Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 598–608 doi: 10.1111/2041-210X.12523

Fast and flexible Bayesian species distribution modelling using Gaussian processes

Nick Golding[1,2]* and Bethan V. Purse[1]

[1]*Centre for Ecology & Hydrology, Crowmarsh Gifford, Wallingford, UK OX10 8BB; and* [2]*Spatial Ecology and Epidemiology Group, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK*

Summary

1. Species distribution modelling (SDM) is widely used in ecology, and predictions of species distributions inform both policy and ecological debates. Therefore, methods with high predictive accuracy and those that enable biological interpretation are preferable. Gaussian processes (GPs) are a highly flexible approach to statistical modelling and have recently been proposed for SDM. GP models fit smooth, but potentially complex response functions that can account for high-dimensional interactions between predictors. We propose fitting GP SDMs using deterministic numerical approximations, rather than Markov chain Monte Carlo methods in order to make GPs more computationally efficient and easy to use.

2. We introduce GP models and their application to SDM, illustrate how ecological knowledge can be incorporated into GP SDMs via Bayesian priors and formulate a simple GP SDM that can be fitted efficiently. This model can be fitted either by learning the hyperparameters or by using a fixed approximation to them. Using a subset of the North American Breeding Bird Survey data set, we compare the out-of-sample predictive accuracy of these models with several commonly used SDM approaches for both presence/absence and presence-only data.

3. Predictive accuracy of GP SDMs fitted by Laplace approximation was greater than boosted regression trees, generalized additive models (GAMs) and logistic regression when trained on presence/absence data and greater than all of these models plus MaxEnt when trained on presence-only data. GP SDMs fitted using a fixed approximation to hyperparameters were no less accurate than those with MAP estimation and on average 70 times faster, equivalent in speed to GAMs.

4. As well as having strong predictive power for this data set, GP SDMs offer a convenient method for incorporating prior knowledge of the species' ecology. By fitting these methods using efficient numerical approximations, they may easily be applied to large data sets and automatically for many species. An R package, GRaF, is provided to enable SDM users to fit GP models.

Key-words: boosted regression trees, Gaussian processes, generalized additive models, MaxEnt, species distribution models

Introduction

Species distribution models (SDMs) quantify the distribution of species using environmental conditions as predictors. Typically, these models use gridded data sets of environmental variables and records of the distribution of a given species 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, Leathwick & Overton 2002; Sinclair, White & Newell 2010; Sinka *et al.* 2010).

A wide range of different statistical models have been suggested for use in 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 & Leathwick 2009). These approaches have a number of features which determine their suitability to model species distributions. These include an ability to represent complex effects of different predictors (such as high-dimensional and nonlinear interactions between predictors) on a species' distribution (Elith *et al.* 2006); susceptibility to overfitting to training data (Wenger & Olden 2012); and the capacity to incorporate existing knowledge of the species' ecology (Murray *et al.* 2009). SDMs are sometimes required to be fitted in large batches for multiple species, for example in order to make predictions of species richness (Ferrier & Guisan 2006) or to understand relative vulnerability of species to environmental change (Huntley *et al.* 2008). Procedures that are computationally efficient are therefore preferable. Additionally, the ability to robustly quantify uncertainty in predictions from an SDM is desirable (Guisan & Zimmermann 2000; Elith, Burgman & Regan 2002).

*Correspondence author. E-mail: nick.golding.research@gmail.com

Gaussian processes (GPs, also referred to as Gaussian random fields) provide a flexible approach to fitting complex statistical models (Rasmussen & Williams 2006) and offer solutions to many of the issues related to SDMs. GPs have seen occasional use in ecology for modelling population dynamics (Patil 2007; Sigourney, Munch & Letcher 2012) and have recently been proposed as an alternative approach for SDM (Vanhatalo, Veneranta & Hudd 2012). Whilst GP models have so far been applied to presence/absence data for SDM, GP models can also be fitted with likelihoods applicable for count and presence-only data (Diggle *et al.* 2013).

Gaussian process models are often fitted via Bayesian inference, typically requiring the use of Markov chain Monte Carlo (MCMC) methods. Whilst MCMC is a useful approach for fitting complex models, it can be very computationally expensive when applied to GP models (Rue, Martino & Chopin 2009) and requires an experienced user to supervise the model fitting process. These limitations make GP SDM models fitted using MCMC infeasible for the many SDM users without experience using MCMC and for applications that require running large batches of models. Computationally efficient deterministic inference procedures such as Laplace approximation and expectation propagation have been developed to overcome these problems for GP models (Rasmussen & Williams 2006). Whereas approximation error in MCMC schemes can be made arbitrarily small by increasing the number of iterations, deterministic approaches are subject to a fixed approximation error that may impinge on predictive accuracy. In addition, many of the proposed methods estimate a fixed set of model hyperparameters [i.e. maximum-likelihood or maximum *a posteriori* (MAP) inference], rather than integrating across full posterior distributions for these hyperparameters as in a full Bayesian analysis. We propose that in spite of these simplifying assumptions, GP models fitted by deterministic approximate inference are a promising method for SDM analyses.

Below, we illustrate how GP SDM models work in both statistical and ecological terms and demonstrate how they provide solutions to some problems commonly encountered in distribution modelling. We then compare their predictive ability with other commonly used approaches in a case study using bird occurrence data from the North American Breeding Bird Survey (BBS). Finally, the advantages and limitations of GPs and potential avenues for future enhancements of the approach as applied to SDMs are discussed.

Statistical explanation of GP models

As with most statistical models, SDMs use data to learn an underlying function which takes as input a set of values of one or more predictors and outputs a single response variable. In the case of SDM, this response variable is typically the probability of presence of the species. Once fitted, this function can be used to convert any set of values of these environmental variables into the expected value of the response variable, enabling prediction.

More formally, for a presence/absence model with binary response vector y indicating species presence or absence at n locations and corresponding n-by-m design matrix $\mathbf{x}$ giving the values of the m predictor variables across these locations ($\mathbf{x}_i$ denoting the ith row of this matrix), we might model the probability of presence p_i for a given location to be some transformation $g()$ (the link function) of an underlying latent variable z_i. We then model z_i to be the output of a function f, evaluated on a corresponding row of the design matrix:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ p_i &= g(z_i) \\ z_i &= f(\mathbf{x}_i) \end{aligned} \qquad \text{eqn 1}$$

The choice of model used to estimate f therefore defines the shape of the resulting function. Figure 1 illustrates using simulated data the modelled response surfaces fitted by a GP model, a generalized additive model (GAM) with univariate smoothers and a boosted regression tree (BRT) model. Whilst the BRT captures the nonlinear (banana-shaped) interaction between the predictors, the estimated surface is highly jagged by comparison with the other models. By contrast, the GAM fits a smooth surface but due its additive structure is unable to capture the nonlinear interaction between the predictors. The GP model fits a smooth surface whilst correctly capturing the interaction.

