Received: 5 July 2021 | Accepted: 6 December 2022

DOI: 10.1111/biom.13816

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Fast Bayesian inference for large occupancy datasets

Alex Diana[1] | **Emily Beth Dennis**[2,1] | **Eleni Matechou**[1] | **Byron John Treharne Morgan**[1]

[1]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK

[2]Butterfly Conservation, Manor Yard, East Lulworth, Wareham, Dorset, UK

Correspondence
Alex Diana, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7FS, UK.
Email: ad625@kent.ac.uk

Abstract

In recent years, the study of species' occurrence has benefited from the increased availability of large-scale citizen-science data. While abundance data from standardized monitoring schemes are biased toward well-studied taxa and locations, opportunistic data are available for many taxonomic groups, from a large number of locations and across long timescales. Hence, these data provide opportunities to measure species' changes in occurrence, particularly through the use of occupancy models, which account for imperfect detection. These opportunistic datasets can be substantially large, numbering hundreds of thousands of sites, and hence present a challenge from a computational perspective, especially within a Bayesian framework. In this paper, we develop a unifying framework for Bayesian inference in occupancy models that account for both spatial and temporal autocorrelation. We make use of the Pólya-Gamma scheme, which allows for fast inference, and incorporate spatio-temporal random effects using Gaussian processes (GPs), for which we consider two efficient approximations: subset of regressors and nearest neighbor GPs. We apply our model to data on two UK butterfly species, one common and widespread and one rare, using records from the Butterflies for the New Millennium database, producing occupancy indices spanning 45 years. Our framework can be applied to a wide range of taxa, providing measures of variation in species' occurrence, which are used to assess biodiversity change.

KEYWORDS

Bayesian analysis, biodiversity change, citizen-science data, occupancy models, pólya-gamma, species distribution models

1 | INTRODUCTION

1.1 | Background and motivation

Robust measures of biodiversity change are vital for monitoring the varying state of species' populations and evaluating progress of conservation actions, for example, toward national and international targets (Butchart et al., 2010). Data from standardized, long-running monitoring schemes are used to produce estimates of species' status and trends, particularly in terms of changes in abundance. However, such data sources are limited taxonomically and

geographically. By their nature of intensive, formal sampling they may be limited in spatial coverage and therefore cannot always be used to appropriately measure changes in species' distributions over time.

Conversely, opportunistic records of occurrence are often and increasingly available in large quantities for extensive geographic areas and time periods, and for a wide variety of taxa. However, opportunistic data are inherently biased (Isaac & Pocock, 2015). Data are typically presence-only, where records only indicate where and when a species is seen rather than including information on non-detection, unless complete lists are recorded. Data recording the distribution of animals and plants are frequently analyzed using occupancy models (MacKenzie et al., 2018), as they allow for imperfect detection. Applying such models to presence-only data requires non-detections to be inferred from the observations of other species (Kéry et al., 2010). Data of this nature are not standardized, and result from the submission and collation of records by citizen scientists who choose where, when, and what to record, but are often available in large quantities. For example, the Global Biodiversity Information Facility (GBIF) consists of more than 2.3 billion occurrence records for at least one million species (GBIF.org, 2022). In the United Kingdom, extensive occurrence data are available for many taxonomic groups, and the Biological Recording Centre (BRC) oversees more than 80 recording schemes (BRC, 2022; Pocock et al., 2015). Such data are commonly used to produce atlases for various taxa (e.g., Blockeel et al., 2014; Randle et al., 2019) and contribute to national biodiversity assessments, for example, the State of Nature report (Hayhow et al., 2019) and government biodiversity indicators (Department for Environment, Food and Rural Affairs, UK, 2020).

In addition to imperfect detection, modeling approaches for occurrence data of this type also need to account for spatial and temporal autocorrelation (Guélat & Kéry, 2018; Strebel et al., 2022). In this case, Bayesian hierarchical models are an appropriate choice, thanks to the available tools for accounting for and inferring the effects of site and time-specific random effects (r.e.). However, Bayesian inference is computationally demanding, in particular when model-fitting involves large numbers of latent variables. Efficient model-fitting is increasingly important with the ongoing growth in the volume of biological recording data, partly due to increasing participation through new technologies and platforms for data submission (August et al., 2015). Fast inference is also motivated by the increasing desire to update species' trend estimates frequently, in order to inform the measuring and reporting of biodiversity change.

1.2 | Current models

One popular form of model describes dynamic occupancy; see, for example, Royle and Dorazio (2008, Chap. 9). This model is designed for data from several years and incorporates parameters representing colonization and extinction. It is therefore mechanistic, with parameters which may assist in the understanding of spatial and temporal changes in the distribution. The basic model may be extended, for example, to allow temporal development to depend upon the status of neighboring sites; see Broms et al. (2016). Typically, these informative, complex models are designed for relatively short studies with small numbers of sites. Both Bayesian and classical inference methods have been used, in the latter case using `unmarked` in R (Fiske & Chandler, 2011); Bayesian inference is discussed in Kéry and Royle (2021, pp. 208, 564). However for large numbers of sites and occasions computing times can be excessive (see van Strien et al., 2013), and other approaches are in current use in these cases.

Alternatively, one can use static models, in which a simple occupancy model is fitted to the data for each year separately, and the site occupancy probability for each year is described by a logistic function of site-specific covariates. This approach was proposed by Dennis et al. (2017); it is fitted using `unmarked` (Fiske & Chandler, 2011) and classical inference. The static model is appreciably faster in execution. However, a drawback of analyzing the data from each year separately arises regarding records from early years, which may not be sufficiently numerous to allow the fitting of a static model in those cases. Similarly, producing occupancy trends for rare or less well-recorded species may not be possible using the static model and there is no sharing of information between years, as each year is modeled separately.

The more recent approaches cast the detection and presence process in a binomial probit or logistic regression framework, taking advantage of fast and efficient Gibbs sampler schemes. Since posterior inference for both the probit and logistic likelihood are analytically intractable, the Gibbs sampler relies on data augmentation schemes to obtain tractable posterior inference. For probit regression, the data augmentation scheme used is due to Albert and Chib (1993), while for logistic regression the scheme is due to Polson et al. (2013). One of the first approaches within the occupancy modeling frameword is due to Dorazio and Rodriguez (2012), who use a probit-regression model formulation for the detection and occupancy probability. Similarly, Johnson et al. (2013) and Hepler and Erhardt (2021) presented a spatial regression model using a conditional autoregressive (CAR) model. All the aforementioned approaches focus on a probit link

function. However, after the introduction of the Pólya-Gamma scheme (PG) of Polson et al. (2013), Clark and Altwegg (2019) have also proposed the use of the logit in occupancy models. The logit link leads to more intuitive interpretation of the regression coefficients in terms of log-odds, and hence is the more natural choice for binary variables, such as site occupancy and detection. Additionally, the PG scheme has been proved to have optimal mixing properties (Choi & Hobert, 2013). Moreover, as mentioned later in Section 6, the PG scheme can be more easily extended to interesting developments, such as variable selection.

Spatial autocorrelation between surveyed sites has typically been incorporated using a CAR process (Mardia, 1988). The CAR prior is usually defined on a lattice, where sites are equally spaced and each site relies on the definition of a neighborhood structure. Therefore, the use of the CAR on irregular site locations entails approximation into a regular grid. For example, Johnson et al. (2013) considered a tessellation of 100 km^2 equally spaced hexagonal survey units and Clark & Altwegg (2019) considered a continuous grid of $5' \times 5'$ cells. However, opportunistic data of the type considered in this paper, because of their nature, are collected at irregular locations and the degree of error by approximating them on a lattice can be considerable.

Temporal autocorrelation has been introduced using a first-order vector autoregressive process (Hepler & Erhardt, 2021), using a spline-basis approach for the spatial effects, whose coefficients follow a time-dependent random walk (Rushing et al., 2019), or using a random walk to describe the changes in occupancy across the years (Outhwaite et al., 2018).

