Predicting habitat suitability for Asian elephants in non-analog ecosystems with Bayesian models

Published in:
Ecological Informatics

DOI:
10.1016/j.ecoinf.2024.102658

Published: 01/09/2024

Please cite the original version:
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, e

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
e Finnish Museum of Natural History LUOMUS, P.O.Box 64, FI-00014, Finland

ARTICLE INFO

Keywords: Bayesian statistical models Species distribution modeling Rewilding Non-analog ecosystems Elephas maximus

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 modeling choices. The results of our analysis suggest that one configuration of Bayesian logistic regression 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.

1. Introduction

Rewilding is a conservation strategy which proposes ecosystem restoration through the reintroduction of extirpated megaflora species (Donlan, 2005; Donlan et al., 2006). While relatively few such wild experiments (Lorimer and Driessen, 2014) have been conducted, several projects have provided insight into the potential for megafauna to reestablish important top-down trophic cascades (Svenning et al., 2016). 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 (Beschta and Ripple, 2016).

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) (Elith and Leathwick, 2009), 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 and other environmental inputs, such as temperature, precipitation, or vegetation. They provide a data-driven assessment of the extent to which the species can thrive in its new range or the ways in which the range can shift over time (Barlow et al., 2021; Eyre et al., 2022; Jarvis and Svenning, 2018).

SDMs are usually statistical, machine learning or process-based models that connect climatic features with species observations or the physiology of the species. These types of models are limiting because they do not model the random 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) (Catalina et al., 2022). 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 basis dimension 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 are possible 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. (Gelman et al., 2020). The Bayesian and non-Bayesian model pairs are chosen so that one represents a less flexible model in terms of possible shapes of its decision boundary (Bayesian logistic regression and MLE logistic regression) and one represents a more flexible 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 (The Intergovernmental Panel on Climate Change, 2024), 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 3000 m MSL, the species’ upper elevation limit (Jarvie and Svenning, 2018).

2. Background on Elephas ecology and range

As a megaherbivore species having the potential to reestablish top-down trophic cascades (Svenning et al., 2016) in trophically downgraded ecosystems (Estes et al., 2011), the Asian elephant has been proposed as a candidate for conservation translocation in South America (Galetti, 2004), North America (Donlan et al., 2006), Europe (Svenning, 2007), Australia (Bowman, 2012), and the tropical Asia-Pacific region (Louys et al., 2014).

According to the current understanding, the genus Elephas originated in Africa during the Pliocene (Sanderson et al., 2016; Zhang, 2020). 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 (Larramendi et al., 2020; Zhang, 2020), the earliest confirmed Elephas species in the most recent phylogenetic analyses (Zhang, 2020) is E. atavus from the Early Pleistocene of East Africa, where it lived in savanna environments and had grass-dominated diet (Cerling et al., 1999). 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 (Sanderson et al., 2010). 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 (Saarinen and Lister, 2023; Sanders, 2023).

The genus Elephas went extinct in Africa during the Pleistocene (e.g., (Sanderson et al., 2010)), but was widespread across Southern Asia from the Pleistocene to the Holocene (e.g., (Kundal et al., 2011)). 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’ to its westernmost occurrence around the Lake Gavur in central Anatolia (Girdland-Flink et al., 2018). The current range of Elephas maximus is limited to parts of India, Sri Lanka and Southeast Asia, including Borneo and Sumatra, and is heavily influenced by the loss of suitable habitat due to land use, as well as other human influence (e.g. (Kundal et al., 2011; Sukumar, 2006)).

Today E. maximus occupies a variety of environments from tropical rainforests in Borneo to seasonally dry grassy woodlands in India and Sri Lanka. It prefers ecotone habitats where the vegetation consists of a mixture of forest, low-growing woody plants, and grasses (Shoshani and Eisenberg, 1982). 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 (Sukumar, 2006). In tropical rainforests the diet can consist almost entirely of browse and fruit (Sukumar, 2006). 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 (Sukumar, 2006).

