Predicting the Habitat Suitability of Asian Elephants in 2070 with Bayesian Models

Ryoko Noda

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 30.5.2022

Supervisor

Prof. Aki Vehtari

Advisor

Michael Francis Mechenich,
Prof. Indrė Žliobaitė

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Ryoko Noda

Title Predicting the Habitat Suitability of Asian Elephants in 2070 with Bayesian Models

Degree programme Computer, Communication and Information Sciences

Major Machine Learning, Data Science and Artificial Intelligence **Code of major** SCI3044

Supervisor Prof. Aki Vehtari

Advisor Michael Francis Mechenich, Prof. Indrė Žliobaitė

Date 30.5.2022 **Number of pages** 63+1 **Language** English

Abstract

Pleistocene rewilding is an ambitious approach to conservation in which extinct megafauna are replaced by extant relatives to fulfill missing ecological roles. While controversial, it has the potential to repair diminishing ecosystems with minimal human intervention. However, careful inspection and simulation is required before the project can take place as this involves introducing species to a new habitat. For this purpose, species distribution models are expected to be a powerful tool as they predict species presence under various climate and habitat conditions.

Therefore, this thesis presents a case study of iterative modeling and Bayesian workflow of species distribution models under a hypothetical Pleistoscene rewilding plan. The aim of this project is to predict a suitable habitat for Asian elephants (*Elephas Maximus*) that remains suitable until the year 2070. All models in the workflow use predicted climate features under the representative concentration pathway 8.5 scenario to produce scores of future habitat suitability (range $[0, 1]$. 0: not suitable/1: suitable).

The iterative model building starts with non-Bayesian machine learning, logistic regression and random forest. These models are then used as benchmarks for two different Bayesian models, a Bayesian generalized linear model and a Bayesian generalized additive model. While building and exploring models, this research explores the effects of feature selection, feature scaling, and for the Bayesian models, the priors and non-linearity settings.

The model exploration process was able to identify a model that gives convincing predictions for present-day and future conditions. The outputs from this model implied that possible rewilding sites would include northern South America, sea-facing regions of east and west Africa, and the north shoreline of Australia.

Keywords Species Distribution Models , Bayesian Modeling , Bayesian Workflow , Machine Learning , Pleistocene Rewilding , Elephas Maximus

Acknowledgements

Many people offered their help in creating this thesis, and I will use this page to share how they contributed.

Professor Aki Vehtari helped me find my thesis topic and agreed to be my supervisor. He was the one who suggested that I look for topics in another university. Without him, I would not have met my advisors. He also helped shape my initially vague thesis topic into the research you see on this document. He was the one who nudged me in the right direction whenever I was stuck on something that I could not quite explain.

Michael Mechenich was the person who started the research on Asian elephant habitat suitability in the University of Helsinki. He let me continue his research from a Bayesian perspective. He provided me with his beautifully aggregated and cleaned dataset that allowed me to start modeling immediately. Additionally, he let me use two of his feature sets he selected in his previous research. I owe the data and the beginning of the thesis research to him. I am also very thankful for the weekly discussions we held as it let me keep pace in my individual work.

Professor Indrė Žliobaitė from the University of Helsinki connected me with Michael. Though I was not from the same university, she gave me frequent support and a chance to present in a seminar. I appreciate all the feedback I received from her and her students and colleagues. They helped me organize my thoughts enough to start writing this thesis.

Professor Jarno Vanhatalo from the University of Helsinki gave me guidance when finding advisors within his university and provided reference articles about Bayesian species distribution models. I deeply appreciate that he took time out of his busy schedule to help me.

Noa Kallioinen gave me a hand when interpreting his new R package under development. It was thanks to him that I found solid proof of the priors overpowering the likelihood.

Professor Juha Saarinen from the University of Helsinki gave me professional insight in deciding whether my best model truly had reasonable outputs. I would not be so sure of the results without him.

And finally, I would like to give thanks to the editors and reviewers of this thesis: Diane Pilkinton-Pihko and Maurice Forget from the writing clinic, my friend and new colleague Frederik Heylen, and Michael Mechenich, Indrė Žliobaitė, and Aki Vehtari again for their feedback.

Otaniemi, 30.5.2022

Ryoko Noda

Contents

Abbreviations and Acronyms

Symbols

$\widehat{R}$	The convergence diagnostic value for Markov chains described by Vehtari et al. [1]
n_eff	effective sample size

Abbreviations

SDM	species distribution model
GLM	generalized linear model
GAM	generalized additive model
AUC	area under the curve
TSS	true skill statistic
MCMC	Markov chain Monte Carlo
HMC	Hamiltonian Monte Carlo
CMIP5	Coupled Model Intercomparison Project Phase 5
ETCCDI	Expert Team on Climate Change Detection and Indices
CCSM4	Community Climate System Model Version 4
RCP	representative concentration pathway
IQR	interquartile range
CV	cross-validation
Coeff	coefficient. Used in plots and tables where space is scarce
NL	non-linearity. Used in plots and tables where space is scarce

1 Introduction

Pleistocene rewilding, or more broadly, trophic rewilding, is a progressive strategy in conservation. It is based on evidence that the ever-decreasing population of large mammals plays a crucial role in sustaining ecosystems. The scheme (re)introduces existing megafauna to areas where the species or its relatives have been extinct. If successful, the introduced species takes over the role of the previously extinct species and creates a top-down trophic interaction, restoring nature and strengthening its biodiversity. The word 'Pleistocene' refers to a geological epoch characterized by large mammals, which are the centerpiece of Pleistocene rewilding.

While controversial, this approach has slowly shifted from fringe science to an option for a conservation project. Though there are still very few cases of replacing completely extinct species with extant relatives [2] [3], there have been rewilding projects that reintroduce species to former habitats [4] [5] [6]. One successful case of reintroducing gray wolves to Yellowstone National Park suppressed the elk population, allowing plants and shrubs to recover after seventy years of overgrazing [7].

However, these projects need to be carefully considered, especially when introducing a species to where they have never lived before. The new range is unlikely to have the same climate as the original habitat. Water may be less accessible or more abundant; there may be more or less vegetation than the species is accustomed to having. Moreover, these factors may change in the future due to climate change. This makes the long-term survival and range shifts of species even more unpredictable.

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

SDMs are usually machine learning or process-based models that connect climatic features with species observations or the physiology of the species. These types of models are problematic because they cannot model the random effects and fluctuations in the environment nor give predictions with uncertainty. 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 be 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 human knowledge about the problem that they are attempting to solve, they can give predictions in the absence of sufficient data. In this case, the predictions are initially closer to human knowledge and will become increasingly data driven as more data arrives.

While some examples of Bayesian SDMs exist [8] [9], there are still very few case studies. Furthermore, the literature that explores basic Bayesian model structures

was mainly published before 2010. This was before the advent of modern Bayesian modeling algorithms, toolkits, and guidelines. Therefore, the aim of this thesis is to iteratively build Bayesian models for an early stage, exploratory Pleistocene rewilding project in order to demonstrate how basic Bayesian species distribution modeling is conducted with recent tools and methods as of 2022.

In this project we attempt to locate feasible rewilding areas for Asian elephants (*Elephas Maximus*) that will remain suitable in the year 2070 even in the worst scenario of climate change. The search area is every land area on Earth, excluding Antarctica.

We explore the answer to this problem by iteratively building a series of models, beginning with two baseline non-Bayesian models and then trying two relatively simple Bayesian models. The pairs of models were chosen so they represent a simple linear model and a more complex non-linear model. For non-Bayesian models, we build logistic regression and random forest. For the Bayesian models, we build a Bayesian generalized linear model (GLM) and a Bayesian generalized additive model (GAM).

While model building, we examine the effects of modeling options that affect the prediction results: feature selection, feature scaling, and for the Bayesian models, choices of priors and non-linearity settings. The process for Bayesian model exploration roughly follows the Bayesian workflow described by Gelman et al. [10]. In addition to describing the results in the traditional way, we present our research in Figure 1 as a suggestion of a novel way to present modeling choices and their results.

This thesis is organized as follows. The Background section introduces the background on SDMs, Bayesian modeling, and its workflow. The Data section presents an overview of our dataset and its sources. In the Methods section, we specify the modeling methods used in the research. Settings and modeling effects specific to Bayesian models are described in the Sampling Settings and Posteriors section. The Results section contains the prediction results of the models. Finally, we end the thesis by explaining the significance of the results in the Discussion section and the Conclusion section.

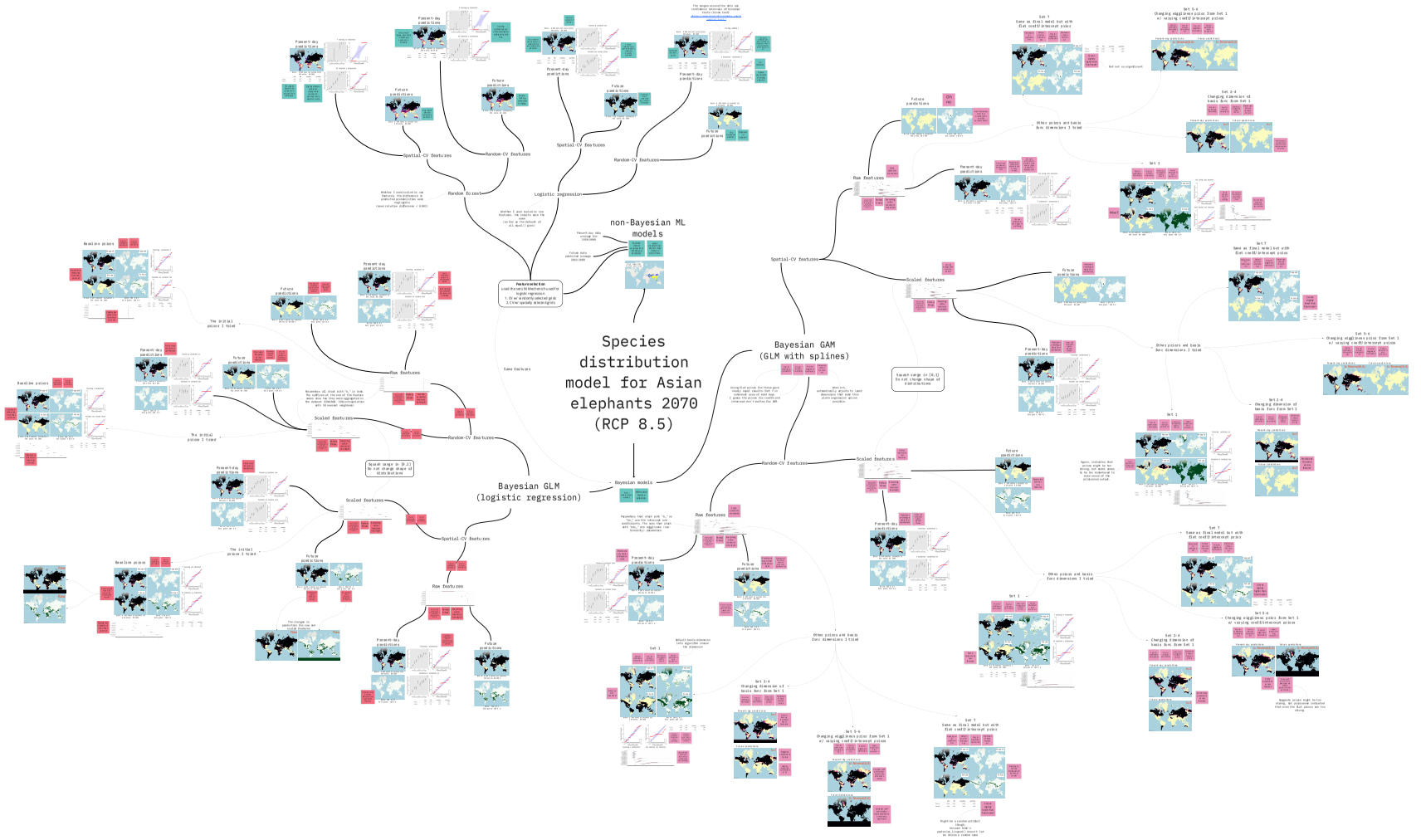


Figure 1: The mind map of this research. An interactive version with working GIF animations can be found at https://miro.com/app/board/uXjVOX_Zhf8=/?share_link_id=937959545296.

2 Background

As stated in the Introduction section, there is limited research on SDMs using Bayesian approaches. The most frequently cited publications are fairly old, published over ten years ago using early Bayesian modeling tools [8] [11]. One reason for this may be that Bayesian statistics tends to be very technical. Until recently, there were no resources that let ecologists easily use or interpret it. However, this situation is beginning to change.

Lately, many tools for Bayesian modeling have become available. Stan is a probabilistic programming language that enables users to fit models using state-of-the-art Bayesian methods [12]. Wrappers such as rstanarm [13] and brms [14] allow access to Stan through traditional R model fitting functions. For Python users, the open-source package PyMC provides features similar to Stan [15].

In addition to these tools, there have been publications that break down the Bayesian modeling workflow in order to provide guidelines to researchers from outside the statistical field. The When-to-worry-and-how-to-Avoid-the-Misuse-of-Bayesian-Statistics-checklist (the WAMBS-checklist) [16] gives a step-by-step guide on how to correct, analyze and report Bayesian models. Gelman et al. published a more detailed walkthrough of the Bayesian modeling process with some use cases [10]. In addition to academic articles, there are textbooks that give readers structured training of Bayesian methods [17] [18].

It is still too early to see whether this trend will increase Bayesian modeling in SDMs, but it is likely that more ecologists will try Bayesian methods as more resources become available. An early adopter may be the new embarcadero package [19] that provides a set of tools for creating SDMs with Bayesian additive regression trees.

2.1 Species Distribution Models

There are mainly two types of species distribution models: mechanistic and correlative. Both predict habitat suitability based on environmental data, but differ in what information they associate to the environment.

Mechanistic models are process-based models that associate a species' physiology to its environment. They are said to be more robust since they model the direct response to an environment, but can only be made for species with sufficient physiological data.

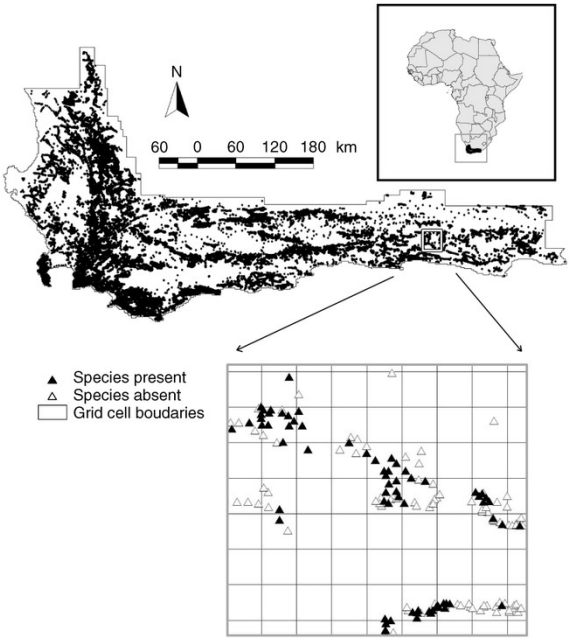
Correlative models, used for this thesis, are machine learning or statistic models that associate known ranges or sighting information of species to its environment. They are more common as range data is easier to collect than physiological data. However, they are influenced by bias in the observation or realized ranges. This implies that if the data is collected only where the field research is convenient or if the current habitat is actually a restricted range due to human activity, it will affect the quality of the predictions.

A correlative SDM is created by first preparing climate, vegetation, or other environment data in the form of regular grids spread across the area of interest. Each

grid cell would have its own distinct set of values for each feature. This type of data is usually obtained as outputs of a climate model [20] or interpolations of weather station data [21] [22].

Once the environment data is ready, species distribution data is mapped into the same grids as the environment data. This may be in the form of binary (presence/absence) raster grids or number of occurrences (sightings or abundance) within each grid. The source data might be gathered through a field research, acquired from a conservation organization, or provided from an expert on the species. An example of observation data mapped in environmental grid cells is shown in Figure 2a.

Next, the environment and species data is used to train or fit a machine learning or statistical model so that it recognizes the patterns that make a species more likely or unlikely to appear in a grid cell. The model's outputs are usually a list of scores or probabilities of habitat suitability or species occurrence, with one value for each input grid cell. When visualizing, these scores are georeferenced as raster grids and overlaid on a map of the target area as shown in Figure 2b.



(a) Example input data with species sightings mapped within grid cells.

(b) Example output data with predicted species occurrence probabilities visualized as raster grids.

Figure 2: Example SDM input and output visualized by Latimer et al. [8].

Evaluating a Species Distribution Model

When evaluating the performance of SDMs, numerical scores are used in addition to visual inspection of raster grids. The most common scores are area under the curve (AUC) [23] and true skill statistic (TSS) [24]. These are both statistics that are calculated by setting a threshold for the output scores that determine a binary classification for each raster grid: suitable habitat or not suitable habitat. The binary classifications are then compared against the true habitat suitability of each grid cell i.e., whether the cell had a species sighting or was within a range specified by experts.

AUC is the area under the receiver-operating characteristic curve, which plots the trade-off between the sensitivity or 'true positive rate', the proportion of suitable habitat grids classified correctly, and the 'false positive rate', the proportion of unsuitable habitat grids incorrectly classified as suitable. The AUC value has a range from 0 to 1 and measures the model's potential discrimination capability, or in other words, how competent a model can be at distinguishing suitable habitats from unsuitable habitats. A score closer to 1 indicates that a model is better at classification.

TSS has different names outside of species distribution modeling; it is also known as Youden's J statistic or informedness. It is calculated for a single, user-defined threshold value as

$$\text{TSS} = \text{sensitivity} + \text{specificity} - 1, \tag{1}$$

where specificity, also known as 'true negative rate', is the proportion of unsuitable habitats classified correctly. This score also has a range from 0 to 1, and a value above 0.75 indicates excellent classification performance.

Challenges of Extrapolation

Correlative SDMs can give quite accurate results when doing short-term predictions for an area close to the training inputs. However, the predictions become erroneous when predicting for the future or for an area far away from the original range. The problem of predicting for these novel conditions is called extrapolation, and it is one of the challenges of SDMs.

Inaccuracy in extrapolation problems are said to be caused by limitations of species range data. Often, the range data is not representative of the true distribution of species given the climate. Species ranges are usually restricted by interactions with other species or humans. Additionally, field research or sighting records tend to concentrate in areas that are easily accessible by humans. As a consequence, the data used in correlative SDMs are typically biased.

Suggested solutions to this challenge include using historical range data, conducting stratified field research, or using generated sighting data in addition to real observations [25]. Even so, these methods are still not enough to mitigate inaccurate predictions in drastic climate change or novel environments. Switching to mechanistic SDMs may be an effective solution, but the cost involved in collecting physiological

data of species makes it unrealistic. Therefore, extrapolation remains an unsolved challenge for SDMs.

2.2 Bayesian Models

Bayesian models are a type of advanced statistical model that predicts or describes an event based on data. They have very similar applications and structure as traditional machine learning models with two additional benefits: they can express uncertainty and incorporate human knowledge. Because of this trait, recent machine learning models are adopting Bayesian methods in their structures.

Regardless of whether it has a Bayesian structure, the objective of fitting a model is to find values for 'parameters', variables within the model structure that determine the relationship between the data and the outcome. However, if a model has Bayesian properties, the fitting process identifies a probability distribution instead of a single value for each parameter. This distribution describes what the values of the parameters could possibly be. This difference influences the outputs of the model so that the prediction is no longer a single prediction value per data point like non-Bayesian methods.

Consequently, fitting Bayesian models produces a distribution on the range where the prediction value is most likely to be with the measure of central tendency as its 'best bet' on a single prediction value. For example, a Bayesian prediction can indicate that the habitat suitability score for a certain location lies between 0.5 and 0.8 90% of the time, but its best single score estimate is the distribution's mean, 0.7. In Bayesian terminology, the former range is called the 'credible interval' and the latter single estimate is called the 'point estimate'.

Parameters within a Bayesian model have initial distributions that are updated with data. These initial distributions are called 'priors' and are user-defined probability distributions that reflect the user's knowledge of the parameters' ranges. Priors can be non-informative, weakly informative, or informative, depending on how certain the user is. Non-informative priors give no human input on the parameter values. Weakly informative priors give just enough input for the model to avoid unrealistic parameter values. Informative priors give a human's estimate of how they think input values should influence the outcome, and are often created from a domain expert's opinion. These priors are used as starting points for the models to find the true parameter distributions, and for models using informative priors, enables models to give feasible predictions in the absence of data.

The process of updating the priors with data, or in other words the process of fitting Bayesian models, is called 'inference' or 'Bayesian inference'. When denoting the vector of parameters as θ and the data as $\mathbf{y}$, this process can be expressed with the Bayes theorem as

$$P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta)}{P(\mathbf{y})}, \tag{2}$$

where $P(\theta)$ is the prior and $P(\theta|\mathbf{y})$ is the prior after it is updated with the data, or in Bayesian terms, the 'posterior'. $P(\mathbf{y}|\theta)$ is called the 'likelihood', and it

is a probability expressing how probable it is that the data was generated by the prior. $P(\mathbf{y})$ is called the 'evidence', and it is the probability of observing the data unconditionally of θ. Theoretically, the evidence is calculated by summing values of $P(\mathbf{y}|\theta)$ across all possible values of θ, but in practice it is indirectly computed through inference algorithms.

There are many ways to do inference, but the two most popular methods are Markov chain Monte Carlo (MCMC) and variational inference. MCMC is a method that simulates a large set of draws directly from the posterior. Variational inference is a method that finds a rough but fast approximation to the posterior distribution. It is used in models where MCMC would be too computationally intensive.

Inference with Markov chain Monte Carlo

This thesis uses a branch of MCMC known as Hamiltonian Monte Carlo (HMC), specifically, the dynamic HMC implemented in the Stan package [26]. HMC and MCMC in general is a computation-heavy but highly efficient simulation method that is utilized when sampling from continuous distributions.

MCMC is a method that approximates the posterior with random simulations in probabilistic space when the posterior distribution is difficult to compute analytically. It sequentially generates random draws through a Markov chain that become ever closer to the true distribution as the process progresses. The distribution of these draws is then used as a proxy to the true distribution $P(\theta|\mathbf{y})$.

Since MCMC generates a chain of samples beginning from a random point, the initial draws from the chain are strongly affected by the starting location. Moreover, in Stan's dynamic HMC, these initial draws are used to automatically test and adjust the settings of the Markov chain simulation, making them even less representative of the true distribution. Therefore, draws from the beginning of the chain are discarded as 'warm-up draws'. Additionally, residual influence of the starting points can be reduced, for example, by using multiple Markov chains instead of one, each with different starting locations. This can also aid in reducing computation time by splitting the chains across multiple cores.

Markov chains that have run long enough in MCMC eventually 'converge' to have the posterior distribution as their stationary distribution. When evaluating whether the chains have converged, users often refer to the trace plot or the $\widehat{R}$ metric proposed by Vehtari et al. [1]. A trace plot, displayed in Figure 3, is a visual representation of the Markov chains overlaid on top of each other. Users may evaluate convergence through these plots by visually examining whether the chains have mixed and whether they concentrate evenly around a range of values. $\widehat{R}$ is a numerical value that measures convergence using the within-chain and between-chain variance. An $\widehat{R}$ value under 1.01 implies that the chains have converged.