Instead, in this paper we model spatial and temporal autocorrelation using a Gaussian process (GP) approach (Rasmussen & Williams, 2006). The advantage of using a GP is that it allows us to naturally model spatial autocorrelation between sites sampled at continuous locations, which is typically the case for opportunistic data, and, in contrast to CAR, allows for a different degree of correlation between sites according to their distance, even if they are neighboring. We also model temporal autocorrelation within a GP framework, and consider an additive structure for the effect of space and time, also known as the *separable* case, and describe how to implement the non-separable case in our framework, where the spatial and temporal r.e. are not a priori independent.

We cast the occupancy and detection process within a logistic regression framework, and take advantage of the efficient PG augmentation scheme (Polson et al., 2013) for inference, which is well-established in the Bayesian literature (Holsclaw et al., 2017; Linderman et al., 2015) but not in the ecological modeling literature, with some recent exceptions (Clark & Altwegg, 2019; Griffin et al., 2020) In addition, we describe and compare different approximations for the GP, subset of regressors (SoR) (Smola & Bartlett, 2001) and nearest neighbor GPs (NNGPs) (Datta et al., 2016), and demonstrate how they can be used within a PG framework.

The new model of this paper responds to the need for a computationally efficient approach to analyze presence–absence data arising from a large number of sites, while accounting for spatial and temporal autocorrelation, and which accommodates species with sparse records by jointly modeling data collected across different years.

1.3 | Paper outline

The model of the paper is described in Section 2. Section 3 discusses the theoretical concepts of our model, such as the PG scheme and the GP approximations. Section 4 presents simulation studies showing comparisons between different spatial approximations. Section 5 applies the new model to two illustrative datasets on UK butterflies. Section 6 discusses possible extensions and the paper ends with discussion in Section 7. Technical details of the MCMC and additional results, including a simulation study demonstrating the importance of accounting for spatial autocorrelation, are provided in the supporting information.

2 | MODEL: BAYESIAN FRAMEWORK AND GAUSSIAN PROCESSES

For any species, observations are collected at S sites and across Y years. A number of observations may be collected at each site and year. This number, which does not need to be defined for the purposes of the model, does not have to be the same for all sites or years and can be equal to 0 for particular pairs of sites and years. We refer to the unique pairs of sites and years with at least one observation as *sampling units* and we index them by $j = 1, \dots, J$. If all sites are sampled in all years, then $J = S \times Y$. We assume that the occupancy status of a site can change between years but not within, which is a standard assumption of similar models for multi-season occupancy data.

We introduce latent variables $z_j, j = 1, \dots, J$, indicating the occupancy status of sampling units, with $z_j = 1$ if sampling unit j is occupied and 0 otherwise. We assume that each sampling unit is occupied with probability ψ_j, that is, $z_j \sim \text{Be}(\psi_j)$. We index the site and year of sampling unit j by s_j and t_j, respectively. Finally, we denote by $\mathbf{x}_s = (x_s^1, x_s^2)$ the location of site s and by w_y the time point of year y. For example, if the data were collected in years 2000, 2001, 2004, and 2005, $(w_1, \dots, w_4) =$

(2000, 2001, 2004, 2005) and $t_j = 1, \dots, 4$ if sampling unit j belongs to years 2000,2001,2004,2005, respectively.

We denote by N the total number of observations across all sampling units and we define y_i, $i = 1, \dots, N$, to be the outcome of observation i, that is, $y_i = 1$ if the species is detected at observation i, and 0 otherwise. Finally, we introduce $k_i \in \{1, \dots, J\}, i = 1 \dots, N$, which indexes the sampling unit of observation i so that if observation i corresponds to sampling unit j, then $k_i = j$. Therefore, if sampling unit j is occupied then $z_{\{i:k_i=j\}} = 1$ and otherwise $z_{\{i:k_i=j\}} = 0$. We account for the probability of a false negative observation but assume that false positive observations do not occur and hence assume that $y_i \sim \text{Be}(p_i z_{k_i})$ with p_i being the probability of detecting the species given presence.

We model the probability of detection p_i as

$$\text{logit}(p_i) = u_{t_{k_i}} + X_i \beta^p, \tag{1}$$

where u_t is a year-specific r.e. with prior distribution $u_t \sim N(\mu_0^p, \sigma_0^p)$, t_{k_i} is the index of the year in which observation i is collected and X_i is the set of covariates for observation i, $i = 1, \dots, N$.

We model the probability that sampling unit j is occupied, ψ_j, as a function of both fixed effects, such as covariates, and r.e., and specifically r.e. that account for temporal autocorrelation between years, spatial autocorrelation between sites and individual variation of sites:

$$\text{logit}(\psi_j) = \mu^\psi + b_{t_j} + a_{s_j} + X_j^C \beta^\psi + \epsilon_{s_j} \tag{2}$$

where μ^ψ is an intercept, b_t is a r.e. for year t, a_s and ϵ_s are r.e. for site s, and X_j^C is the set of covariates for sampling unit j. The site-specific random effects $(\epsilon_1, \dots, \epsilon_S)$ are modeled as independent random variables $\epsilon_s \sim N(0, \sigma_\epsilon^2)$, while the rest of the r.e. are defined below using GPs.

2.1 | Gaussian processes

To define a distribution for the r.e. b and a, we introduce the concept of GPs (Rasmussen & Williams, 2006). Given a general covariance function $k(\xi_i, \xi_j)$, we define the entries of the covariance matrix between the sets of points $\xi^1 = (\xi_1^1, \dots, \xi_n^1)$ and $\xi^2 = (\xi_1^2, \dots, \xi_m^2)$, $K(\xi^1, \xi^2)$, as $\{K(\xi^1, \xi^2)\}_{i,j} = k(\xi_i^1, \xi_j^2)$. If $\xi^1 = \xi^2$, we simplify the notation $K(\xi^1, \xi^1)$ to $K(\xi^1)$ and we might omit ξ if the dependency is clear. A function f has a GP prior distribution if, for every combination of values $\xi_1, \dots, \xi_n$, it holds that $(\eta_1, \dots, \eta_n) \sim \text{N}(0, K(\xi_1, \dots, \xi_n))$, where $\eta_i = f(\xi_i)$. In this paper, we consider the exponential covariance function $k(\xi_i, \xi_j) = \sigma^2 \text{e}^{-\frac{|\xi_i - \xi_j|^2}{l^2}}$, where σ tunes the overall variability of the GP and l tunes the correlation between points, and we write the related covariance matrix as $K_{l,\sigma}$. The points $\xi_1, \dots, \xi_n$ are called *support points*. Although, in general, the GP is defined for a function with an infinite number of support points, in our case, we apply the GP on a function defined on a finite number of points, as we explain below, and hence this is simply equivalent to assuming a multivariate normal distribution on $(\eta_1, \dots, \eta_n) = (f(\xi_1), \dots, f(\xi_n))$. The advantage of GPs is that posterior inference is analytically tractable. If a prior $\eta \sim N(0, K)$ is used with a likelihood $y \sim N(\eta, \sigma^2 I)$, the posterior distribution $p(\eta|y)$ has the form $N(\frac{1}{\sigma^2}(K^{-1} + \sigma^{-2} I)^{-1} y, (K^{-1} + \sigma^{-2} I)^{-1})$. The posterior distribution at new points is also readily available.

To account for temporal correlation, we assume that the year-specific r.e. $\mathbf{b} = (b_1, \dots, b_Y)$ are distributed according to a GP with parameters (l_T, σ_T) and support points $(w_1, \dots, w_Y)$, which corresponds to assuming that $(b_1, \dots, b_Y) \sim N(0, K_{l_T, \sigma_T}(w_1, \dots, w_Y))$. Similarly, we account for spatial autocorrelation by assuming that the $\mathbf{a} = (a_1, \dots, a_S)$ are distributed according to a GP with parameters (l_S, σ_S) and support points the locations $(\mathbf{x}_1, \dots, \mathbf{x}_S)$ of the sites, which corresponds to assuming that $(a_1, \dots, a_S) \sim N(0, K_{l_S, \sigma_S}(\mathbf{x}_1, \dots, \mathbf{x}_S))$.

