Title: **Gaussian processes with numerical approximations for fast and flexible Bayesian species distribution modelling**

Running title: **Gaussian process species distribution modelling**

Word count: **6,139**

Authors:

Nick Golding

Spatial Ecology and Epidemiology Group,

Department of Zoology,

University of Oxford

and

Centre for Ecology & Hydrology,

Crowmarsh Gifford,

Wallingford

Bethan V. Purse

Centre for Ecology & Hydrology,

Crowmarsh Gifford,

Wallingford

Corresponding author email: nick.golding@zoo.ox.ac.uk

Abstract

1. Species distribution modelling (SDM) is very widely used in ecology and predictions from these models often inform both policy and ecological debates. It is therefore important to use methods with high predictive accuracy that permit sources of bias to be taken into account and enable biological interpretation.

2. Gaussian processes (GPs) are a highly flexible approach to statistical modelling and have recently been proposed for SDM. Typically, these models require computationally intensive Markov chain Monte carlo (MCMC) methods for inference, making them unsuitable for many SDM applications. We propose fitting GP distribution models using numerical approximations instead of MCMC in order to overcome these hurdles and make GPs more widely applicable. We provide an intuitive introduction to GPs for SDM and demonstrate how the approach can be used to account for common sources of bias in species distribution data. Using a dataset of 227 terrestrial vascular plant distributions within the UK we compare the predictive accuracy of GP models with those of widely used species distribution models.

3. Predictive accuracy of GP models fitted using numerical approxima-

tions was consistently higher than Boosted regression trees and Generalized additive models when trained on presence-absence plant data and greater both of these models plus MaxEnt when trained on presence-only data. This result was consistent under both randomly-stratified and geographically stratified training and evaluation datasets.

4. As well as offering grater predictive power than existing methods, GP models offer users the ability to account for imperfect occurence records, incorporate prior knowledge of the species' ecology and automatically estimate prediction uncertainty. We provide an open source R package GRaF, to allow ecologists to implement these models.

Keywords: Gaussian processes, Gaussian random fields, Boosted regression trees, MaxEnt, Generalized additive models, measurement error, Laplace approximation, R package

Introduction

Species distribution models (SDMs), in their basic form, attempt to model the distribution of species using environmental conditions as predictors. Typically these models make use of records of the distribution of the species in question and gridded datasets of environmental variables to generate maps of the species' predicted distribution. In recent years SDMs have become some of the most widely used methods in ecology (Elith & Leathwick, 2009), providing essential tools for both theoretical and applied research. Among other applications, SDMs are used to investigate drivers of global biodiversity patterns and to guide conservation policy and public health interventions (Lehmann *et al.*, 2002; Sinclair *et al.*, 2010; Sinka *et al.*, 2010).

A wide range of different approaches has been suggested for SDMs, ranging from relatively simple 'envelope' models and commonly used statistical methods such as logistic regression to more complex methods such as those developed in the field of machine learning (Elith *et al.*, 2006). These approaches have a number of features which determine their suitability to model species distributions including:

Predictive performance. Predictive accuracy is likely to depend on a number of factors, amongst which the ability to model complex (non-

linear) effects of and interactions between drivers of species distributions seems to be particularly important (Elith *et al.*, 2006). Preventing the model from overfitting to the training data (modelling random noise, rather than the true ecological response) will also increase predictive performance when the model is applied to new datasets drawn from different time periods or geographical regions (Wenger & Olden, 2012);

Imperfect data. SDMs are often applied to distribution records opportunistically collated from a variety of different sources, rather than from planned surveys. Such datasets are prone to various sources of error, such as observation bias, a lack of absence records and uncertainty in the location or reliability of individual records (Newbold, 2010; Elith *et al.*, 2010a). Failure to account for these sources of error can lead to biased model predictions.

Predicted distribution maps are often needed for species where few occurrence data are available (Pearson *et al.*, 2006). In these cases it may be useful to augment these limited data with existing knowledge of the species' ecology (Murray *et al.*, 2009).

Computational efficiency. SDMs are often required to be run in

batch operations, where the distributions of multiple species are modelled at once. For routine analyses such as these, SDMs which are relatively quick to run and do not require much user input are advantageous.

Prediction uncertainty. SDM predictions typically represent the model's *best guess* at the species' distribution, given the occurrence and environmental data available. These predictions are usually subject to multiple sources of uncertainty in the data and model parameters. It is therefore desirable to provide maps quantifying uncertainty in these predictions (Elith *et al.*, 2002; Guisan & Zimmermann, 2000).

Gaussian processes (GPs; also referred to as Gaussian random fields) provide a flexible approach to fitting complex statistical models (Rasmussen & Willians, 2006) and they have seen occasional use in ecology for modelling population dynamics (Patil, 2007; Sigourney *et al.*, 2012). GPs have recently been proposed as an alternative approach for fitting flexible, non-linear species distribution models (Vanhatalo *et al.*, 2012).

Fitting GP models typically requires the use of computationally expensive Markov chain Monte Carlo (MCMC) methods. Whilst MCMC is a useful approach to fitting complex models it can be very time consuming and requires an experienced user to supervise the model fitting process.

These limitations make models fitted using MCMC infeasible for the many SDM users lacking experience with MCMC and for applications which require running of large batches of models, such as for making predictions of species richness (Ferrier & Guisan, 2006). We propose that GP models fitted using efficient numerical approximations can overcome these drawbacks and provide a solution to a number of issues inherent in species distribution modelling.

Below, we illustrate how GP models work, demonstrate how they provide solutions to some problems commonly encountered in SDM and compare their predictive ability with other commonly used approaches on a large dataset of known vascular plant occurrence records. The advantages and limitations of GPs and potential avenues for future enhancements of the approach as applied to SDM are discussed. A software package GRaF for the statistical programming language R (R Development Core Team, 2012) is also provided to allow ecologists to employ these methods for fitting SDMs.

Gaussian process models

While most statistical models attempt to describe the relationship between covariates and the response variable by parameterising some equation (e.g. linear regression), GP models instead describe this relationship based on

an assumption that observations with similar covariate values will yield a similar response value. The model then uses the available data to construct a normally distributed, correlated 'process' of variables which give rise to the observed response variable.

The approach of fitting models based on similarity (or dissimilarity) between sites, rather than fitting directly to the covariates, is shared by a number of related 'kernel methods' such as kernel regression (used in Generalised dissimilarity matrix models (Ferrier, 2002; Ferrier *et al.*, 2002)), kernel support vector machines (Evgeniou *et al.*, 2005) and kriging (**?**). The flexibility of explicit Bayesian statistical treatment of GP models enables a number of useful model-fitting procedures and extensions (such as accounting for uncertainty in covariates, and incorporating prior ecological knowledge, as detailed below) which would be difficult to implement for many of these other methods.