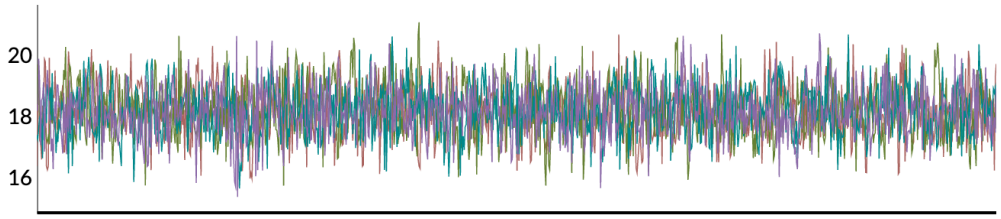


Figure 3: An example of a well-behaved trace plot output from the ShinyStan package [27].

2.3 Bayesian Workflow

Bayesian workflow refers to the practical process of finding a Bayesian model that best fits the modeling objective. This involves choosing and designing a model, fitting it, diagnosing problems, evaluating, modifying a model, and exploring other possible models. In publications it is usually introduced as a guideline or a case study.

In their publication, Gelman et al. give a thorough walkthrough of the modeling process, including example cases and common mistakes [10]. They describe the workflow in Figure 4 as an iterative process of picking one candidate model, fitting and troubleshooting it, evaluating, and then either adjusting the original model or switching to another model. Unlike the linear workflow described in textbooks, they promote exploring many different models to find the one that best fits the data.

Until recently, there were very few resources that describe this workflow. Consequently, this limited access to Bayesian modeling. Only statisticians who already had a background in Bayesian statistics were able to navigate the process. Lately, however, there have been efforts to explain this process to people in other fields. Most of these are still provided through academic publications or modeling tool providers, but there is an increasing number of casual blog posts and articles available. Most likely this trend will continue and there will be more applications of Bayesian models in the future.

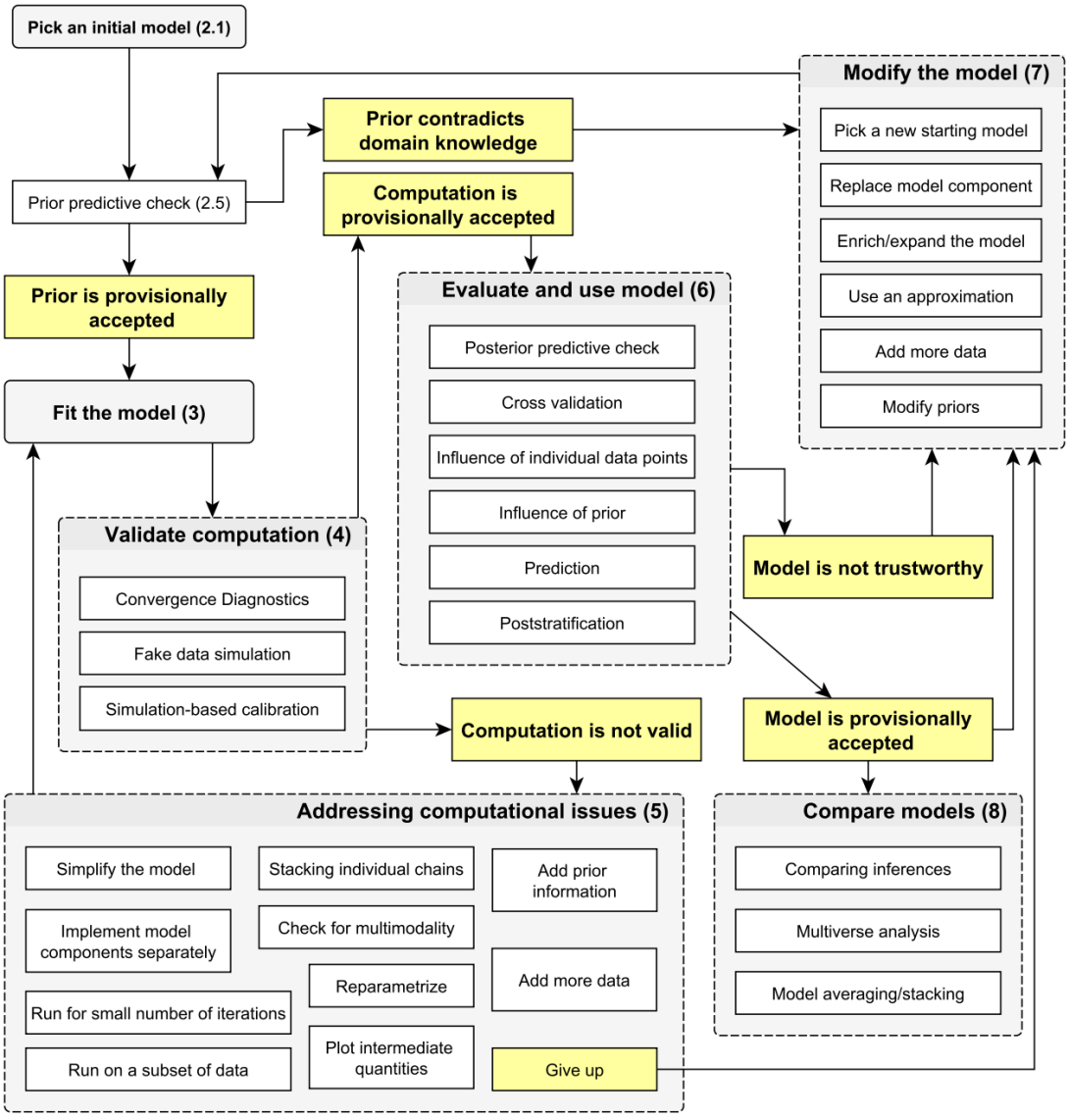


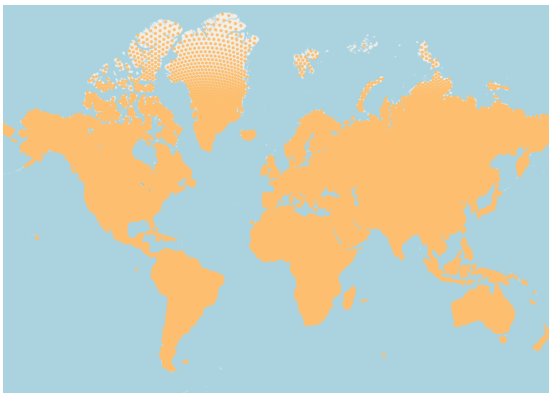
Figure 4: The Bayesian workflow diagram as shown in Gelman et al.'s publication [10].

3 The Data

The models used in this research use climate data as input features and estimated species range data for the labels. These are from three different sources, two for the climate and one for the species distribution. All features and labels used for the SDMs are mapped to a regular polygon grid that covers the Earth's surface.

The training dataset for all models are present-day features and labels in areas for which we have range information or areas for which pseudo-absence labels were assigned. Details can be found in the Species Distribution section. The target prediction area is all land areas on Earth in present-day and future conditions. Excluding the present-day areas that we used for training, these do not have label assignments. The training and target prediction areas are visualized in Figures 5a and 5b respectively.

(a) The training area with the present natural range from the PHYLACINE database (positives) [28] in yellow and pseudo-absence areas (negatives) in blue.



(b) The test area in light orange.

Figure 5: Training area and target prediction areas in this thesis visualized as centroids of grids. The centroids are visible as discrete points only around the arctic circle since we used a high-resolution grid.

3.1 The Discrete Global Grid System

A discrete global grid system is a set of regions that partition the Earth's surface into smaller subsections. In other words, it is the 'grid' that the climate and species distribution data are mapped to.

The most common discrete global grid system is the square grid system based on latitudes and longitudes, but for building our species distribution models we used an equal-area hexagonal grid system. While not as popular as square grid systems, this type of grid can partition a spherical surface into equal-sized cells with minimal error. As a result, these have a statistical advantage compared to other grid types.

The exact grid system that is used in our data is ISEA3H09, a family of the ISEA3H grid system shown in Figure 6. ISEA3H09 has fine grid cells where the

distance between the centroids of the hexagons are approximately 50 kilometers at any given region on Earth.

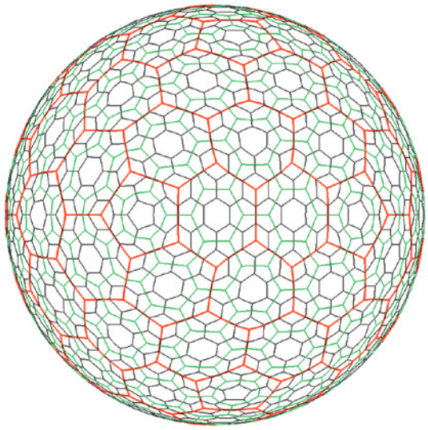


Figure 6: The ISEA3H discrete global grid system as shown in Sahr et al.'s article [29].

3.2 Climate Datasets

The climate dataset used in the thesis comprises the climate extreme indices from Coupled Model Intercomparison Project Phase 5 (CMIP5) and bioclimatic variables from WorldClim (available from https://climate-modelling.canada.ca/data/climdex/climdex.shtml and https://www.worldclim.org/). From both sources we chose the average data from the years 1950-2000 to represent the present-day climate and the average predicted data for the years 2061-2080 to represent the expected climate in 2070.

In CMIP5, multiple climate models were used in an ensemble to predict climate extreme indices defined by the Expert Team on Climate Change Detection and Indices (ETCCDI) [30] [31]. Out of the component models, we chose present-day and future outputs from the Community Climate System Model Version 4 (CCSM4) [20] for this research.

The present-day WorldClim bioclimatic variables are high-resolution weather station data interpolated using thin-plate smoothing splines [21]. The future bioclimatic variables are downscaled outputs from CCSM4 adjusted using the present-day, interpolated weather station data as the baseline climate. While an updated version was available [22], the older version of WorldClim, WorldClim 1.4, was selected so that the source climate model (CCSM4) matches the ETCCDI climate extreme indices.

For all future prediction data we selected the representative concentration pathway 8.5 (RCP8.5) scenario , also known as the 'business as usual' scenario [32]. 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 present the worst case scenario of global warming.

Distributions of the Climate Data

One important characteristic to note about the climate data is the difference in values between the training data and the test data (future data and data from present-day areas not used for training). Some climate features had very different training data and test data distributions, suggesting that finding suitable habitats for elephants would be an extreme extrapolation problem. Examples of the difference between the distributions are shown in Figure 7.

Another characteristic to note is the skew of the data distributions. For some features, either the training or test data was skewed in one direction. This can be observed in Figure 7 for the future TN10P values and in the distributions of RX1DAY.

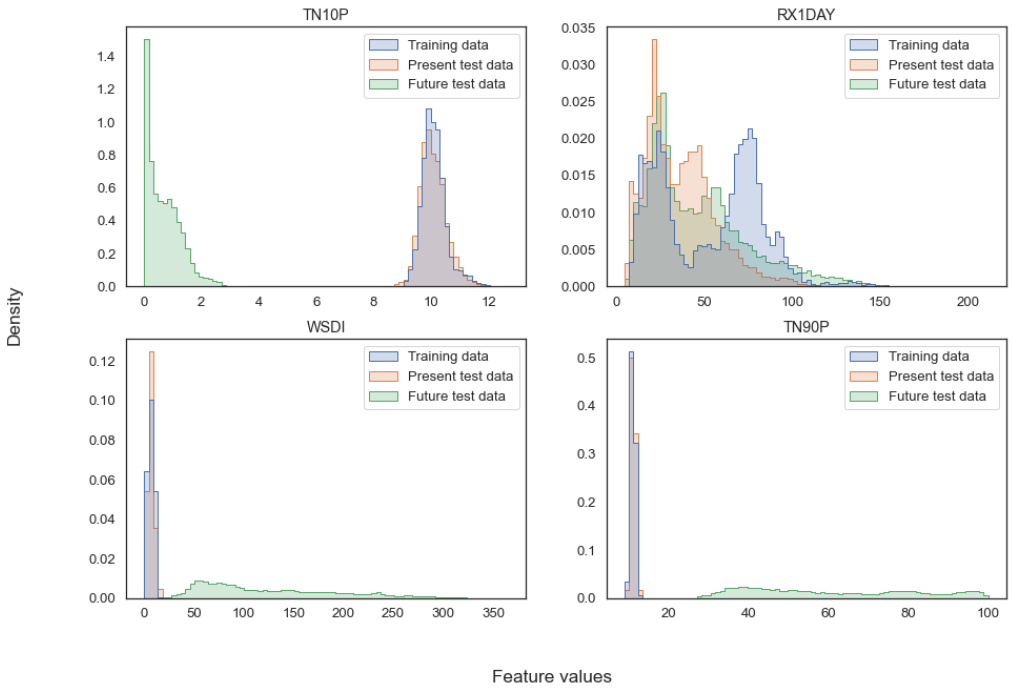


Figure 7: Some examples of the difference between distributions of the training and test features. The definitions of the features can be found in Tables 1 and 2.

3.3 Species Distribution

The species distribution data in this research is the Asian elephant range data from the PHYLACINE database [28] with an additional unsuitable habitat area designed by Michael Mechenich, one of the advisors of this thesis [33].

PHYLACINE is a collection of phylogenies, multiple types of range data, traits, and threat status for mammals that have lived since the last interglacial period. The range data we chose from PHYLACINE is the 'present natural range', the yellow

area in Figure 5a that represents areas that are estimated to have been habitats for Asian elephants if not for human activity. The grid cells included within this range are labeled as positives, or areas where Asian elephants are able to survive.

Since the models in this thesis also required unsuitable habitat areas as examples for training, there was a need to add a synthetic area where elephants could not survive. This 'pseudo-absence' area was designed as the land area that elephants could have migrated to but were not included in the PHYLACINE range. As shown in Figure 5a, this is the land area surrounding the present natural range that excludes high mountain ranges where elephants cannot stay. The grid cells included within this range are labeled as negatives, or areas where Asian elephants cannot survive. The pseudo-absence area was constructed so that the number of cells with negative labels is approximately the same as the number of cells with positive labels. After adding the pseudo-absence area, the total grid cells with habitat suitability labels were 7,331, 3,765 of which were positive.

4 Methods

This section introduces the models, model design choices, and evaluation methods used during the iterative model building workflow of this research.

The model building process begins with two non-Bayesian models, logistic regression and random forest. These are used as baselines for two Bayesian models, Bayesian GLM and Bayesian GAM. The version of Bayesian GLM used in this project refers to the direct Bayesian counterpart of logistic regression. Bayesian GAM is not a Bayesian complement of random forest, but is a non-linear model of comparable complexity. The details of each model are explained within the subsections of this section.

Aside from the model types, we explored various modeling options listed below and their effects on the prediction outcome.

- Feature selection
- Feature scaling
- Priors (Bayesian models only)
- Basis dimension (Bayesian GAM only)

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 [39]. For numerical scores we present AUC and maximum attainable TSS, along with sensitivity and specificity values to make the interpretation of the TSS score easier.

4.1 Logistic Regression

Logistic regression is a machine learning algorithm that is used to predict binary outcomes. Its outputs are either designed to be binary labels or a probability of the outcome being positive. For the baseline model, we used the glm function of R's stat package and set the output to be the probability of a grid cell being a suitable habitat for Asian elephants.

In essence, logistic regression is a version of linear regression. If we denote the probability of a data point being positive to be $P(y = 1)$, and the input features as $x_1,\ x_2,\ \cdots,\ x_k$, it is fitting a linear hyperplane as shown below

$$\log \frac{P(y=1)}{1-P(y=1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{3}$$

where the coefficients and intercept $\beta_0,\ \beta_1,\ \cdots,\ \beta_k$ are the parameters for the model's algorithm to find. The response variable, $\log(P(y = 1)/(1 - P(y = 1)))$, is called the 'log-odds', and it is an intermediate value that needs to be converted to the final output.

To transform the log-odds into a probability, logistic regression uses the inverse logit function, substituting the log-odds in place of z shown below. The end result would be $P(y = 1)$.

$$f(z) = \frac{1}{1 - e^{-z}} \tag{4}$$

To sum up, logistic regression is simply an algorithm to find a set of coefficients and intercept $\beta_0,\ \beta_1,\ \cdots,\ \beta_k$ that best fits the probability of the outcome being positive according to the examples given in the training data. Equations 3 and 4 can be summarized as follows as one equation that the model needs to fit.

$$P(y = 1) = \frac{1}{1 - \exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)\}} \tag{5}$$

When one needs probabilities as the prediction, the output of Equation 5 can be used directly. If binary predictions are required, The user can specify a threshold value between 0 and 1 that classifies the predictions as positive or negative. Outputs predicted above the value are classified as positives and others as negatives. The most simple threshold value is 0.5, instructing the model to classify outputs as positive when the probability of being positive is over 50%.

4.2 Random Forest

Random forest is a relatively complex non-linear machine learning model. It is an ensemble of multiple simpler models, decision trees, and their outputs are aggregated at the end to produce the final prediction result.

As its name implies, decision trees are models of tree-shaped, flowchart-like structures. They represent a series of sequential decisions where each node of the tree is a binary test made on the attributes (e.g., temperature above 25 degrees or not) to reach an output similar to the training data. In the fitting phase, a decision tree machine learning model identifies the optimal nodes and the conditions of their tests that best describe the data. Once trained, decision trees can provide complex predictions but is prone to overfitting, especially when the tree is large.

Random forest is a collection of small decision trees, each trained on a random subset of the input features. When predicting, the outputs from the decision trees are either averaged or 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. A diagram of the algorithm is presented in Figure 8.

The baseline random forest models of this thesis were designed to match the implementation of a previous research by Mechenich [33]. We used the average predictions of 100 random forests with 100 decision trees each for the final prediction outputs. The extra layer of averaging was added to mitigate the variations in prediction values caused by random seed selection.

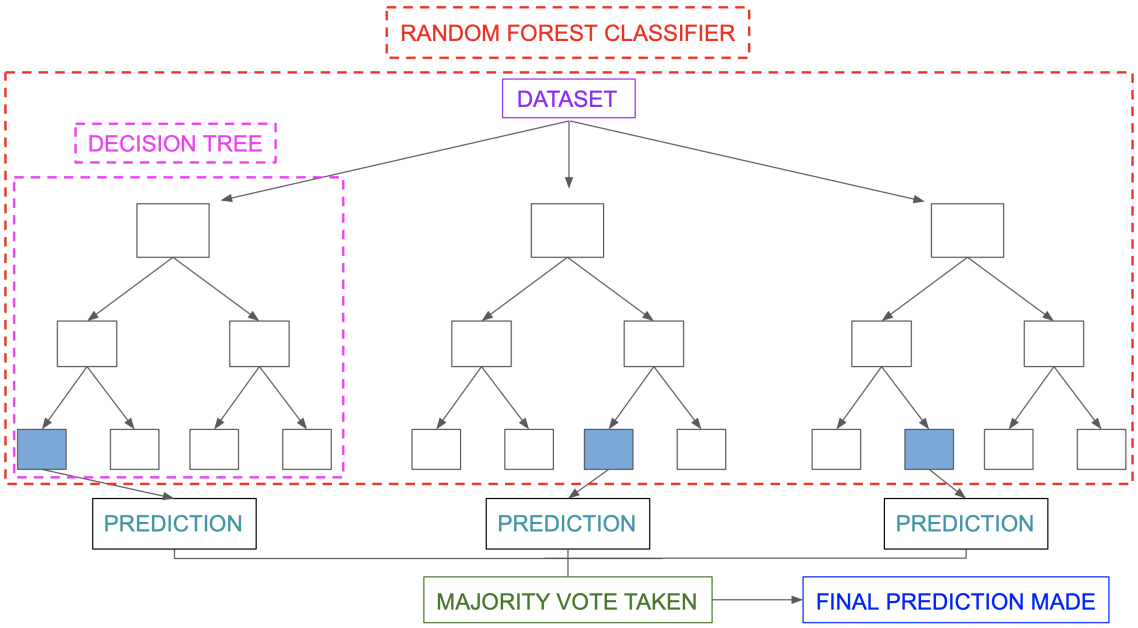


Figure 8: A visualization of the random forest algorithm from an article from RPubs [40].

4.3 Bayesian GLM

Bayesian GLM refers to a series of multiple versions of Bayesian linear regression that uses intermediate functions and transformations to model complex relationships between the input variables and outcome. The version that we used has exactly the same model structure and equations as logistic regression. The only difference is the coefficients and intercept of Equation 3, which are distributions of possible values instead of single estimates. In MCMC, these are histograms of the draws from the posterior as shown in Figure 9.

When making predictions, sets of coefficients and intercepts are drawn from these histograms to create multiple versions of the regression hyperplane. These create a distribution of multiple realizations of predictions from which the point prediction and credible interval are determined.

All predictions within this thesis are made with 500 draws from the posteriors. 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 priors we experimented with are introduced later in this section.

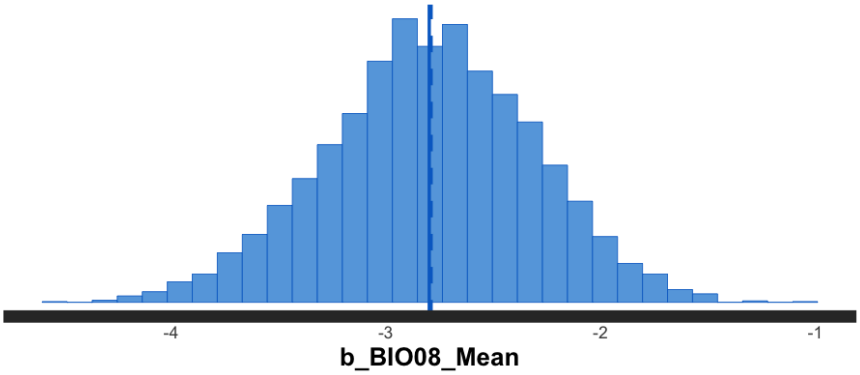


Figure 9: Example histogram of posterior draws of a parameter from the ShinyStan package [27].

4.4 Bayesian GAM

Bayesian GAM and has a very similar structure as Bayesian GLM and logistic regression, but has a different approach for modeling the log-odds. As shown below, its log-odds equation uses flexible functions of the input features $s_1(x_1),\ s_2(x_2),\ \cdots,\ s_k(x_k)$ instead of using the features directly. The rest of the process is the same as Bayesian GLM, substituting the log-odds in Equation 4 to obtain the predicted probability distributions.

$$\log \frac{P(y=1)}{1-P(y=1)} = \beta_0 + \beta_1 s_1(x_1) + \beta_2 s_2(x_2) + \cdots + \beta_k s_k(x_k), \tag{6}$$

In the brms implementation of Bayesian GAMs, users are able to set priors for the distributions of the coefficients, intercept, and an additional parameter for each flexible function that determines their non-linearity. For the flexible functions we chose smooths using thin-plate regression spline [41] as a basis, which are explained in the Modeling Options subsection.

As in Bayesian GLM, our predictions are from 500 draws from the posteriors, reported using medians as point predictions accompanied with the IQR of the distributions. Different priors and complexity settings were tested for this model, which are presented in later sections.

4.5 Modeling Options

Feature Selection

Feature selection plays a crucial role in any machine learning or statistical model. To test the influence of feature selection on our models, we used two subsets of features chosen in Mechenich's research [33]. These we name 'random-CV features' and 'spatial-CV features' in this thesis after the cross-validation (CV) method in which they were selected. The lists of features and their short descriptions are presented in Table 1 and Table 2.

Both random-CV features and spatial-CV features were selected through performance evaluations for a logistic regression model. Random-CV features were selected through a typical CV where data points were randomly assigned to a fold. Spatial-CV features were selected through a CV where the folds are divided by geographical locations as shown in Figure 10. The Results for both CV methods were evaluated with multiple fold patterns to obtain the final feature set.