2.2 | Comparison between GP and CAR

As mentioned in the introduction, a popular alternative to the GP prior for modeling temporal or spatial autocorrelations is the conditionally autoregressive (CAR) prior (Besag & Kooperberg, 1995). The CAR prior is defined conditionally on a *neighborhood* structure for the observations. Given a neighborhood matrix, W, where $W_{ij} = 1$ if the observations a_i and a_j are in the same neighborhood and 0 otherwise, and a spatial dependence parameter ρ, the CAR model defines a prior for the vector $\mathbf{a} = (a_1, \dots, a_S)$ by defining a prior on the full conditional distributions $a_i | a_{-i} \sim N(\rho \frac{\sum_{j \neq i} w_{ij} a_j}{d_i}, \frac{\sigma^2}{d_i})$, where d_i is the number of elements in the neighborhood of i. Therefore, the conditional mean of the ith observation is a weighted average of the observations in its neighborhood, that is, the observations j for which $W_{ij} = 1$. It follows from these assumptions that $\mathbf{a} \sim N(0, Q^{-1})$, with the precision matrix $Q = \frac{1}{\sigma^2}(D - \rho W)$, where D is a diagonal matrix with entries $D_{ii} = d_i$. Since D and W are sparse, this prior leads to sparse precision matrices (but dense covariance matrices in general), with non-neighboring elements having entry 0 in the precision matrix, but

not necessarily in the covariance matrix. This leads to the CAR being more computationally efficient than the GP, since the precision matrix of the GP is in general not sparse except in the case of the Laplace kernel $k(\xi_1, \xi_2) = a\exp(b|\xi_1 - \xi_2|)$. However, the CAR assigns equal correlation to elements in the same neighborhood, irrespective of their actual distance. We note that extensions to irregular locations do exist (Rue & Held, 2005), but they are mathematically more challenging and have not been considered in an occupancy framework. On the other hand, GPs account for irregular locations, but are computationally more expensive, and approximation methods have to be considered when the number of observations, in this case sites or times, are large.

2.3 | Hierarchical structure

The following hierarchical structure completes the definition of our model, including the prior distributions of all parameters, where $i = 1, \dots, N$, $j = 1, \dots, J$, $t = t_1, \dots, t_J$, and $s = 1, \dots, S$,

$$\begin{cases} y_i \sim \text{Be}(p_i z_{k_i}) & z_j \sim \text{Be}(\psi_j) \\ \text{logit}(\psi_j) = \mu^\psi + b_{t_j} + a_{s_j} + X_j^C \beta^\psi + \epsilon_{s_j}, & \mu^\psi \sim N(\mu_0^\psi, \sigma_0^\psi), \quad \beta^\psi \sim N(0, \phi^\psi I) \\ & \epsilon_s \sim N(0, \sigma_\epsilon^2) \quad \sigma_\epsilon^2 \sim \text{IG}(a_\epsilon, b_\epsilon) \\ \text{logit}(p_i) = u_{t_{k_i}} + X_i \beta^p, & u_t \sim N(\mu_0^p, \sigma_0^p), \quad \beta^p \sim N(0, \phi^p I) \\ (b_1, \dots, b_Y) \sim N(0, K_{l_T, \sigma_T}(w_1, \dots, w_Y)), & \sigma_T \sim \text{IG}(a_{\sigma_b}, b_{\sigma_b}), \quad l_T \sim \text{Gamma}(a_{l_T}, b_{l_T}), \\ (a_1, \dots, a_S) \sim N(0, K_{l_S, \sigma_S}(\mathbf{x}_1, \dots, \mathbf{x}_S)), & \sigma_S \sim \text{IG}(a_{\sigma_S}, b_{\sigma_S}), \quad l_S \sim \text{Gamma}(a_{l_S}, b_{l_S}). \end{cases} \tag{3}$$

3 | THEORY

In this section, we define the basic building blocks of our inference strategy. First, we describe the PG scheme, which is a data augmentation scheme used to obtain analytically tractable posterior distributions in a logistic regression setting. Next, we define the GP approximation chosen to efficiently model autocorrelation between a large number of observations.

3.1 | Pólya-Gamma scheme

A random variable w has a PG distribution, $w \sim \text{PG}(d, c)$ if $w = \frac{1}{2\pi^2}\sum_{k=1}^{\infty} \frac{g_k}{(k-\frac{1}{2})^2 + \frac{c^2}{4\pi^2}}$, where $g_k \sim \text{Gamma}(d, 1)$. According to the PG scheme, given a set of n observations $y_i \sim \text{Binomial}(d_i, p_i)$, where $\text{logit}(p_i) = X_i\beta$, a Gibbs sampler scheme for β is available by introducing a set of random variables ω_i, such that $\omega_i \sim \text{PG}(d_i, 0)$. More specifically, assuming prior distribution $\beta \sim N(b, B)$, the full conditional distributions used for the Gibbs sampler are

$$(\omega_i|\beta) \sim \text{PG}(d_i, X_i\beta) \qquad i = 1, \dots, n \tag{4}$$

$$(\beta|y, \omega) \sim N((X^T\Omega X + B^{-1})^{-1}(X^T k + B^{-1}b), (X^T\Omega X + B^{-1})^{-1}), \tag{5}$$

where $\Omega = \text{diag}(\omega_1, \dots, \omega_n)$ and $k = (y_1 - \frac{d_1}{2}, \dots, y_n - \frac{d_n}{2})$. Polson et al. (2013) described an efficient algorithm to sample a PG r.v. that does not require truncating the infinite sum in the definition of the PG distribution. We use the PG scheme to sample jointly from the posterior distribution of the parameters (u_t, β^p) and $(\mu^\psi, b_t, a_s, \beta^\psi)$ in Equations (1) and (2), respectively.

3.2 | Spatial approximations

As explained in Section 2.1, in the context of continuous observations, inference using GPs relies on factorization of the $S \times S$ matrix $(K^{-1} + \sigma^{-2}I)$, where S is the number of observations, whereas when using the PG scheme to model binary observations, we need to factorize the matrix $(K^{-1} + X^T\Omega X)$ in Equation (5), since the prior covariance matrix B of the spatial r.e. in Equation (5) corresponds to the GP matrix K. If the number of points, in our case the number of sites, is large ($\approx 10^6$ in the case study), it becomes computationally prohibitive to obtain the factorization of the $S \times S$ matrix. Therefore, approximations of the GP have to be considered. There is a large literature on approximation methods for GPs, so we do not aim to give a comprehensive review here but instead focus on two popular types of approximations: low-rank approxima-

tions and sparse approximation methods. For an extensive review, we refer the reader to Liu et al. (2020). In the simulation study in Section 4.1, we compare a method from the class of low-rank approximations, the SoR, and a method from the class of sparse approximations, the NNGP. We also consider a very basic approximation, in which the initial GP on S locations is approximated by introducing another GP computed on a smaller number of support points $(\tilde{x}_1, \dots, \tilde{x}_M)$, where $M \ll S$, with respective values $(\tilde{a}_1, \dots, \tilde{a}_M)$, and replacing each original value a_j with the value of its closest support point $\tilde{a}_{\tilde{j}}$. We term this approximation the closest point (CP) and note that its complexity is $O(M^3 + S)$.