The GP approach is widely used in the field of model-based geostatistics (Diggle & Ribeiro Jr, 2007), where the covariates are the geographic location of the observation and the GP therefore models spatial or temporal correlation. In applying GPs to species distribution modelling, we take this concept and apply it to *environmental* covariates to model the response variable - probability of occurrence.

A statistical explanation of how such a GP model may be formulated and fitted is given in Appendix S1. Here we provide a more intuitive illustration of how GP models differ from other SDMs, using as an example the effect of temperature on the probability of presence of a hypothetical species.

Covariance function

In order to construct the GP we first calculate the environmental distances between observations. In our example the environmental distances are simply the difference in temperature between each pair of sites (Fig. 1a). In a model with more covariates we would calculate the multi-dimensional Euclidean distance between pairs of sites in environmental space. We convert these distances into expected correlations in probability of presence between sites using a *covariance function*. There are a number of different covariance functions that we could use, but a good choice is the squared-exponential covariance function since it is easy to parameterise and produces ecologically plausible smooth curves (Rasmussen & Willians, 2006). As well as the distances between observations, we must supply the covariance function with a *lengthscale* parameter for each environmental covariate in the model. These lengthscales dictate how the correlation between probabilities of occurrence at pairs of observations decays with environmental distance; and therefore

the complexity of the fitted response curves. Fig. 1b shows how this function converts temperature difference to expected correlation given three different lengthscales. Assuming a lengthscale of one degree celsius, the expected correlation between two observations with a one degree difference in temperature is around 0.6, whereas with a difference of two degrees this drops to around 0.14. With a longer (higher valued) lengthscale, these expected correlations will be higher, resulting in a less complex fitted line (Fig. 2). In practice these lengthscales do not need to be specified in advance as they can be estimated from the data, though we may wish to inform the model of how likely different lengthscales are for the species being modelled.

Mean function

In addition to these expected correlations, we specify a *mean function*: an initial estimate of how the response variable changes with the covariates which is later updated by the model fitting procedure. If nothing is known about how the probability of presence of a species responds to the temperature gradient, a flat mean function may be used, which assumes an equal probability of presence at all temperatures. If we have some prior knowledge that the species is more likely to be present at low temperatures than at high temperatures, we can incorporate this information into the model. For

example, the mean function could be a linear model (with fixed parameters) relating temperature to probability of presence.

Fig. 2 demonstrates the effects of these two different mean functions on our model, with varying lengthscales. We can see from this illustration that where there are a sufficient number of observations, the mean function has little effect on the fitted line, but where there are few data points, such as toward the limits of the recorded temperature range, the mean function determines the shape of the fitted response.

GP species distribution models

Model structure

Machine learning algorithms such as boosted regression trees (BRT; Elith *et al.* (2008a)) have been shown to perform very well at predicting species distributions, likely due to their ability to fit complex and highly non-linear responses to environmental covariates (Elith *et al.*, 2006).

A drawback of BRT and similar methods is that they fit 'jerky' and biologically implausible predictive responses which may contribute to their tendency to overfitting to training data (i.e. they fit to noise in the data as well as the species' distribution, Wenger & Olden (2012)). By comparison, more traditional approaches such as univariate generalized additive models (GAM; Hastie & Tibshirani (1986)) fit more biologically realistic smooth functions. Whilst some implementations of GAM can fit multivariate smoothers (Wood, 2011), they perform poorly in more than a few dimensions so are unable to account for complex interactions.

GPs (using the squared exponential covariance function) offer a attractive solution to this trade-off between model flexibility and ecological realism by allowing for interactions between and highly non-linear effects of covariates, whilst fitting biologically plausible smooth predictive surfaces

(see Fig. 3).

Uncertainty in occurrence data

Modellers often want to make high resolution predictions from high resolution gridded environmental data but are hampered by the low resolution of the species occurrence data. This mismatch in resolution is problematic when the modeller needs to extract covariate values corresponding to each record, since there will be a number of different covariate values from which to choose. Thus the problem of uncertainty in the record location can be more usefully considered as uncertainty in the measured value of the environmental covariate for each occurrence record. A simple approach to this problem is to fit the model using the mean of the covariate values. Whilst straightforward to implement, this approach ignores the uncertainty in the covariate and can lead to regression dilution, which dampens the apparent effect of a covariate on the species distribution.

The problem of regression dilution has been well studied in statistics (Frost & Thompson, 2000) and measurement error models have been proposed to deal with this bias in SDMs (McInerny *et al.*, 2011) and other ecological models (McNamara & Harding, 2004). As well as the mean of the environmental covariates for each record, measurement error models use an

estimate of the error variance (which in the case of spatial uncertainty may be calculated directly from the multiple covariate values for each record) and use this information to correct for the regression dilution effect. Whereas most measurement error models are fitted by MCMC, we can fit these models in the numerical approximation GP framework at negligible computational cost by accounting for this error directly within the covariance function (see Fig. 4). This approach is detailed in full in (Dallaire *et al.*, 2011).

GP models also allow users to provide regression weights to individual records in order to account for the variable reliability of different records or to account for observation bias (Phillips *et al.*, 2009) in a similar way to the case weights and bias grids available in other SDMs (Elith *et al.*, 2010a).

Incorporating prior ecological knowledge

Often when modelling species distributions with few occurrence data there are other forms of information about the ecology of the target, or similar, species which could be used to improve the model. For example experimental studies may have demonstrated a relationship between the species' ability to persist and some environmental gradient. In such cases it may be desirable to incorporate this prior knowledge of the species ecology into the model to augment the occurrence data. Unfortunately this is not easily accomplished

in many current SDMs.

Bayesian statistical inference provides a convenient way of incorporating prior information of this sort into statistical models and has become increasingly popular in ecology (see e.g. McCarthy (2007)). In a Bayesian model the user specifies a prior probability distribution over each model parameter, representing their existing knowledge about what values of the parameter are likely. The model then compares this prior probability with the parameter estimate suggested by the data and produces a form of weighted average over the two; the posterior distribution. As the amount of data available to the model increases, the impact of the prior on the posterior diminishes.

The GP framework allows the user to incorporate ecological knowledge into distribution models by manipulating two Bayesian priors: the mean function and the lengthscale hyperprior. The mean function acts as a prior over the whole model and can be used to incorporate specific knowledge of the species' response to environmental gradients. The lengthscale hyperprior determines how likely different lengthscales are and can be used to inform the model how rapidly probability of presence is likely to change with different values of the environmental covariates.

In the absence of any prior information, a flat mean function can be used, as in Fig. 2. Similarly a flat lengthscale hyperprior could be used, indicating

that all levels of complexity in the fitted response are *a priori* assumed to be equally likely. By default the R package GRaF which we provide uses a flat mean function but an informative lengthscale hyperprior which represents ecologically plausible response curves (detailed in Appendix S1).