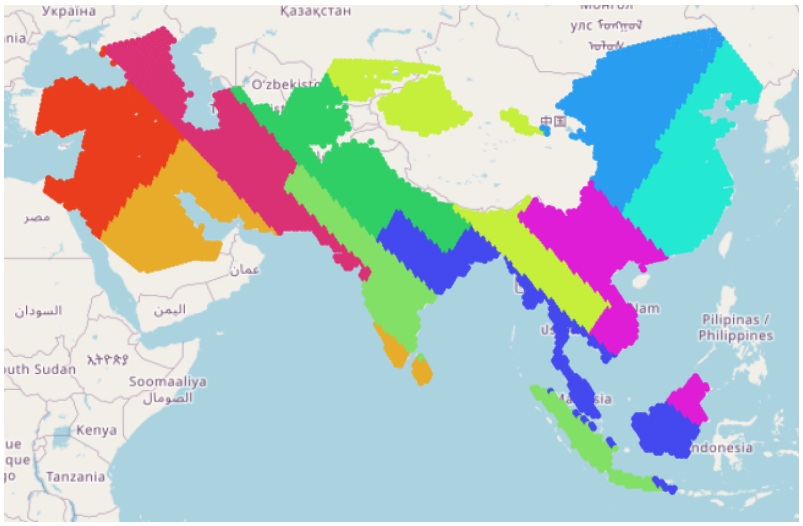


Figure 10: One spatial-CV fold pattern used in feature selection. This particular fold was also utilized in the spatial-CV evaluation for the models in this thesis.

Feature Scaling

For our dataset, feature scaling refers to preprocessing the input features so that it has range [0,1] but retains the original shape of its distribution. This is done through min-max scaling, shown below for an arbitrary feature x.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{7}$$

Scaling features is known to improve prediction performance for machine learning models, but for Bayesian models, it has an additional effect. Shrinking or expanding the range of the features will change its relative power against the priors. This will make the likelihood interact differently with the priors and may affect the prediction results.

Therefore, we tried both raw and scaled versions of features when modeling to see the scaling's effects. This means that there are four feature sets to fit on per model type (raw random-CV features, scaled random-CV features, raw spatial-CV features, and scaled spatial-CV features).

Priors

When combining the priors with the data via the likelihood, the choice of the priors can influence the prediction outcome. A tighter prior regulates the posterior from following the likelihood too closely. But if a prior is too restrictive, it prevents the posterior from reflecting true patterns in the data. For these reasons, we needed to be careful when choosing priors.

When modeling the Bayesian models for the thesis, we started with default weakly informative priors. Then for the Bayesian GLM models, we tested whether the priors are too strong against the likelihood with the priorsense package [42], which automates the diagnosis by using slightly perturbed versions of the prior and likelihood distributions. After obtaining the diagnosis, we adjusted the scales of the priors according to the outputs.

For the GAM models, however, the process was not as simple. We attempted to use priorsense for these as well, but for every model, importance sampling, the backbone method of priorsense, failed when testing the likelihood. The tool did produce outputs that indicated extremely high likelihood sensitivity, but since the values were most likely untrustworthy, we decided not to rely on priorsense for these models. Instead, we used the traditional method of testing different priors and examining the outputs.

The list of initial and final priors used for Bayesian GLM and Bayesian GAM can be found on Table 3.

Basis Dimension

The smooth functions we chose in Equation 6 for Bayesian GAM uses an eigende-composed approximation of a thin-plate regression spline.

The original thin-plate regression spline is a penalized, piecewise non-linear function composed of smaller non-linear segments connected together at each data point (or a subset of them). This regression method allows users to model highly non-linear patterns while adjusting the penalization to avoid overfitting. However, a piecewise regression function requires substantial computing power to handle, which is why the version implemented in brms uses an approximation.

The approximation of the thin-plate regression 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 basis dimension is automatically adjusted. However, the user can set an upper bound on it through an argument of the smoothing function. We tried adjusting this upper bound in addition to the prior distribution settings to restrict the complexity of the smooth functions in Bayesian GAM. The initial and final values of k are shown with the priors in Table 3.

4.6 Details of Quantitative Evaluation

Calculating Numerical Scores

For evaluation, 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 5a were compared to their true labels to compute the values.

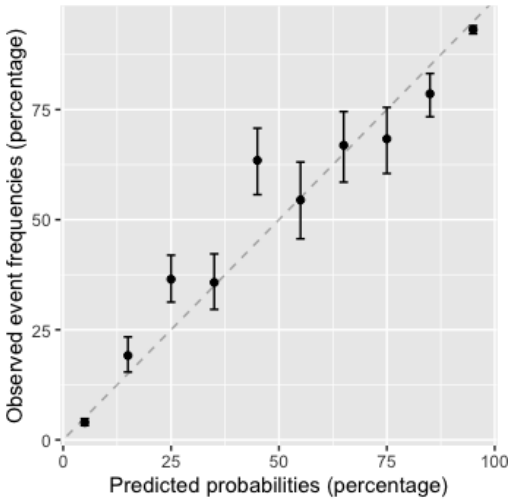
The validation scores of the models were calculated through a 10-fold cross-validation using one of the fold patterns of spatial-CV feature selection shown in Figure 10. The scores are the average of the 10 validation folds.

Calibration Plots

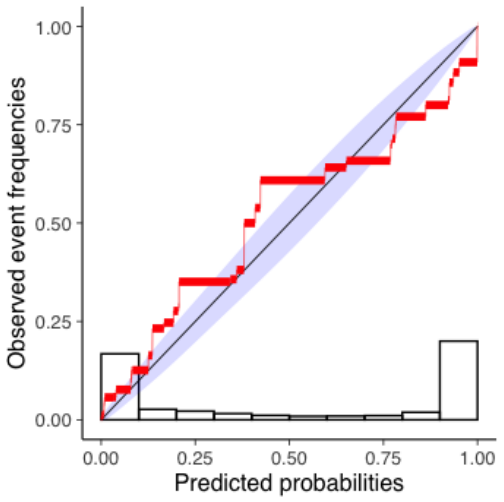
Calibration plots or reliability diagrams are visualizations that diagnose whether the outputs of a model are well calibrated probabilities. This is necessary because even though logistic regression, GAM, and other models are mathematically structured to model probabilities, they cannot guarantee that their outputs are well-calibrated probabilities, especially when the model overfits or underfits the data.

The traditional 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.

In this thesis we present the traditional calibration plot from the caret package [43] and a new version by Dimitriadis et al. [44]. This new method uses non-parametric



(a) A traditional calibration plot. In this implementation of R's caret package, the calibrations are plotted with uncertainty intervals around the observed proportion of positives. The x-axis and y-axis are plotted in percentages.



(b) The new calibration plot from Dimitriadis et al.'s package [44]. In this version, the observed proportion of positives are plotted as uneven horizontal line segments. The shaded area along the diagonal can be interpreted as a confidence interval of a perfectly calibrated plot. The histogram at the bottom shows the distribution of predicted values.The x-axis and y-axis are plotted in decimals.

Figure 11: Examples of calibration plots shown in this thesis.

isotonic regression to overcome the weakness of traditional calibration plots where bin selection can influence the plot's interpretation. It automatically creates a calibration curve for a given prediction outcome, therefore eliminating the need for bin settings. Examples of the two calibration plots are shown in Figures 11a and 11b.

Both the calibrations for the training data and the validation folds are shown in the Results section. The calibration for the validation folds were created by appending the predictions for all folds together to approximate how calibrated the model can be when extrapolating.

4.7 Details of Qualitative Evaluation

When examining the predictions on QGIS, we mainly focused on the point predictions since most of the IQR outputs did not give revealing input. We based our decision of what is a realistic and convincing output on our knowledge of present-day climate and whether that is likely a suitable habitat for both present-day and future conditions. Since 2070 is less than fifty years away, we assumed that there would not be sufficient time for vegetation to change dramatically in any given area. For example, even if the predicted future climate suggests that a former tundra or desert can turn into forests in 2070, we would not expect a forest to be growing there. This assumption discredits models that predict areas such as the far north, the Sahara, and the Arabian peninsula as suitable habitats for Asian elephants.

For models that had convincing outputs, we did another qualitative evaluation by comparing the predictions to former habitats of extinct elephant relatives. These are dispersed over multiple land areas. North America was once home to multiple elephant relatives and ancestors [34]. Other close relatives were found near the Tigris-Euphrates river system [35]. Still other relative species were discovered in Africa, Europe, and a wide area of South America [36] [38] [37]. These ranges may still have an unfulfilled niche in the ecosystem that Asian elephants can fill, and therefore models that predict suitable habitats for these areas will have additional credibility.

Table 1: List of random-CV features.

Feature name	**Source**	**Definition**
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.
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 maximum temperature is larger than the 90th percentile of historical data.

Table 2: List of spatial-CV features.

Feature name	**Source**	**Definition**
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.
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 maximum temperature is larger than the 90th percentile of historical data.

Table 3: Initial and final adjusted prior and basis dimension (for GAM) settings for Bayesian models. Settings that are not changed from the initial set are denoted by '-'. 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. When setting it to 1 in brms, it forces the splines to use the least possible basis dimension value.

Model type	**Feature set**	**Scaling**	**Setting type**		**Initial**	**After adjustment**
GLM	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)

5 Sampling Settings and Posteriors

5.1 Sampling Settings

All Bayesian models in this project used dynamic HMC in Stan through the brms package for inference. The sampling specifications used for the models are presented in Table 4.

Table 4: Sampling settings used in Bayesian models.

Model type	GLM	GAM
Total iterations	2,000	3,000
Warm-up draws	1,000	1,500
Chains	4	4
Target average acceptance probability	0.99	0.99
Maximum tree depth	11 (raw features) 10 (scaled features)	13

The sampling settings for brms and Stan are interpreted as follows. Total iterations is the number of simulated parameter values per chain including the warm-up draws. For example, the Bayesian GLM models allocated 1,000 draws out of 2,000 to the warm-up, yielding 1,000 post-warmup values for one chain. This was done for four chains, producing a total of 4,000 values after the warm-up phase. The Bayesian GLM models used these 4,000 draws for inference. Similarly, our Bayesian GAM models used 6,000 post-warmup draws for parameter estimation and prediction. These settings were chosen to keep the $\widehat{R}$ value under 1.01.

The target average acceptance probability is a value between 0 and 1, and increasing the value will indirectly make the HMC simulation more fine-grained. We increased its value to reduce divergent transitions, which are undesirable occurrences where the simulation does not match its theoretical state. For the inference in this project, increasing the target average acceptance probability to 0.99 was enough to eliminate divergences.

Maximum tree depth is a setting specific to the type of HMC we used, which is a setting that adjusts the trade-off between the simulation runtime and sampling efficiency. A larger value allows the algorithm to generate a set of values that are representative of the posterior with fewer draws but will also increase the computation time. In brms and Stan, a warning is issued whenever the sampling method sacrificed sampling efficiency in order to decrease runtime. Since the default setting resulted in too many of these warnings for most models, especially for those fit on raw features, we adjusted the settings until the number of maximum tree depth warnings were under 5% of the full sample size excluding warm-up draws.

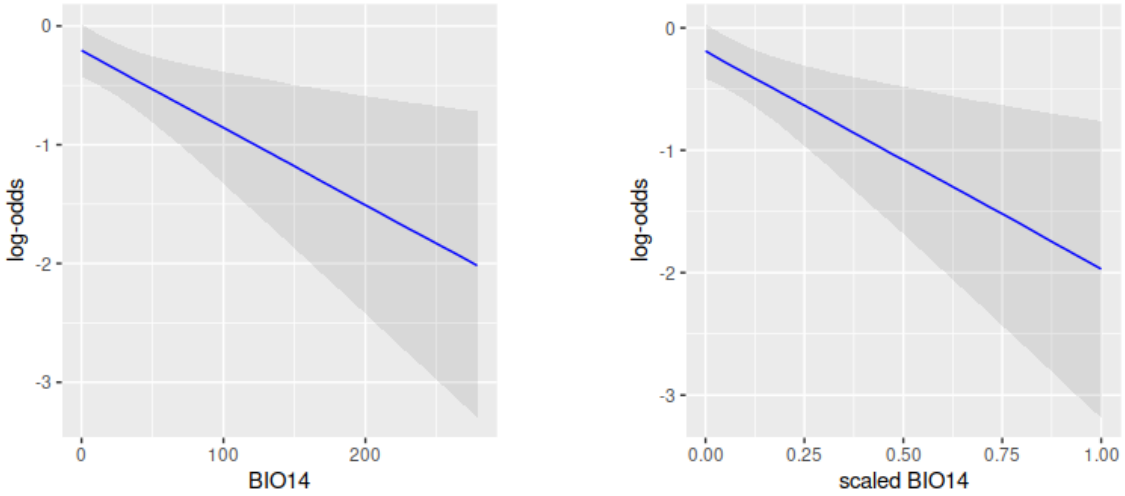
The overview of the inference results are presented in the following sections. The summary statistics of the posteriors for the models listed in Table 3 can be found in

the Appendix.

5.2 Posteriors of Bayesian GLM

For Bayesian GLM, scaling the features had the largest influence on the posterior distributions. When modeling with raw features, every posterior except the intercept had a tight distribution very close to zero. When scaling the features before modeling, all posteriors became distributed evenly around the range [-20, 20]. We did not observe this kind of effect when making the priors wider, which did not change the posteriors significantly. The effects of scaling the features and making the priors wider is shown in Figure 13.

It seems that scaling had such a significant effect on the posteriors because the model created similar hyperplanes for Equation 3 regardless of scaling. As an example we plot the conditional effects of the feature BIO14 on the log-odds on one model while holding the other features constant at their means in Figure 12. The plots are very similar whether we fit the raw feature with range [0, 278.7] or the scaled feature with range [0,1]. This indicates that when doing inference for the raw features, the model created a posterior that has nearly the same response as the scaled features when multiplied with the feature values, consequently making the coefficients for the raw features concentrated tightly around smaller values.



(a) The conditional effect of raw BIO14.

(b) The conditional effect of scaled BIO14.

Figure 12: Conditional effects of raw and scaled BIO14 on the log-odds for the models fit on random-CV features. The lines are very similar whether the feature is scaled or not.

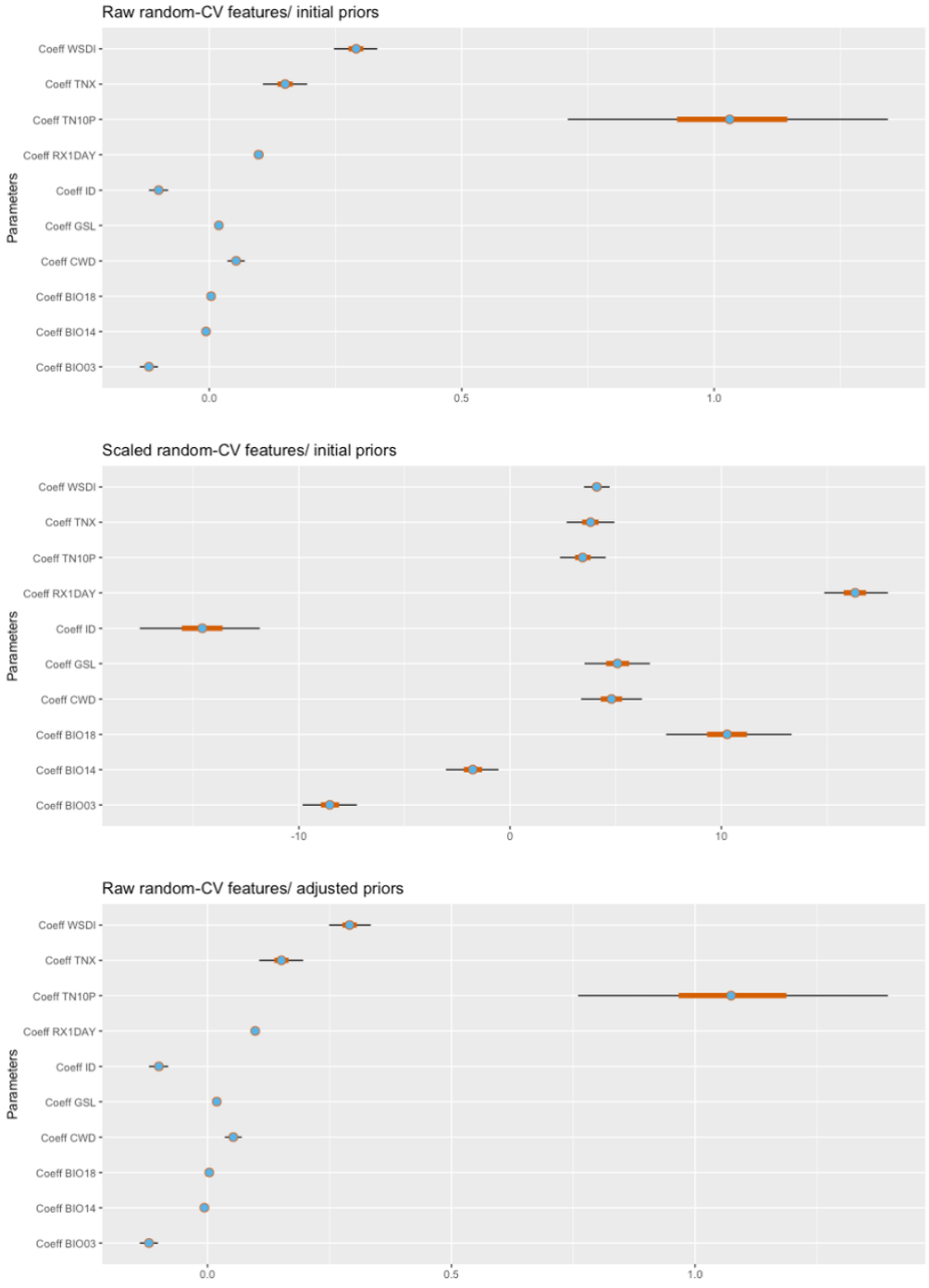


Figure 13: The effects of scaling and changing priors on the posteriors for the Bayesian GLM model using random-CV features. The pale blue point is the median, the thick orange line the 50% credible interval, and the thin black line the 95% credible interval. The abbreviation Coeff is short for coefficient. The posteriors for the intercepts are omitted so they do not dwarf the other distributions in the top and bottom plots. Top: Raw features with the initial prior set. Middle: Scaled features with the initial prior set. Bottom: Raw features with the adjusted, wider prior set.

5.3 Posteriors of Bayesian GAM

For GAM, the factor that most influenced the prediction results was the non-linearity parameter. This parameter controls the non-linearity, or wiggliness, of the flexible functions in Equation 6. The farther away from zero this parameter is, the more non-linear the functions get.

When using the initial prior setting, a Student's t-distribution centered around zero, some posteriors of the non-linearity parameter were in unrealistically high ranges. As shown in Figure 15, the posterior of BIO18 overshoots 100, which is suspiciously too non-linear. Unlike Bayesian GLM, scaling the parameters did not have an effect on the posteriors.

The other parameters, the coefficients and intercept of Equation 6, seemed to be more of a supplement that models whatever the non-linearity parameter could not express. Because of that, the coefficients often changed their polarity when restricting the non-linearity. This is a behavior that could not be observed for Bayesian GLM, whose posteriors did not change their polarity when scaling or changing the priors.

Additionally, when there is a tight restriction on the non-linearity, changing the prior distributions of the coefficients and intercept did not have an effect on the posterior. This can be observed in the Appendix section on Tables A5, A6, A7, and A8.

Also, when checking for convergence, we observed that some posteriors for GAM are skewed. This could also be observed in the trace plots for which we show an example on Figure 14. This might mean that the posteriors have not converged yet, but since the inference process for the more complex Bayesian GAM models failed to run in reasonable time, we could not confirm the effects of increasing the number of iterations beyond 3,000.

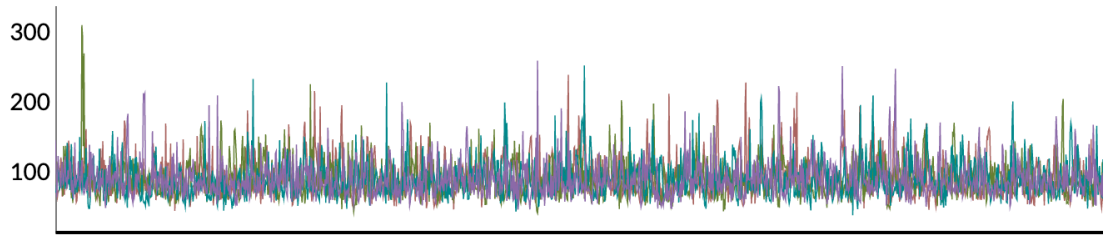


Figure 14: One of the skewed trace plots observed in Bayesian GAM. Unlike the well-behaved trace plot in Figure 3, it has traces protruding above.

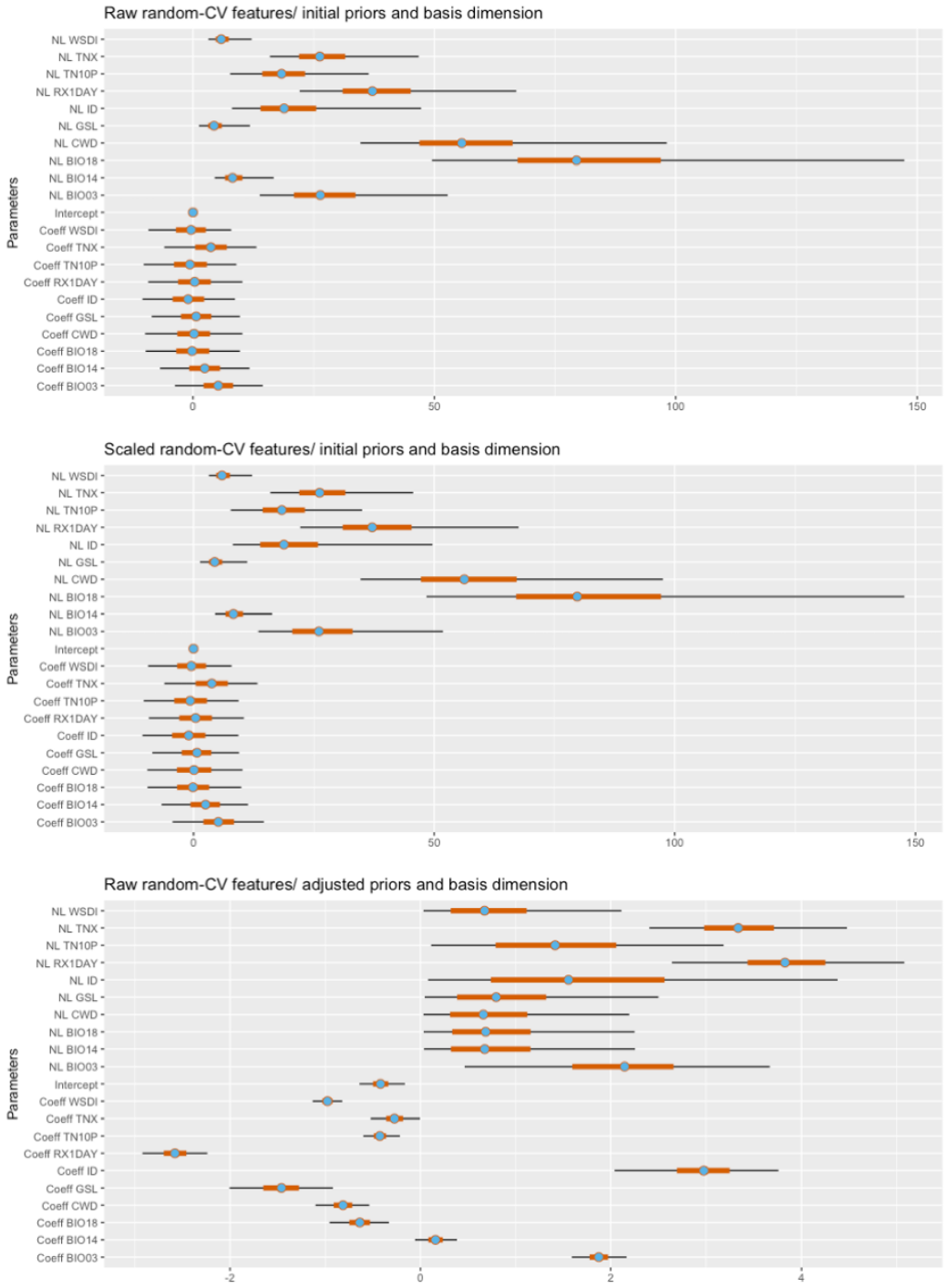


Figure 15: The effects of scaling and changing the priors and basis dimension on the posteriors for the Bayesian GAM model using random-CV features. The abbreviation NL is short for non-linearity. Top: Raw features with the initial priors and basis dimension. Middle: Scaled features with the initial priors and basis dimension. Bottom: Raw features with the adjusted, tighter priors and basis dimension.