Many SDMs attempt to specify this function between predictors and the response variable by fitting a parametric equation. Once the 'best' set of equation parameters has been learned from the data, the function can then be completed with new predictor values to make predictions. Instead of modelling this function by learning parameters of a fixed equation, GP models instead consider the unknown function itself to be a single realization of an underlying stochastic process, which is modelled as a GP.

DEFINING A GP

Under a GP model for this latent function $f()$, each latent variable z_i is considered to follow a Gaussian distribution with mean μ_i and variance σ_i^2. These response variables are not independent, however, but are subject to some pairwise correlation $\mathbf{C}_{ij}$ between any two responses z_i and z_j. With more than two dependent variables, this structure is more concisely expressed as a multivariate Gaussian distribution with vector response variable z, vector mean μ and symmetric, square covariance matrix Σ. Σ can be constructed from the vector of marginal variance parameters σ^2 and symmetric positive definite correlation matrix $\mathbf{C}$ of dimension n, following the identity $\Sigma = \text{diag}(\sigma)\mathbf{C}\text{diag}(\sigma)$'.

Since this model will likely have far more correlation parameters than observations ($\mathbf{C}$ having $n(n-1)/2$ unique elements), learning each of these correlation parameters independently is not feasible. Instead, the covariances Σ_{ij} between pairs of response variables are themselvesmodelled as a parametric function – the covariance function $k()$ – of corresponding predictor values $\mathbf{x}_i$ and $\mathbf{x}_j$, given some parameters θ_k. Similarly, the mean vector μ can also be defined by

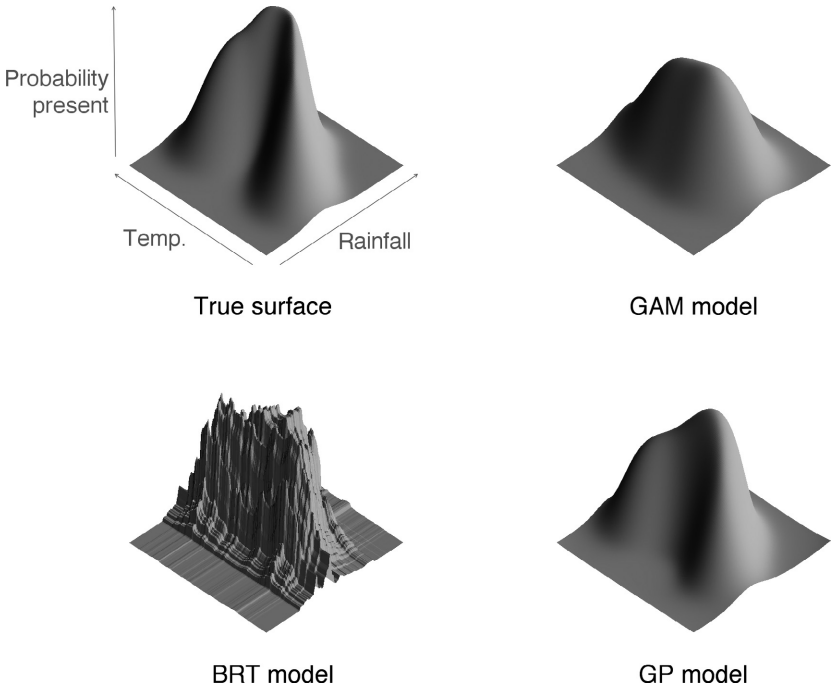

Fig. 1. Predictive surfaces fitted by boosted regression trees (BRT), a generalized additive model with univariate smoothers (GAM) and a GP model to simulated data with a strong nonlinear interaction. The true surface represents the probability of presence of a hypothetical species in response to temperature and rainfall. Models were fitted to 1000 random presence/absence observations drawn from the true probability surface (a mixture of Gaussians).

a parametric function – the mean function $m()$ – over $\mathbf{x}_i$ given parameters θ_m. When evaluated on a data set of fixed size, the GP is therefore defined as the following multivariate normal distribution:

$$\begin{aligned} z &\sim N(\mu, \Sigma) \\ \mu &= m(\mathbf{x}; \theta_m) \\ \Sigma &= k(\mathbf{x}, \mathbf{x}'; \theta_k) \end{aligned} \qquad \text{eqn 2}$$

However, since the elements of this distribution are controlled by functions defined over a continuous space (the support of $\mathbf{x}$), each draw from the GP can be considered a function of $\mathbf{x}$. Hence, GPs are often referred to as infinite-dimensional Gaussian distributions, or probability distributions over functions. The GP structure can therefore be written more concisely for the continuous case [following the notation of Rasmussen & Williams (2006)] as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \qquad \text{eqn 3}$$

A SIMPLE GP SDM

Throughout the remainder of this paper, we consider the following relatively simple GP model applicable to presence/absence SDM:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ p_i &= \Phi(z_i) \\ z &= f(\mathbf{x}) \\ f(\mathbf{x}) &\sim \mathcal{GP}(m(\mathbf{x}), K_{\text{se}}(\mathbf{x}, \mathbf{x}')), \end{aligned} \qquad \text{eqn 4}$$

where $\Phi(.)$ denotes the probit link function (the logit link could also be used), $m()$ is a mean function and $k_{\text{se}}()$ is the squared-exponential covariance function with vector parameter l of length m giving the characteristic lengthscales for each predictor. The parameters in l are considered hyperparameters of the model. We define $k_{\text{se}}()$ as:

$$\begin{aligned} k_{\text{se}}(\mathbf{x}, \mathbf{x}) &= \exp\left(-\frac{\mathbf{r}^2}{2}\right) \\ \mathbf{r} &= \sqrt{\sum_{j=1}^{m}\left(\frac{\mathbf{x}_j - \mathbf{x}'_j}{l_j^2}\right)^2} \end{aligned} \qquad \text{eqn 5}$$

Note that the squared-exponential kernel commonly also has a variance parameter; however, the marginal variances of z are not identifiable from Bernoulli-distributed data and the variance is therefore fixed at one. There is a wide array of different covariance functions that we could use here, but we opt for the squared-exponential covariance function since it is easy to parameterize and produces ecologically plausible smooth curves (Rasmussen & Williams 2006).

