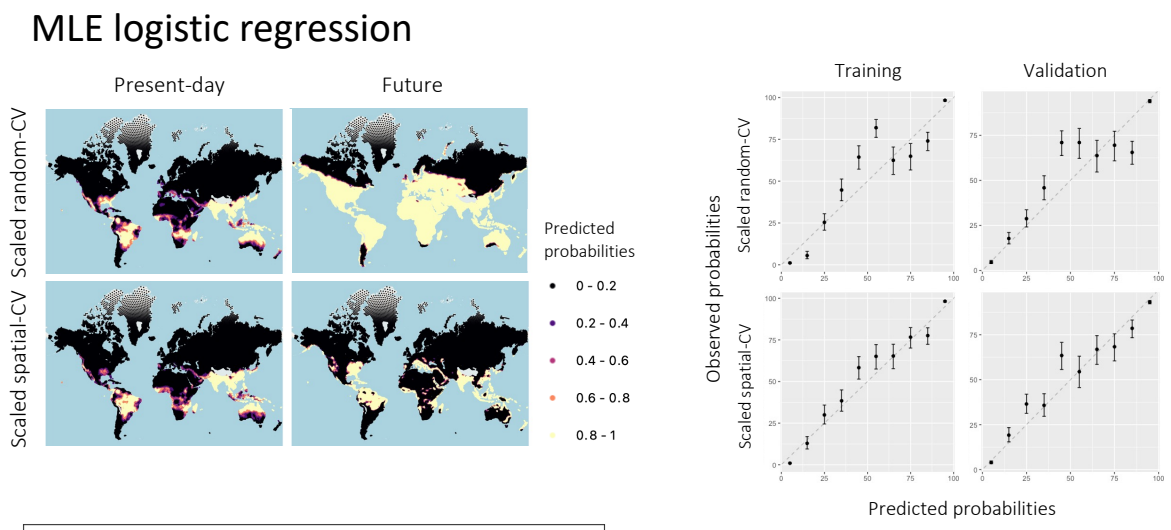


Figure A.10: An example of the posteriors we observed for the initial prior and basis dimension settings for Bayesian GAM. This particular one corresponds to the model fit on scaled random-CV features (the top row of Figure A.9). The distributions starting with the abbreviation 'NL' are posteriors that represent the non-linearity of the splines in Equation 3. The other distributions are posteriors for the coefficients and intercept of the same equation. We observed from this plot that the posterior for the non-linearity is too excessive.



Score type	Random-CV features		Spatial-CV features	
	Training	Validation	Training	Validation
AUC	0.98	0.92	0.98	0.91
TSS	0.85	0.79	0.84	0.77

Figure B.11: The prediction outputs, calibration plots, and numerical scores for MLE logistic regression models with scaled features. This is the baseline model for the results presented in Figure 3, but is reported here instead of the main section since the visualizations were nearly identical.

doi:10.1007/s004420050869.

[22] J. Saarinen, A. M. Lister, Fluctuating climate and dietary innovation drove ratcheted evolution of proboscidean dental traits, Nature Ecology & Evolution 7 (9) (2023) 1490–1502. doi:10.1038/s41559-023-02151-4.

[23] W. J. Sanders, Evolution and Fossil Record of African Proboscidea, CRC Press, Taylor & Francis Group, 2023, p. 346.

[24] Y. Kundal, S. N. Kundal, Elephas cf. E. Maximus Indicus (Elephantidae, Mammalia) from the Post Siwalik Deposits of Jammu Province, Jammu and Kashmir, India, Vertebrata Palasiatica 49 (2011) 348–361.

[25] L. Girdland-Flink, E. Albayrak, A. M. Lister, Genetic insight into an extinct population of Asian elephants (Elephas Maximus) in the Near East, Open Quaternary 4 (2018) 1–9. doi:10.5334/oq.36.