Low-rank methods approximate the covariance matrix K as $\Lambda^T \tilde{K} \Lambda$, where $\tilde{K}$ is an $M \times M$ matrix and Λ is an $M \times S$ matrix, where $M \ll S$. The Woodbury identity matrix can then be used to replace the inversion of the $S \times S$ matrix K^{-1} with the $M \times M$ matrix $\tilde{K}^{-1}$. One of the most popular approximations in this class is the SoR method, which consists of using the degenerate covariance function $k_{\text{SOR}}(x, y) = K(x, \mathbf{x}^\star)(\underbrace{K(\mathbf{x}^\star, \mathbf{x}^\star)}_{\tilde{K}})^{-1} K(\mathbf{x}^\star, y)$, where $K(x, y)$ is a covariance function, defined in Section 2.1, and $\mathbf{x}^\star$ is a set of M points, called *inducing points*. A useful alternative representation of the SoR is to express the vector of effects $a \sim N(0, K(x, x))$ as $\tilde{a} = \underbrace{K(x, \mathbf{x}^\star)\tilde{K}^{-1}}_{K^\star} a^\star$, where $a^\star \sim N(0, \tilde{K})$ is a vector of lower dimension M following an exact GP prior. Using this representation, inference can be performed as in a standard regression model, where $K^\star$ is the design matrix and $a^\star$ is the M-dimensional vector of regression coefficients. This leads to the posterior precision matrix $\sigma^{-2}(K^\star)^T K^\star + \tilde{K}^{-1}$ in the context of continuous observations and, from Equation (5), $(K^\star)^T \Omega K^\star + \tilde{K}^{-1}$ in the context of the PG scheme.

However, if in the continuous case $(K^\star)^T K^\star$ needs to be precomputed only once, in the PG case $(K^\star)^T \Omega K^\star$ has to computed for each new draw of Ω from Equation (4). Since the computation of $(K^\star)^T \Omega K^\star$ has complexity $O(SM^2)$, which is much greater than the cost $O(M^3)$ to factorize the precision matrix, this becomes the dominant calculation and the SoR method quickly becomes unfeasible. To avoid this drawback, we propose to replace the full design matrix $K^\star$ of dimension $S \times M$ with a smaller design matrix of dimension $S \times M^\star$, by taking the $M^\star$ biggest components of each row (or, equivalently, by considering only the $M^\star$ CPs between all the M support points). This approximation, which we term approximated SoR (ASoR), has reduced complexity $O(M^3 + SM^{\star 2})$. We note that the CP approximation can be seen as a special case of the ASoR, where $M^\star = 1$ and $K^\star(x, x^\star) \equiv 1$.

Sparse approximation methods rely on obtaining a sparse approximation of the precision matrix K^{-1} by zeroing some of its elements, so that fast methods for factorizing sparse matrices can be employed. This approach is closely related to working with a Gaussian Markov random field (GMRF) and Rue and Tjelmeland (2002) have proposed embedding the irregular point locations in a regular lattice and approximating the GP with a GMRF.

A related method is the NNGP (Datta et al., 2016). The idea of the NNGP is to replace the conditional distributions $a_i | a_1, \dots, a_{i-1}$, with an approximation using only the m closest neighbors in the previous $i-1$ observations, $a_i | a_{i_1^\star}, \dots, a_{i_m^\star}$. Using this approximation, the precision matrix K^{-1} can be expressed as $(I - A)D(I - A)^T$, where A is a sparse triangular matrix and D is diagonal, hence the product is also a sparse matrix. For more details on inference with NNGP, we refer the reader to Finley et al. (2019). We note that the complexity of the NNGP is not known in general as it depends on the sparsity pattern of the matrix A. The NNGP is also a GMRF, as the full conditional of each observation depends only on the value of the observations in its neighborhood.

4 | SIMULATION STUDIES

4.1 | Approximation of Gaussian processes

We performed a simulation study to compare the three GP approximation methods described in Section 3.2: the CP, the ASoR, and the NNGP. We ran the ASoR method by choosing $M^\star = 10$, since we observed that the performance was very similar to the standard SoR method if $M^\star$ was chosen as large as 10. To perform the simulation study, we generated data over 10 years on $S = 4900$ sites, spread uniformly over a unit square. We performed 15 runs, where for each run we fitted the model on 70% of the sites, chosen at random, predicted the spatial pattern a_s on the remaining sites and computed the mean absolute error between the true values $\bar{a}_s$ and the posterior means $\hat{a}_s$. To tune the CP and the ASoR method, we varied the number of inducing points M. The inducing points were chosen by taking a uniform grid of points across the unit square and varying the grid step. To tune the NNGP, we varied the number of neighbors considered. Results are shown in Figure 1. The ASoR and the NNGP method clearly outperform the CP as far as mean absolute error is concerned, although the CP is generally faster as expected. The NNGP and the ASoR obtained comparable performances in terms of both computational time and predictive power.

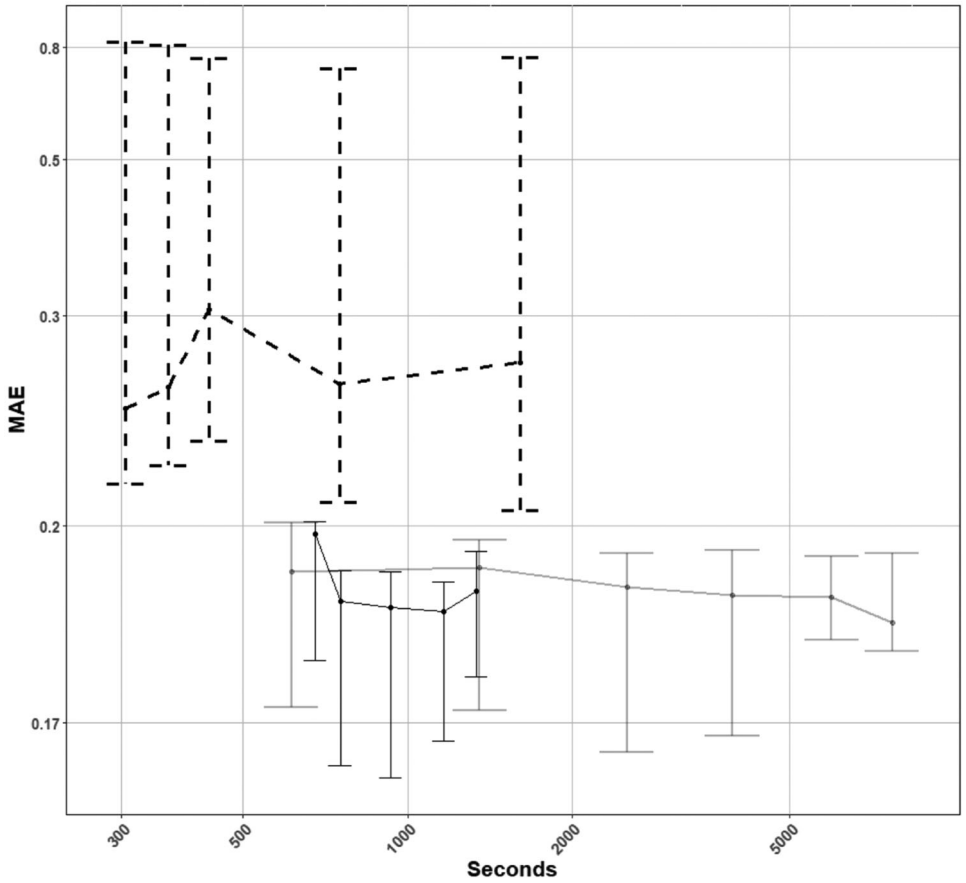

FIGURE 1 Comparison between the Gaussian process (GP) approximation methods showing the relationship between computational time and mean absolute error (MAE) for each method. The *x*-axis represents the computational time and the *y*-axis represents the mean absolute error. The results of the ASoR are shown by the solid black line, the results of the NNGP by the solid grey line, and the results of the CP by the dashed line. For the CP, we used grid steps {0.2, 0.175, 0.15, 0.125, 0.1}, for the ASoR, we used grid steps {0.3, 0.25, 0.2, 0.175, 0.15}, while for the NNGP we used as number of neighbors {5, 10, 15, 20, 25, 30}

Therefore, we recommend either the NNGP or the ASoR, since the CP approximation has been found to be too crude, and consider the ASoR for the case studies presented in this paper.

5 | CASE STUDIES

We applied our model to data for two UK butterfly species: Ringlet (*Aphantopus hyperantus*) and Duke of Burgundy (*Hamearis lucina*). In doing so, we demonstrate the performance of the new model for both a common, widespread species (Ringlet) and a rare, range-restricted species (Duke of Burgundy).