Uncertainty in model predictions

As with any model, predictions from SDMs are uncertain estimates of the probability of presence of the species. Where these predictions are to be used for some practical purpose it would be beneficial to provide maps representing the uncertainty in the predicted distribution map (Elith *et al.*, 2002). Such maps allow users to determine how much confidence they can place in a given prediction, information which is especially valuable where the predictions have policy implications.

SDM uncertainty estimates can be produced by bootstrapping data (Elith *et al.*, 2002) though this requires models to be run many hundreds of times and can therefore be computationally prohibitive. GP models automatically produce estimates of uncertainty in model predictions, without the need for bootstrapping procedures, since these are calculated directly from the estimated posterior distribution of the model. Fig. 5 illustrates predictions and associated uncertainty estimates from a GP distribution model of a plant

297 species, Bog Myrtle (*Myrica gale*), in the UK.

Comparison of GPs with existing SDMs

We compared the predictive ability of GPs (fitted using GRaF) with commonly used approaches for modelling species distributions from both presence/absence and presence-only data under at both interpolation and extrapolation tasks. All model fitting, predictions and statistical analyses were performed in R version 3.0.0. R code used to carry out these analyses and to plot all figures in this manuscript are provided in Appendix S5. The dataset used to perform these comparisons is available to download from Figshare [DOI/URL tbc].

Methods

Data

Gridded presence/absence maps of native terrestrial vascular plant species of Great Britain at 10 km resolution were obtained from the New Atlas of the British and Irish Flora (Preston *et al.*, 2002). Of the 1335 distributions in the original dataset, a subset of 227 species of different genera was selected for the model comparison. Criteria for selection of these species are outlined in Appendix S3. The distribution of British plant species is well characterized at this spatial scale, so records were assumed to represent known presence or absence which is essential to compare models fairly. The

227 plant species selected had a wide range of prevalences (proportion of grid squares occupied; ranging from 0.075 to 0.925, median 0.394), a factor known to influence the accuracy of SDMs (McPherson *et al.*, 2004) and inhabited a wide range of habitats (see Appendix S3).

Gridded maps of 10 indices of environmental conditions were used as covariates for model fitting and prediction. These were derived from a time series of satellite images of the UK by subsequent Fourier decomposition and principal components analysis to produce variables representing the major axes of environmental variation in the UK. The advantage of these abstract indices is that they compress a large amount of information on conditions and seasonality (the ten principal components explain 90% of variation in the Fourier variables) into relatively few orthogonal variables which enable us to make accurate predictions (Dormann *et al.*, 2008). Whilst they are difficult to to interpret biologically, the first three indices broadly correspond to gradients from arable land to pasture, lowlands to uplands, and urban areas to rural areas respectively. Details of this dataset and how it was produced are given in Appendix S3. All models were fitted using the full set of 10 covariates.

A total of 2774 10 km grid squares contained both distribution data and environmental data and were used to compare the different modelling

approaches.

Presence/absence models

For the presence/absence comparison we compared GPs with BRT and GAM. For each of the 227 species 300 grid squares (10.8% of the 2774 available records) were used to train each of the three models. This number of observations was selected as a practical compromise between having a sufficient training set, and the need to have sufficient presence and absence records in the training and evaluation datasets whilst accommodating the very rare and very common species.

Model predictions for the remaining 2474 grid squares were then used to compare the predictive ability of the models. GP models were fitted using GRaF 0.1-12 (Golding, 2013), optimising the lengthscale parameters and otherwise using default settings. BRT models were fitted using the gbm package version 2.1 (Ridgeway, 2013) with 5-fold cross validation, a tree complexity of 5, a learning rate of 0.001 and a minimum of 1000 trees were fitted (in accordance with Elith *et al.* (2008a)). The optimal number of trees in the final BRT model was selected from the cross-validation folds using the gbm.perf function. GAM models were fitted using the gam package version 1.09 (Hastie, 2011) with univariate spline smoothers for each covariate and

default settings.

Presence-only models

For the presence-only data we compared GPs with MaxEnt (Phillips *et al.*, 2006), one of the most widely used approaches for modelling species distributions (Yackulic *et al.*, 2012). Since absence data is not available in presence-only datasets, it is necessary to augment the presence data with a set of 'background data' against which the models may contrast the environmental signature of the presence data. For each species 100 presence records were selected from the dataset along with a further 1000 background points selected from the remaining data. As for the presence/absence case, this number of sampling points was a compromise between accommodating a range of species prevalences and the number of available records; though real-world applications of presence-only SDM typically use similar numbers of occurrence points (studies reviewed by Yackulic *et al.* (2012) used a median of 146 occurrences).

GP and MaxEnt models were fitted to these 1100 data points and used to predict the relative probability of presence at the remaining 1674 grid squares not used to fit the models. GP models were all fitted as for presence-absence data and MaxEnt models were fitted using dismo version 0.8-11

(Hijmans *et al.*, 2012) with default settings.

Defining the evaluation dataset

SDMs are widely used both for interpolation ('filling in' the species distribution in the region from which occurrence data is available) and extrapolation (prediction to geographically distinct regions) (Elith & Leathwick, 2009). Previous comparisons of SDM approaches have focussed the interpolation capacity of these models by selecting an evaluation set at random from the available data (Elith *et al.*, 2006). This procedure is unlikely to accurately reflect the extrapolation capacity of the various models and where the available data are spatially dense, validation statistics calculated using this procedure are likely to be artificially inflated as a result of spatial autocorrelation between the training and evaluation sets (Wenger & Olden, 2012). We evaluated the interpolation and extrapolation capacities of the different approaches by running model comparisons using both a randomly stratified and a geographically stratified dataset for the presence/absence and presence-only test sets.

For the randomly stratified tests, training and evaluation data were sampled at random from the dataset, subject to the constraint that at least ten presence and ten absence points were available in the evaluation set and the

same minimum number present in the training set for the presence/absence comparison.

For the geographically stratified tests, a disc with a radius of 250km was used to divide the training and evaluation datasets. All grid cells falling within this region were used as evaluation data. Grid cells outside both this disc, and an additional buffer region of 20km around the disc were used as the pool from which to draw the training data. Cells falling within the buffer region were discarded. For each species and test (presence/absence or presence-only) the disc was centred on a different randomly selected grid cell, subject to the constraint that at least ten presence and ten absence points were available in the evaluation set and in the training dataset for the presence-absence test. This procedure was carried out using a rejection algorithm (provided as R code in the supplementary material), incorporating functions from the sperrorest R package version 0.2-1 (Brenning, 2012).

Validation statistics

When fitting presence-absence statistical models to presence-background data, the background data is implicitly assumed to represent absence of the species. As a result, these models are subject to prediction error caused by 'contamination' of background points with those in which the species is

in fact present (Ward *et al.*, 2009; Elith *et al.*, 2010b). MaxEnt, by contrast, explicitly considers this data as background data and is unaffected by this form of bias. However, since the prevalence of each species in the study area is unidentifiable from such data (Ward *et al.* (2009); Phillips & Elith (2013) - unlike in the presence-absence case), predictions from both MaxEnt and presence-absence models applied to presence-only data, are not of the absolute probability of presence of the species, but only an uncalibrated or *relative* probability of presence (Elith *et al.*, 2010b).