The simple model above could easily be extended to different SDM use cases by modifying the likelihood term. For example, a Poisson point process likelihood for presence-only data could be used instead (Warton & Shepherd 2010), corresponding to a log-Gaussian Cox process spatial model (Diggle *et al.* 2013) but with the function modelled by the GP acting over environmental rather than geographic space.

INFERENCE ON GPS

The hierarchical and correlated nature of the GP model means that inference is somewhat more involved than for other statistical models. The inference procedure can be considered in two distinct parts: inference over the latent variables z given a set of kernel hyperparameters and inference over the hyperparameters themselves.

Inference over latent variables

Given a fixed set of hyperparameters l, the GP can be considered a prior distribution over the latent parameters z. Learning z, given these hyperparameters and the observed data y (responses) and $\mathbf{x}$ (predictors), is therefore carried out by applying Bayes theorem. The parameters of the posterior distribution over z can then be estimated for any $\mathbf{x}$ to yield the posterior distribution over the corresponding elements of z.

If y follows a Gaussian distribution, it can be considered a direct observation of $f(\mathbf{x})$ and an analytic solution to the GP posterior can be computed [details are given in Rasmussen & Williams (2006)]. Where y follows some non-Gaussian distribution, as in the GP SDM case, no analytical solution is available and inference must be carried out by some approximation scheme. Several efficient procedures have been developed for inference in this case, including Laplace approximation (used throughout this paper), expectation propagation and variational inference (Rasmussen & Williams 2006; Hensman, Fusi & Lawrence 2013).

If the GP SDM is fitted with fixed hyperparameters determined by the user, this amounts to a fully Bayesian model with the lengthscales defining the model prior. These lengthscales may be used to incorporate ecological knowledge into the SDM, and their behaviour is illustrated below. Rather than specifying them *a priori*, it may instead be preferable to learn these hyperparameters from the data.

Inference over hyperparameters

If the lengthscales are learned from the data, the GP then becomes a layer in a hierarchical model, rather than a prior. Such a hierarchical model may fitted in either a maximum-likelihood, a fully Bayesian or a partially Bayesian (MAP) framework.

Whilst the likelihood for the latent variables $p(y|z)$ is trivial to compute, calculation of the likelihood for the lengthscale hyperparameters $p(y|l)$ requires marginalization of the latent variables z:

$$p(y|l) = \int p(y|z)p(z|l)\mathrm{d}z \qquad \text{eqn 6}$$

As with inference for the latent variables, the marginal likelihood is analytically tractable for the Gaussian likelihood, but must be approximated in the GP SDM case, using the numerical approximations to the posterior density of z. This marginalization is analogous to restricted maximum-likelihood estimation in generalized mixed-effects models and has a similar penalization effect to other approaches such as lasso or ridge regression.

Maximum-likelihood inference can therefore be carried out by numerical optimization of the marginal likelihood to identify an optimal set of hyperparameters. Alternatively, Bayesian inference over this hierarchical model may then be carried out with the specification of further prior distributions over the hyperparameters. A fully Bayesian treatment of this model would necessitate consideration of the probability distribution over these hyperparameters and integrating out this distribution when making predictions. This was the approach taken by Vanhatalo, Veneranta & Hudd (2012) in their paper applying a Bayesian GP model to SDM. Unfortunately, this requires computationally intensive inference procedures such as MCMC, which we aim to avoid. Integrated nested Laplace approximation (Rue, Martino & Chopin 2009) may provide an efficient procedure for Bayesian inference on GP models in certain circumstances, though available software (Lindgren & Rue 2015; Rue *et al.* 2015) does not enable users to fit the high-dimensional GPs required for a GP SDM analysis, only allowing for two-dimensional GPs.

Alternatively, MAP inference (maximizing the marginal posterior density of the model) enables the incorporation of prior knowledge over hyperparameters whilst being much less computationally intensive than fully Bayesian inference. In the empirical case study below, we compare GP SDMs with fixed lengthscales and with lengthscales selected by MAP inference.

Ecological explanation of GP models

Fitting and interpreting any SDM require an understanding of how the different model components relate to the ecology of a given species. Next, we provide a more intuitive illustration of how inference for GP models differs from other SDMs, using as an example the effect of temperature on the probability of presence of a hypothetical species.

COVARIANCE FUNCTION

The range of function shapes allowed by the GP depends on the covariance function, which relates environmental values to expected correlations in the response variables at different locations. As with most widely used covariance functions, the squared-exponential covariance function calculates these covariances as a function of the multivariate Euclidean distance between predictor values at these locations. In order to construct the GP, the first step is therefore to calculate the environmental distances between observations. In our example, the environmental distances are simply the difference in temperature between each pair of sites (Fig. 2a). The lengthscales of the kernel function then dictate how the correlation between probabilities of occurrence at pairs of observations decays with the environmental distance between them and therefore the complexity of the fitted response curves. Figure 2b illustrates how a squared-exponential covariance 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. 3).

In a practical application of a GP model, these lengthscales may either be specified in advance or may be estimated from the data. In either case, the lengthscale parameters provide a

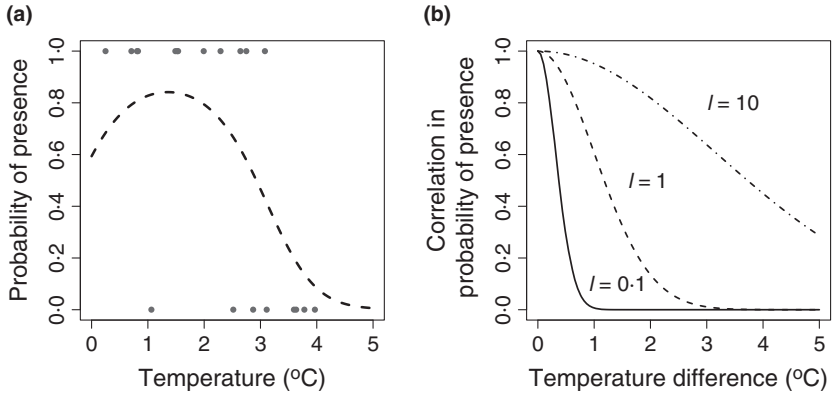

Fig. 2. Illustration of the covariance function using synthetic data: (a) observed presence/absence data (points) and the true underlying probability of presence as a function of temperature (dashed line); (b) correlation between probability of presence at different sites, calculated from temperature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