Butterfly data were collated through the Butterflies for the New Millennium (BNM) recording scheme run by Butterfly Conservation, using records collected between 1970 and 2014, during which the database consisted of over 11 million records of UK butterflies (Fox et al., 2015). BNM data were restricted to records with an exact date and location of 1 km resolution or finer. For each of the two species, records were then filtered to months within which records of the focal species had been recorded, and observations of other species used to form detection histories (Kéry et al., 2010). Thus for Ringlet, the dataset featured > 2 million records from 140,887 unique 1 km^2 (defined as sites), of which Ringlet had been recorded at 47,561 sites from 218,225 detections. Conversely, the dataset for Duke of Burgundy consisted of approximately 1.5 million records from 128,197 sites (1 km^2), of which Duke of Burgundy had been recorded at 747 sites from 6,584 detections. On a machine equipped with an Intel Core i7-10610@1.8 GHz with 16 GB of RAM, the model took 19 h to run on each dataset for 15×10^4 burn-in iterations and 25×10^4 iterations.

For both species, we considered the interactions between year and easting and between year and northing as covariates for occupancy probability. For the detection probability, we considered as covariates the relative list length and the first three powers of the Julian date. The relative list length is obtained by dividing the list length,

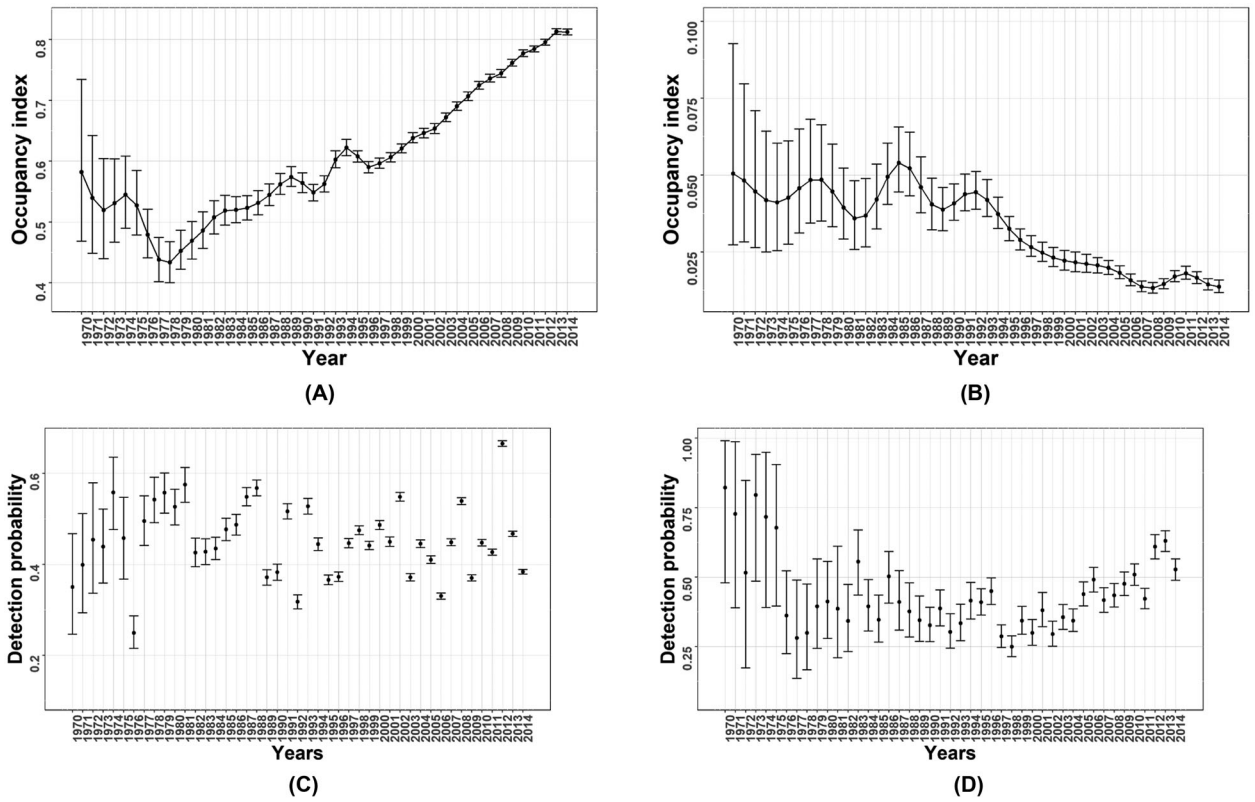

FIGURE 2 Top row: 95% posterior credible interval (PCI) of the occupancy index for (A) Ringlet and (B) Duke of Burgundy. Bottom row: 95% PCIs of the year-specific detection probabilities at the average value of the list length covariate for (C) Ringlet and (D) Duke of Burgundy. The dots represent the posterior medians. Note that different scales are used for the two species

which is the number of species recorded for a given site/date (Szabo et al., 2010; van Strien et al., 2013), by the maximum recorded list length in a neighboring area of 50 km. All covariates were standardized to have zero mean and unit variance. We do not consider the main effects for year or easting/northing, since the effects of year and space on the probability of occupancy are already accounted for in the processes b_t and a_s and therefore adding these main effects would lead to confounding between the spatial r.e. and the fixed-effect covariates (Hodges & Reich, 2010; Reich et al., 2006). Finally, we employ the ASoR approximation defined in Section 3.2 with inducing points taken on a grid of 20 km width on the study area, which corresponds to $M = 909$ inducing points.

For each species, we calculate the yearly occupancy index (Dennis et al., 2017) at each MCMC iteration using $I_t^{(l)} = \frac{1}{S}\sum_{j=1}^{S}\psi_{t,j}^{(l)}$, where $\psi_{t,j}^{(l)}$ is the occupancy probability at site s and year t for iteration l. Posterior summaries of the occupancy index for both species are shown in Figure 2, and support previous findings suggesting that Ringlet has increased in occurrence since 1970, whereas Duke of Burgundy has seen a reduction in occurrence (Fox et al., 2015). The indices for both species show increasing precision with time, reflecting an increase in underlying recording effort (Dennis et al., 2017), which is also a feature for other taxa (Isaac & Pocock, 2015).

The estimated occupancy probabilities for the two species are mapped over space for selected years in Figure 3. Note that the map for the Duke of Burgundy has been zoomed in to the part of the country where the species can be found, due to its restricted range. These patterns are consistent with what is known, namely that Ringlet has been expanding in range and now occupies most of the UK, with the exception of Northern Scotland and a small portion of northern England, whereas Duke of Burgundy has been contracting in range and can now only be found at a very small number of locations.

Ringlet has been shown to have increased in both range and abundance (Fox et al., 2015), which is a likely response to recent climate change (Mason et al., 2015). Duke of Burgundy is one of the UK's most threatened species (Fox et al., 2011), with long-term declines in both abundance and distribution (Fox et al., 2015), but as seen in Figure 2 the decline in occurrence appears to have stabilized in more recent years, which may be due to conservation efforts (Ellis et al., 2012).