6 Results

This section presents the outputs of the models: logistic regression, random forest, Bayesian GLM, and Bayesian GAM. Since there is limited space available, not every result is presented within this section. For the full results, we refer readers to Figure 1 and the Appendix section.

6.1 Logistic Regression

The only modeling option that had an impact on the logistic regression models was feature selection. Scaling the features had no effect on the outcome and the predicted values remained the same as the raw features. Therefore, the results for the scaled features are not presented in this thesis.

Both sets of features had high numerical scores on both the training data and validation folds. The calibration plots, however, showed that using random-CV features makes the models ill-calibrated compared to spatial-CV features. These quantitative evaluation results are shown on Table 5 and Figure 16.

When predicting for the present-day climate, both feature sets gave similar predictions, though models fit on spatial-CV features were slightly more pessimistic. However, the feature sets had a significant effect on the future predictions.

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 models fit on spatial-CV features were more conservative, either expanding or contracting suitable areas that they had already been predicted as suitable habitats for present-day conditions. These predictions are presented in Figure 17.

Table 5: The scores for the logistic regression models fit on raw feature values. As defined in the Feature Selection subsection, random-CV features are features selected based on the performance on conventional CV and spatial-CV features are features selected based on the performance on a geographically divided CV.

Score type	**Random-CV features**		**Spatial-CV features**	
	Training	**Validation**	**Training**	**Validation**
AUC	0.98	0.92	0.97	0.91
TSS	0.85	0.79	0.84	0.77
sensitivity	0.94	0.86	0.94	0.89
specificity	0.90	0.93	0.90	0.88

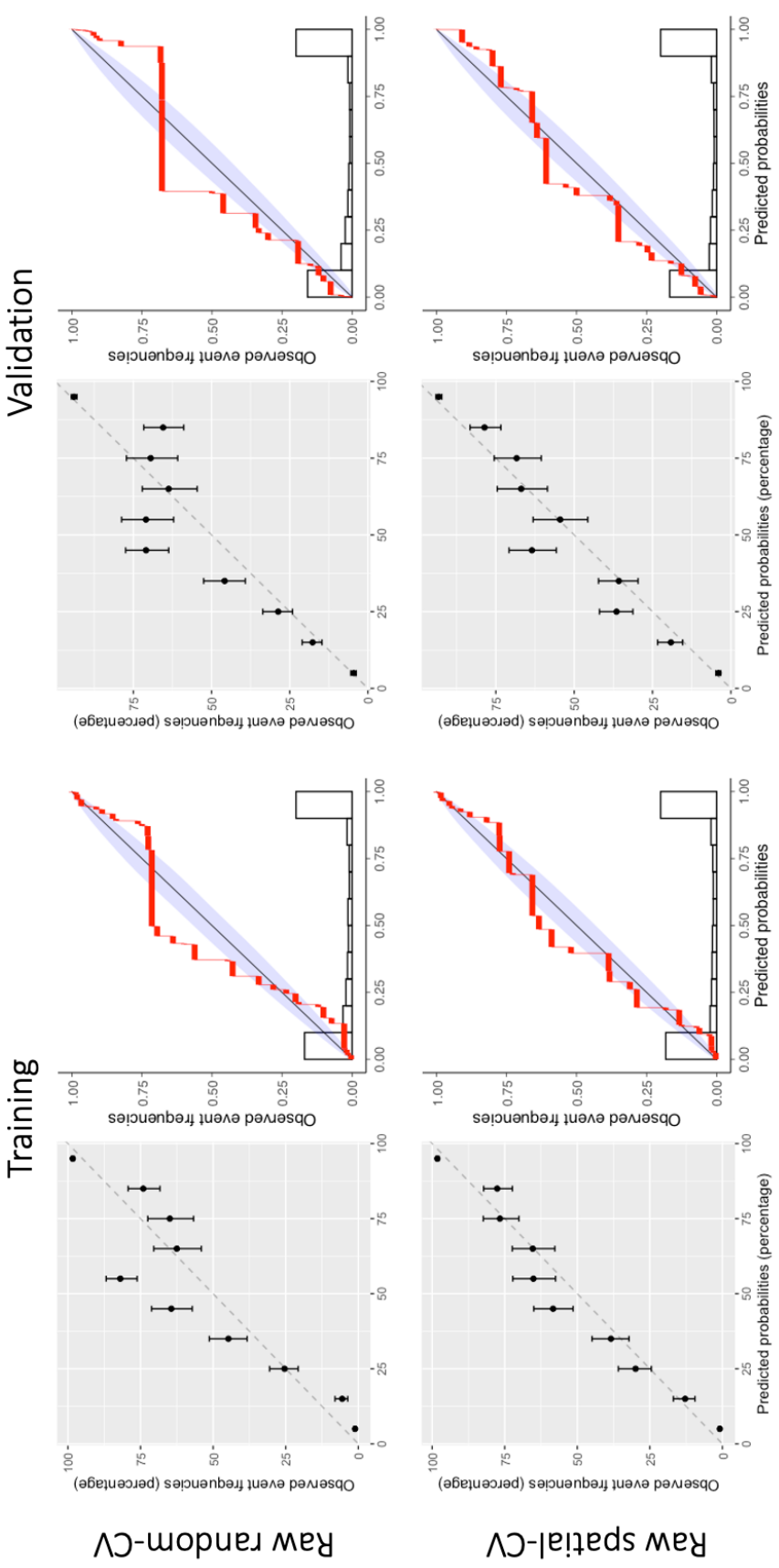


Figure 16: Top row: Calibration plots for the logistic regression model trained on raw random-CV features. The left two are for the training data and the right two are for the validation folds. Bottom row: Calibration plots for the logistic regression model trained on raw spatial-CV features.

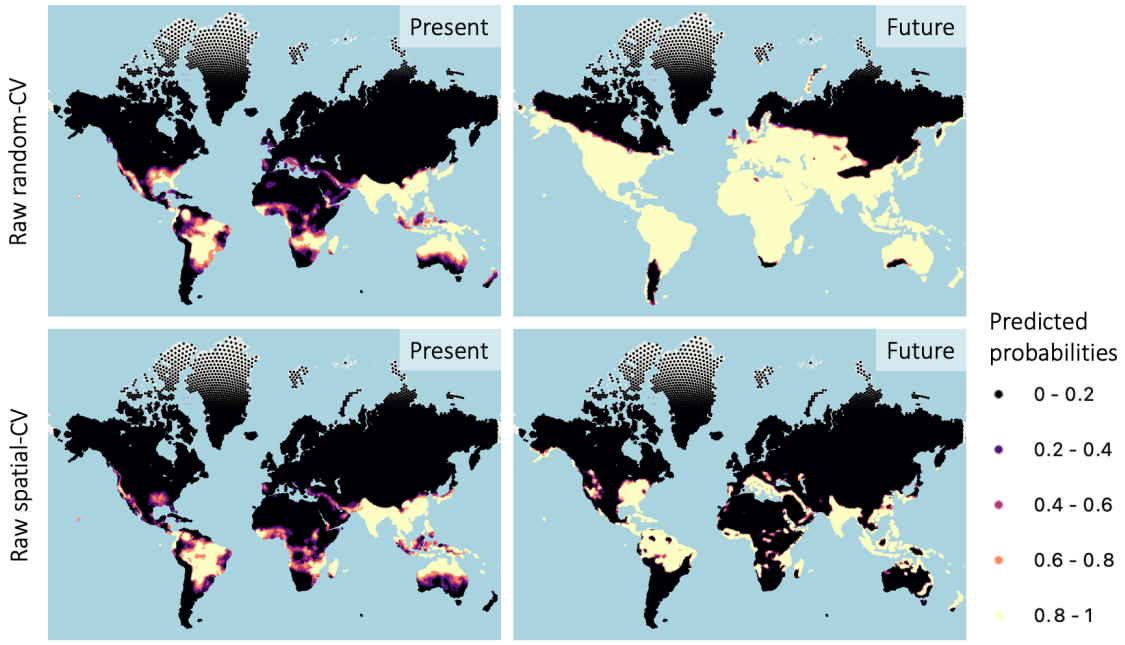


Figure 17: Top row: Predictions for the logistic regression model trained on raw random-CV features. Bottom row: Predictions for the logistic regression model trained on raw spatial-CV features.

6.2 Random Forest

For the random forest models, neither feature selection nor feature scaling had a substantial effect. The numerical scores and calibration plots were similar between models fit on raw and scaled features and the difference between predictions were not as dramatic as logistic regression.

One distinctly different output for the random forest models was their calibration plots. The training calibrations implied that the models had found parameters that could almost completely separate the positives from the negatives. 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 may have been an early warning of the problem we later encountered in Bayesian GAM.

The outputs of random forest are shown in Table 6 and Figures 18 and 19. The effects of scaling were negligible, hence the results of the models fit on scaled features are not presented in this report.

Table 6: The scores for the random forest models fit on raw feature values.

Score type	**Random-CV features**		**Spatial-CV features**	
	Training	**Validation**	**Training**	**Validation**
AUC	1.0	0.92	1.0	0.93
TSS	0.96	0.76	0.97	0.80
sensitivity	0.98	0.83	0.98	0.84
specificity	0.98	0.94	0.98	0.96

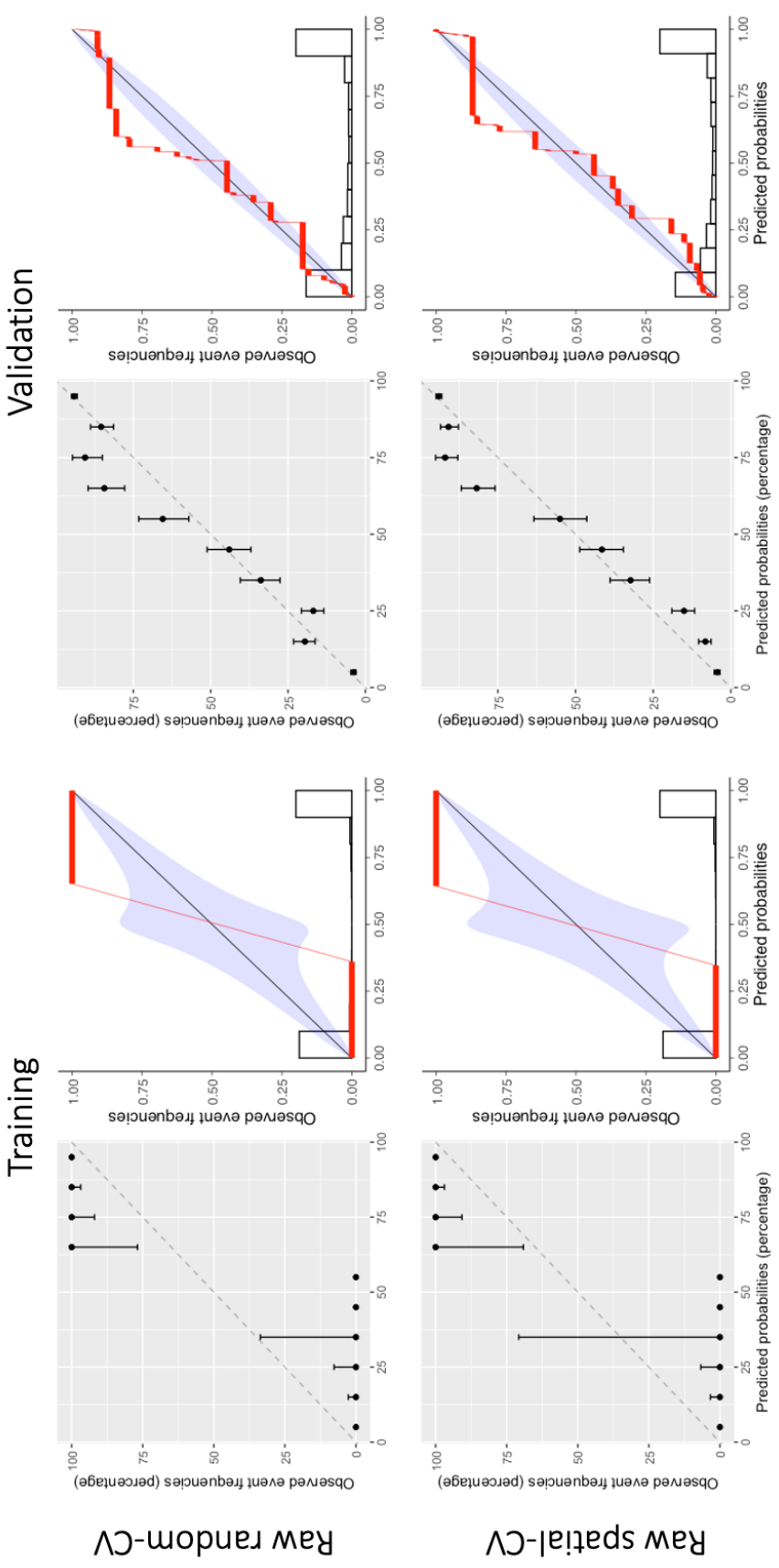


Figure 18: Top row: Calibration plots for the random forest model trained on raw random-CV features. The left two are for the training data and the right two are for the validation folds. Bottom row: Calibration plots for the random forest model trained on raw spatial-CV features.

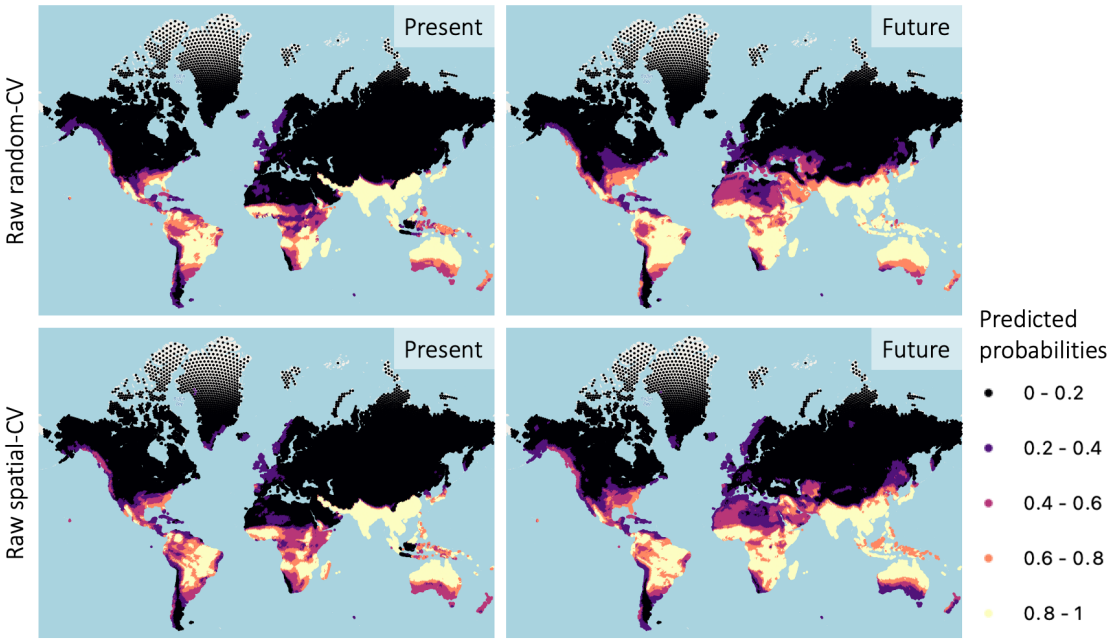


Figure 19: Top row: Predictions for the random forest model trained on raw random-CV features. Bottom row: Predictions for the random forest model trained on raw spatial-CV features.

6.3 Bayesian GLM

For Bayesian GLM, the initial priors were slightly wide normal distributions centered around zero. This gives no prior belief on the effects of the features while restricting any single feature from overpowering the others. The prior for the intercept was set wider than the coefficients because it had no direct interaction with the features.

The Bayesian GLM models gave point predictions, scores, and calibration plots that were similar to logistic regression. The IQR of the predictions were usually larger near intermediate point prediction outputs. The future predictions from models trained on spatial-CV features were an exception to this, having larger IQRs in some areas where they predicted a high or low probability. These can be observed in Table 7 and Figures 21, 22, A1, A2, and A3.

One notable difference from logistic regression was the effect of scaling the spatial-CV features. As shown in the top row of Figure 20, the model fit on the scaled spatial-CV features predicted slightly pessimistic future habitat suitability compared to the model fit on raw features.

This effect of scaling implied that the priors might be overpowering the features with smaller values, so we used the priorsense package to test whether the priors were too strong. Priorsense indeed suggested that the priors are too restrictive for some models, so we switched to larger scales for those models for which priorsense issued warning notifications. The messages we based our adjustments on are shown in Table A9.

An interesting property we noticed about the priors is that we needed to widen all of them simultaneously instead of adjusting the ones that priorsense warned about. The priors seemed to interact with each other so that making only select priors wider seemed to make others relatively more restrictive. If this was not the expected behavior of Bayesian models, it could be related to the inherent interdependence of the climatic features e.g., seasonality influences temperature and temperature influences precipitation.

Scaling spatial-CV features still had a small effect on the prediction outcome after adjusting the priors, but the difference was not as large. This is depicted in the bottom row of Figure 20.

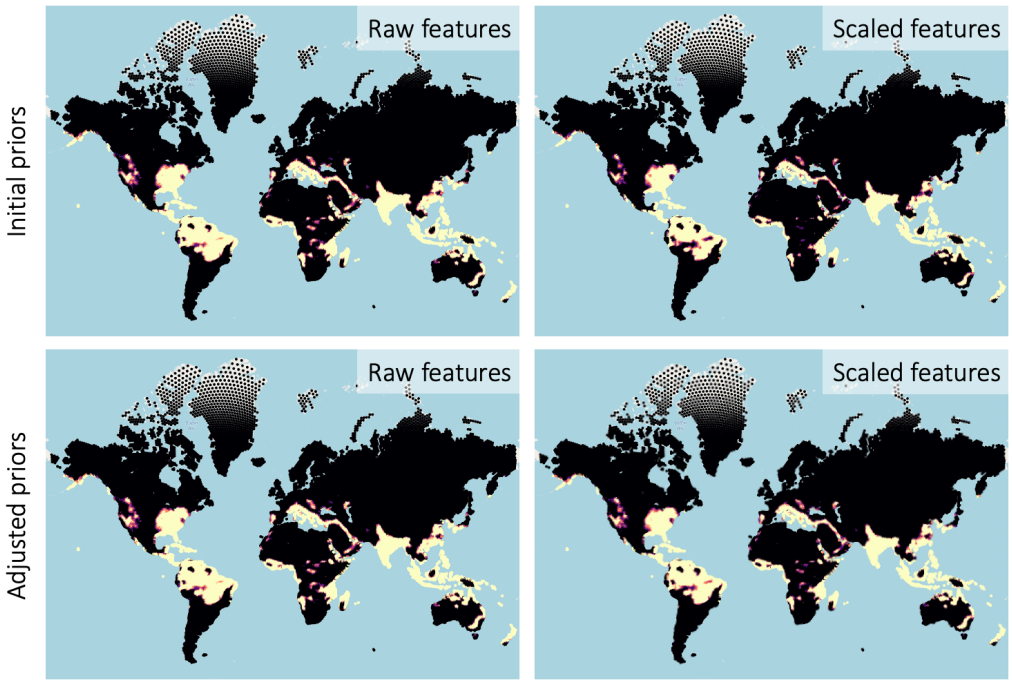


Figure 20: The difference in future predictions for raw and scaled spatial-CV features. Top row: The future predictions before the priors were adjusted. The model fit on scaled features has a more pessimistic outcome. Bottom row: The future predictions after prior adjustment. The difference seen in the top row is less pronounced.

Table 7: The scores for the Bayesian GLM models. There are no scores after prior adjustment for the model fit on raw spatial-CV features since their priors did not need to be changed.

Prior set	**Scaling**	**Score type**	**Random-CV features**		**Spatial-CV features**	
			Training	**Validation**	**Training**	**Validation**
initial	yes	AUC	0.98	0.92	0.98	0.91
		TSS	0.85	0.80	0.83	0.77
		sensitivity	0.94	0.86	0.93	0.88
		specificity	0.91	0.93	0.90	0.89
	no	AUC	0.98	0.92	0.98	0.91
		TSS	0.85	0.79	0.84	0.77
		sensitivity	0.94	0.87	0.93	0.88
		specificity	0.91	0.93	0.90	0.89
adjusted	yes	AUC	0.98	0.92	0.98	0.91
		TSS	0.85	0.79	0.84	0.77
		sensitivity	0.94	0.86	0.93	0.88
		specificity	0.91	0.93	0.90	0.89
	no	AUC	0.98	0.92	-	-
		TSS	0.85	0.79	-	-
		sensitivity	0.94	0.87	-	-
		specificity	0.91	0.93	-	-

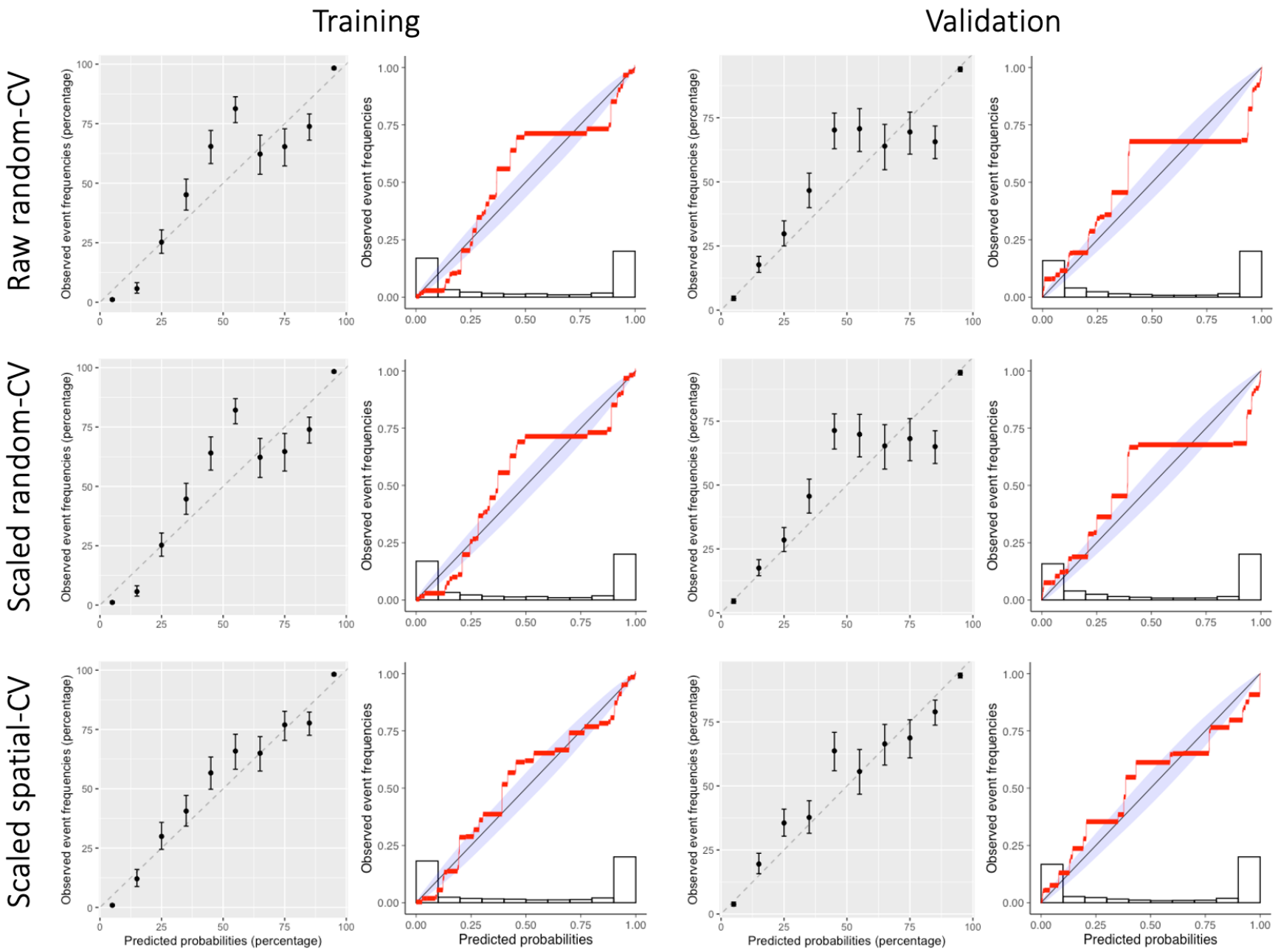


Figure 21: Calibration plots for the Bayesian GLM models using the adjusted prior settings. The calibrations for the initial prior settings including the model fit on raw spatial-CV features (which did not need adjustment) are shown in Figures A1 and A2.