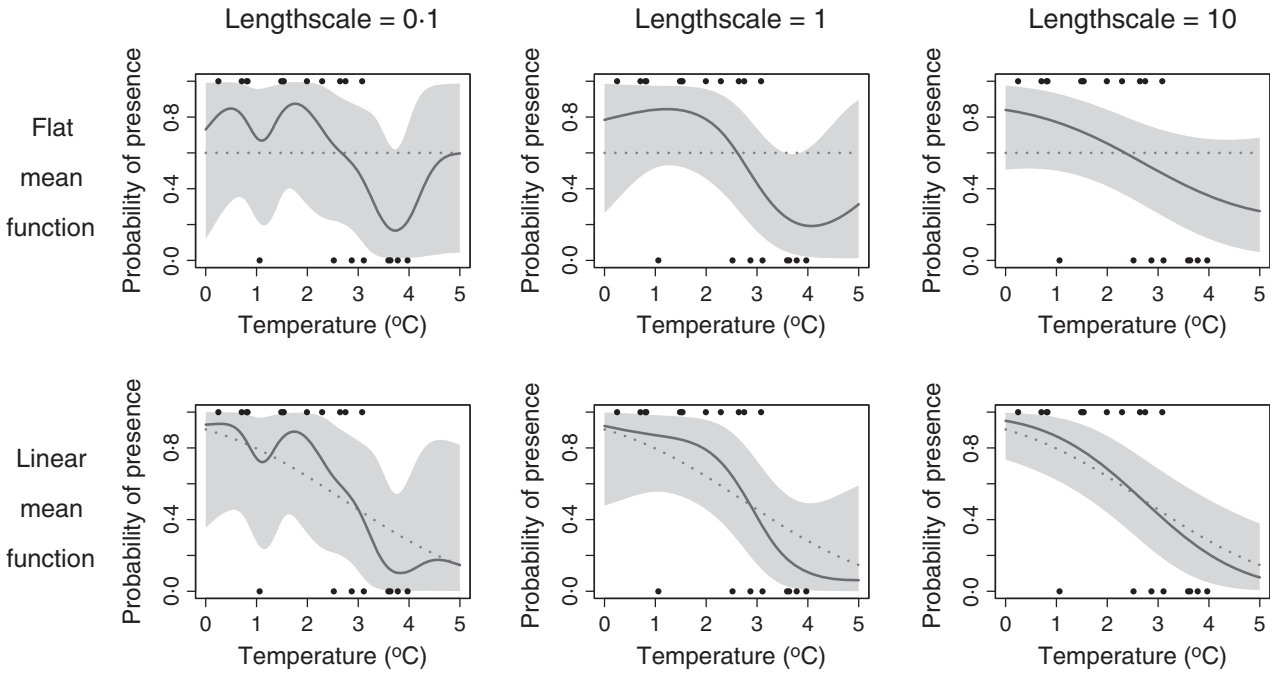

Fig. 3. Effects of the mean function and lengthscale on the fitted GP model. Shown are the observed data (points), the value of the mean function (dotted line), the probability of presence predicted by the GP model (solid line) and associated 95% credible intervals for this prediction (shaded grey area). 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).

mechanism by which to inform the model of the ecology of the species being modelled. If the distribution of the species is known *a priori* to depend strongly on a particular predictor, then specifying a short lengthscale will incorporate this knowledge in the model, without a need to specify a particular functional form for the relationship.

MEAN FUNCTION

In addition to these expected correlations, we provide the GP model with a *mean function*: an initial 'best guess' at the true function defining how the species' probability of presence depends on the predictors. If nothing is known about how the probability of presence of a species responds to an environmental gradient, such as the temperature gradient in our example, it is possible to instead use a flat mean function, which assumes an equal probability of presence regardless of the values of predictors. 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 relating temperature to probability of presence.

Figure 3 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 towards the limits of the recorded temperature range, the mean function determines the shape of the fitted response.

Advantages of GP SDMs

MODEL STRUCTURE

Machine learning algorithms such as BRTs (Elith, Leathwick & Hastie 2008) have been shown to perform particularly well at predicting species distributions, likely due to their ability to fit complex responses to multiple environmental predictors (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 (Wenger & Olden 2012). By comparison, more traditional approaches such as univariate GAMs (Hastie & Tibshirani 1986) fit more biologically realistic smooth functions. Whilst some implementations of GAM can fit multivariate smoothers that may represent such high-dimensional interactions, the high number of degrees of freedom in these models means they can only be fitted to very large data sets (Wood 2011). GPs offer an attractive solution to this trade-off between model flexibility and ecological realism by allowing for complex interactions between predictors whilst fitting biologically plausible smooth predictive surfaces (see Fig. 1).

INCORPORATING PRIOR ECOLOGICAL KNOWLEDGE

In many SDM analyses, the modeller has some prior knowledge of the species' ecology (such as a preference for some environmental conditions) that it may be advantageous to incorporate into the model. Bayesian statistical inference provides a convenient way of incorporating external information of this sort into statistical ecological models via prior distributions (McCarthy 2007).

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 predictors. In the absence of any prior information, non-informative priors may be used.

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, allowing users to determine how much confidence they can place in a given prediction (Elith, Burgman & Regan 2002). For some SDM methods, including generalized linear model (GLM) and GAM, estimates of prediction uncertainty can be obtained analytically and therefore with minimal computational cost. For some of the more modern (and best performing) methods, such as BRT and MaxEnt, analytic uncertainty estimates are not available. For these methods, uncertainty estimates can be produced by bootstrapping data (Elith, Burgman & Regan 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 represented by the estimated posterior distribution of the model. The second and third panels of Fig. 4 illustrate mapped predictions and associated uncertainty estimates from a GP distribution model of a bird species, the Mourning warbler (*Geothlypis philadelphia*), in North America. The predicted distribution map indicates the 'best guess' prediction of probability of presence – the posterior mode of the predictive