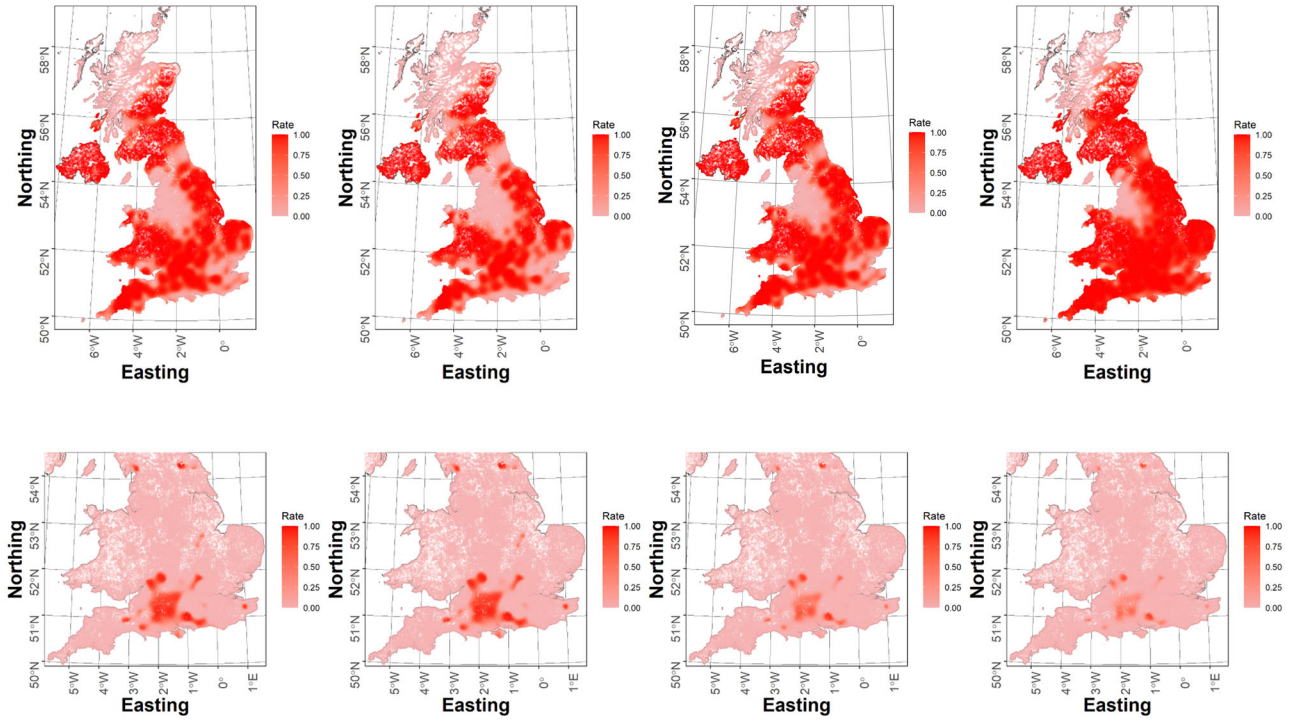

FIGURE 3 Posterior medians of the site-specific occupancy probabilities for Ringlet (top row) and Duke of Burgundy (bottom row) for 1970 (first column), 1985 (second column), 2000 (third column), and 2014 (fourth column). White areas represent parts of the country with no records of any butterfly species. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

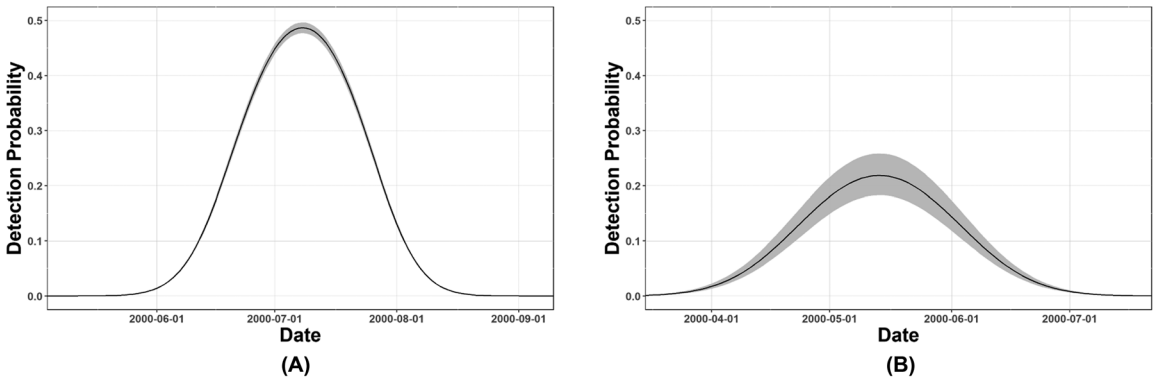

FIGURE 4 Posterior median and 95% posterior credible interval (PCI) of the detection probability p across the year for the Ringlet (A) and Duke of Burgundy (B) in the year 2000, at the average value of the relative list length. The black line represents the posterior median. We note that we have plotted only one year as the coefficients of Julian date are constant across time and hence the trend in other years is simply a shifted version

Relative list length has a positive effect on detection probability with 95% posterior credible interval (PCI) (1.085, 1.098) and (0.797, 0.866) for Ringlet and Duke of Burgundy, respectively. The PCIs of the year-specific detection probabilities are shown in Figure 2. Interestingly, detection probabilities for Ringlet appear relatively stable over time, whereas estimated detection probabilities for Duke of Burgundy may have increased slightly, possibly due to increases in recorder effort to observe this rare, but also diminutive, species. In Figure 4, we show the posterior summaries of detection probability at each time t of the year, p_t, for both species, where it can be seen that the detection probability is extremely low outside the summer months corresponding to each species' flight period. However, it is important to consider that in our model we assume that occupancy status of sites does not change

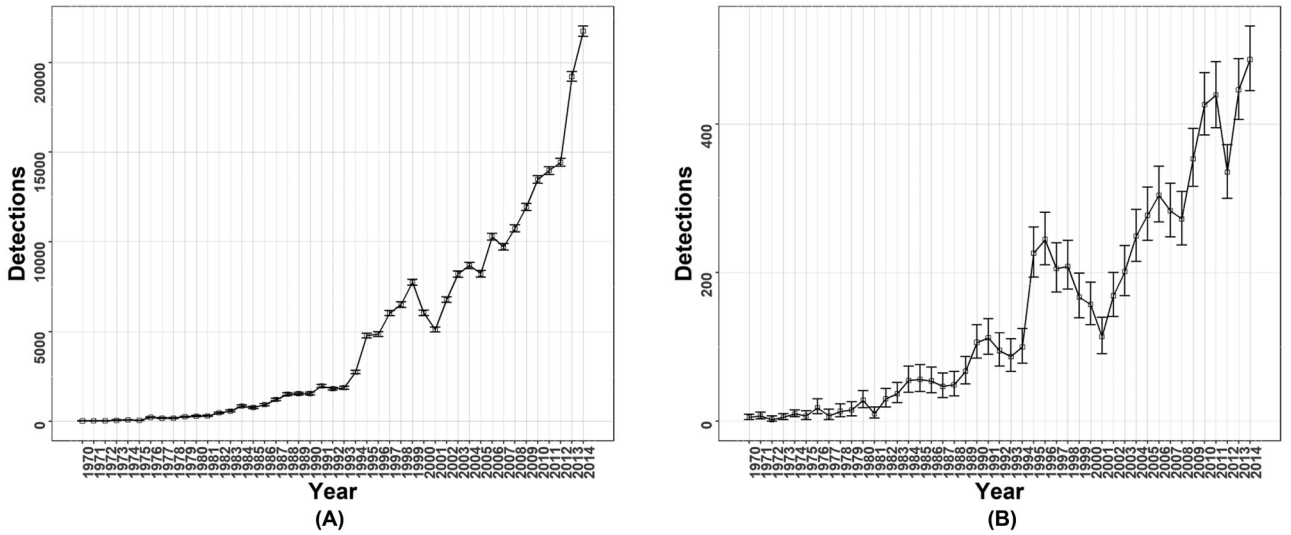

FIGURE 5 Goodness of fit for yearly detections for (A) Ringlet and (B) Duke of Burgundy. The squares represent the true values, while the error bars represent the 95% posterior credible interval (PCI) of the test statistics

during a year, even though butterflies obviously do not fly throughout the year. Therefore, the probability of detection at time p_t in our model can be interpreted instead as the product $p_0 d_t$, where d_t is the probability of butterflies of the species flying at time t and p_0 is the probability of detecting at least one butterfly of that species, conditional on the species flying at time t, with the latter usually considered as the detection probability.

Convergence has been checked by monitoring traceplots from single chains, which we have reported in the supporting information.