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 (3000 m MSL (Jarvie and Svenning, 2018)) 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 Fig. 1a and b 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 (Birch et al., 2007). However, in developing our SDMs we utilized the ISEA3H DGGS (Sahr et al., 2003), a hexagonal grid system. The cells of this system are equal-area and maximally in every habitat, as well as other human influence (e.g. (Kundal et al., 2011; Sukumar, 2006)).

Today E. maximus occupies a variety of environments from tropical rainforests in Borneo to seasonally dry grassy woodlands in India and Sri Lanka. It prefers ecotone habitats where the vegetation consists of a mixture of forest, low-growing woody plants, and grasses (Shoshani and Eisenberg, 1982). 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 (Sukumar, 2006). In tropical rainforests the diet can consist almost entirely of browse and fruit (Sukumar, 2006). 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 (Sukumar, 2006).

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 (3000 m MSL (Jarvie and Svenning, 2018)) 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 Fig. 1a and b 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 (Birch et al., 2007). However, in developing our SDMs we utilized the ISEA3H DGGS (Sahr et al., 2003), a hexagonal grid system. The cells of this system are equal-area and maximally in every habitat, as well as other human influence (e.g. (Kundal et al., 2011; Sukumar, 2006)).

We used climate and species distribution data from the Eco-ISEA3H
The centroids of the hexagonal cells is approximately 50 km globally. The Phase 5 (CMIP5) global climate models for which these indices were and Indices (ETCCDI), and the 19 bioclimatic variables from WorldClim extremes indices defined by the Expert Team on Climate Change Detection - inter-centroid distances. Because a high-resolution grid was used, centroids are visible as discrete points only at high latitudes, where the map projection exaggerates the centroids. Because a high-resolution grid was used, centroids are visible as discrete points only at high latitudes, where the map projection exaggerates the inter-centroid distances.

spatial database (Mechenich, 2022; Mechenich and Zliobaitė, 2023) at resolution nine, a high-resolution grid in which the distance between the centroids of the hexagonal cells is approximately 50 km globally. The dataset is available via (Mechenich, 2022).

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 (Hijmans et al., 2005). 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) (Gent et al., 2011), one of a number of Coupled Model Intercomparison Project Phase 5 (CMIP5) global climate models for which these indices were calculated (Sillmann et al., 2013a; Sillmann et al., 2013b). Present-day WorldClim bioclimatic variables were derived from monthly temperature and precipitation, interpolated from weather station observations using thin-plate smoothing splines (Hijmans et al., 2005). 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 (Riahi et al., 2011). 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 (Faurby et al., 2018; Faurby et al., 2020). The present-natural range, mapped in yellow in Fig. 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 (Faurby and Araújo, 2018).

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 (Barbet-Massin et al., 2012). 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 Fig. 1a, this pseudo-absence buffer surrounds the present-natural range, excluding areas of high topographic relief (above 3000 m 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 7331, 3765 of which were marked as the species being present.

4. Statistical models

For comparative analysis of alternative computational modeling choices, we used model pairs that vary in their level of flexibility in the Bayesian and non-Bayesian frameworks.

4.1. The less flexible 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 \( x_1, x_2, \ldots, x_k \), it is modeling the logit of the probability with a linear hyperplane as

\[
\log \left( \frac{P(y = 1)}{1 - P(y = 1)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{1}
\]

where the coefficients and intercept \( \beta_0, \beta_1, \ldots, \beta_k \) 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
\[
P(y = 1) = \frac{1}{1 + e^{-z}}.
\]

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. The more flexible models

4.2.1. Baseline model: random forest

Random forest (Breiman, 2001) 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., whether the mean annual temperature is above 25 °C). 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 (Mechenich et al., 2024). 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:
\[
\log\frac{P(y = 1)}{1 - P(y = 1)} = \beta_0 + \sigma_1 s_1(x_1) + \cdots + \sigma_k s_k(x_k),
\]

where \(\sigma_k\) are weights of non-linear functions, and each non-linear function \(s_k(x_k)\) is presented as weighted sum of simpler basis functions \(b_{k,l}\) as
\[
s_k(x_k) = \sum_{l=1}^{L} \beta_{k,l} b_{k,l}(x_k).
\]

In the brms (P.-C., 2017) package, the package we used, the default basis functions are regularized thin plate splines. We explored different configurations of the basis function complexity, which we explain more in Section 5.2.4.