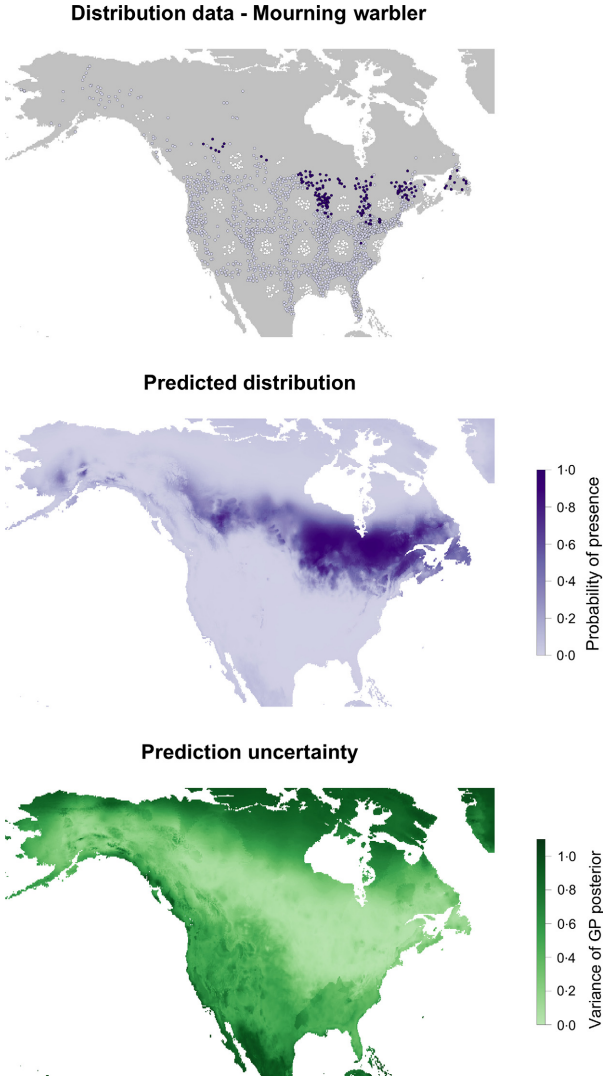

Fig. 4. Distribution data and predicted distribution of the Mourning warbler *Geothlypis philadelphia* in North America. From top: distribution data from the Breeding Bird Survey 2011, training locations in colour (dark purple – presence, light purple – absence) and evaluation locations in white; predicted probability of presence (posterior mode) from a GP-fixed model fitted to the Bioclim predictor data set, as in the model comparison; uncertainty in this prediction represented as the marginal variance of the posterior distribution of the underlying function inferred by the GP model [i.e. the variance of z in eqn (4)].

distribution for each pixel. The prediction uncertainty map quantifies model uncertainty in these predictions measured as the variance of the estimated function (z in eqn 4) for the set of predictors given in each pixel.

Comparison of GPs with existing SDMs

We compared the predictive ability of GPs fitted using Laplace approximation, an efficient approximate inference method, with commonly used approaches for modelling species distributions from both presence/absence and presence-only data. All model fitting, predictions and statistical analyses were performed in R version 3.0.0 (R Core Team, 2014). R code used to carry out these analyses and to plot all figures in this manuscript is provided as an open-access code repository at https://github.com/goldingn/gp_sdm_paper.

METHODS

Gaussian process SDM

We fitted the GP SDM given in eqn (4) using two different approaches for determining the kernel lengthscales: estimating the MAP values, and using a fixed values under based on the values of predictors in the presence and absence training data. Whilst the MAP estimate would be expected to provide a better prediction, it requires repeated model fitting to determine the optimal parameters and is therefore considerably more computationally expensive. For the GP models with MAP estimation of lengthscale hyperparameters (hereafter referred to as GP-MAP), we used independent log-Gaussian priors on each lengthscale with mean of log(10) and a standard deviation of 1. As we scale predictors to have a standard deviation of one during model fitting, this prior is insensitive to varying scales of measurement. This prior restricts the optimization to ecologically plausible lengthscales by penalizing highly complex GP surfaces and also forces the posterior density to be convex (the likelihood surface alone is not guaranteed to be convex). This prior and the range of lengthscales it suggests are illustrated in the supplementary material. We then estimate the MAP values of the lengthscales by performing numerical optimization using the limited-memory variant of the BFGS algorithm (Byrd *et al.* 1995), implemented using R's optim function) and an analytic solution to the gradient of the marginal posterior density. The MAP estimates for the lengthscales are then used to fit the final GP model.

For GP models fitted using a fixed approximation to the lengthscale hyperparameters (hereafter GP-fixed), we calculate the lengthscale for each predictor as the ratio of the standard deviation of the predictor at presence records to the standard deviation at all records, multiplied by eight. This simple metric gives an indication of the difference in predictor values between presence and absence locations and therefore the likely utility of the predictor. The decision of a scaling factor of eight was arbitrarily chosen to reflect broadly realistic lengthscales (those leading to few inflection points in random functions drawn from the GP), prior to the model evaluation experiments performed here. This metric could undoubtedly be improved, but that is beyond the scope of this paper.

Data

Training and evaluation data sets of the distributions of North American bird species were obtained from a subset of the North American BBS in 2011 (Sauer *et al.* 2014) and a set of eight minimally correlated Bioclim climatic variables (Hijmans *et al.* 2005) as predictors. This data set was an exact replication of that used in the Joint SDM comparison of Harris (2015). Full details of this model evaluation data set are given in that paper, and the R code used to construct it is provided by Harris in a code repository at https://github.com/davharris/mistnet. These occurrence data comprised the presence or absence of 370 bird species at each of 2768 survey routes, divided into 2467 training and 301 evaluation locations using a disc-based spatial stratification procedure. This procedure was applied to remove spatial autocorrelation between training and evaluation data, which may inflate the validation statistics of SDMs if not accounted for Wenger & Olden (2012). The locations of these training and evaluation locations are shown in the first panel of Fig. 4, with the training-set presences and absences of *G. philadelphia* also indicated. The eight Bioclim variables selected by Harris (2015) are detailed in Appendix A of the supporting information for that paper.

Presence/absence models

For the presence/absence comparison, we compared GP-MAP and GP-fixed with BRT, GAM and GLMs. Each of these models was fitted for each bird species using occurrence data at the training routes and all eight environmental predictors. Each model was then used to predict the species' probability of presence at each of the evaluation routes. The predictive accuracy of each model was quantified as the log-likelihood of the withheld data from the predictions of each model, a measure that assesses the calibration accuracy of the predicted probability of presence (Lawson *et al.* 2014).

Presence-only models

In order to assess predictive capacity in presence-only analyses, we simulated presence-only data by subsetting the data set used in the presence/absence comparison. We simulated opportunistic presence records for each species by randomly selecting 146 of the training routes reporting presence of the given species. This number corresponds to the median number of occurrence records in 108 published presence-only SDM analyses reviewed by Yackulic *et al.* (2012) and is therefore intended to be representative of standard presence-only SDM practice. We omitted all species with fewer than 146 presence records in the training set leaving a total of, coincidentally, 146 species for use in the presence-only comparison. The remaining 2321 training routes were used as background records, regardless of whether the species had been in fact been reported there. As survey effort is broadly similar across all survey routes, this

presence-only data set does not suffer from the observation bias that typically affects presence-only SDM analyses, and corresponds to the target-group approach to selecting background points suggested by Phillips *et al.* (2009).