5.1 | Goodness of fit

To check the goodness of fit of the model, we have also performed posterior predictive checks using two test statistics: the number of yearly detections across all sites, $T_t^1(y) = \sum_{\substack{k_i=j \\ t_j=t}} y_i$ and the number of detections in a given region r, $T_r^2(y) = \sum_{\substack{k_i=j \\ s_j \in r}} y_i$. We have compared the true value of the statistic in each case with the 95% PCI of the posterior predictive distribution of the test statistic, $T(\tilde{y})$, where $\tilde{y}$ has distribution $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$. We note that draws $\tilde{y}_1, \ldots, \tilde{y}_l$ from $p(\tilde{y}|y)$ can easily be obtained by sampling at each step of the MCMC $\tilde{y} \sim p(y|\bar{\theta})$, where $\bar{\theta}$ is the value of the parameters at the each iteration. For the test statistics $T_r^2(y)$, we took as region the patches used for the spatial approximation.

The resulting goodness of fit plots for both datasets are shown in Figures 5 and 6. Figure 5 shows that the model properly accounts for the variation across years for both species. It is worth noting that we also ran the model with a constant detection probability across years and the PCIs of $T_t^1(y)$ did not always contain the true values, suggesting that the fit of the model is not as good in that case. We show plots of the goodness of fit for the model with constant detection probability in Figure 2 of the supporting information. The lack of fit of $T_r^2(y)$ (i.e. seen in Figure 6) is likely a suggestion that detection probability exhibits variation across space as well as time. However as commented earlier, we do not model variation of the detection probability across space since we already model spatial variation of the occupancy probability, and modeling the spatial variation of both quantities could lead to unidentifiability issues between the two. We note that using list length instead of relative list length causes a bias in the goodness of fit and leads to the number of detections in the north being consistently underestimated. The cause of the bias is that since fewer butterfly species inhabit the North of the UK, observers in the North are penalized with respect to the ones in the South as it is more difficult further north to detect a large number of species, and hence their capabilities are underestimated compared to observers in the South.

6 | POTENTIAL EXTENSIONS

We model temporal and spatial r.e. as additive independent effects, as shown in Equation (2). To allow for interaction between time and space, we can define a GP prior jointly over time and space in the following way. Formally, we introduce $S \times Y$ r.e., $\{c_{ys}\}_{y=1,\ldots,Y,s=1,\ldots,S}$, where c_{ys} is the r.e. for year y and site s, and assume a GP prior

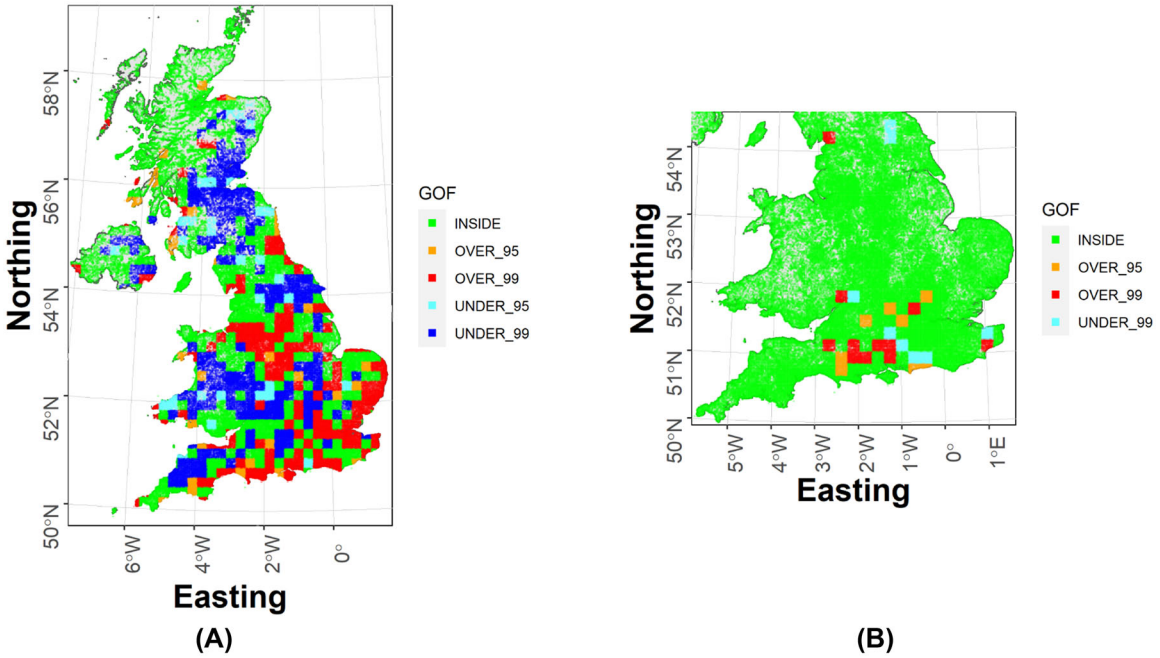

FIGURE 6 Goodness of fit for space detections for (A) Ringlet and (B) Duke of Burgundy. Different colors identify where the true statistic is inside the 95% posterior credible interval (PCI), above and below the 99% PCI, and between the 95% and 99% PCI. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

distribution with support points $(w_y, \mathbf{x}_s)_{y=1,\dots,Y,s=1,\dots,S}$, such that $(c_{11}, \dots, c_{SY}) \sim N(0, K)$, where $K((\omega_{y_1}, \mathbf{x}_{s_1}), (\omega_{y_2}, \mathbf{x}_{s_2}))$ depends on the distance between the time–space points $(\omega_{y_1}, \mathbf{x}_{s_1})$ and $(\omega_{y_2}, \mathbf{x}_{s_2})$. Similar approaches have been proposed. For example, Datta et al. (2016) proposed to use NNGP to assume nonseparable covariance matrices in a GP framework while obtaining scalable computations. We note that our additive modeling approach arises if $K((\omega_{y_1}, \mathbf{x}_{s_1}), (\omega_{y_2}, \mathbf{x}_{s_2})) = K_{l_T,\sigma_T}(\omega_{y_1}, \omega_{y_2}) + K_{l_S,\sigma_S}(\mathbf{x}_{s_1}, \mathbf{x}_{s_2})$. In the non-separable case, it is paramount to use approximations such as the ones described in Section 3.2, as the covariance matrix is of dimension $SY \times SY$. For example, in the non-separable case, the complexity of the CP method is $O(M^3 + SY)$, while the complexity of the ASoR is $O(M^3 + M^\star S^2 Y^2)$. Moreover, M should be chosen bigger in the nonseparable case as the grid is used to approximate time and space together.

In addition to the optimal mixing properties, another advantage of the PG scheme is that it allows efficient variable selection, as performed in Griffin et al. (2020), since the logistic regression equations for the detection and presence processes can be cast in the linear regression framework using the PG augmentation. Although not considered in this paper, which aims to introduce the new modeling framework, we note that this approach can be used to perform variable selection on the occupancy and detection probabilities if a number of covariates are available as potential predictors for either of the two processes. Finally, we note that the PG scheme is easily parallelisable with respect to the variables ω_i in Equation (4), which would bring further computational advantages for large datasets.

7 | DISCUSSION

We proposed a unifying Bayesian framework for modeling large occupancy datasets, while accounting for spatio-temporal autocorrelation and for the effect of covariates on the probabilities of occupancy and detection. We employed and developed a number of algorithms and approximations for fast inference, even for very large datasets, and we used simulation to assess the performance of our new models.

We compared two popular approximation methods, a low-rank approximation and a sparse approximation method, according to computational time and predictive power. We found that although the methods have very different theoretical biases, they tend to perform similarly in the context of occupancy modeling. We note that the NNGP approximation has been also considered within an occupancy model framework in a recent paper by Doser et al. (2022).

Our model incorporates both time and space, and the results for the two case studies are in accord with what is known for the species involved. The spatial maps of Figure 3 demonstrate how the distribution of each species

changes over time, similarly to what is shown in Dennis et al. (2017).

We have illustrated how goodness of fit can be routinely studied. It was interesting to note the differences between the magnitudes of detection probability for the two species, and this highlights the potential of using the model for further investigation of this poorly understood aspect of citizen-science occupancy modeling.