4.3. Implementation

Modeling was done with various packages within R (R Core Team, 2022). For modeling MLE logistic regression we used the glm function of R’s stat package, for random forest we used the randomForest package (Liaw and Wiener, 2002), and for the Bayesian models we used the brms package (P.-C., 2017).

5. Analysis protocol

5.1. The iterative modeling process

Our process of finding the best model roughly follows the Bayesian workflow described by Gelman et al. (Gelman et al., 2020). This is an iterative modeling process in which one systemically repeats the cycle of inspecting the outputs of a model, analyzing the effects of the model and its configurations, and adjusting the configurations or switching to another model type according to the observations. The cycle is repeated multiple times while recording the observations and results.

In our case, the possible configurations we explored were these modeling options: two sets of features selected from a larger pool of available features (method in Section 5.2.1), whether or not to scale the features, choice of prior strength, and for Bayesian GAM only, the basis dimension. For the priors, we were partially able to utilize the priorsense package (Kallioinen et al., 2021) to optimize the exploration of different configurations. For the rest of the settings however, we kept to the repeated iterations of the Bayesian workflow. Our workflow results are the visualizations and scores presented in Figs. 3, 4, and B.11, and the observations are the discussions and analysis in this text: what we can interpret from the scores and visualizations and diagnoses of problems, if any.

5.2. Modeling options

5.2.1. Feature selection

SDM development requires two feature selection steps, namely (1) removing highly collinear features (Dormann et al., 2013), and (2) identifying a minimum subset of available features with greatest ecological relevance and explanatory power (Elith and Leathwick, 2009). First, collinear features were removed from the initial set of 46 climatic predictors via the vifcor algorithm (Naimi et al., 2014); 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 (Hastie et al., 2020), 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 (Roberts et al., 2017), in which the full training dataset was partitioned into spatial blocks like those shown in Fig. 2.

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 (Mechenich et al., 2024). 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.

5.2.2. 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 each have range $[0,1]$ but retain the shape of the original distribution. This is done through min-max scaling

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}.$$  \hspace{1cm} (5)

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

5.2.3. Selecting priors

We started by assigning weak priors to Bayesian models. Then for Bayesian logistic regression, we used the priorsense package 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 GAMs, 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.

5.2.4. GAM basis dimension

The flexible functions we chose in Eq. (3) for Bayesian GAM uses an eigendecomposed approximation of a regularized thin-plate regression spline (Wood, 2003).

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 analyzed the sensitivity of the results with regard to this argument by first utilizing the automatic adjustment and then manually setting upper bounds for it according to the prediction outputs. In our case the automated basis dimension selection yielded highly overfit results (presented in Fig. A.9), so in response we set increasingly restrictive basis dimensions until we reached the lowest possible value for $k$ allowed by brms, which is $k = 1$. With $k = 1$ the model includes for each feature a linear term and a non-linear term which is close to quadratic. However, even with this restricted non-linearity, the non-linear model was overfit compared to having only the linear terms. The initial and final values of $k$ are shown with the priors in Table 2.

5.3. 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 in QGIS (QGIS Development Team, 2024). For numerical scores, we present area under the curve (AUC) (Bradley, 1997) and maximum attainable true skill statistic (TSS) (Allouche et al., 2006).