Gaussian process-MAP and GP-fixed models were again compared with BRT model, GAM and GLM as well as with MaxEnt (Phillips, Anderson & Schapire 2006). The GP model, BRT model, GAM and GLM were all fitted to these presence-only data by naively considering the background data as true absences. MaxEnt models were trained on the same data set, treating these background records as background records.

Models trained on these presence-only data were used to make predictions for the 301 evaluation routes, and these predictions were compared with the true presence/absence data available at these locations. Since the prevalence of each species in the study area is unidentifiable from presence-only data (Ward *et al.* 2009; Phillips & Elith 2013), predictions from both MaxEnt and presence/absence models applied to presence-only data are not estimates of the absolute probability of presence of the species, but only an uncalibrated or *relative* probability of presence (Elith *et al.* 2010). Since likelihood-based and other calibration-sensitive metrics are not appropriate for assessing relative probabilities such as these, we instead calculated the Area Under the Receiver Operating Statistic Curve statistic [AUC; calculated using the PROC R package – version 1.6.0.1; Robin *et al.* (2011)] for each model's predictions for each species. Whilst use of AUC in applied SDM studies has been criticized (Lobo, Jiménez-Valverde & Real 2008), it provides a reliable measure of model discrimination capacity when comparing a continuous estimate of relative probability of presence against true presence/absence data (Lawson *et al.* 2014).

Model fitting

Gaussian process-MAP and GP-fixed models were fitted using the function `graf` in the GRaF R package version 0.1-14 (Golding 2013) setting the argument `opt.l` to `TRUE` in the former case and `FALSE` in the latter. BRT models were fitted using the gbm R package version 2.1 (Ridgeway 2013) with fivefold cross-validation, a tree complexity of 5, a learning rate of 0·001 and a minimum of 1000 trees [in accordance with Elith, Leathwick & Hastie (2008)]. The optimal number of trees in the final BRT model was selected from the cross-validation folds using the gbm.perf function. GAMs were fitted using the mgcv package version 1.8-3 (Wood 2011) with univariate thin-plate regression spline smoothers for each predictor using default estimation of the dimension of the smooth terms and implementing covariate selection by penalization, using the argument `select = TRUE`. GLMs were fitted using the glm function in R with a binomial likelihood and logistic link function (i.e. logistic regression) with linear terms for each predictor. Covariate selection was carried out by backward stepwise selection to minimize the Akaike Information Criterion (Akaike 1973) of the final model. Whilst other methods of covariate selection (such as penalization) may be preferable to stepwise selection for GLMs, these are not as widely used as stepwise methods, and our aim here was to compare the performance of GP models with standard approaches. MaxEnt models were fitted using dismo version 0.8-11 (Hijmans *et al.* 2012). Each model was fitted using all eight predictors, and all other settings were held at their software defaults for each model.

Statistical analysis

Validation statistics for each species/model combination were analysed by linear 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 predictors were the SDM model type (modelled as a fixed effect) and 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 residuals of this model were assessed to ensure that the residuals were normally distributed with homogeneous variances. The statistical significance of differences between mean model validation statistics for each model was assessed by *t*-tests on coefficients for the SDM model type.

To allow users to assess overall goodness-of-fit of the presence/absence models, we also calculated the proportion of null deviance explained by these models. For each species in the presence/absence data set, the null model predicted the probability of presence at each test location to be equal to the species' prevalence in the training set.

In the presence/absence comparison, 23 (1·26%) models had markedly poor validation scores, with negative log-likelihood scores more than three times higher than the negative log-likelihood score of a prevalence-only null model. These models comprised of 22 GAMs and 1 GLM fitted to species for which training set prevalences were low (all 0·08 or lower). Given that each of these models predicted unreasonably low probabilities of presence in the test set (much lower than the species' training set prevalence), it seems likely that they would have been rejected by an SDM modeller in an applied analysis. The validation scores for these 23 models represented clear outliers in the evaluation data set and violated the normality assumption of the mixed-effects model. We therefore removed these 23 models and their validation scores from the presence/absence results data set prior to statistical analysis (but retained other models for these species). Exclusion of these models resulted in improved mean prediction metrics for GAMs and GLMs. However as the validation scores for these model types 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 visualize the expected predictive capacity from each SDM whilst removing species-level effects. The marginal statistics indicate the expected differences in model performance, whilst integrating out the effects

of species-specific ecology or of the species' prevalence (McPherson, Jetz & Rogers 2004).

Results

Averaging across all species, both GP models made more accurate predictions to the withheld, geographically stratified data than all other models for both presence/absence and presence-only comparisons (Fig. 5). Whilst GP-MAP models had higher validation statistics than GP-fixed models in both the presence/absence (log-likelihood 0·061 ± 0·176 SE higher, $t_{1453} = 0{\cdot}35$, $P < 0{\cdot}7278$) and presence-only (AUC 0·001 ± 0·002 SE higher, $t_{725} = 0{\cdot}64$, $P < 0{\cdot}5231$) experiments, these were not statistically significant, indicating that the fixed lengthscales provided a good approximation to the full model for this data set.

The expected log-likelihood for a GP-fixed model fitted to presence/absence data for an average species was estimated to be 3·8 (±0·348 SE, $t_{1453} = 10{\cdot}91$, $P < 0{\cdot}0001$) higher than for BRT, 11·404 (±0·996 SE, $t_{1453} = 11{\cdot}45$, $P < 0{\cdot}0001$) higher than for GAM and 10·704 (±0·741 SE, $t_{1453} = 14{\cdot}44$, $P < 0{\cdot}0001$) higher than for GLMs.

The expected AUC score for a GP-fixed model fitted to presence-only data for an average species was 0·885 (±0·007 SE). This was marginally (0·005 ± 0·002 SE, $t_{725} = 2{\cdot}85$, $P < 0{\cdot}0045$) higher than for BRT models; higher than for MaxEnt (0·01 ± 0·002 SE, $t_{725} = 5{\cdot}22$, $P < 0{\cdot}0001$) and GAMs (0·013 ± 0·003 SE, $t_{725} = 4{\cdot}36$, $P < 0{\cdot}0001$); and markedly higher than for GLMs (0·034 ± 0·004 SE, $t_{725} = 9{\cdot}36$, $P < 0{\cdot}0001$).

In the presence/absence comparison, GP-MAP models explained an average of 53·13% (±1·161%SE) and GP-fixed models 53·11% (±1·146% SE) of null deviance in probability of presence in the evaluation set, compared to 47·53 (±1·312% SE) for BRT, 44·58 (±1·278% SE) for GLM and 44·34 (±1·481% SE) for GAM.