We therefore assessed predictions from presence-absence and presence-only models using different validation statistics. For presence-absence models calculated the log-likelihood of the withheld data from the predictions of each model. The Log-likelihood accurately assesses the quality of probabilistic predictions for predicting the true probability of presence and is preferable where it can be applied (Lawson *et al.*, 2013).

For presence-only models we calculated the Area Under the Receiver Operating Statistic Curve (AUC; calculated using the pROC R package - version 1.6.0.1; Robin *et al.* (2011)) which assesses the ability of each each model to correctly rank sites in order of probability of presence and is therefore preferable where predictions are of the relative probability of presence (Lawson *et al.*, 2013).

For both statistics higher scores indicate a better fit to the data.

Statistical analysis

These validation statistics were analysed by mixed-effects regression, implemented using the nlme R package version 3.1-113 (Pinheiro *et al.*, 2012). In each regression the response variable was the metric of predictive performance (log-likelihood or AUC) and the covariates were the SDM model type (modelled as a fixed effect) and plant species (modelled as a random effect in order to account for the nested study design). As the residual variances differed between model types, a separate variance parameter was estimated for each SDM model type. The statistical significance of differences between model validation statistics was assessed by t-tests on coefficients for the SDM model type.

In the geographically-stratified presence-absence model comparison, GAM models for 11 species had very low log-likelihood scores (all less than -3000) and had to be omitted to enable the mixed effects regression model to converge. These low scores are indicative of a failure of the GAM fitting algorithm to converge to a sensible result, and would likely be rejected by an SDM user during the model fitting process in a real-world application of the method. Exclusion of these models resulted in higher average prediction

metrics for GAMs; however as the validation statistics for GAM models in this comparison were markedly lower than other models, the results of the comparison are still robust.

Marginal validation statistic scores were calculated from the residuals of null models with an intercept term and random effects terms for plant species, but no fixed effect of model type. These marginal statistics enable us to visualise the expected predictive capacity from each SDM whilst removing species-level effects, essentially representing likely model performance expected for a 'typical' species in the dataset, averaging out the effects of species-specific ecology or of the species' prevalence.

Results

GP models made more accurate predictions to the withheld data than other models for both presence-absence and presence-only data and both under randomly and geographically stratified training/evaluation sets (Fig. 6).

Random stratification

Predictive log-likelihoods for presence/absence GP models under random stratification were 34.95 ($\pm$ 1.68 SE, $t_{452} = 20.82$, p<0.0001) higher than for BRT and 157.89 ($\pm$ 6.49 SE, $t_{452} = 24.35$, p<0.0001) higher than for

GAM models.

Similarly, presence-only GP models had an average AUC score of 0.81; 0.024 ($\pm$ 0.002 SE, $t_{678} = 14.58$, p<0.0001) higher than for BRT models; 0.019 ($\pm$ 0.001 SE, $t_{678} = 17.81$, p<0.0001) higher than for GAM models and 0.018 ($\pm$ 0.001 SE, $t_{678} = 14.08$, p<0.0001) higher than for MaxEnt models.

GP models explained an average of 28.6% of null deviance for the presence/absence evaluation sets, compared with 25.8% for BRT and 0.17% for GAM models.

Geographic stratification

Predictive log-likelihoods for presence/absence GP models under geographic stratification were 16.31 ($\pm$ 5.08 SE, $t_{441} = 3.21$, p<0.0014) higher than for BRT and 231.46 ($\pm$ 22.50 SE, $t_{441} = 10.29$, p<0.0001) higher than for GAM models.

Under geographic stratification, presence-only GP models had an average AUC score of 0.68; 0.018 ($\pm$ 0.004 SE, $t_{678} = 4.38$, p<0.0001) higher than for BRT models; 0.008 ($\pm$ 0.003 SE, $t_{678} = 2.31$, p<0.0212) higher than for GAM models and 0.021 ($\pm$ 0.004 SE, $t_{678} = 5.83$, p<0.0001) higher than for MaxEnt models.

Under the more stringent geographic stratification test, all models explained less of the null deviance in the presence/anbsence evaluation sets than under the random stratification, with GP models explaining an average of 11.6%, BRT models 9.3% and GAM models 3.6%.

Discussion

SDM comparison

In our comparison GP SDMs clearly outperformed a number of popular SDM approaches, including BRT which has been shown to be one of the best performing of existing SDM approaches (Elith *et al.*, 2008b).

In this comparison, we fitted each model following best-practice guidelines where available and default settings otherwise. We also compared models across a very large dataset of species distributions with varying ecologies and prevalences. We therefore consider this to be a fair comparison of the SDMs considered. However further comparisons of these methods using different datasets and by different modellers, would be useful in order to evaluate the performance of GP species distribution models across a broader swathe of SDM applications.

Advantages of a Bayesian approach

The ability of GP models to incorporate prior ecological knowledge and to account for uncertainty in occurrence locations stems from the use of a Bayesian statistical approach. These features are likely to prove useful where there are few occurrence records (Murray *et al.*, 2009) and where

there is a well-understood environmental driver of a species' distribution, such as a temperature limit on the distribution of a pathogen (Gething *et al.*, 2011). This method of incorporating prior knowledge could also be used to integrate process-based ecological models (Dormann *et al.*, 2012) with the more commonly used correlative SDMs.

By taking a Bayesian approach, GPs can produce estimates of uncertainty in model predictions, by considering a probability distribution over the predicted values (accounting for uncertainty in the shape of the GP) rather than making a single 'best guess' prediction, as is the case with most SDMs.

Whilst predictions from GPs fitted using the numerical approximation procedures described here (and implemented in the GRaF R package) account for uncertainty in the shape of the GP (allowing us to produce credible intervals around species response curves), they do not account for uncertainty in the lengthscale hyperparameters (which control how dynamic the effects of environmental covariates are on probability of presence), but are conditional on an optimum estimate calculated from the dataset. Conditional posterior predictions are likely to satisfy most SDM users' requirements since most widely used SDMs do not provide any measure of prediction uncertainty, let alone accounting for uncertainty in hyperparameters. If

required, predictions accounting for uncertainty in hyperparameters can be approximated by numerical integration (e.g. a deterministic algorithm as in (Rue *et al.*, 2009) or by Monte Carlo). Such a procedure will inevitably be more computationally intensive, but is likely to be far more efficient than alternative approaches such as MCMC.

Computational efficiency

GP models fitted using numerical approximations are reasonably computationally efficient, with model fitting times to the 300 presence-absence datasets in our comparison similar to those of BRT models (GP: mean 39.95 seconds ± 9.32 SD cf. BRT: 10.81 seconds ± 0.89 SD, n = 227). GPs are fairly efficient for datasets of up to a few thousand data points and running our implementation on a desktop computer is likely to be sufficiently fast for the majority of users.