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 7331 grid cells. Their outputs for the area shown in Fig. 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 Fig. 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 (Cawley and Talbot, 2010; Smialowski et al., 2010). 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 modeling approaches, and to assess all feature sets via spatial CV (which accounts for the spatial autocorrelation present in environmental datasets (Roberts et al., 2017)), 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. For example, if a perfectly calibrated model predicts the habitat suitability of a set of spatial grid cells to be 70%, 70% of those cells should actually be suitable and 30% unsuitable. A

### Table 1

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Feature Name</th>
<th>Source</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>random-CV</td>
<td>BIO03</td>
<td>WorldClim</td>
<td>Isotherality: Mean of monthly (maximum - minimum temperature) divided by the annual temperature range.</td>
</tr>
<tr>
<td></td>
<td>TN10P</td>
<td>ETCDDI</td>
<td>Proportion of days when the minimum temperature is lower than the 10th percentile of historical data.</td>
</tr>
<tr>
<td></td>
<td>GSL</td>
<td>ETCDDI</td>
<td>Growing season length.</td>
</tr>
<tr>
<td></td>
<td>TXN</td>
<td>ETCDDI</td>
<td>Maximum daily minimum temperature.</td>
</tr>
<tr>
<td></td>
<td>BIO08</td>
<td>WorldClim</td>
<td>Mean temperature of the wettest quarter.</td>
</tr>
<tr>
<td>spatial-CV</td>
<td>BIO02</td>
<td>WorldClim</td>
<td>Mean diurnal range: Mean of monthly difference between the maximum and minimum temperature.</td>
</tr>
<tr>
<td></td>
<td>TN90P</td>
<td>ETCDDI</td>
<td>Proportion of days when the minimum temperature is higher than the 90th percentile of historical data.</td>
</tr>
<tr>
<td></td>
<td>ID</td>
<td>ETCDDI</td>
<td>Icing days.</td>
</tr>
<tr>
<td></td>
<td>BIO14</td>
<td>WorldClim</td>
<td>Precipitation of the driest month.</td>
</tr>
<tr>
<td></td>
<td>BIO18</td>
<td>WorldClim</td>
<td>Precipitation of the warmest quarter.</td>
</tr>
<tr>
<td></td>
<td>CWD</td>
<td>ETCDDI</td>
<td>Maximum length of wet spell.</td>
</tr>
<tr>
<td>common for both sets</td>
<td>RX1DAYS</td>
<td>ETCDDI</td>
<td>Monthly maximum one-day precipitation.</td>
</tr>
<tr>
<td></td>
<td>WSDI</td>
<td>ETCDDI</td>
<td>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.</td>
</tr>
</tbody>
</table>
priors were used for Bayesian models. However, since this was a logistic regression save for slight details in the QGIS visualizations and fit on spatial-CV features had more conservative point predictions but point predictions did not have clear presences or absences. The models areas such as the Sahara desert and the southern parts of Alaska and the generous future suitable habitat area for Asian elephants, including conditions. The models fit on the random-CV features predicted a very almost identical to Bayesian logistic regression.

Table 2

<table>
<thead>
<tr>
<th>Model type</th>
<th>Feature set</th>
<th>Scaling</th>
<th>Setting type</th>
<th>Initial</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>logistic regression</td>
<td>random-CV</td>
<td>yes</td>
<td>priors</td>
<td>coefficients</td>
<td>Normal(0,5)</td>
</tr>
<tr>
<td></td>
<td>spatial-CV</td>
<td>yes</td>
<td>priors</td>
<td>intercept</td>
<td>Normal(0,10)</td>
</tr>
<tr>
<td>GAM</td>
<td>all feature sets and scaling</td>
<td></td>
<td>priors</td>
<td>coefficients</td>
<td>Normal(0,5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>non-linearity</td>
<td>Student-t(3,0,2.5)</td>
<td>Normal(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>intercept</td>
<td>Normal(0,10)</td>
<td>Normal(0,20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>argument</td>
<td>basis dimension (k)</td>
<td>–1 (default)</td>
</tr>
</tbody>
</table>

 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.

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 final prior and basis dimension settings for the Bayesian models. We present the summary of our observations 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.

6.1. The less flexible models - Bayesian and MLE logistic regression

The numerical scores and visualizations for these models are presented in Figs. 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 a 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 Fig. B.11.

6.2. The more flexible models - Bayesian GAM and random forest

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

6.2.1. Bayesian GAM

The final prior and basis dimension settings for Bayesian GAMs (in the column ‘Final’ on Table 2) placed strong restrictions on the models’ non-linearity, which was controlled by a subset of priors and the basis dimension of the splines. 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 Elephas maximus. It seemed that restricting the priors and the basis dimension 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 flexible for the patterns in the data, and caused a problem called ‘complete separation’, described later in Section 7.2.