Generalized linear models took the least time to run (<0·1 s on average per species for both presence-only and presence/absence models, inclusive of the stepwise selection procedure) followed by MaxEnt (0·5 s), GP-fixed (2·1 and 2·3 s), GAM (1·6 and 4·6 s) and BRT (15·8 and 18·6 s), with GP-MAP models taking the most time (149·6 and 168·4 s) – around 70 times longer than GP-fixed.

Discussion

In our comparison, GP SDMs 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, Leathwick & Hastie 2008). For this comparison, we used an existing data set with predetermined training and validation sets and fitted each model following best-practice guidelines where available and default settings otherwise. The large number of species considered and the use of an evaluation data set from a previous study by other researchers make this a fair, albeit preliminary assessment of the performance of GP SDMs vs. other modelling approaches. However, as each of these models can be tweaked by an expert user to improve their performance, our results are unlikely to be representative of the best possible implementation of any of the models. A more robust future assessment of these models could be obtained by their implementation by different modellers on a range of different data sets, as in Elith, Leathwick & Hastie (2008). An assessment of how the utility of GP SDMs and alternative

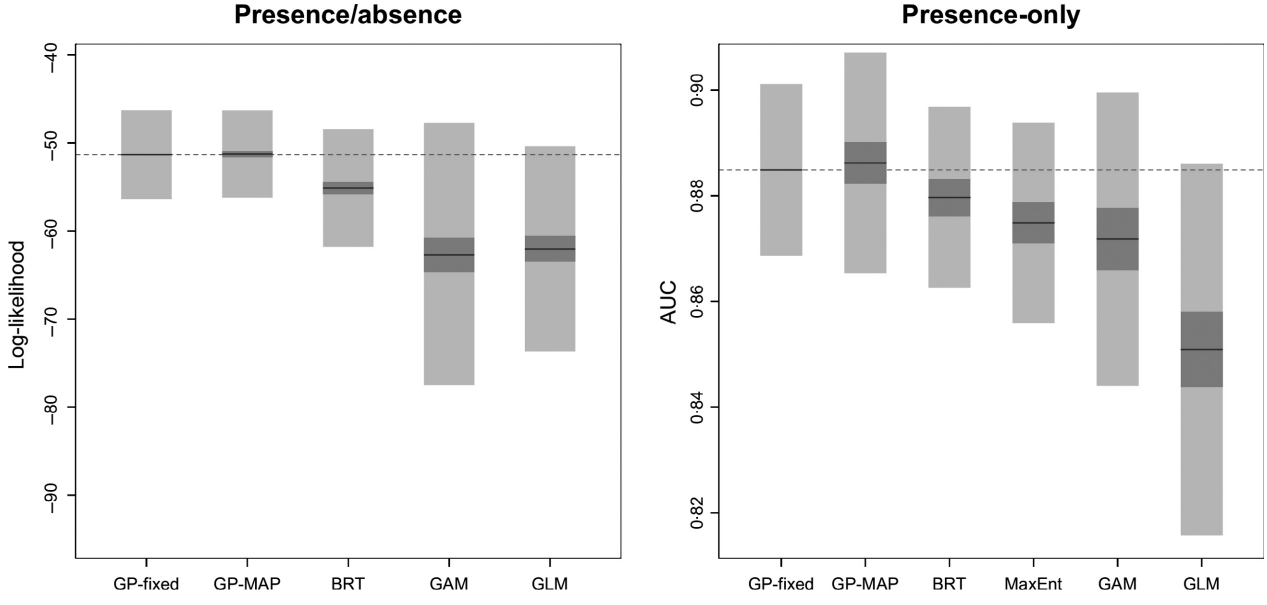

Fig. 5. Marginal validation statistics for model predictions to withheld training sets for presence/absence and presence-only data with two types of cross-validation. Centre lines give the means of the marginal validation statistic and light grey boxes give ±1 standard deviation of the marginal statistics, as an indication of the likely differences in performance of each model on an 'average' species in the plant data set. Dark grey boxes give ±1 standard deviation of the estimated difference in the mean of the statistic for each model from the mean statistic for the GP-fixed model (indicated by the horizontal dashed line), as a visual representation of the statistical tests carried out. Higher log-likelihoods and higher AUCs indicate more accurate predictions to the evaluation set.

SDMs varies according to the spatial scale, sample size and number and type of predictors would also be beneficial.

Our preliminary model comparison did not evaluate explicitly how well GP models deal with collinearity in predictors – a common issue in SDM. Like BRT, but unlike GLMs, MaxEnt or the additive GAMs, GP models can model high-dimensional interactions between predictors and thus can explicitly model a ridged surface reflecting the collinearity between any two predictors. Furthermore, the squared-exponential covariance function optimizes a lengthscale for each predictor by the marginal likelihood, so that uninformative predictors are automatically assigned large lengthscales (Rasmussen & Williams 2006). These features suggest that GP SDMs might perform well when faced with collinearity, but this should be assessed empirically in a future study.

The way in which Bayesian GP models incorporate prior ecological knowledge, via a prior estimate of the modelled function, and their rate of response to environmental gradients, seems particularly well suited to SDM applications. Such an approach is likely to be useful in cases where few occurrence records (Murray *et al*. 2009) are available but where the effects of environmental drivers are well understood, such as temperature limits on pathogen distributions (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. For example, a mechanistic model of how a species responds to one or more key predictors (such as temperature) could be fitted independently, then used as the mean function of a GP model to learn a correlative relationship with a wider range of predictors. The resulting model would retain the ability of the mechanistic model to extrapolate to future environmental scenarios or new regions, whilst still accounting for predictors whose impacts on the species' distribution are harder to describe mechanistically.

Similarly to GLMs and GAMs, GP models produce estimates of uncertainty in model predictions without the need for bootstrapping procedures as is the case with BRT, MaxEnt and other popular SDM approaches. Whilst predictions from GPs fitted using the numerical approximation procedures described here account for uncertainty in the shape of the response curve, they do not account for uncertainty in the lengthscale hyperparameters to be incorporated into this uncertainty estimate. If required, predictions accounting for uncertainty in hyperparameters can be approximated by numerical integration (e.g. a deterministic algorithm as in Rue, Martino & Chopin (2009) or by Monte Carlo). Such a procedure will inevitably be more computationally intensive, but is still likely to be far more efficient than alternative approaches such as MCMC.