This computational efficiency is achieved by carrying out an efficient numerical approximation routine (by default Laplace approximation) to fit the model to the data. Predictions resulting from these models are therefore subject to a degree of approximation error which may limit their performance. In Appendix S2 we discuss this issue further and compare the accuracy and of these numerical approximations with those of an MCMC

approach which would be computationally and technically prohibitive for the vast majority SDM applications, by users without MCMC experience and limited computing resources.

A downside to GP models is that in the naive case they scale cubicly with the size of the dataset (due to multiple matrix decompositions of $\mathcal{O}(n^3)$ complexity), so for very large data sets, GP models can be disproportionately slow. For users who wish to fit GP models to very large datasets efficiently, substantial speed-ups can be achieved by exploiting parallel computing. Our R implementation GRaF uses R's base functions for linear algebra (which in turn call third-party linear algebra libraries) and consequently these computations can be efficiently parallelised, without any additional coding, simply by linking R to a parallel linear algebra library (see e.g. Schmidberger *et al.* (2009)). Planned updates to GRaF include optional sparse approximation methods (see e.g. Vanhatalo *et al.* (2010)) which should enable major improvements in computational efficiency for large models.

Model complexity

GP models can account for highly complex interactions between covariates, a feature which probably contributes greatly to their strong predictive performance. When using complex models such as this it is important to avoid

overfitting to the training data - finding patterns in random noise and therefore producing biased predictions to new datasets. Common approaches to dealing with this problem include covariate selection procedures, which seek to find a parsimonious subset of the available covariates which explain the data well; and regularization (e.g. the 'lasso' which is used in MaxEnt, among others (Tibshirani, 1996)) which is used during model fitting to include only the most important covariates.

It is common to tune hyperparameters of GP models using the model marginal likelihood (Rasmussen & Willians (2006); this is the approach taken in GRaF implementation) an approach which automatically reaches an optimal trade off between model fit and complexity. GPs therefore reduce the influence of environmental covariates with little explanatory power and prevent model overfitting, without requiring the user to carry out a covariate selection procedure.

A downside of fitting models with high-dimensional and non-linear interactions is that it becomes harder for the user to interpret the relationship between the species and its environment. This problem is not unique to GPs, but is an inevitable trade-off when modelling complex data.

The GRaF R package provides functions to help users to interrogate GP models. The effects of individual covariates on probability of presence

589 can be visualised, as can two-way interactions between covariates. Mod-
590 els can also be compared using deviance information criteria (Spiegelhalter
591 *et al.* (2002); an equivalent of commonly used information criteria, such as
592 Akaike's (Akaike, 1973), for hierarchical models) to quantify the relative
593 importance of covariates in driving the species' distribution. A worked ex-
594 ample modelling the distribution of Bog Myrtle and demonstrating these
595 features is provided in Appendix S4.

596 **Future development of the GRaF R package**

597 Though GRaF is currently designed to fit the kinds of SDMs that are most
598 commonly used at present, the approach could be extended in a number
599 of ways. Since latent GP models include, as a special case, a large sub-
600 set of generalised linear mixed-effects models (Rue *et al.*, 2009), GRaF
601 could easily be extended to model abundance data, nested study designs
602 and spatio-temporal autocorrelation. Multiple-response GP models could
603 also be implemented to allow users to fit models for whole communities of
604 species, parameterising and accounting for correlations between their dis-
605 tributions (Wisz *et al.*, 2013; Kissling *et al.*, 2011). GraFs open source R
606 implementation will facilitate these extensions.

Acknowledgements

The authors acknowledge funding from the NERC Centre for Ecology & Hydrology (CEH) Environmental Change Integrating Fund Programme. We thank Chris Preston and David Roy from the Biological Records Centre, CEH, Wallingford and the Botanical Society of the British Isles for providing access to the plant atlas data and kindly agreeing to its dissemination with this paper. We thank the maintainers of the CEH NEMESIS computing cluster and David Rogers, Miles Nunn, Luigi Sedda, David Harris, Bob O'Hara and Marianne Sinka who all provided helpful comments on the manuscript.

References

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. B.N. Petrov & F. Caski, eds., *Proceedings of the Second International Symposium on Information Theory*, pp. 267 – 281. Budapest.

Brenning, A. (2012) Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the r package 'sperrorest'. *IEEE International Symposium on Geoscience and Remote Sensing IGARSS*. In press.

Dallaire, P., Besse, C. & Chaib-draa, B. (2011) An approximate inference with Gaussian process to latent functions from uncertain data. *Neurocomputing*, **74**, 1945–1955.

Diggle, P.J. & Ribeiro Jr, P.J. (2007) *Model-based geostatistics*. Springer, New York.

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S. & Schröder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H.,

635 Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. &
636 Singer, A. (2012) Correlation and process in species distribution models:
637 bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

638 Elith, J., Burgman, M.a. & Regan, H.M. (2002) Mapping epistemic uncer-
639 tainties and vague concepts in predictions of species distribution. *Ecolog-*
640 *ical Modelling*, **157**, 313–329.

641 Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan,
642 A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J.,
643 Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M.,
644 Nakazawa, Y., Overton, J.M.M., Townsend Peterson, A., Phillips, S.J.,
645 Richardson, K., Scachetti-Pereira, R., Schapire, Robert, E., Soberón, J.,
646 Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods im-
647 prove prediction of species distributions from occurrence data. *Ecography*,
648 **29**, 129–151.

649 Elith, J., Kearney, M. & Phillips, S.J. (2010a) The art of modelling range-
650 shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

651 Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological
652 Explanation and Prediction Across Space and Time. *Annual Review of*
653 *Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Leathwick, J.R., Hastie, T. & R. Leathwick, J. (2008a) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Elith, J., Leathwick, J.R., Hastie, T. & R. Leathwick, J. (2008b) Elith, Leathwick & Hastie A working guide to boosted regression trees - Online Appendices Page 1. *Journal of Animal Ecology*, **77**, 802–13.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2010b) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Evgeniou, T., Micchelli, C. & Pontil, M. (2005) Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, **6**, 615–637.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Ferrier, S., Watson, G., Pearce, J.L., Drielsma, M. & Manion, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity & Conservation*, **11**, 2275–2307.

672 Frost, C. & Thompson, S.G. (2000) Correcting for regression dilution bias:
673 comparison of methods for a single predictor variable. *Journal of the*
674 *Royal Statistical Society: Series A (Statistics in Society)*, **163**, 173–189.

675 Gething, P.W., Van Boeckel, T.P., Smith, D.L., Guerra, C.a., Patil, A.P.,
676 Snow, R.W. & Hay, S.I. (2011) Modelling the global constraints of temper-
677 ature on transmission of Plasmodium falciparum and P. vivax. *Parasites*
678 *& Vectors*, **4**, 92.