[26] R. Sukumar, A brief review of the status, distribution and biology of wild Asian elephants Elephas Maximus, International Zoo Yearbook 40 (1) (2006) 1–8. doi:10.1111/j.1748-1090.2006.00001.x.

[27] J. Shoshani, J. F. Eisenberg, Elephas maximus, Mammalian Species (182) (1982) 1–8. arXiv:https://academic.oup.com/mspecies/article-pdf/doi/10.2307/3504045/8070918/182-1.pdf, doi:10.2307/3504045.
URL https://doi.org/10.2307/3504045

[28] C. P. D. Birch, S. P. Oom, J. A. Beecham, Rectangular and hexagonal grids used for observation, experiment and simulation in ecology, Ecological Modelling 206 (3-4) (2007) 347–359. doi:10.1016/j.ecolmodel.2007.03.041.

[29] K. Sahr, D. White, A. J. Kimerling, Geodesic discrete global grid systems, Cartography and Geographic Information Science 30 (2) (2003) 121–134. doi:10.1559/152304003100011090.

[30] M. F. Mechenich, Eco-ISEA3H: A spatial database of Earth's climate and biogeography, Fairdata.fi (2022). doi:10.23729/37d3e51e-3bf0-453a-a2ab-ed1a935ccaf8.

[31] M. F. Mechenich, I. Žliobaitė, Eco-ISEA3H, a machine learning ready spatial database for ecometric and species distribution modeling, Scientific Data 10 (2023) 77. doi:10.1038/s41597-023-01966-x.

[32] S. Faurby, M. Davis, R. O. Pedersen, S. D. Schowanek, A. Antonelli, J.-C. Svenning, PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology, Ecology 99 (11) (2018) 2626. doi:10.1002/ecy.2443.

[33] S. Faurby, R. O. Pedersen, M. Davis, S. D. Schowanek, S. Jarvie, A. Antonelli, J.-C. Svenning, PHYLACINE 1.2.1: An update to the Phylogenetic Atlas of Mammal Macroecology, Zenodo (2020). doi:10.5281/zenodo.3690867.

[34] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas, International Journal of Climatology 25 (15) (2005) 1965–1978. doi:10.1002/joc.1276.

[35] P. R. Gent, G. Danabasoglu, L. J. Donner, M. M. Holland, E. C. Hunke, S. R. Jayne, D. M. Lawrence, R. B. Neale, P. J. Rasch, M. Vertenstein, P. H. Worley, Z.-L. Yang, M. Zhang, The Community Climate System Model version 4, Journal of Climate 24 (19) (2011) 4973–4991. doi:10.1175/2011JCLI4083.1.

[36] J. Sillmann, V. V. Kharin, X. Zhang, F. W. Zwiers, D. Bronaugh, Climate Extremes Indices in the CMIP5 Multimodel Ensemble: Part 1. Model Evaluation in the Present Climate, Journal of Geophysical Research: Atmospheres 118 (4) (2013) 1716–1733. doi:10.1002/jgrd.50203.

[37] J. Sillmann, V. V. Kharin, F. W. Zwiers, X. Zhang, D. Bronaugh, Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. future climate projections, Journal of Geophysical Research: Atmospheres 118 (6) (2013) 2473–2493. doi:10.1002/jgrd.50188.

[38] K. Riahi, S. Rao, V. Krey, C. Cho, V. Chirkov, G. Fischer, G. Kindermann, N. Nakicenovic, P. Rafaj, RCP 8.5-a scenario of comparatively high greenhouse gas emissions, Climatic Change 109 (1-2) (2011) 33–57. doi:10.1007/s10584-011-0149-y.

[39] S. Faurby, M. B. Arajo, Anthropogenic range contractions bias species climate change forecasts, Nature Climate Change 8 (3) (2018) 252–256, 10.1038/s41558-018-0089-x.