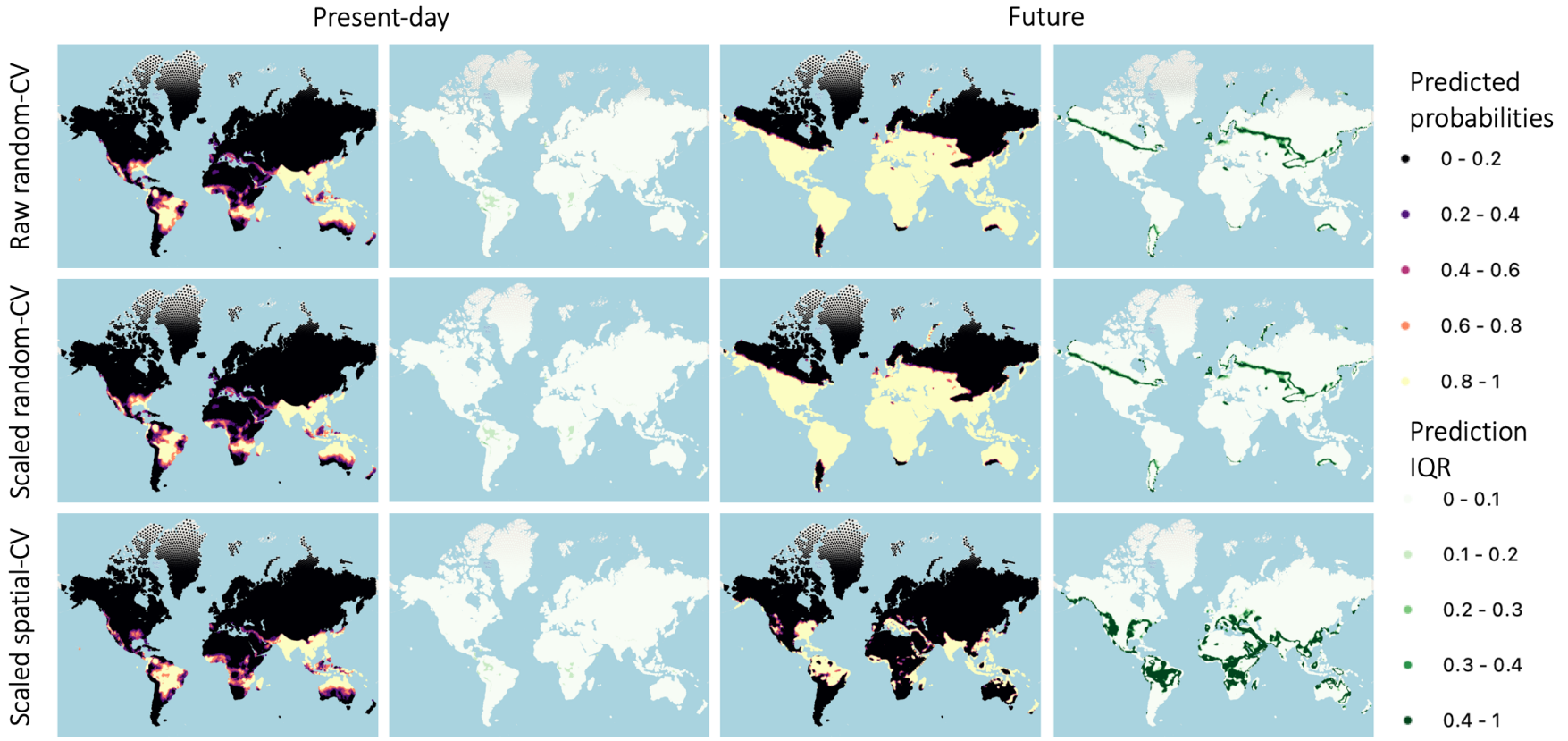


Figure 22: Predictions of the Bayesian GLM models using the adjusted prior settings. The predictions for the initial prior settings including the model fit on raw spatial-CV features are shown in Figure A3.

6.4 Bayesian GAM

Compared to Bayesian GLM, the priors of Bayesian GAM models have a complex relationship. The intercept and coefficients of Equation 6 combine with the non-linearity priors to create a joint effect, making their relationship with the features difficult to predict beforehand. Therefore, the initial priors for the intercept and coefficients were set to be equal to the initial priors of Bayesian GLM. The initial non-linearity priors were the default settings of brms, moderately wide Student's t-distributions centered around zero. At first we felt that this was appropriate since this prior gave an opportunity for the non-linearity parameters to drift away from zero, which represented a linear relationship with the features. The basis dimension settings were also set to default at the beginning of the iterative modeling.

However, the initial priors and basis dimension resulted in severe overfitting as demonstrated in Figure A6. For all feature sets, the point predictions became extreme and unrealistic for future conditions and present-day areas not included in the training data. The IQR of the predictions showed high variability in present-day Greenland and in multiple regions in the future. In particular, the models fit on spatial-CV features had very wide IQRs for every prediction output.

The symptoms of overfitting could be observed in both the TSS scores and the calibration plots. The TSS values were exceptionally high for the training scores while being suboptimal for the validation scores. Likewise, the training calibrations were well-aligned with the diagonal while the validation calibrations were very misaligned.

This output demanded an investigation of the cause of overfitting, so we examined the posterior distributions. From the top and middle plot of Figure 15 we could interpret that the non-linearity of the functions in Equation 6 was too excessive, so we explored two methods to restrict it: selecting a tighter non-linearity prior and restricting the basis dimension.

We first experimented by tightening the non-linearity prior to a standard normal distribution. After observing that this had adverse effects, we tested an improper flat prior for the coefficients and intercept to confirm whether loosening the restrictions for other parameters would alleviate the problem. This however, did not improve the predictions, so we changed the non-linearity prior back to the initial distribution and examined outputs while restricting the basis dimension. This only resulted in more unconvincing predictions. Finally, we restricted both the non-linearity prior and basis dimension settings and found that this restricted the model enough to obtain results similar to logistic regression and Bayesian GLM for the models fit on random-CV features. As an example, our observations for the model fit on raw random-CV features are presented in Figure 23.

As demonstrated in Figure 23, making the individual settings more restrictive did change the outputs, but it did not make the predictions less extreme. However, when restricting both the non-linearity priors and the basis dimension, we managed to tame the models to some extent.

However, though restricting the settings resulted in predictions similar to Bayesian GLM for random-CV features, this was not so for the spatial-CV features. The models fit on spatial-CV features failed to give convincing predictions even when

restricting the non-linearity, but became less overfit according to the scores and calibration plots. These effects can be observed in Table 8 and Figures 24, 25, 26, A4, A5, and A6.

The settings other than feature selection, the non-linearity priors, and basis dimension had less significant effects. Feature scaling had only a slight influence on the outcome. And as discussed in the Sampling Settings and Posteriors section, the distributions of the intercept and coefficient priors had negligible effects on the posteriors and consequently the outcome once the non-linearity was restricted. This is the reason why there are both normal and flat priors specified for the adjusted settings in Table 3; the two settings we explored did not make a difference.

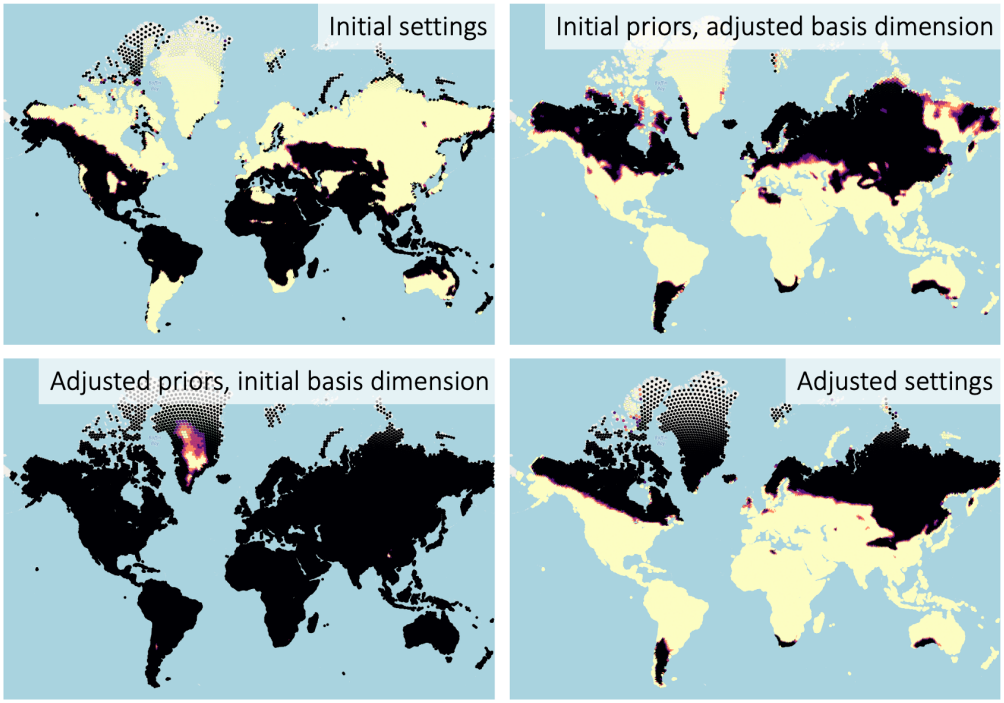


Figure 23: An example of changing the settings from the initial prior and basis dimension (top left). Neither changing only the basis dimension (top right) nor only the priors (bottom left) could restrict the non-linearity. Changing both restricted the non-linearity enough so that the predictions resembled the outputs from Bayesian GLM (bottom right). All plots are future predictions of models trained on raw random-CV features.

Table 8: The scores for the Bayesian GAM models.

Prior set	**Scaling**	**Score type**	**Random-CV features**		**Spatial-CV features**	
			Training	**Validation**	**Training**	**Validation**
initial	yes	AUC	0.99	0.87	1.0	0.88
		TSS	0.92	0.64	0.94	0.68
		sensitivity	0.96	0.81	0.97	0.83
		specificity	0.96	0.83	0.98	0.85
	no	AUC	0.99	0.87	1.0	0.88
		TSS	0.92	0.64	0.94	0.68
		sensitivity	0.96	0.81	0.97	0.83
		specificity	0.96	0.83	0.97	0.85
adjusted	yes	AUC	0.98	0.89	0.98	0.90
		TSS	0.85	0.73	0.86	0.75
		sensitivity	0.95	0.84	0.93	0.89
		specificity	0.90	0.89	0.93	0.86
	no	AUC	0.98	0.89	0.98	0.90
		TSS	0.85	0.73	0.86	0.75
		sensitivity	0.95	0.84	0.93	0.89
		specificity	0.91	0.90	0.93	0.86

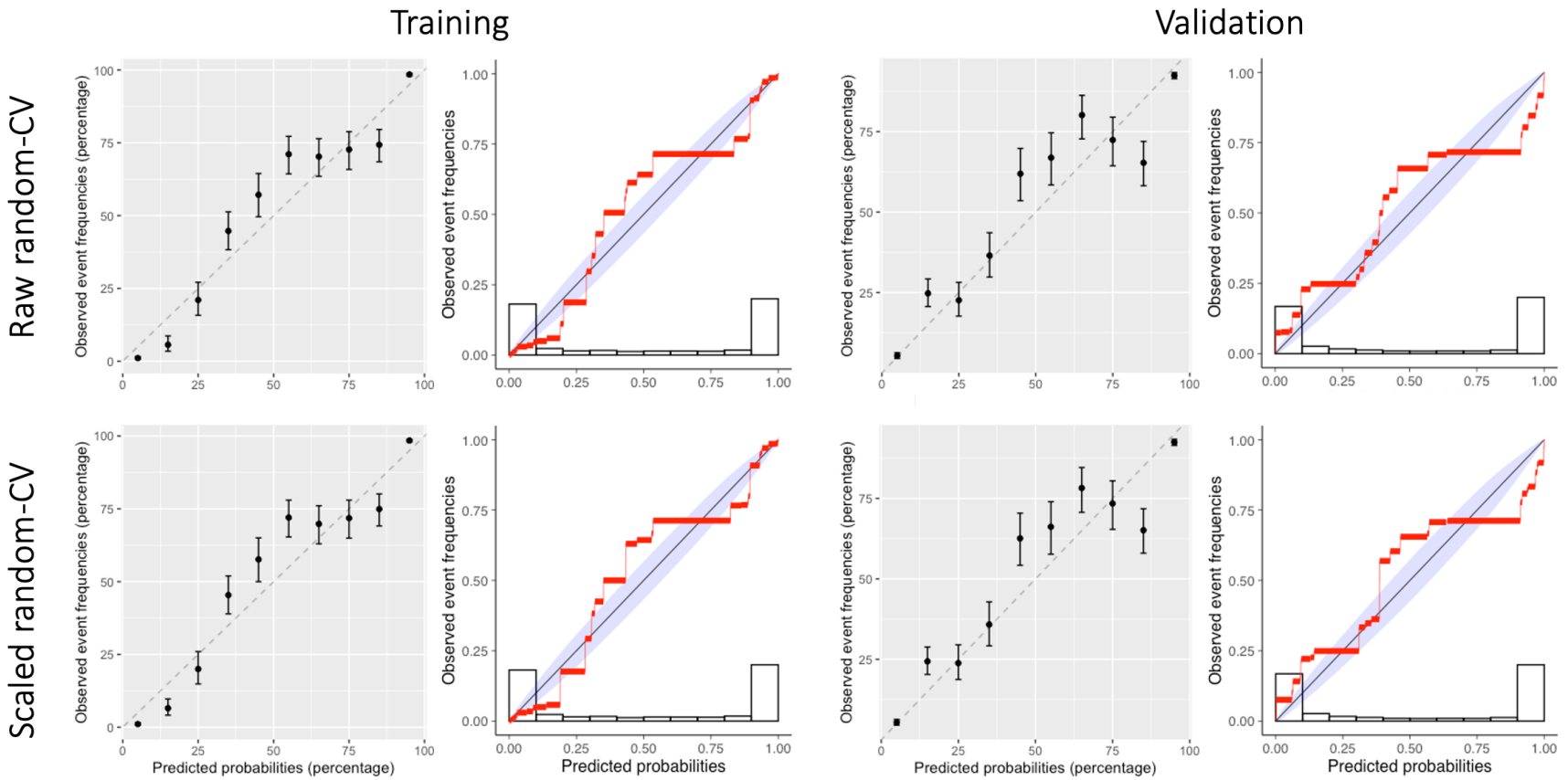


Figure 24: Calibration plots of the Bayesian GAM models fit on random-CV features using the adjusted prior and basis dimension settings. The calibration plots for the models fit on spatial-CV features are shown in Figure 25.

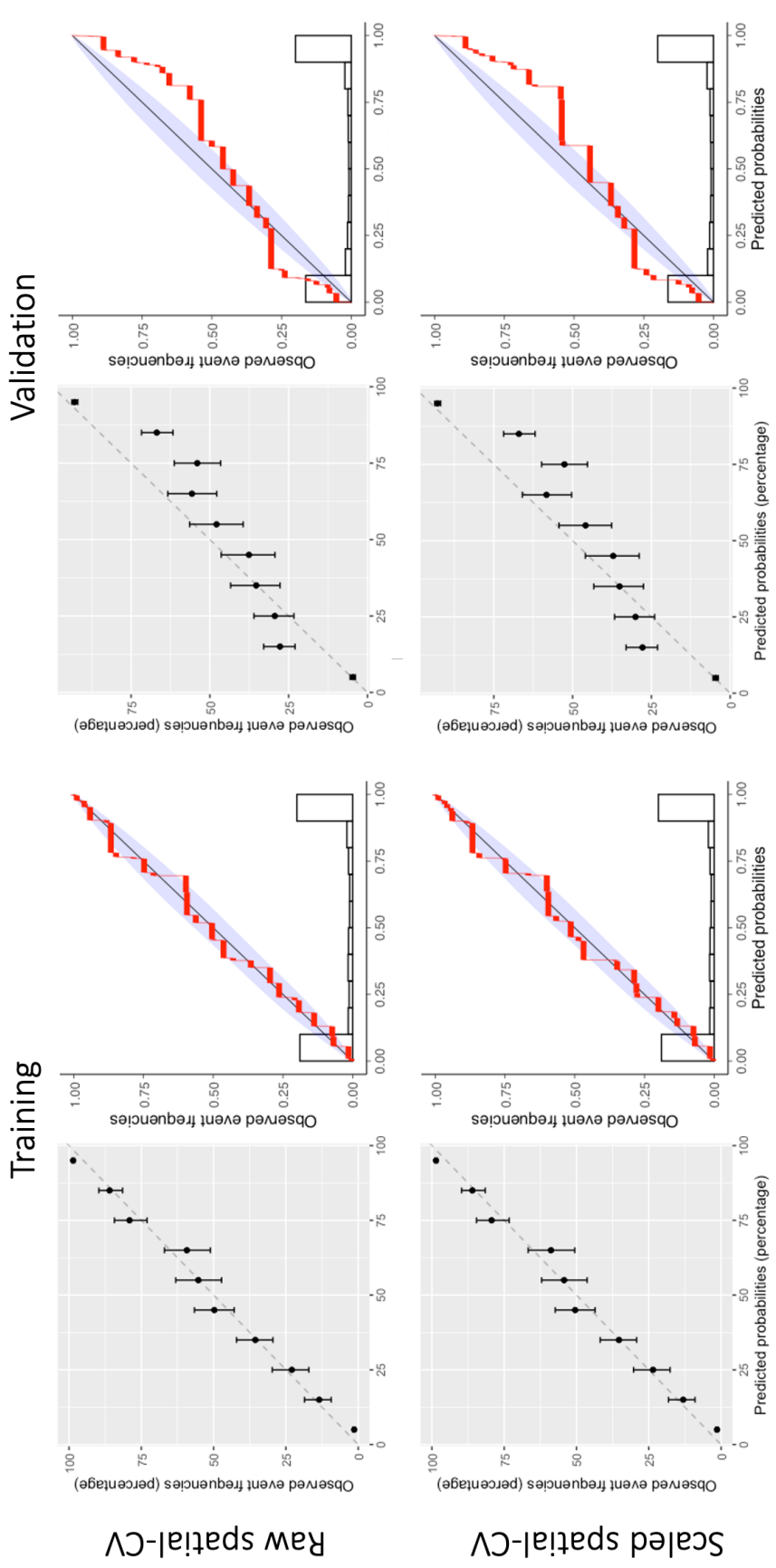


Figure 25: Calibration plots of the Bayesian GAM models fit on spatial-CV features using the adjusted prior and basis dimension settings. The calibration plots for the models fit on random-CV features are shown in Figure 24.

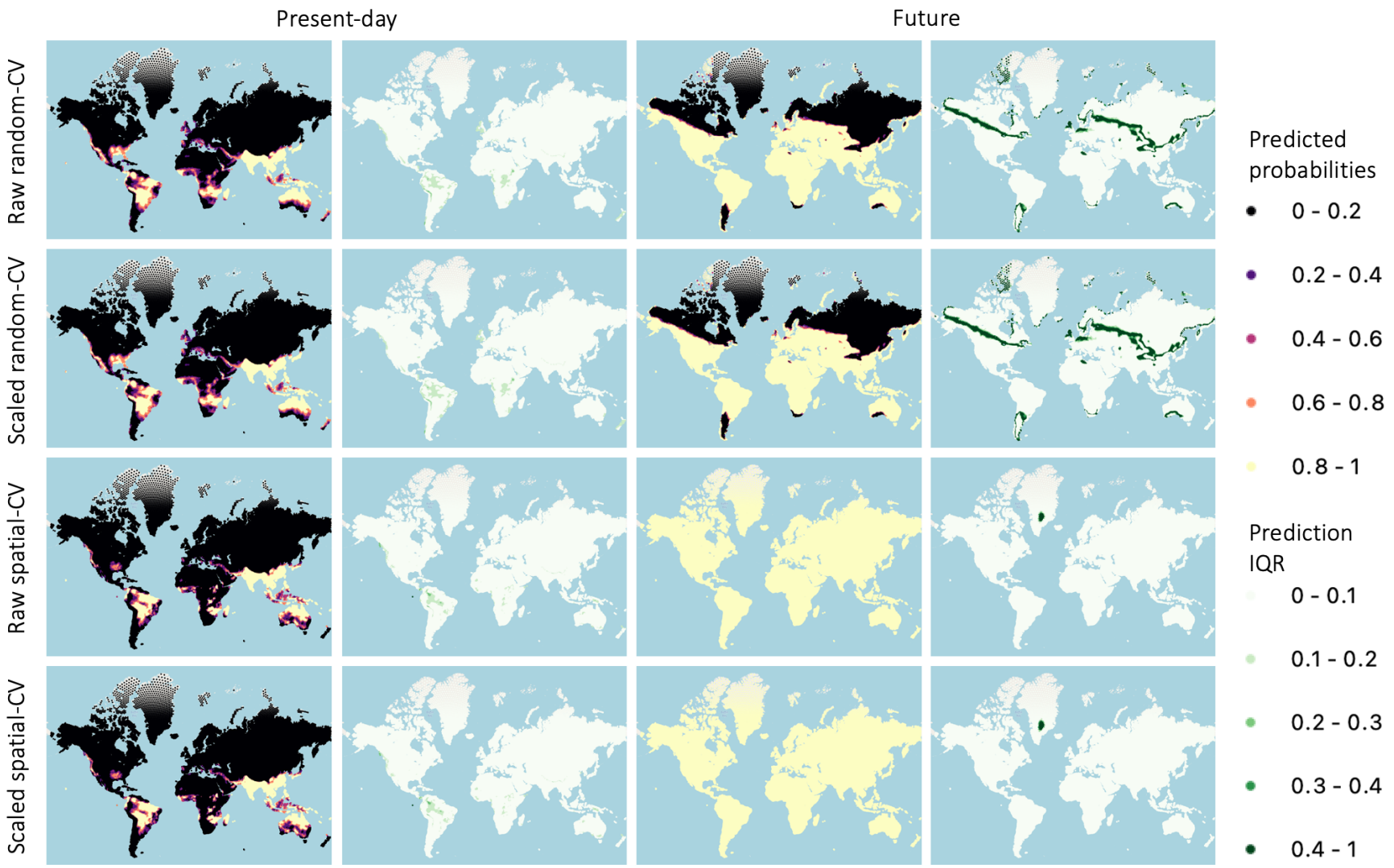


Figure 26: Predictions of the Bayesian GAM models using the adjusted prior and basis dimension settings.