679 Golding, N. (2013) *GRaF: Species distribution modelling using latent Gaus-*
680 *sian random fields*. R package version 0.1-0.

681 Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution
682 models in ecology. *Ecological modelling*, **135**, 147–186.

683 Hastie, T. (2011) *gam: Generalized Additive Models*. R package version
684 1.06.2.

685 Hastie, T. & Tibshirani, R. (1986) Generalized additive models. *Statistical*
686 *Science*, **1**, 297–318.

687 Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2012) *dismo: Species*
688 *distribution modeling*. R package version 0.7-17.

689 Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McIn-

690 erny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M.,
691 Singer, A., Svenning, J.C., Zimmermann, N.E. & OHara, R.B. (2011) To-
692 wards novel approaches to modelling biotic interactions in multispecies
693 assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–
694 2178.

695 Lawson, C.R., Hodgson, J.A., Wilson, R.J. & Richards, S.A. (2013) Preva-
696 lence, thresholds and the performance of presence-absence models. *Meth-*
697 *ods in Ecology and Evolution*, **4**.

698 Lehmann, A., Leathwick, J. & Overton, J. (2002) Assessing New Zealand
699 fern diversity from spatial predictions of species assemblages. *Biodiversity*
700 *& Conservation*, pp. 2217–2238.

701 McCarthy, M. (2007) *Bayesian Methods for Ecology*. Cambridge University
702 Press, Cambridge.

703 McInerny, G.J., Purves, D.W. & McIntyre, K.M. (2011) Fine-scale environ-
704 mental variation in species distribution modelling : regression dilution ,
705 latent variables and neighbourly advice. *Methods in Ecology and Evolu-*
706 *tion*, **2**, 248–257.

707 McNamara, J.M. & Harding, K.C. (2004) Measurement error and estimates
708 of population extinction risk. *Ecology Letters*, **7**, 16–20.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, pp. 811–823.

Murray, J.V., Goldizen, A.W., OLeary, R.A., McAlpine, C.A., Possingham, H.P. & Choy, S.L. (2009) 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-wallabies Petrogale penicillata. *Journal of Applied Ecology*, **46**, 842–851.

Newbold, T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, **34**, 3–22.

Patil, A. (2007) *Bayesian Nonparametrics for Inference of Ecological Dynamics*. Ph.D. thesis, University of California Santa Cruz.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Townsend Peterson, A. (2006) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.

Phillips, S.J., Anderson, R.P. & Schapire, Robert, E. (2006) Maximum en-

tropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J.R. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–97.

Phillips, S.J. & Elith, J. (2013) On Estimating Probability of Presence from Use-Availability or Presence-Background Data. *Ecology*.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2012) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-106.

Preston, C., Pearman, D. & Dines, T. (2002) *New atlas of the British and Irish flora: an atlas of the vascular plants of Britain, Ireland, the Isle of Man and the Channel Islands*. Oxford University Press, Oxford.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasmussen, C. & Willians, C. (2006) *Gaussian processes for machine learning*. MIT Press.

745 Ridgeway, G. (2013) *gbm: Generalized Boosted Regression Models*. R pack-
746 age version 2.1.

747 Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C.
748 & Mller, M. (2011) proc: an open-source package for r and s+ to analyze
749 and compare roc curves. *BMC Bioinformatics*, **12**, 77.

750 Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference
751 for latent Gaussian models by using integrated nested Laplace approx-
752 imations. *Journal of the Royal Statistical Society: Series B (Statistical*
753 *Methodology)*, **71**, 319–392.

754 Schmidberger, M., Tierney, L. & Mansmann, U. (2009) State of the Art in
755 Parallel Computing with R. *Journal of Statistical Software*, **31**.

756 Sigourney, D.B., Munch, S.B. & Letcher, B.H. (2012) Combining a Bayesian
757 nonparametric method with a hierarchical framework to estimate individ-
758 ual and temporal variation in growth. *Ecological Modelling*, **247**, 125–134.

759 Sinclair, S.J.S., White, M.M.D. & Newell, G.R. (2010) How useful are species
760 distribution models for managing biodiversity under future climates. *Ecol-*
761 *ogy and Society*, **15**.

762 Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., Hem-

763 ingway, J., Patil, A.P., Temperley, W.H., Gething, P.W., Kabaria, C.W.,
764 Okara, R.M., Van Boeckel, T.P., Godfray, H.C.J., Harbach, R.E. & Hay,
765 S.I. (2010) The dominant Anopheles vectors of human malaria in Africa,
766 Europe and the Middle East: occurrence data, distribution maps and
767 bionomic precis. *Parasites & Vectors*, **3**, 117.

768 Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. (2002)
769 Bayesian measures of model complexity and fit. *Journal of the Royal*
770 *Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

771 Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Jour-*
772 *nal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**,
773 267–288.

774 Vanhatalo, J., Pietiläinen, V. & Vehtari, A. (2010) Approximate infer-
775 ence for disease mapping with sparse Gaussian processes. *Statistics in*
776 *Medicine*, **29**, 1580–607.

777 Vanhatalo, J., Veneranta, L. & Hudd, R. (2012) Species distribution mod-
778 eling with Gaussian processes: A case study with the youngest stages
779 of sea spawning whitefish (Coregonus lavaretus L. s.l.) larvae. *Ecological*
780 *Modelling*, **228**, 49–58.

781 Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-
782 only data and the em algorithm. *Biometrics*, **65**, 554–63.

783 Wenger, S.J. & Olden, J.D. (2012) Assessing transferability of ecological
784 models: an underappreciated aspect of statistical validation. *Methods in*
785 *Ecology and Evolution*, **3**, 260–267.

786 Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard,
787 C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A.,
788 Heikkinen, R.K., Hø ye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nils-
789 son, M.C., Normand, S., Ockinger, E., Schmidt, N.M., Termansen, M.,
790 Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.C. (2013) The
791 role of biotic interactions in shaping distributions and realised assemblages
792 of species: implications for species distribution modelling. *Biological Re-*
793 *views of the Cambridge Philosophical Society*, **88**, 15–30.

794 Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal
795 likelihood estimation of semiparametric generalized linear models. *Journal*
796 *of the Royal Statistical Society (B)*, **73**, 3–36.