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 Figs. 3 and B.11 suggested that the models fit on random-CV features underfit the true pattern, since the predictions undershot 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. MLE logistic regression was not affected by feature scaling.

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. 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 Fig. 4. Feature scaling did not have an influence on the random forest models, which is expected as decision-tree based models should not be affected by it (nevertheless we report it here for consistency).

Bayesian GAM initially showed signs of extreme overfitting, which is discussed in Appendix A. Attempts to restrict the non-linearity through the priors and basis dimension 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 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 final 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.

The random forest models and Bayesian GAMs overfitting the true pattern may be an example of the effect illustrated by Fourcade et al. (Fourcade et al., 2018). When training flexible models on spatially autocorrelated predictors, the models can fit patterns that are complete...
nonsense - even ones that are created from paintings.

7.2. Analysis of anomalies in model predictions

By visualizing the conditional effects of individual features with brms’s \texttt{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

Fig. 4. The prediction outputs, calibration plots, and numerical scores for the Bayesian GAMs and random forest models with scaled features. The Bayesian models were fit using the final priors and basis dimensions in Table 2.
The first pattern is the high non-linearity and uncertainty in the conditional effects, shown in Fig. 5a.

The second pattern can be observed in Figs. 5b, 6, and the left part of Fig. 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 Fig. 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 Fig. 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 Fig. 5a, the model ignoring the effects of ID in Fig. 5b, and additionally, the unusually large and variable posteriors for the Bayesian GAMs we present in Fig. 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 Fig. 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 the final prior settings. The outputs of this model are shown in Fig. 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 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
South America, South-Eastern Africa, Madagascar (except the arid geographic areas outside the historical range are located in Northern forest habitats of this species in Asia. The predicted currently suitable the distribution of biomes that are broadly similar to the woodland and savannas during the Plio-Pleistocene in Africa, such as increased hyp sodonty and lamellar count (e.g. Cerling et al., 1999; Saarinen and Nottomastodon 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 (Svenning, 2007) 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 (Girdland-Flink et al., 2018). 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 our predictions in this study are entirely based on climatic variables and they do not take into account other ecological 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 (Young et al., 2005). 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 (Cerling et al., 1999; Sukumar, 1995; Sukumar, 2006). 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 (Guldemond et al., 2017). *E. maximus* has not been shown to significantly affect vegetation structure in its habitats, except in artificially high population densities (Fernando and Leimbruger, 2011; Sukumar, 2006). 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 different adaptations and evidence of niche partitioning, especially in diet (Calandra et al., 2008).

*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 the 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 (Sukumar, 2006). 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 case of potential rewilding. 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 magnitudes 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 are 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 modeling dataset. This resonates with the widely acknowledged challenge of modeling potentially non-analog ecosystems (Williams and Jackson, 2007) 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 model fit on scaled spatial-CV features with the final prior settings is the most reliable model given the data, our iterative modeling process, and historical habitats of proboscidean species (outputs shown in Fig. 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
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is made available on the Github repository mentioned in the manuscript. Other than that, it is also available at https://doi.org/10.23729/37d3e51e-3bf0-453a-a2ab-ed1a935ccaf8.

Appendix A. The Bayesian iterative modeling process

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 Fig. 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.

Fig. 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.

A.2. Bayesian GAM

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

Examining the posteriors (example in Fig. A.10), we observed that the non-linearity of the functions in Eq. (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 Fig. 4.
Fig. A.9. Some examples of the overfit predictions from the Bayesian GAMs using the initial prior and basis dimension settings.

Fig. 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 Fig. A.9). The distributions starting with the abbreviation ‘NL’ are posteriors that represent the non-linearity of the splines in Eq. (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.

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 Fig. B.11 for transparency.
MLE logistic regression

Fig. 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 Fig. 3, but is reported here instead of the main section since the visualizations were nearly identical.

References