7 Discussion

7.1 Modeling Decisions and Outcomes

This research iteratively assessed four model types, logistic regression, random forest, Bayesian GLM, and Bayesian GAM with combinations of different modeling options. This subsection will summarize and add details to the findings presented in the Results section.

Logistic regression and Bayesian GLM had the most simple structure, hence it never showed signs of overfitting. Rather, the calibration plots in Figures 16, 21, A1 and A2 suggested that models fit on random-CV features underfit the true pattern since the predictions undershoot the actual observed frequencies around the 0.5 bin and then overshoot them around the 0.8 bin.

Feature selection had an effect on the future predictions of both the logistic regression and Bayesian GLM. Models fit on random-CV features gave a very optimistic outlook while the models fit on spatial-CV features had a more conservative view. Feature scaling did not influence logistic regression but affected Bayesian GLM by changing the relative relationship of the data against the priors. This also suggested that prior selection had an effect on Bayesian GLM, which we confirmed by observing that wider priors made the predictions from models fit on scaled features closer to the models fit on raw features. However, the effects of feature scaling and prior selection only seemed to apply to models fit on spatial-CV features, which may indicate that the influence of the priors on the outputs depends on the features used for modeling.

Random forest is a complex non-linear model but has a built-in structure that prevents overfitting. This structure could not regulate the models enough for our datasets, however, since the training calibrations implied that the models were overfit to the extent where they could almost completely separate the positives from the negatives. Surprisingly, 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 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 19. Also, like logistic regression, feature scaling did not have an influence on the random forest models.

Bayesian GAM showed signs of extreme overfitting in their TSS scores and calibration plots. The other numerical scores did have lower values for validation than the other models, but since they were not as sensitive as TSS, were not critical factors in the diagnosis. The predictions were incomprehensible to humans, predicting high habitat suitability for areas that are unlikely to support elephant populations such as Greenland, Russia, the Sahara region, and the northernmost Canadian islands. Attempts to restrict the non-linearity either had unexpected effects or made the outputs closer to Bayesian GLM (for random-CV features).

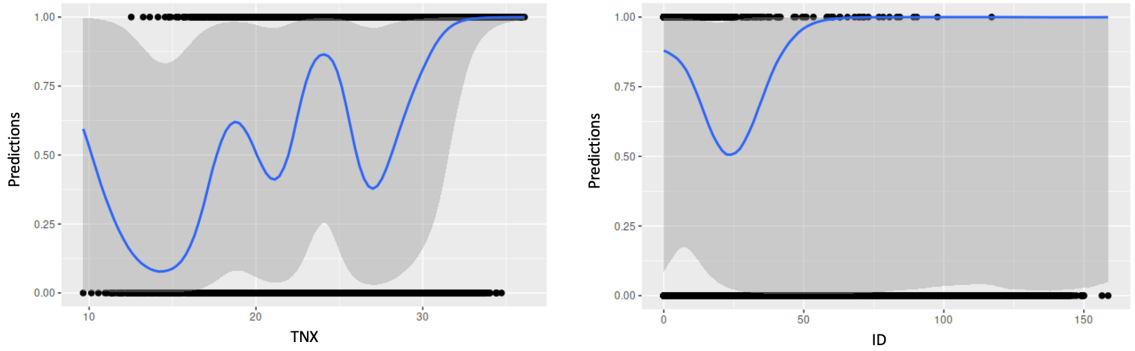
Aside from feature selection, the effects of modeling decisions were either unclear or unpredictable for Bayesian GAM. Feature scaling did not have a large effect, though it did slightly alter the predictions for some prior and basis dimension combinations.

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 GLM 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.

7.2 The Cause of Misbehaved Models

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 top row of Figure A6).

The first pattern is the excessive non-linearity in the conditional effects shown in Figure 27a. The conditional effect follows a pattern that is not apparent to the human eye.



(a) The conditional effect of TNX. The effect is too non-linear and does not effectively model the data points on the left side of the plot.

(b) The conditional effect of ID. The effect is too variable and on average, does not follow the trend the data points indicate.

Figure 27: 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.

The second pattern can be observed in Figures 27b, 28, and the left part of Figure 27a. Some features or ranges within features seem to be ignored completely in the model so that their individual contributions are not reflected in the output. Though these are examples from a complex model, this pattern was also observed in less flexible models and model settings.

The third pattern is visualized in Figure 29, where the feature values of present-day and future Greenland lie outside of the training data's range and has unrealistic contributions to the predictions. Though the severity differed, similar patterns were observed in all Bayesian models, GLM or GAM.

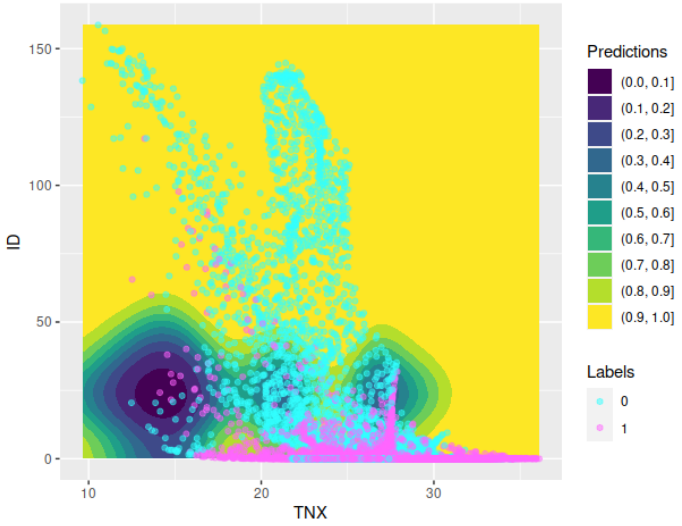


Figure 28: The interaction effect of TNX and ID on habitat suitability from the same model as Figures 27a and 27b. The other features are held constant at their mean values. The overlaid dots are the data points colored by their labels (0: habitat unsuitable, 1: habitat suitable).

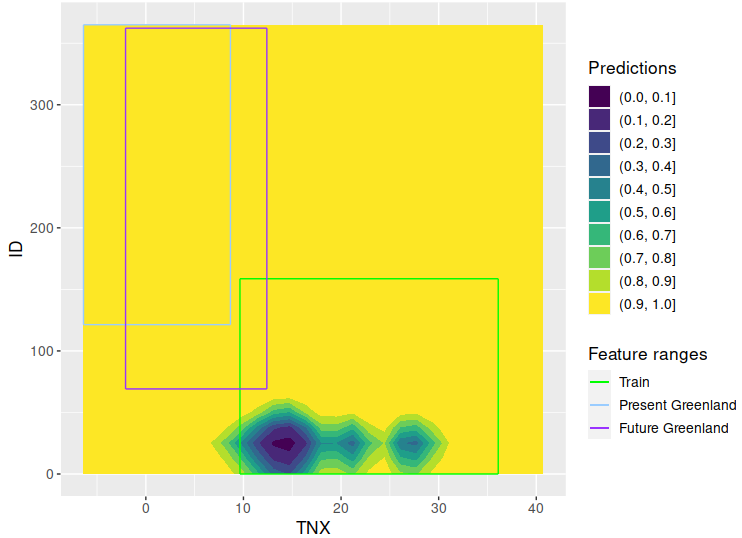


Figure 29: The same interaction effect as Figure 28 plotted on a range wider than that of the training data. The hollow rectangles overlaid on the plot are ranges of the data points that belong to the training data, present-day Greenland, and future Greenland.

The first and second pattern and perhaps the random forest calibrations in Figure 18 may indicate a problem called 'complete separation', where the model finds a parameter combination that can perfectly separate the positives or the negatives. Once this happens, the maximum likelihood estimate of at least one parameter would be either minus infinity or infinity. A Bayesian model does not use a maximum likelihood estimate, but is still influenced by complete separation. A complete separation in a Bayesian model is characterized by weak likelihoods that cannot influence the priors or unrealistic posteriors. These effects of complete separation appears to explain the conditional effects of TNX following a seemingly nonexistent pattern in Figure 27a, the model ignoring the effects of icing days in Figure 27b, and additionally, the unusually large and variable posteriors in Figure 15.

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 they predict novel ranges. Therefore, the two features in Figure 29 shows 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 on 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. Additionally, the third pattern might be mitigated by enhancing the training data with feature values that are physiologically not acceptable for Asian elephants.

7.3 The Most Reliable Model

From what we observed, the most reliable model given the dataset, the model types, and the modeling options we tried would be the Bayesian GLM model fit on scaled spatial-CV features with adjusted priors. The outputs of this model are shown in the third row of Figure 22 as well as in a larger plot below in Figure 30. This model had fairly good validation scores with the most calibrated predictions. Additionally, the scaled features and the simple model structure allows straightforward interpretations of posteriors. Above all, this model had the most convincing future predictions.

While the model did have unrealistic predictions on the south coast of Alaska, most of the areas predicted as suitable habitats by this model in future conditions were not only intuitive to the human eye, but overlapped with areas that extinct families of elephants once inhabited. As stated in the Details of Qualitative Evaluation subsection, elephant relatives and ancestors were dispersed in areas including North and South America, Europe, Africa, and Asia. Of course, this may be a coincidence since the model was simply modeling the distribution patterns of Asian elephants and never had the distributions of its relatives as input. Nevertheless, it might be interesting in a future, more paleontology-oriented research to investigate whether this truly came about by chance or if not, whether there are common climatic features

that characterizes suitable habitats for elephant families. Similarly, the regions that changed from suitable habitats to unsuitable habitats in the predictions may present another topic for research. These regions include large areas of Myanmar, Thailand, Laos, Cambodia, Vietnam, southeastern China, and the coastal area of Pakistan and Iran.

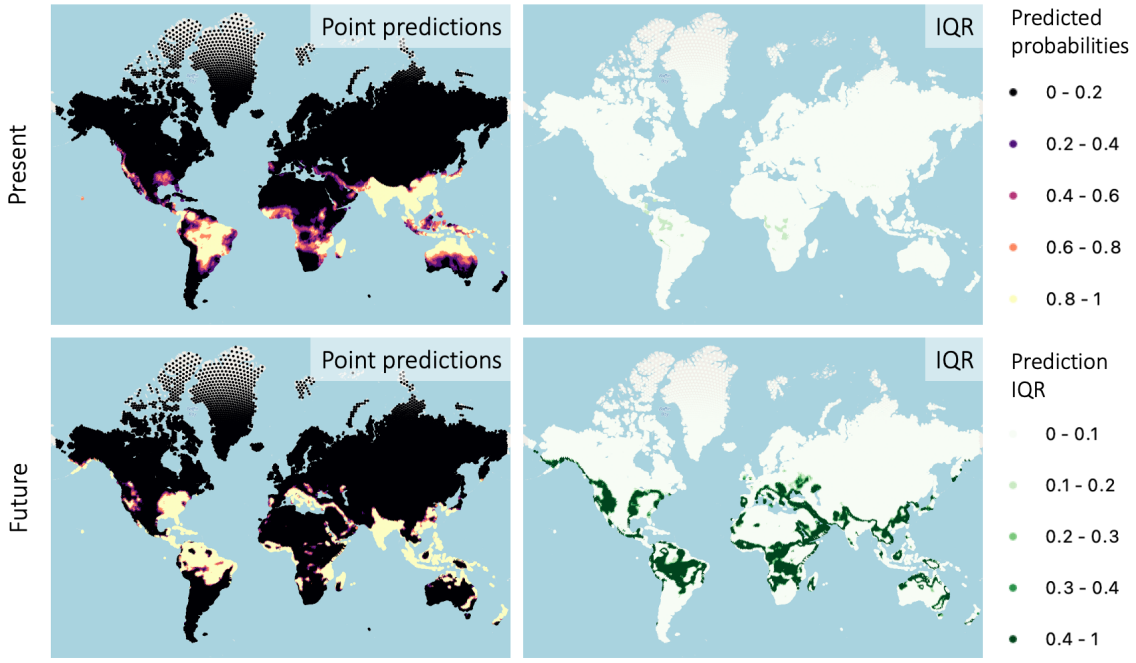


Figure 30: Outputs from the model that gave the most convincing predictions among the models explored within this thesis. A smaller version of the same plots are shown in the third row of Figure 22.

Since the Bayesian GLM had the best predictions, one could argue that logistic regression could have the same outputs while requiring less time and computation to fit. However, Bayesian models have advantages that were not fully explored within the scope of this thesis. First, users can specify informative priors in place of the weakly informative priors we used to further refine the predictions. In this way the model can incorporate an expert's insight in addition to what the data indicates. The second advantage is the uncertainty in the predictions, which were presented as IQRs in plots but never explored in depth within this thesis. These intervals provide additional information to the point predictions by indicating areas where the model cannot provide a confident prediction. This information can be used for further investigation or to avoid areas with high uncertainty.

8 Conclusion

This thesis presented the iterative modeling process for Bayesian models for a hypothetical Pleistocene rewilding project. While exploring different models, we tested different modeling options for both Bayesian models and baseline non-Bayesian models. 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 other modeling options.

The more complex Bayesian model, Bayesian GAM, showed symptoms of misbehavior that may have been partially caused by complete separation. Bayesian GLM, being a simpler model, did not exhibit this problem but did give unlikely predictions depending on the features selected for modeling. All models seemed to have at least some issues when extrapolating or predicting outside of the range of the training data.

The purpose of fitting the models was to identify a model that provides convincing predictions to assist in finding rewilding sites for Asian elephants that remain suitable in the year 2070. After quantitative evaluation, visual inspection of the outputs, and comparison against the ranges of extinct elephant species, we concluded that the Bayesian GLM fit on scaled spatial-CV features with adjusted, wider priors (Figure 30) is the most reliable model given the data and our iterative modeling process. 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 unsuitable in the RCP 8.5 scenario.

Of course, even if there was a model that gave convincing predictions, this research is limited in that it fit models on only climatic features concentrated in a limited area. Furthermore, 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. Therefore, enhancing the dataset and exploring informative priors would be good topics to address in future research.

Although this project has its limitations, it provides beneficial outputs and findings to multiple research areas. For the Bayesian community, it provides an example of the Bayesian process while suggesting a novel method of presenting the iterative modeling process in Figure 1. For the ecology-paleontology community, it provides an example of how to fit and interpret Bayesian models while suggesting areas for enhancement. Should it be required for a future project, the codes created in the research are available at https://github.com/RyokoNod/sdm-asian-elephants . Additionally, for informally exploring the outputs of models, an interactive version of Figure 1 is available at https://miro.com/app/board/uXjVOX_Zhf8=/?share_link_id=937959545296.

References

[1] Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion). Bayesian Analysis. 2021;16(2). Available from: doi: 10.1214/20-ba1221

[2] Zimov S. Pleistocene Park: Return of the Mammoth's Ecosystem. Science. 2005;308(5723):796-798. Available from: doi: 10.1126/science.1113442

[3] Lorimer J, Driessen CPG. Experiments with the wild at the Oostvaardersplassen. Ecos. 2014;35(3/4):44-52.

[4] Yellowstone Wolf Project - Yellowstone Forever [Internet]. Yellowstone Forever. [cited 17 February 2022]. Available from: https://www.yellowstone.org/wolf-project/

[5] MEXICO TORTOISE PROJECTS [Internet]. Tucsonherpsociety.org. [cited 17 February 2022]. Available from: https://tucsonherpsociety.org/projects/mexican-tortoise-project/

[6] Donlan CJ, Berger J, Bock C, Bock J, Burney D, Estes J et al. Pleistocene Rewilding: An Optimistic Agenda for Twenty-First Century Conservation. The American Naturalist. 2006;168(5):660-681.

[7] Beschta R, Ripple W. Riparian vegetation recovery in Yellowstone: The first two decades after wolf reintroduction. Biological Conservation. 2016;198:93-103.

[8] Latimer A, Wu S, Gelfand A, Silander Jr. J. Building Statistical Models To Analyze Species Distributions. Ecological Applications. 2006;16(1):33-50.

[9] Vanhatalo J, Veneranta L, Hudd R. Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (Coregonus lavaretus L. s.l.) larvae. Ecological Modelling. 2012;228:49-58.

[10] Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, et al. Bayesian Workflow. arxiv.org [Internet]. 2020 Nov 3 [cited 22 February 2022]; Available from: https://arxiv.org/abs/2011.01808

[11] Murray JV, Goldizen AW, O'Leary RA, McAlpine CA, Possingham HP, Choy SL. How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabiesPetrogale penicillata. Journal of Applied Ecology. 2009 Aug;46(4):842–51.

[12] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. Journal of Statistical Software [Internet]. 2017 Jan 11;76(1):1–32. Available from: https://www.jstatsoft.org/article/view/v076i01

[13] Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1. 2020 Jun 13;2(1).

[14] Bürkner P-C. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software [Internet]. Vol 80:1. 2017. Available at: https://www.jstatsoft.org/article/view/v080i01

[15] Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science. 2016 Apr 6;2:e55.

[16] Depaoli S, van de Schoot R. Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. Psychological Methods. 2017 Jun;22(2):240–61.

[17] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Boca Raton, Fl: Chapman & Hall/Crc; 2014.

[18] McElreath R. Statistical rethinking : a Bayesian course with examples in R and Stan. Boca Raton ; London ; New York: Chapman & Hall/Crc; 2020.

[19] Carlson CJ. embarcadero: Species distribution modelling with Bayesian additive regression trees in r. Price S, editor. Methods in Ecology and Evolution. 2020 Apr 16;11(7):850–8.

[20] Gent PR, Danabasoglu G, Donner LJ, Holland MM, Hunke EC, Jayne SR, et al. The Community Climate System Model Version 4. Journal of Climate. 2011 Oct 1;24(19):4973–91.

[21] Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology. 2005 Nov 30;25(15):1965–78.

[22] Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology. 2017 May 15;37(12):4302–15.

[23] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997 Jul;30(7):1145–59.

[24] Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology. 2006 Sep 12;43(6):1223–32.

[25] Richmond OMW, McEntee JP, Hijmans RJ, Brashares JS. Is the Climate Right for Pleistocene Rewilding? Using Species Distribution Models to Extrapolate Climatic Suitability for Mammals across Continents. Merenlender AM, editor. PLoS ONE. 2010 Sep 22;5(9):e12899.

[26] Stan Development Team. 2019. Stan Modeling Language Users Guide and Reference Manual, Version 2.29. https://mc-stan.org

[27] Stan Development Team. shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models. R package version 2.4.0. 2017. Available at: http://mc-stan.org/.

[28] Faurby S, Davis M, Pedersen RØ, Schowanek SD, Antonelli1 A, Svenning J. PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology. Ecology. 2018 Oct;99(11):2626–6.

[29] Sahr K, White D, Kimerling AJ. Geodesic Discrete Global Grid Systems. Cartography and Geographic Information Science. 2003 Jan;30(2):121–34.

[30] Sillmann J, Kharin VV, Zhang X, Zwiers FW, Bronaugh D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. Journal of Geophysical Research: Atmospheres. 2013 Feb 27;118(4):1716–33.

[31] Sillmann J, Kharin VV, Zwiers FW, Zhang X, Bronaugh D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. Journal of Geophysical Research: Atmospheres. 2013 Mar 25;118(6):2473–93.

[32] Riahi K, Rao S, Krey V, Cho C, Chirkov V, Fischer G, et al. RCP 8.5—A scenario of comparatively high greenhouse gas emissions. Climatic Change. 2011 Aug 13;109(1-2):33–57.

[33] Mechenich M. Machine Learning Methods for Bridging the Paleontological-Ecological Divide [Doctoral dissertation (unpublished)]. University of Helsinki. No date.

[34] Donlan J. Re-wilding North America. Nature. 2005 Aug;436(7053):913–4.

[35] Lister AM, Dirks W, Assaf A, Chazan M, Goldberg P, Applbaum YH, et al. New fossil remains of Elephas from the southern Levant: Implications for the evolutionary history of the Asian elephant. Palaeogeography, Palaeoclimatology, Palaeoecology. 2013 Sep;386:119–30.

[36] Smith KM, Stynder DD. Biogeography and molar morphology of Pleistocene African elephants: new evidence from Elandsfontein, Western Cape Province, South Africa. Quaternary Science Reviews. 2015 May;115:101–11.

[37] Mothé D, dos Santos Avilla L, Asevedo L, Borges-Silva L, Rosas M, Labarca-Encina R, et al. Sixty years after "The mastodonts of Brazil": The state of the art of South American proboscideans (Proboscidea, Gomphotheriidae). Quaternary International. 2017 Jul;443(A):52–64.

[38] Gagliardi F, Maridet O, Becker D. The record of Deinotheriidae from the Miocene of the Swiss Jura Mountains (Jura Canton, Switzerland). 2020. Preprint bioRxiv.

[39] QGIS Development Team. QGIS Geographic Information System. QGIS Association; 2022.

[40] Chatterjee D. Random Forest Using R [Internet]. RPubs. Available from: https://rpubs.com/Diptirtha99/677810

[41] Wood SN. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003 Feb;65(1):95–114.

[42] Kallioinen N, Paananen T, Bürkner P-C, Vehtari, A. Detecting and diagnosing prior and likelihood sensitivity with power-scaling [Preprint]. 2021. Preprint arXiv:2107.14054

[43] Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software [Internet]. 2008;28(5). Available from: https://www.jstatsoft.org/article/view/v028i05

[44] Dimitriadis T, Gneiting T, Jordan AI. Stable reliability diagrams for probabilistic classifiers. Proceedings of the National Academy of Sciences. 2021 Feb 17;118(8).

[45] Gelman A, Hill J, Vehtari A. Regression and other stories [Internet]. Cambridge Cambridge University Press; 2020 [cited 2022 Apr 11]. Available from: https://users.aalto.fi/ ave/ROS.pdf

A Appendix

Table A1: Posterior summary statistics of Bayesian GLM fit on raw random-CV features. n_eff represents the effective sample size, which is an indicator of the number of independent samples from the posterior distribution.