797 Yackulic, C.B., Chandler, R.B., Zipkin, E.F., Royle, J.A., Nichols, J.D.,
798 Campbell Grant, E.H. & Veran, S. (2012) Presence-only modelling using

MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Figures

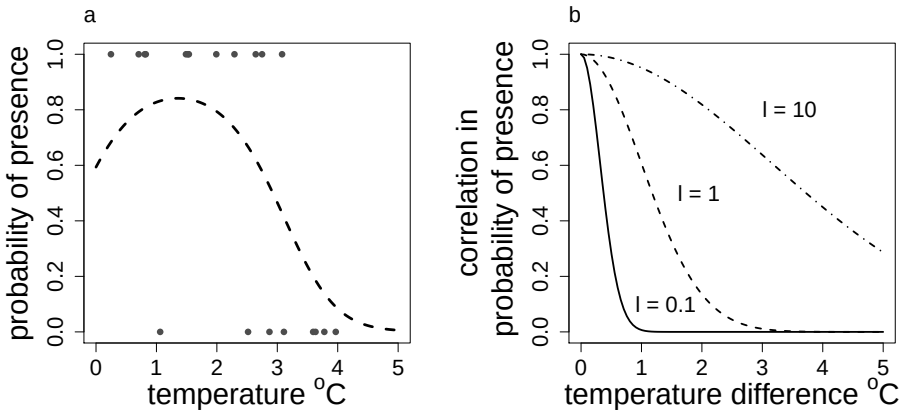

Fig. 1: Illustration of the covariance function using synthetic data: (a) observed presence-absence data (points) and the true underlying probability of presence as a function of temperature (dashed line); (b) correlation between probability of presence at different sites, calculated from temperature difference between these sites using the covariance function with three different lengthscale parameters (discussed in the text). Models fitted to the data using these three lengthscales are illustrated in Fig. 2.

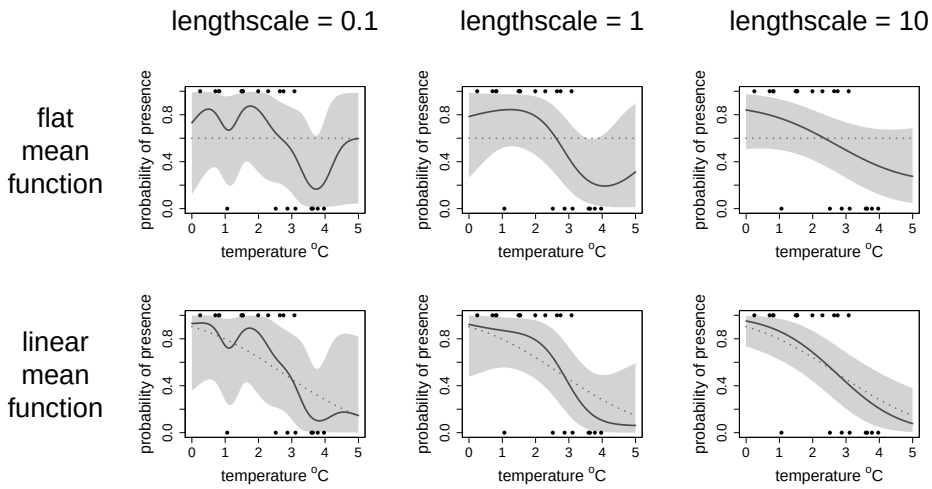

Fig. 2: Effects of the mean function and lengthscale on the fitted GP model. Shown are the observed data (points), the value of the mean function (dotted line), the probability of presence predicted by the GP model (solid line) and associated 95% credible intervals for this prediction (shaded grey area). Models are fitted with either the default flat mean function at the mean probability of presence (upper row) or a mean function representing some prior knowledge about how probability of presence relates to temperature, as described in the text (lower row).

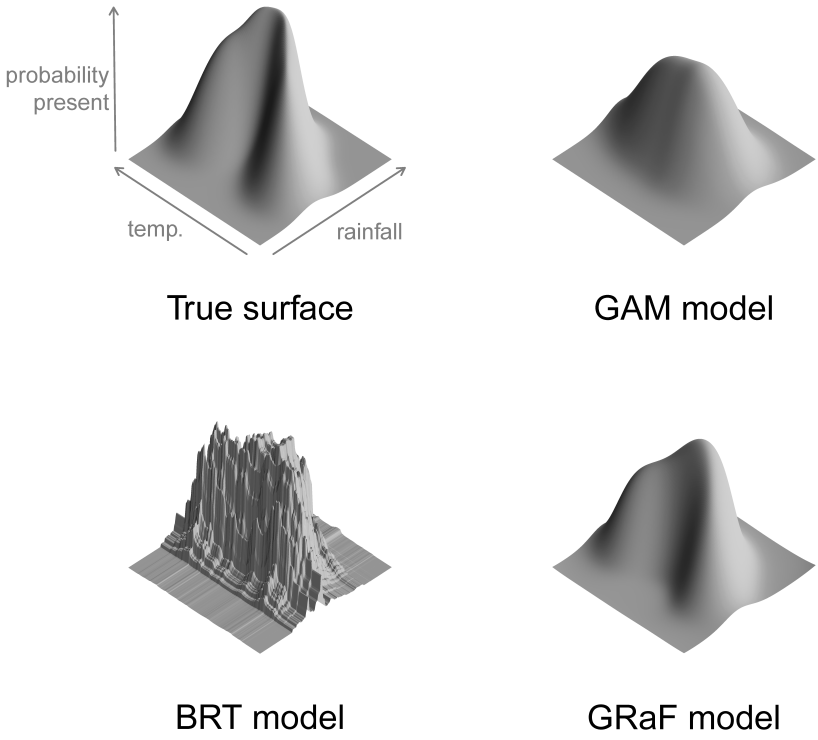

Fig. 3: Predictive surfaces fitted by boosted regression trees (BRT), a generalized additive model with univariate smoothers (GAM) and a GP model to simulated data with a strong non-linear interaction. The true surface represents the probability of presence of a hypothetical species in response to temperature and rainfall. Models were fitted to 1000 random presence/absence observations drawn from the true probability surface (a mixture of Gaussians).

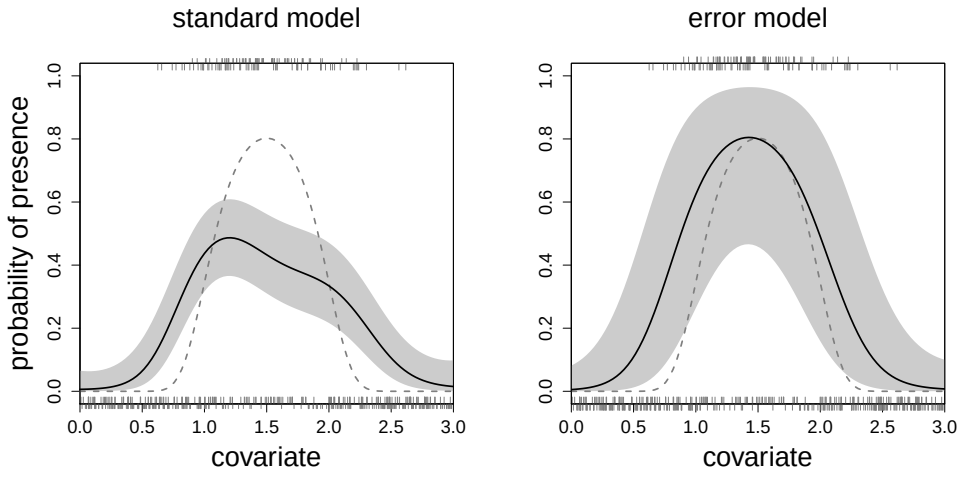

Fig. 4: Comparison of predictive surfaces fitted by GP models to simulated data with the covariate measured under error either ignoring measurement error (standard model) or accounting for it (error model). Solid lines give predictions and shaded regions represent 95% credible intervals. Three hundred presence-absence points were generated from the true model (dashed line) given the correct value of the covariate (tick marks outside box, presence on top line and absence on bottom line). Normally distributed random noise was then added to simulate covariates measured under error (tick marks insode box). Both models were then fitted to these data with measurement error. The error model was provided with the correct standard deviation of the error (0.5) whilst the standard model ignored the error.