[40] M. Barbet-Massin, F. Jiguet, C. H. Albert, W. Thuiller, Selecting pseudo-absences for species distribution models: how, where and how many?, Methods in Ecology and Evolution 3 (2) (2012) 327–338.

[41] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[42] M. F. Mechenich, J. Saarinen, I. Žliobaitė, Evaluating species distribution models in new environmental settings (In Preparation).

[43] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2022).
URL https://www.R-project.org/

[44] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (3) (2002) 18–22.
URL https://CRAN.R-project.org/doc/Rnews/

[45] P.-C. Brkner, brms: An R Package for Bayesian Multilevel Models Using Stan, Journal of Statistical Software 80 (1) (2017). doi:10.18637/jss.v080.i01.

[46] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carr, J. R. Garca Marquz, B. Gruber, B. Lafourcade, P. J. Leit£o, T. Mnkemller, C. McClean, P. E. Osborne, B. Reineking, B. Schrder, A. K. Skidmore, D. Zurell, S. Lautenbach, Collinearity: A review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (1) (2013) 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

[47] B. Naimi, N. A. S. Hamm, T. A. Groen, A. K. Skidmore, A. G. Toxopeus, Where is positional uncertainty a problem for species distribution modelling?, Ecography 37 (2) (2014) 191–203. doi:10.1111/j.1600-0587.2013.00205.x.

[48] T. Hastie, R. Tibshirani, R. Tibshirani, Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons, Statistical Science 35 (4) (2020) 579–592. doi:10.1214/19-STS733.

[49] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (8) (2017) 913–929. doi:10.1111/ecog.02881.

[50] N. Kallioinen, T. Paananen, P.-C. Brkner, A. Vehtari, Detecting and diagnosing prior and likelihood sensitivity with power-scaling (2021). doi:10.48550/ARXIV.2107.14054.
URL https://arxiv.org/abs/2107.14054

[51] S. N. Wood, Thin plate regression splines, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 (1) (2003) 95–114. doi:10.1111/1467-9868.00374.

[52] QGIS Development Team, QGIS Geographic Information System, QGIS Association (2022).
URL https://www.qgis.org

[53] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159. doi:10.1016/s0031-3203(96)00142-2.

[54] O. ALLOUCHE, A. TSOAR, R. KADMON, Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (tss), Journal of Applied Ecology 43 (6) (2006) 1223–1232. doi:10.1111/j.1365-2664.2006.01214.x.

[55] G. C. Cawley, N. L. C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research 11 (70) (2010) 2079–2107.
URL http://jmlr.org/papers/v11/cawley10a.html

[56] P. Smialowski, D. Frishman, S. Kramer, Pitfalls of supervised feature selection, Bioinformatics 26 (3) (2010) 440–443. doi:10.1093/bioinformatics/btp621.

[57] T. P. Young, T. M. Palmer, M. E. Gadd, Competition and compensation among cattle, zebras, and elephants in a semi-arid savanna in laikipia, kenya, Biological Conservation 122 (2) (2005) 351–359. doi:10.1016/j.biocon.2004.08.007.

[58] R. R. R. Sukumar, Elephant foraging : is browse or grass more important?, A Week With Elephants (1995).

[59] R. A. Guldemond, A. Purdon, R. J. van Aarde, A systematic review of elephant impact across africa, PLOS ONE 12 (6) (2017). doi:10.1371/journal.pone.0178935.

[60] P. Fernando, P. Leimbruger, Asian elephants and seasonally dry forests, Smithsonian Institution Scholarly Press, 2011.

[61] I. Calandra, U. B. Ghlich, G. Merceron, How could sympatric megaherbivores coexist? example of niche partitioning within a proboscidean community from the miocene of europe, Naturwissenschaften 95 (9) (2008) 831–838. doi:10.1007/s00114-008-0391-y.

[62] J. Williams, S. Jackson, Novel climates, no-analog communities, and ecological surprises, Frontiers in Ecology and Environment 5 (9) (2007) 475–482.