Prior set	**Parameters**	**n_eff**	$\widehat{\mathbf{R}}$	**Percentiles**				
				2.5	**25**	**50**	**75**	**97.5**
Initial	Intercept	2,786	1	-26.4	-24.2	-23	-21.9	-19.7
	Coeff BIO03	2,195	1	-0.1	-0.1	-0.1	-0.1	-0.1
	Coeff TN10P	2,697	1	0.7	0.9	1	1.1	1.3
	Coeff GSL	2,372	1	0	0	0	0	0
	Coeff TNX	2,806	1	0.1	0.1	0.2	0.2	0.2
	Coeff ID	2,695	1	-0.1	-0.1	-0.1	-0.1	-0.1
	Coeff BIO14	2,915	1	0	0	0	0	0
	Coeff BIO18	3,750	1	0	0	0	0	0
	Coeff CWD	2,344	1	0	0	0.1	0.1	0.1
	Coeff RX1DAY	3,353	1	0.1	0.1	0.1	0.1	0.1
	Coeff WSDI	3,321	1	0.2	0.3	0.3	0.3	0.3
Adjusted	Intercept	2,202	1	-13	-12.1	-11.7	-11.3	-10.5
	Coeff BIO03	2,053	1	-9.8	-9	-8.5	-8.1	-7.3
	Coeff TN10P	3,424	1	2.4	3.1	3.4	3.8	4.5
	Coeff GSL	1,757	1	3.5	4.6	5.1	5.6	6.6
	Coeff TNX	2,525	1	2.7	3.4	3.8	4.2	4.9
	Coeff ID	2,400	1	-17.5	-15.5	-14.6	-13.6	-11.8
	Coeff BIO14	2,192	1	-3	-2.2	-1.8	-1.3	-0.5
	Coeff BIO18	2,330	1	7.4	9.3	10.3	11.2	13.3
	Coeff CWD	2,544	1	3.4	4.3	4.8	5.3	6.3
	Coeff RX1DAY	3,043	1	14.9	15.8	16.3	16.8	17.9
	Coeff WSDI	3,272	1	3.5	3.9	4.1	4.3	4.7

Table A2: Posterior summary statistics of Bayesian GLM fit on scaled random-CV features.

Prior set	Parameters	n_eff	$\widehat{R}$	Percentiles				
				2.5	25	50	75	97.5
Initial	Intercept	2,202	1	-13	-12.1	-11.7	-11.3	-10.5
	Coeff BIO03	2,053	1	-9.8	-9	-8.5	-8.1	-7.3
	Coeff TN10P	3,424	1	2.4	3.1	3.4	3.8	4.5
	Coeff GSL	1,757	1	3.5	4.6	5.1	5.6	6.6
	Coeff TNX	2,525	1	2.7	3.4	3.8	4.2	4.9
	Coeff ID	2,400	1	-17.5	-15.5	-14.6	-13.6	-11.8
	Coeff BIO14	2,192	1	-3	-2.2	-1.8	-1.3	-0.5
	Coeff BIO18	2,330	1	7.4	9.3	10.3	11.2	13.3
	Coeff CWD	2,544	1	3.4	4.3	4.8	5.3	6.3
	Coeff RX1DAY	3,043	1	14.9	15.8	16.3	16.8	17.9
	Coeff WSDI	3,272	1	3.5	3.9	4.1	4.3	4.7
Adjusted	Intercept	1,880	1	-6	-4.1	-3.1	-2.1	-0.3
	Coeff BIO03	1,785	1	-0.1	-0.1	-0.1	-0.1	0
	Coeff TN10P	1,790	1	0.3	0.3	0.3	0.3	0.3
	Coeff GSL	2,420	1	-0.5	-0.4	-0.4	-0.4	-0.3
	Coeff TNX	2,110	1	-1.1	-1	-0.9	-0.9	-0.7
	Coeff ID	2,505	1	-0.1	-0.1	-0.1	-0.1	-0.1
	Coeff BIO14	3,232	1	0	0	0	0	0
	Coeff BIO18	4,000	1	0	0	0	0	0
	Coeff CWD	2,666	1	0	0	0	0	0
	Coeff RX1DAY	2,156	1	0.1	0.1	0.1	0.1	0.1
	Coeff WSDI	2,574	1	0.3	0.3	0.3	0.3	0.4

Table A3: Posterior summary statistics of Bayesian GLM fit on raw spatial-CV features. There are no adjusted priors for this feature set.

Prior set	Parameters	n_eff	$\widehat{R}$	Percentiles				
				2.5	25	50	75	97.5
Initial	Intercept	2,786	1	-26.4	-24.2	-23	-21.9	-19.7
	Coeff BIO08	2,195	1	-0.1	-0.1	-0.1	-0.1	-0.1
	Coeff TXX	2,697	1	0.7	0.9	1	1.1	1.3
	Coeff BIO02	2,372	1	0	0	0	0	0
	Coeff TN90P	2,806	1	0.1	0.1	0.2	0.2	0.2
	Coeff ID	2,695	1	-0.1	-0.1	-0.1	-0.1	-0.1
	Coeff BIO14	2,915	1	0	0	0	0	0
	Coeff BIO18	3,750	1	0	0	0	0	0
	Coeff CWD	2,344	1	0	0	0.1	0.1	0.1
	Coeff RX1DAY	3,353	1	0.1	0.1	0.1	0.1	0.1
	Coeff WSDI	3,321	1	0.2	0.3	0.3	0.3	0.3

Table A4: Posterior summary statistics of Bayesian GLM fit on scaled spatial-CV features.

Prior set	**Parameters**	**n_eff**	$\widehat{\mathbf{R}}$	**Percentiles**				
				2.5	**25**	**50**	**75**	**97.5**
Initial	Intercept	2,467	1	-5.8	-5.1	-4.8	-4.5	-3.9
	Coeff BIO08	2,351	1	-3.8	-3.1	-2.8	-2.5	-1.9
	Coeff TXX	2,202	1	6.9	7.7	8.1	8.5	9.3
	Coeff BIO02	2,840	1	-6	-5.5	-5.2	-5	-4.4
	Coeff TN90P	2,732	1	-4.3	-3.8	-3.5	-3.2	-2.7
	Coeff ID	2,645	1	-19.2	-17.8	-17	-16.3	-15
	Coeff BIO14	2,689	1	-7.1	-6.4	-6.1	-5.8	-5.2
	Coeff BIO18	2,711	1	9.5	11.3	12.2	13.2	15
	Coeff CWD	2,725	1	0.6	1.3	1.7	2	2.8
	Coeff RX1DAY	2,531	1	16.4	17.5	18.1	18.6	19.8
	Coeff WSDI	2,913	1	4	4.4	4.7	4.9	5.4
Adjusted	Intercept	2,319	1	-5.9	-5.3	-4.9	-4.6	-4
	Coeff BIO08	1,777	1	-4.1	-3.5	-3.1	-2.8	-2.1
	Coeff TXX	1,891	1	7.1	8	8.4	8.9	9.6
	Coeff BIO02	2,670	1	-6.1	-5.5	-5.3	-5	-4.5
	Coeff TN90P	2,582	1	-4.2	-3.7	-3.4	-3.1	-2.6
	Coeff ID	1,852	1	-20.3	-18.8	-18	-17.2	-15.9
	Coeff BIO14	2,321	1	-7.3	-6.7	-6.4	-6.1	-5.5
	Coeff BIO18	2,523	1	10.1	12.1	13.1	14.1	16
	Coeff CWD	2,673	1	0.5	1.2	1.6	2	2.7
	Coeff RX1DAY	2,388	1	17	18.1	18.7	19.3	20.5
	Coeff WSDI	2,677	1	3.9	4.4	4.7	4.9	5.4

Table A5: Posterior summary statistics of Bayesian GAM fit on raw random-CV features. Only the values for the settings where the priors for the intercept and coefficients are normal distributions are shown in the bottom half since the settings with flat priors had similar values.

Prior/basis dimension set	**Parameters**	**n_eff**	$\widehat{\mathbf{R}}$	**Percentiles**				
				2.5	**25**	**50**	**75**	**97.5**
Initial	Intercept	4,804	1	-0.5	-0.2	0	0.2	0.5
	Coeff BIO03	6,000	1	-3.7	2.2	5.2	8.3	14.4
	Coeff TN10P	6,000	1	-10.2	-3.9	-0.6	2.9	9
	Coeff GSL	6,000	1	-8.6	-2.5	0.7	3.8	9.7
	Coeff TNX	6,000	1	-5.9	0.4	3.7	7	13.2
	Coeff ID	6,000	1	-10.5	-4.2	-1	2.3	8.7
	Coeff BIO14	6,000	1	-6.8	-0.8	2.4	5.6	11.7
	Coeff BIO18	6,000	1	-9.8	-3.5	-0.2	3.4	9.7
	Coeff CWD	6,000	1	-9.9	-3.2	0.2	3.5	10.2
	Coeff RX1DAY	6,000	1	-9.3	-3.1	0.3	3.7	10.2
	Coeff WSDI	6,000	1	-9.2	-3.5	-0.4	2.6	7.9
	NL BIO03	2,971	1	13.8	20.9	26.3	33.6	52.8
	NL TN10P	2,322	1	7.7	14.3	18.3	23.2	36.4
	NL GSL	3,371	1	1.2	3.1	4.3	6	11.8
	NL TNX	2,430	1	15.9	21.9	26.2	31.5	46.7
	NL ID	2,955	1	8.1	13.9	18.8	25.5	47.2
	NL BIO14	3,872	1	4.5	6.6	8.2	10.2	16.7
	NL BIO18	2,456	1	49.5	67.2	79.5	96.9	147.3
	NL CWD	1,937	1	34.6	46.9	55.7	66.2	98.2
	NL RX1DAY	2,954	1	22.1	31	37.1	45.1	67
	NL WSDI	3,390	1	3.2	4.7	5.8	7.4	12.1
Adjusted	Intercept	3,356	1	-0.6	-0.5	-0.4	-0.3	-0.2
	Coeff BIO03	5,005	1	1.6	1.8	1.9	2	2.2
	Coeff TN10P	4,555	1	-0.6	-0.5	-0.4	-0.4	-0.2
	Coeff GSL	3,545	1	-2	-1.7	-1.5	-1.3	-0.9
	Coeff TNX	2,703	1	-0.5	-0.4	-0.3	-0.2	0
	Coeff ID	3,049	1	2	2.7	3	3.2	3.8
	Coeff BIO14	3,973	1	-0.1	0.1	0.2	0.2	0.4
	Coeff BIO18	5,182	1	-1	-0.7	-0.6	-0.5	-0.3
	Coeff CWD	4,872	1	-1.1	-0.9	-0.8	-0.7	-0.5
	Coeff RX1DAY	5,156	1	-2.9	-2.7	-2.6	-2.5	-2.2
	Coeff WSDI	6,000	1	-1.1	-1	-1	-0.9	-0.8
	NL BIO03	3,116	1	0.5	1.6	2.1	2.7	3.7
	NL TN10P	3,508	1	0.1	0.8	1.4	2.1	3.2
	NL GSL	6,000	1	0	0.4	0.8	1.3	2.5
	NL TNX	4,938	1	2.4	3	3.3	3.7	4.5
	NL ID	2,859	1	0.1	0.7	1.6	2.6	4.4
	NL BIO14	6,000	1	0	0.3	0.7	1.2	2.3
	NL BIO18	6,000	1	0	0.3	0.7	1.2	2.3
	NL CWD	6,000	1	0	0.3	0.7	1.1	2.2
	NL RX1DAY	6,000	1	2.6	3.4	3.8	4.3	5.1
	NL WSDI	6,000	1	0	0.3	0.7	1.1	2.1

Table A6: Posterior summary statistics of Bayesian GAM fit on scaled random-CV features. Like Table A5, only the values for the adjusted settings with normal priors are shown in the bottom half.

Prior/basis dimension set	**Parameters**	**n_eff**	$\widehat{\mathbf{R}}$	**Percentiles**				
				2.5	**25**	**50**	**75**	**97.5**
Initial	Intercept	4,452	1	-0.5	-0.2	0	0.2	0.5
	Coeff BIO03	6,000	1	-4.3	2	5.1	8.4	14.6
	Coeff TN10P	6,000	1	-10.4	-4	-0.7	2.8	9.4
	Coeff GSL	6,000	1	-8.6	-2.4	0.7	3.7	9.5
	Coeff TNX	6,000	1	-6.1	0.4	3.8	7.1	13.3
	Coeff ID	6,000	1	-10.6	-4.4	-1	2.4	9.3
	Coeff BIO14	6,000	1	-6.6	-0.6	2.5	5.5	11.3
	Coeff BIO18	6,000	1	-9.6	-3.4	-0.1	3.2	9.9
	Coeff CWD	6,000	1	-9.6	-3.5	0.1	3.7	10.2
	Coeff RX1DAY	6,000	1	-9.3	-2.9	0.4	3.8	10.5
	Coeff WSDI	6,000	1	-9.4	-3.4	-0.4	2.6	7.9
	NL BIO03	2,339	1	13.4	20.5	26	33	51.8
	NL TN10P	2,835	1	7.7	14.3	18.3	23.1	35
	NL GSL	2,800	1	1.3	3.2	4.4	5.9	11.2
	NL TNX	2,075	1	15.9	21.9	26.2	31.5	45.6
	NL ID	2,902	1	8.1	13.8	18.7	25.8	49.6
	NL BIO14	3,279	1	4.5	6.6	8.3	10.3	16.3
	NL BIO18	2,214	1	48.4	67	79.8	97.2	147.7
	NL CWD	2,421	1	34.7	47.2	56.2	67.1	97.6
	NL RX1DAY	2,484	1	22.1	30.9	37.1	45.3	67.5
	NL WSDI	3,734	1	3.2	4.7	5.9	7.5	12.2
Adjusted	Intercept	3,064	1	-0.6	-0.5	-0.4	-0.3	-0.2
	Coeff BIO03	4,264	1	1.6	1.8	1.9	2	2.2
	Coeff TN10P	4,398	1	-0.6	-0.5	-0.4	-0.4	-0.2
	Coeff GSL	3,879	1	-2	-1.6	-1.5	-1.3	-0.9
	Coeff TNX	3,538	1	-0.5	-0.4	-0.3	-0.2	0
	Coeff ID	2,525	1	2.1	2.7	3	3.2	3.7
	Coeff BIO14	3,631	1	-0.1	0.1	0.2	0.2	0.4
	Coeff BIO18	5,398	1	-0.9	-0.7	-0.6	-0.5	-0.3
	Coeff CWD	4,525	1	-1.1	-0.9	-0.8	-0.7	-0.5
	Coeff RX1DAY	5,549	1	-2.9	-2.7	-2.6	-2.5	-2.2
	Coeff WSDI	4,849	1	-1.1	-1	-1	-0.9	-0.8
	NL BIO03	2,956	1	0.3	1.6	2.1	2.6	3.6
	NL TN10P	3,262	1	0.1	0.8	1.4	2.1	3.3
	NL GSL	6,000	1	0	0.4	0.8	1.3	2.5
	NL TNX	5,428	1	2.4	3	3.3	3.7	4.5
	NL ID	2,498	1	0.1	0.7	1.5	2.5	4.5
	NL BIO14	6,000	1	0	0.3	0.7	1.2	2.3
	NL BIO18	6,000	1	0	0.3	0.7	1.2	2.2
	NL CWD	6,000	1	0	0.3	0.7	1.1	2.2
	NL RX1DAY	6,000	1	2.7	3.4	3.8	4.3	5.1
	NL WSDI	6,000	1	0	0.3	0.7	1.2	2.1

Table A7: Posterior summary statistics of Bayesian GAM fit on raw spatial-CV features. Like Table A5, only the values for the adjusted settings with normal priors are shown in the bottom half.

Prior/basis dimension set	**Parameters**	**n_eff**	$\widehat{\mathbf{R}}$	**Percentiles**				
				2.5	**25**	**50**	**75**	**97.5**
Initial	Intercept	4,402	1	-0.6	-0.2	0	0.2	0.6
	Coeff BIO08	6,000	1	-11.1	-5.8	-3.1	-0.2	6.2
	Coeff TXX	6,000	1	-3.5	2.9	6.2	9.6	15.9
	Coeff BIO02	6,000	1	-18.6	-13	-10	-7	-0.9
	Coeff TN90P	6,000	1	-11.1	-4.6	-1.1	2.3	8.6
	Coeff ID	6,000	1	-9.9	-3.6	-0.4	3	9.4
	Coeff BIO14	6,000	1	-6.3	0	3.1	6.1	11.7
	Coeff BIO18	6,000	1	-9.7	-3.5	-0.1	3.4	9.8
	Coeff CWD	6,000	1	-10	-3.4	0	3.3	10
	Coeff RX1DAY	6,000	1	-8.5	-2.2	1.1	4.4	11.1
	Coeff WSDI	6,000	1	-10.4	-4.7	-1.6	1.5	7.2
	NL BIO08	1,842	1	0.3	2.8	5.6	8.8	16.3
	NL TXX	2,068	1	14	18.4	21.5	25.3	36.2
	NL BIO02	2,275	1	6.7	10.6	13.3	16.7	25.4
	NL TN90P	2,441	1	10.8	15.8	19.4	24.2	36.8
	NL ID	2,738	1	25.3	38.3	48.5	61.5	103.3
	NL BIO14	3,435	1	3.6	5.5	7	8.9	15
	NL BIO18	2,433	1	58.8	78.8	94.1	112.5	166.6
	NL CWD	1,769	1	43.4	57.5	67.2	79.9	115.9
	NL RX1DAY	1,977	1	37.4	51.5	61.4	73.2	106.3
	NL WSDI	1,870	1	3.3	6.6	9.6	13.2	22.3
Adjusted	Intercept	2,334	1	-0.5	-0.3	-0.2	-0.1	0.1
	Coeff BIO08	3,445	1	0.8	0.9	1	1.1	1.3
	Coeff TXX	2,915	1	-2.7	-2.5	-2.4	-2.3	-2.1
	Coeff BIO02	4,738	1	1.2	1.3	1.4	1.4	1.6
	Coeff TN90P	3,516	1	1.2	1.3	1.4	1.5	1.6
	Coeff ID	1,971	1	1.7	2.5	2.8	3.1	3.5
	Coeff BIO14	4,055	1	0.4	0.5	0.5	0.6	0.7
	Coeff BIO18	4,788	1	-1.3	-1.1	-1	-0.9	-0.7
	Coeff CWD	3,512	1	-0.5	-0.3	-0.3	-0.2	0
	Coeff RX1DAY	4,992	1	-3.6	-3.4	-3.2	-3.1	-2.9
	Coeff WSDI	4,512	1	-1.1	-1	-1	-0.9	-0.8
	NL BIO08	6,000	1	0	0.3	0.7	1.2	2.3
	NL TXX	3,021	1	0.1	0.6	1.1	1.6	2.6
	NL BIO02	3,554	1	0.1	0.6	1	1.5	2.5
	NL TN90P	4,221	1	4	4.6	5	5.3	6.1
	NL ID	2,003	1	0.1	0.9	1.9	3	5
	NL BIO14	6,000	1	0	0.4	0.8	1.4	2.5
	NL BIO18	6,000	1	0	0.3	0.7	1.1	2.3
	NL CWD	4,819	1	0.1	0.5	1.1	1.7	2.9
	NL RX1DAY	4,863	1	2.5	3.4	3.9	4.3	5.2
	NL WSDI	6,000	1	0.6	1.3	1.7	2.1	3

Table A8: Posterior summary statistics of Bayesian GAM fit on scaled spatial-CV features. Like Table A5, only the values for the adjusted settings with normal priors are shown in the bottom half.

Prior/basis dimension set	**Parameters**	**n_eff**	$\widehat{\mathbf{R}}$	**Percentiles**				
				2.5	**25**	**50**	**75**	**97.5**
Initial	Intercept	6,000	1	-0.6	-0.2	0	0.2	0.6
	Coeff BIO08	6,000	1	-11.3	-5.9	-3.1	-0.2	6.2
	Coeff TXX	6,000	1	-3.3	2.9	6.2	9.6	15.8
	Coeff BIO02	6,000	1	-18.7	-13	-10	-6.9	-1
	Coeff TN90P	6,000	1	-10.8	-4.6	-1.2	2.2	8.5
	Coeff ID	6,000	1	-9.9	-3.5	-0.2	3	9.3
	Coeff BIO14	6,000	1	-6.1	-0.2	2.9	6	11.8
	Coeff BIO18	6,000	1	-10.1	-3.3	-0.1	3.2	9.7
	Coeff CWD	6,000	1	-10.2	-3.3	0	3.4	10.1
	Coeff RX1DAY	6,000	1	-8.7	-2.3	1.1	4.6	10.9
	Coeff WSDI	6,000	1	-10.2	-4.6	-1.6	1.4	7.4
	NL BIO08	1,784	1	0.3	2.9	5.6	8.8	16.6
	NL TXX	2,584	1	14.1	18.4	21.7	25.5	36.9
	NL BIO02	2,291	1	6.7	10.6	13.3	16.8	25.8
	NL TN90P	2,764	1	11	16	19.5	24.2	36.5
	NL ID	3,043	1	25.6	38.7	48.8	62.1	103.7
	NL BIO14	3,451	1	3.6	5.5	7	9	15.3
	NL BIO18	2,662	1	60	79.7	94.7	113.4	170.7
	NL CWD	2,134	1	42.9	57.3	67.4	80	115.1
	NL RX1DAY	2,061	1	38.4	52	61.5	74.1	108.9
	NL WSDI	1,836	1	3.3	6.6	9.8	13.4	22.5
Adjusted	Intercept	2,335	1	-0.5	-0.3	-0.2	-0.1	0.1
	Coeff BIO08	4,035	1	0.8	0.9	1	1.1	1.3
	Coeff TXX	3,205	1	-2.7	-2.5	-2.4	-2.3	-2.1
	Coeff BIO02	4,606	1	1.2	1.3	1.4	1.4	1.6
	Coeff TN90P	3,659	1	1.2	1.3	1.4	1.5	1.6
	Coeff ID	2,044	1	1.6	2.5	2.8	3.1	3.5
	Coeff BIO14	4,860	1	0.4	0.5	0.6	0.6	0.7
	Coeff BIO18	5,360	1	-1.3	-1.1	-1	-0.9	-0.7
	Coeff CWD	4,510	1	-0.5	-0.3	-0.3	-0.2	0
	Coeff RX1DAY	5,212	1	-3.6	-3.4	-3.2	-3.1	-2.9
	Coeff WSDI	4,610	1	-1.1	-1	-1	-0.9	-0.8
	NL BIO08	6,000	1	0	0.3	0.7	1.2	2.2
	NL TXX	3,056	1	0.1	0.6	1	1.5	2.5
	NL BIO02	4,474	1	0.1	0.7	1	1.5	2.5
	NL TN90P	4,667	1	4	4.6	5	5.4	6.1
	NL ID	2,136	1	0.1	1	2	3.1	5
	NL BIO14	6,000	1	0	0.4	0.8	1.4	2.5
	NL BIO18	6,000	1	0	0.3	0.7	1.2	2.2
	NL CWD	4,721	1	0	0.5	1	1.7	3
	NL RX1DAY	6,000	1	2.5	3.4	3.9	4.3	5.3
	NL WSDI	6,000	1	0.6	1.3	1.7	2.1	3

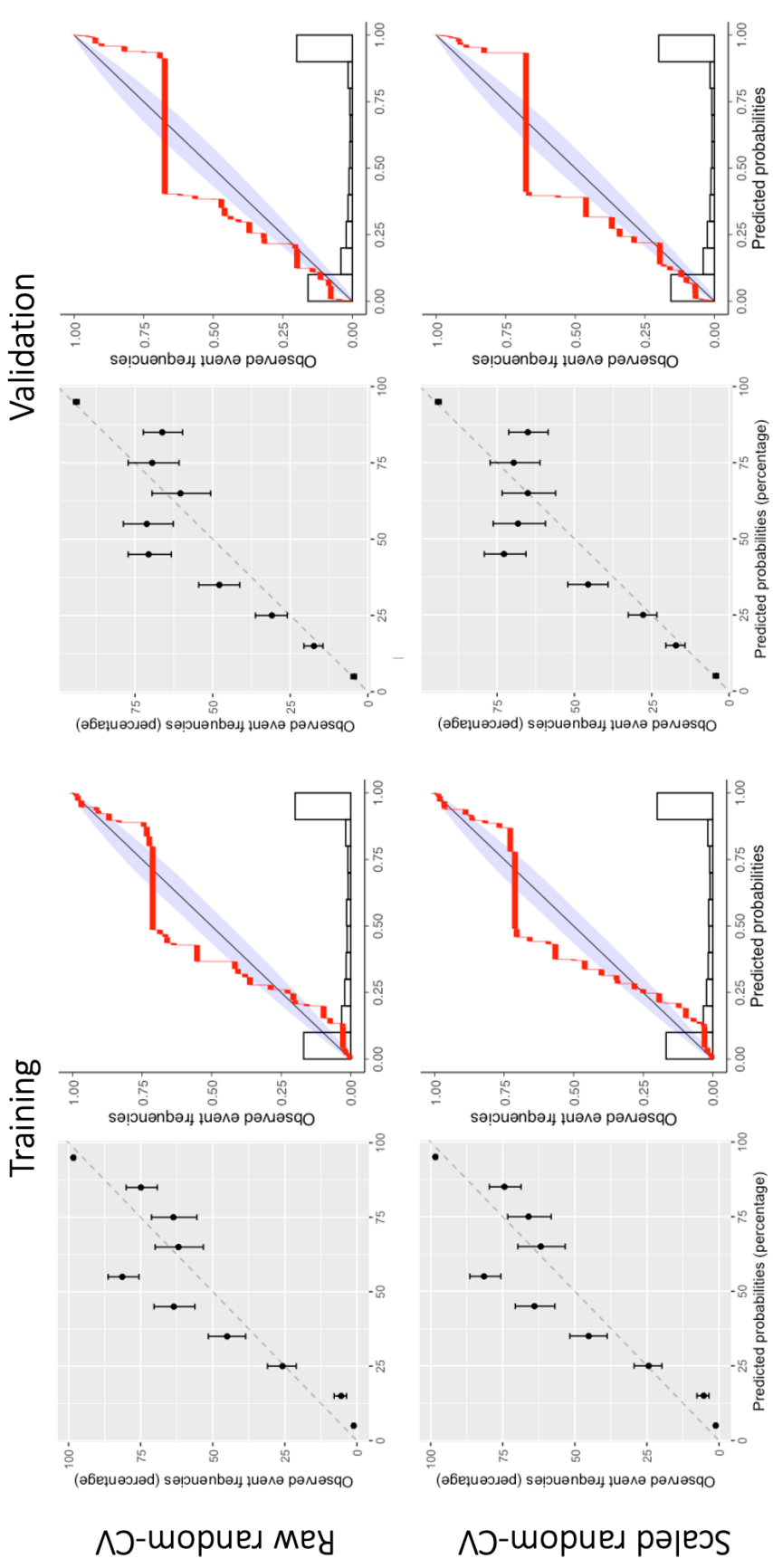


Figure A1: Calibration plots of the Bayesian GLM models fit on random-CV features using the initial prior settings. The calibration plots for the models fit on spatial-CV features are shown in Figure A2.