Gaussian process models fitted using Laplace approximation are reasonably computationally efficient, with the GP-fixed models taking an equivalent time to GAMs, and less than BRT models. GP-MAP models took much longer, although run times of 2–3 min per species are likely to be acceptable for many applications. A downside to GP models is that in the naive case (using full-rank covariance matrices, as here), they scale cubically with the size of the data set [due to multiple matrix decompositions of $\mathcal{O}(n^3)$ complexity], so for very large data sets, GP models can be disproportionately slow. For example, on a single CPU of the computer used for model comparisons, GP-fixed takes around 0·14 s to fit to a data set with 1000 observations, for 10 000 observations that rises to 140 s, and for 100 000 (an implausibly large number for most SDM analyses), it would take around 38 h. For users who wish to fit GP models to large data sets efficiently, substantial speed-ups can be achieved via parallel computing, by using linear algebra routines optimized for multi-core machines. Far greater increases in computational efficiency could be achieved by implementing sparse GP models in addition to these approximate inference methods (see e.g. Vanhatalo, Pietiläinen & Vehtari 2010). Whilst this would enable major improvements in computational efficiency for large models, it would also entail additional approximation error.

The GP model we evaluated is one of the simplest GP models possible. The wide array of different covariance functions, likelihoods and inference methods that have been developed within the machine learning community provide an array of potential extensions and improvements to the GP model presented here. These include multi-output GP models which could be used for joint SDM (Alvarez & Lawrence 2009; Pollock *et al*. 2014), compositional covariance functions that extract simple relationships from complex signals (Duvenaud *et al*. 2013) and the tools to easily integrate SDMs with geostatistical and time-series models (Paciorek 2003; Diggle *et al*. 2013). Continuing evaluation and development of GP models for SDM therefore has the potential to vastly improve our capacity to predict and understand species distributions.

Acknowledgements

This work was funded by the NERC Centre for Ecology and Hydrology National Capability Allocation. We thank David Rogers, Miles Nunn, Luigi Sedda, Dave Harris, Bob O'Hara, Marianne Sinka, an associate editor and two reviewers who all provided helpful comments on the manuscript.

Data accessibility

The full North American Breeding Bird Survey data set can be downloaded at https://www.pwrc.usgs.gov/bbs, and the Bioclim data sets can be downloaded at http://www.worldclim.org/bioclim. The code used by Harris (2015) to download the relevant subsets of both of these data sets is available at https://github.com/davharris/mistnet. The code used to reproduce all analyses and figures in this manuscript is available at https://github.com/goldingn/gp_sdm_paper.

Conflict of interest

The authors declare that they have no conflict of interests.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory* (eds B.N. Petrov & F. Caski), pp. 267–281. Akademiai Kiado, Budapest.

Alvarez, M. & Lawrence, N.D. (2009) Sparse convolved Gaussian processes for multi-output regression. *Advances in Neural Information Processing Systems* (NIPS 21), pp. 57–64.

Byrd, R.H., Lu, P., Nocedal, J. & Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**, 1190–1208.

Diggle, P.J., Moraga, P., Rowlingson, B. & Taylor, B.M. (2013) Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, **28**, 542–563.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H., Hartig, F. *et al*. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Duvenaud, D., Lloyd, J.R., Grosse, R., Tenenbaum, J.B. & Ghahramani, Z. (2013) Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint* arXiv:1302.4922.

Elith, J., Burgman, M.A. & Regan, H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–329.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

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

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A. *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

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

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

Gething, P.W., Van Boeckel, T.P., Smith, D.L., Guerra, C.A., Patil, A.P., Snow, R.W. & Hay, S.I. (2011) Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. *Parasites & Vectors*, **4**, 4–15.

Golding, N. (2013) *GRaF: species distribution modelling using latent Gaussian random fields*. R package version 0.1-0.

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

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, **6**, 465–473.

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

Hensman, J., Fusi, N. & Lawrence, N. (2013) Gaussian processes for big data. *Proceedings of UAI*, **29**, 282–290.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

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

Huntley, B., Collingham, Y.C., Willis, S.G. & Green, R.E. (2008) Potential impacts of climatic change on European breeding birds. *PLoS Biology*, **3**, 1–11.

Lawson, C.R., Hodgson, J.A., Wilson, R.J. & Richards, S.A. (2014) Prevalence, thresholds and the performance of presence–absence models. *Methods in Ecology and Evolution*, **5**, 54–64.

Lehmann, A., Leathwick, J.R. & Overton, J.M. (2002) Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity & Conservation*, **11**, 2217–2238.

Lindgren, F. & Rue, H. (2015) Bayesian spatial modelling with R – INLA. *Journal of Statistical Software*, **63**, 1–25.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

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

McPherson, J.A.N.A., 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*, **41**, 811–823.

Murray, J.V., Goldizen, A.W., O'Leary, 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.

Paciorek, C.J. (2003) Nonstationary Gaussian processes for regression and spatial modelling. Thesis, 6. doi: 10.1371/journal.pone.0019736.

Patil, A. (2007) Bayesian nonparametrics for inference of ecological dynamics. Ph.D. thesis, University of California Santa Cruz.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J. & Elith, J. (2013) On estimating probability of presence from use-availability or presence-background data. *Ecology*, **94**, 1409–1419.

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**, 181–197.

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.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., Hara, R.B.O., Parris, K.M., Vesk, P.A. & Mccarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397–406.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C.E. & Williams, C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Ridgeway, G. (2013) *gbm: generalized boosted regression models*. R Package Version 2.1.

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

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

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A. & Krianski, E.T. (2015) *INLA: Functions Which Allow to Perform Full Bayesian Analysis of Latent Gaussian Models Using Integrated Nested Laplace Approximation*. R package version 0.0-1420281647.

Sauer, J.R., Hines, J.E., Fallon, J., Pardieck, K.L., Ziolkowski Jr, D.J. & Link, W.A. (2014) *The North American Breeding Bird Survey, Results and Analysis 1966–2013. Version 01.30.2015*. USGS Patuxent Wildlife Research Center, Laurel, MD.

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

Sinclair, S.J., White, M.D. & Newell, G.R. (2010) How useful are species distribution models for managing biodiversity under future climates. *Ecology and Society*, **15**, 8–20.

Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., Hemingway, J. *et al*. (2010) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors*, **3**, 117.

Vanhatalo, J., Pietiläinen, V. & Vehtari, A. (2010) Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, **29**, 1580–1607.

Vanhatalo, J., Veneranta, L. & Hudd, R. (2012) 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*, **228**, 49–58.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *The Annals of Applied Statistics*, **4**, 1383–1402.

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

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

Yackulic, C.B., Chandler, R.B., Zipkin, E.F., Royle, J.A., Nichols, J.D., 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.

Received 24 July 2015; accepted 13 November 2015
Handling Editor: David Warton