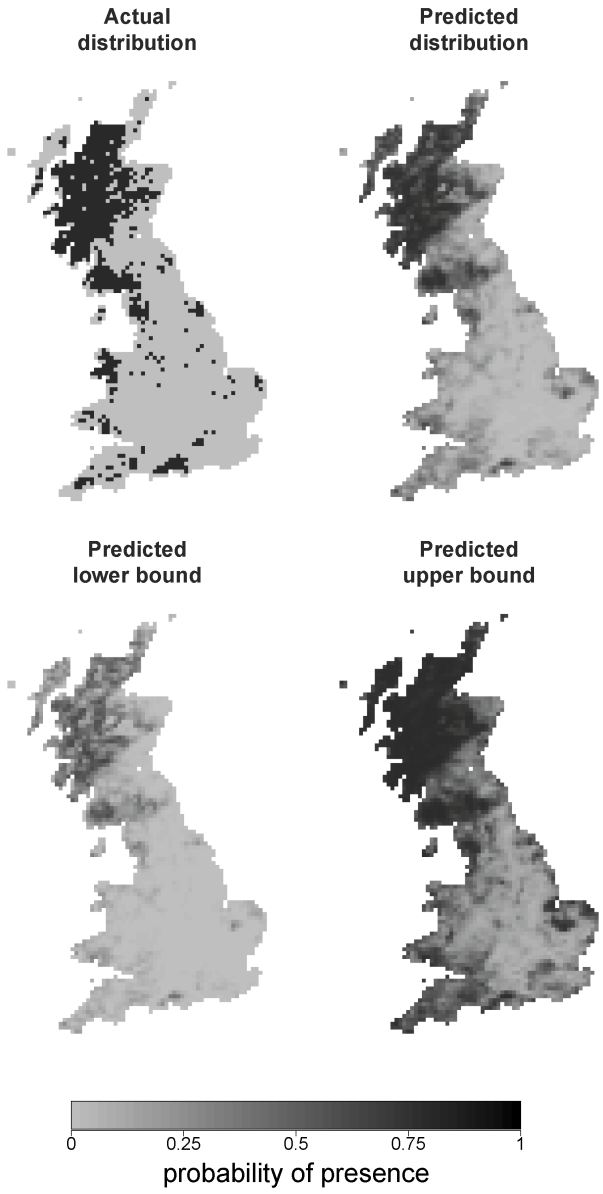

Fig. 5: True and predicted distributions of Bog myrtle (*Myrica gale*; prevalence 0.27) in Great Britain. Shown are the true distribution, the predicted distribution from a GP model (the Maximum *a posteriori* prediction), and lower and upper bounds on this prediction, representing our uncertainty in it (95% credible intervals, automatically generated by the GP). The GP model was fitted to 300 presence-absence data points - around 10% of the dataset.

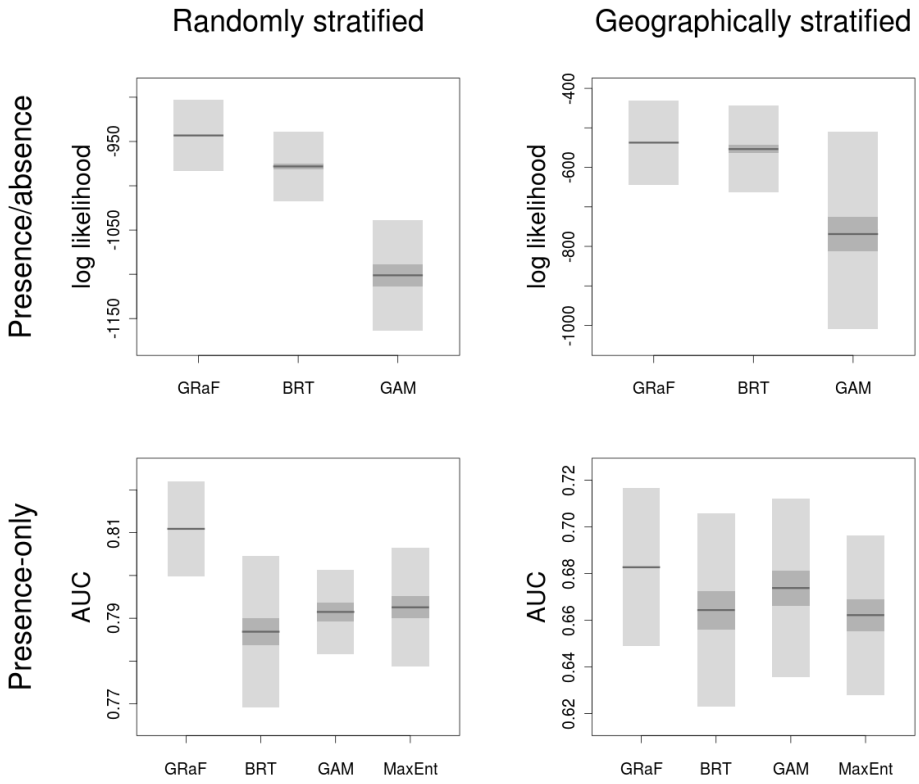

Fig. 6: Marginal validation statistics for model predictions to withheld training sets for presence/absence and presence-only data with two types of cross-validation. Centre lines give the means of the marginal validation statistic and light grey boxes give ±1 standard deviation of the marginal statistics, as an indication of the likely differences in performance of each model on an 'average' species in the plant dataset. Dark grey boxes give ±1 standard deviation of the estimated difference in the mean of the statistic for each model from the statistic for GRaF, as a visual representation of the statistical tests carried out. As GRaF was used as the contrast for these tests, no dark grey box is presented for this model. Higher log-likelihoods and higher AUCs indicate more accurate predictions to the evaluation set.

Supporting Information

Appendix S1. Statistical explanation and specification

Appendix S2. Assessment of approximation error

Appendix S3. Explanation of data for SDM comparison

Appendix S4. Demonstration of GRaF R package

Appendix S5. R code and data to compare SDMs and reproduce figures