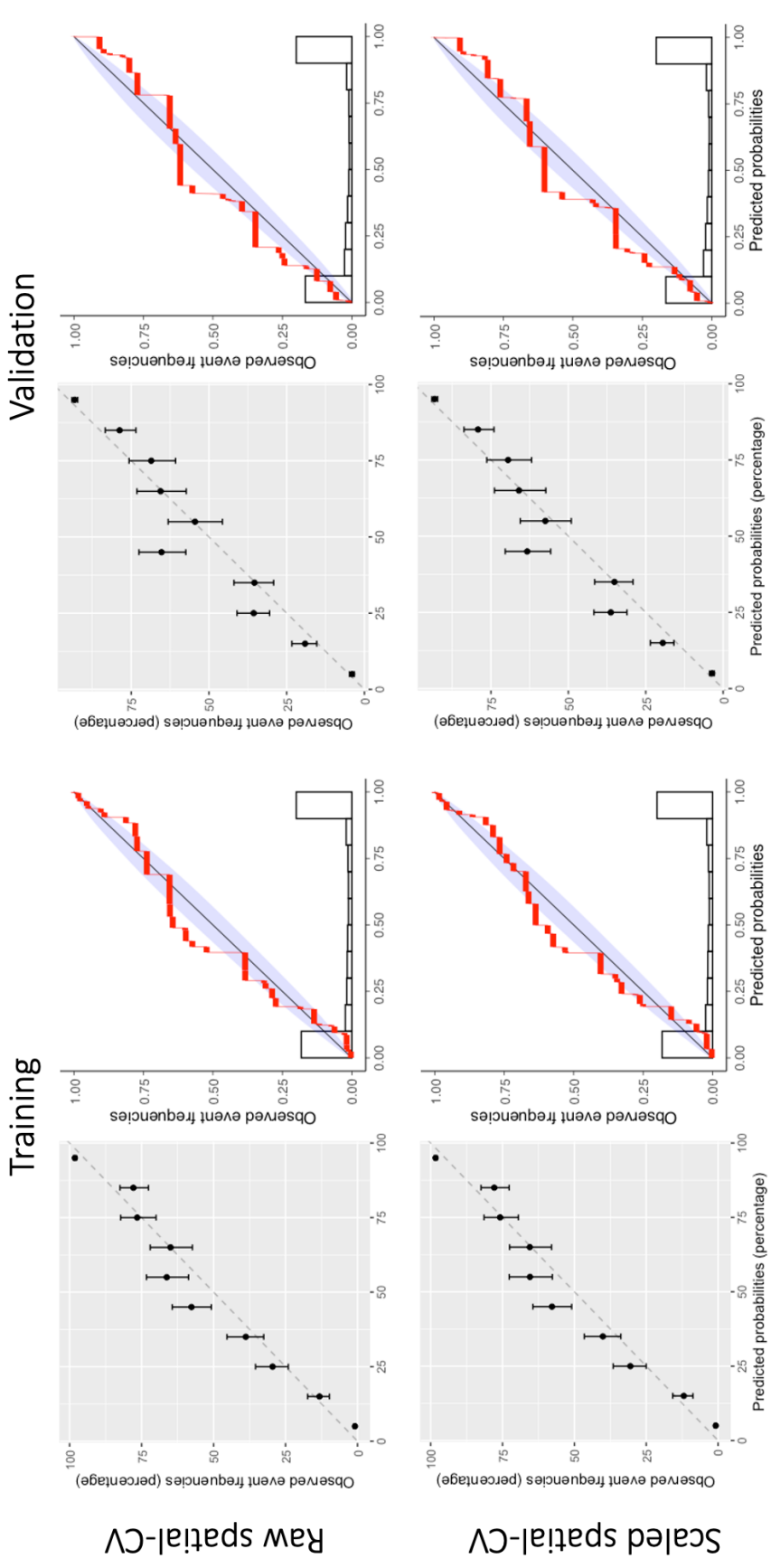


Figure A2: Calibration plots of the Bayesian GLM models fit on spatial-CV features using the initial prior settings. The calibration plots for the models fit on random-CV features are shown in Figure A1.

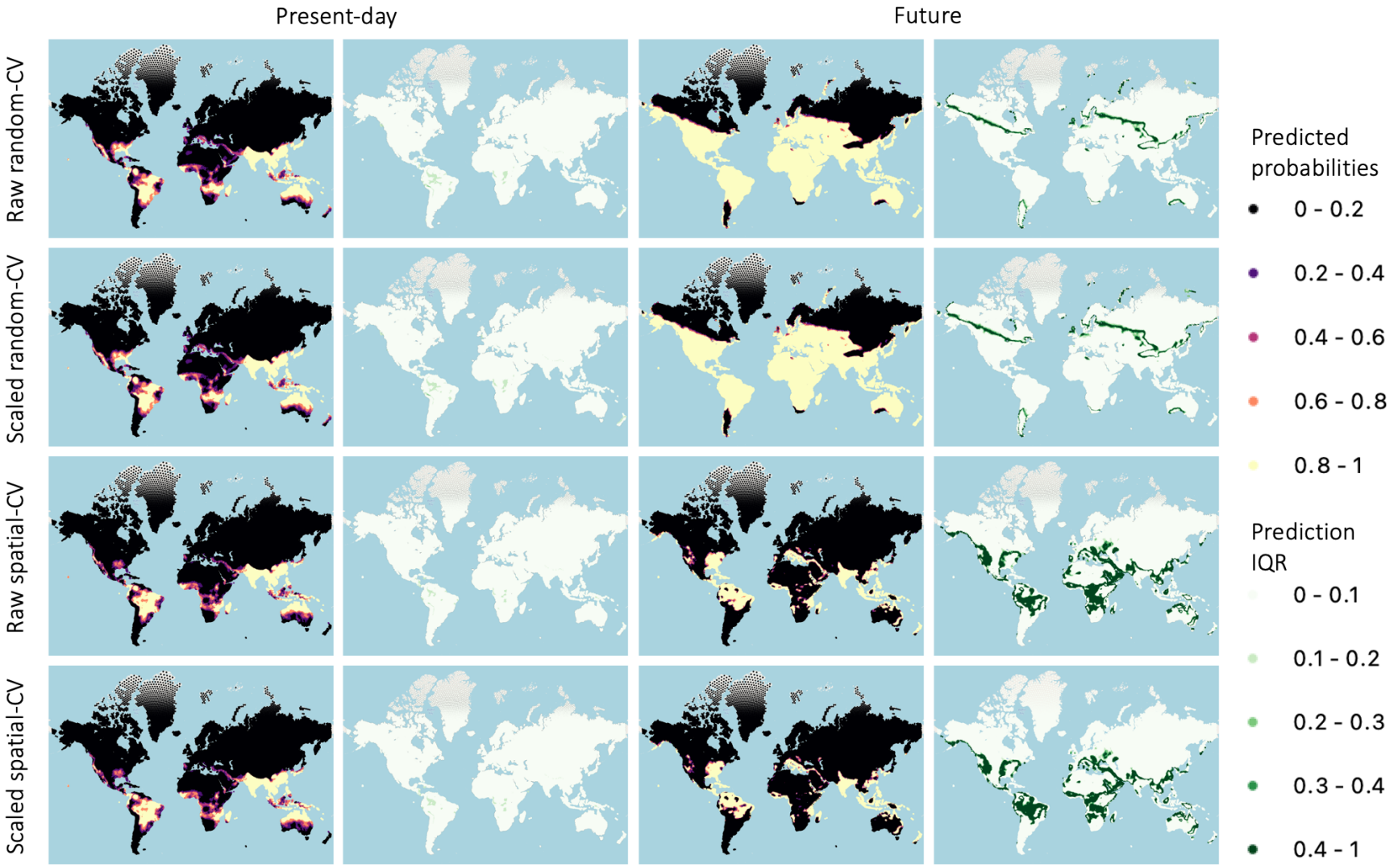


Figure A3: Predictions of the Bayesian GLM models using the initial prior settings.

Table A9: Diagnosis messages from the priorsense package for the Bayesian GLM models with initial prior sets. The priors of the models that had warning outputs were adjusted.

Feature set	**Parameters**	**Initial diagnosis**	
		Raw features	**Scaled features**
random-CV	Intercept	-	weak likelihood
	Coeff BIO03	-	-
	Coeff TN10P	-	-
	Coeff GSL	-	-
	Coeff TNX	-	-
	Coeff ID	-	prior-data conflict
	Coeff BIO14	prior-data conflict	-
	Coeff BIO18	-	prior-data conflict
	Coeff CWD	prior-data conflict	prior-data conflict
	Coeff RX1DAY	-	prior-data conflict
	Coeff WSDI	-	prior-data conflict
spatial-CV	Intercept	-	-
	Coeff BIO08	-	prior-data conflict
	Coeff TXX	-	prior-data conflict
	Coeff BIO02	-	prior-data conflict
	Coeff TN90P	-	-
	Coeff ID	-	prior-data conflict
	Coeff BIO14	-	prior-data conflict
	Coeff BIO18	-	prior-data conflict
	Coeff CWD	-	-
	Coeff RX1DAY	-	prior-data conflict
	Coeff WSDI	-	-

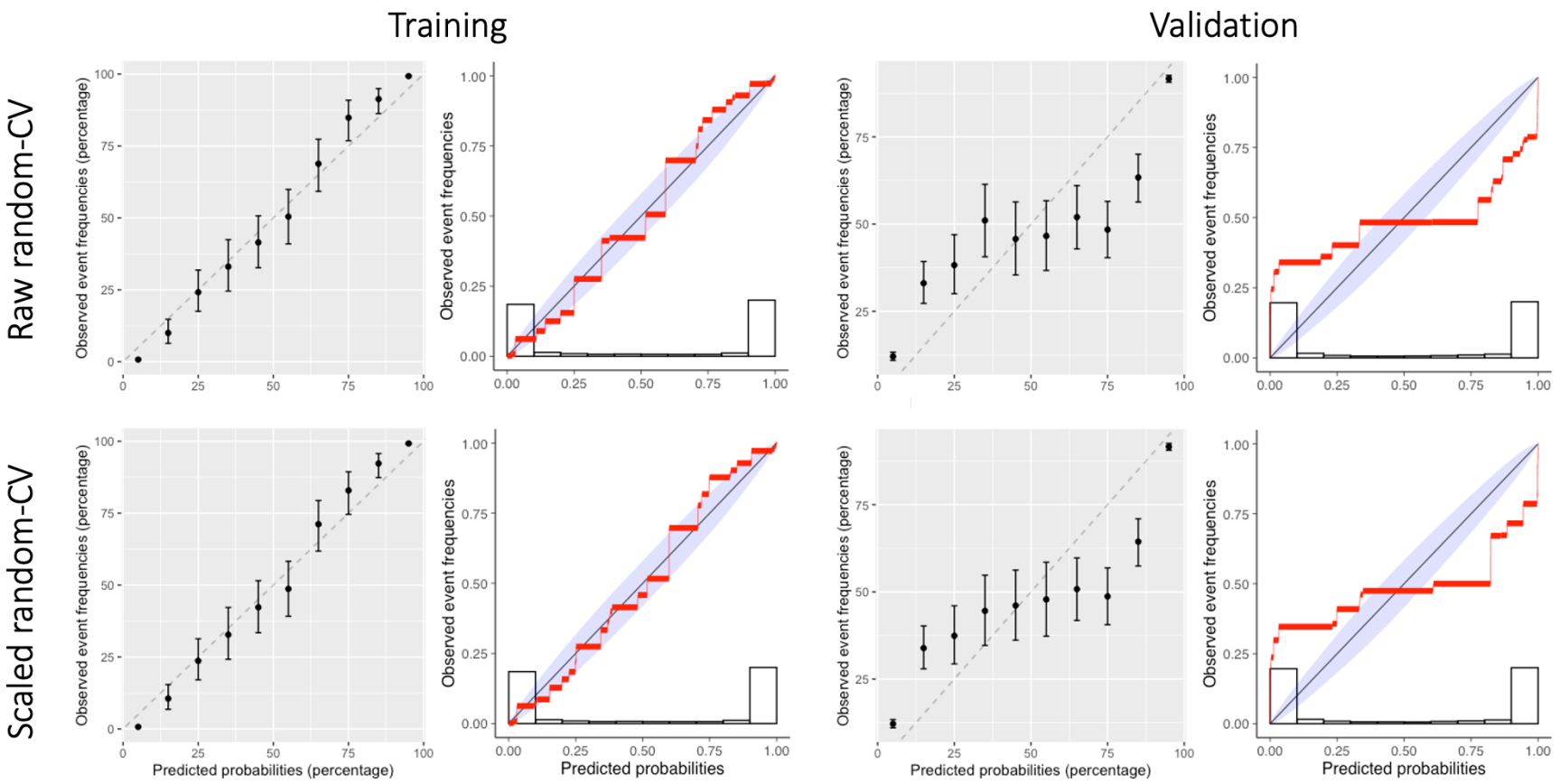


Figure A4: Calibration plots of the Bayesian GAM models fit on random-CV features using the initial prior and basis dimension settings. The calibration plots for the models fit on spatial-CV features are shown in Figure A5.

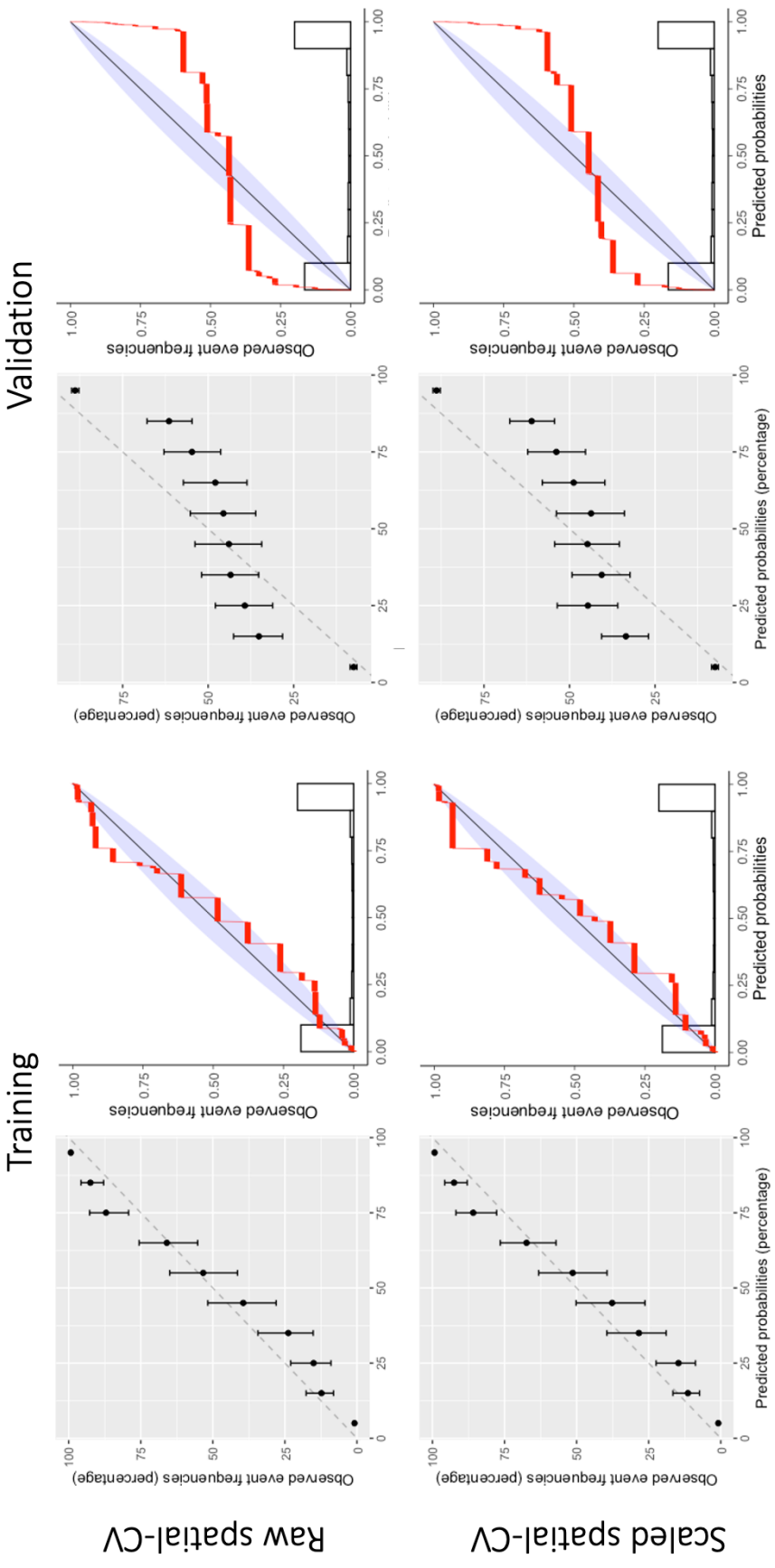


Figure A5: Calibration plots of the Bayesian GAM models fit on spatial-CV features using the initial prior and basis dimension settings. The calibration plots for the models fit on random-CV features are shown in Figure A4.

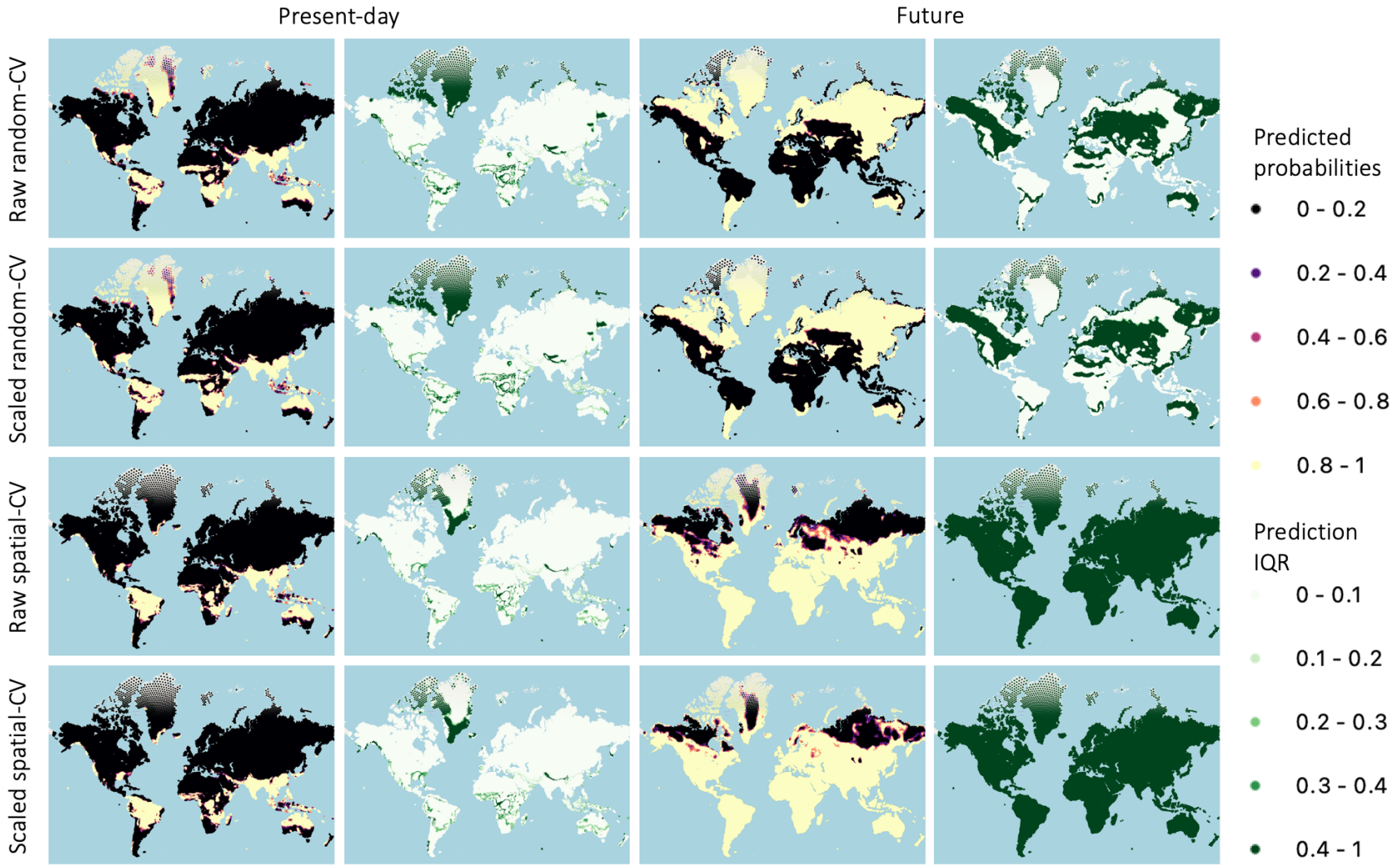


Figure A6: Predictions of the Bayesian GAM models using the initial prior and basis dimension settings.