As with all models, several assumptions are made on how the probabilities of species' presence and the probability of detection vary across sites or years. The validity of results will depend on how realistic these assumptions are and the general appropriateness of the model for the data at hand.

ACKNOWLEDGMENTS

We are very grateful to all of the volunteers who have contributed to the Butterflies for the New Millennium project, which is run by Butterfly Conservation with support from Natural England. We thank Richard Fox, the AE and two anonymous referees for their insightful comments that greatly improved this paper. Byron Morgan was supported by a Leverhulme research fellowship. This work was partly funded by NERC grant NE/T010045/1.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are available from the corresponding author upon reasonable request.

ORCID

Alex Diana https://orcid.org/0000-0002-8130-2988
Eleni Matechou https://orcid.org/0000-0003-3626-844X
Byron John Treharne Morgan https://orcid.org/0000-0002-5465-8006

REFERENCES

Albert, J.H. & Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679.

August, T., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T. & Jepson, P. (2015) Emerging technologies for biological recording. *Biological Journal of the Linnean Society*, 115(3), 731–749.

Besag, J. & Kooperberg, C. (1995) On conditional and intrinsic autoregressions. *Biometrika*, 82(4), 733–746.

Blockeel, T.L., Bosanquet, S.D.S., Hill, M.O. & Preston, C.D. (2014) *Atlas of British & Irish bryophytes*. Newbury: Pisces Publications.

BRC (2022) Biological Records Centre Home Page. Available at: www.brc.ac.uk (Accessed 19th December 2022).

Broms, K.M., Hooten, M.B., Johnson, D.S., Altwegg, R. & Conquest, L.L. (2016) Dynamic occupancy models for explicit colonization processes. *Ecology*, 97, 194–204.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., et al. (2010) Global biodiversity: indicators of recent declines. *Science*, 328(5982), 1164–1168.

Choi, H.M. & Hobert, J.P. (2013) The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7, 2054–2064.

Clark, A.E. & Altwegg, R. (2019) Efficient Bayesian analysis of occupancy models with logit link functions. *Ecology and Evolution*, 9(2), 756–768.

Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E. (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514), 800–812.

Datta, A., Banerjee, S., Finley, A.O., Hamm, N.A. & Schaap, M. (2016) Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, 10(3), 1286–1316.

Dennis, E.B., Morgan, B.J.T., Freeman, S.N., Ridout, M.S., Brereton, T.M., Fox, R., et al. (2017) Efficient occupancy model-fitting for extensive citizen-science data. *PLoS ONE*, 12, e0174433, https://doi.org/10.1371/journal.pone.0174433

Department for Environment, Food and Rural Affairs, UK (2020) UK Biodiversity Indicators 2020.

Diana, A. (2022) *FastOccupancy package*. Available at: https://github.com/alexdiana1992/FastOccupancy (accessed July 5, 2021).

Dorazio, R.M. & Rodriguez, D.T. (2012) A Gibbs sampler for Bayesian analysis of site-occupancy data. *Methods in Ecology and Evolution*, 3(6), 1093–1098.

Doser, J.W., Leuenberger, W., Sillett, T.S., Hallworth, M.T. & Zipkin, E.F. (2022) Integrated community occupancy models: a framework to assess occurrence and biodiversity dynamics using multiple data sources. *Methods in Ecology and Evolution*, 13(4), 919–932.

Ellis, S., Bourn, N.A.D. & Bulman, C.R. (2012) *Landscape-scale conservation for butterflies and moths: lessons from the UK*. Wareham, Dorset: Butterfly Conservation.

Finley, A.O., Datta, A., Cook, B.D., Morton, D.C., Andersen, H.E. & Banerjee, S. (2019) Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2), 401–414.

Fiske, I. & Chandler, R.B. (2011) `unmarked`: an R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43(10), 1–23.

Fox, R., Brereton, T.M., Asher, J., August, T.A., Botham, M.S., Bourn, N.A.D., et al. (2015) *The State of UK's Butterflies 2015*. Wareham, Dorset: Butterfly Conservation and the Centre for Ecology & Hydrology.

Fox, R., Warren, M.S., Brereton, T.M., Roy, D.B. & Robinson, A. (2011) A new red list of British butterflies. *Insect Conservation and Diversity*, 4, 159–172.

GBIF.org (2022) GBIF Home Page. Available at: https://www.gbif.org [Accessed 19th December 2022].

Griffin, J.E., Matechou, E., Buxton, A.S., Bormpoudakis, D. & Griffiths, R.A. (2020) Modelling environmental DNA data: Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(2), 377–392.

Guélat, J. & Kéry, M. (2018) Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, 9(6), 1614–1625.

Hayhow, D.B., Eaton, M.A., Stanbury, A.J., Burns, F., Kirby, W.B., Bailey, N., et al. (2019) *The State of Nature*. The State of Nature Partnership.

Hepler, S.A. & Erhardt, R.J. (2021) A spatiotemporal model for multivariate occupancy data. *Environmetrics*, 32(2), e2657.

Hodges, J.S. & Reich, B.J. (2010) Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4), 325–334.

Holsclaw, T., Greene, A.M., Robertson, A.W., Smyth, P., et al. (2017) Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, 11(1), 393–426.

Isaac, N.J.B. & Pocock, M.J.O. (2015) Bias and information in biological records. *Biological Journal of the Linnean Society*, 115(3), 522–531.

Johnson, D.S., Conn, P.B., Hooten, M.B., Ray, J.C. & Pond, B.A. (2013) Spatial occupancy models for large datasets. *Ecology*, 94(4), 801–808.

Kéry, M. & Royle, J.A. (2021) *Applied hierarchical modeling in ecology*, Vol. 2. London: Academic Press.

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Haefliger, G., et al. (2010) Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, 24(5), 1388–1397.

Linderman, S.W., Johnson, M.J. & Adams, R.P. (2015) Dependent multinomial models made easy: stick breaking with the Pólya-Gamma augmentation. *arXiv:1506.05843*.

Liu, H., Ong, Y.-S., Shen, X. & Cai, J. (2020) When Gaussian process meets big data: a review of scalable GPS. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11), 4405–4423.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., & Hines, J.E. (2018) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*, 2nd edition, New York: Academic Press.

Mardia, K.V. (1988) Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2), 265–284.

Mason, S.C., Palmer, G., Fox, R., Gillings, S., Hill, J.K., Thomas, C.D. & Oliver, T.H. (2015) Geographical range margins of many taxonomic groups continue to shift polewards. *Biological Journal of the Linnean Society*, 115(3), 586–597.

Outhwaite, C.L., Chandler, R.E., Powney, G.D., Collen, B., Gregory, R.D., & Isaac, N.J.B. (2018) Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data. *Ecological Indicators*, 93, 333–343.

Pocock, M.J., Roy, H.E., Preston, C.D. & Roy, D.B. (2015) The Biological Records Centre: a pioneer of citizen science. *Biological Journal of the Linnean Society*, 115(3), 475–493.

Polson, N.G., Scott, J.G. & Windle, J. (2013) Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504), 1339–1349.

Randle, Z., Evans-Hill, L.J., Parsons, M.S., Tyner, A., Bourn, N.A.D., Davis, A.M., et al. (2019) *Atlas of Britain & Ireland's Larger Moths*. Newbury: Pisces Publications.

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

Reich, B.J., Hodges, J.S. & Zadnik, V. (2006) Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4), 1197–1206.

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical modeling and inference in ecology*. Amsterdam: Academic Press.

Rue, H. & Held, L. (2005) *Gaussian Markov random fields: theory and applications*. New York: Chapman and Hall.

Rue, H. & Tjelmeland, H. (2002) Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1), 31–49.

Rushing, C.S., Royle, J.A., Ziolkowski, D.J. & Pardieck, K.L. (2019) Modeling spatially and temporally complex range dynamics when detection is imperfect. *Scientific Reports*, 9(1), 1–9.

Smola, A.J. & Bartlett, P. (2001) Sparse greedy Gaussian process regression. *Advances in Neural Information Processing Systems*, 14, 619–625.

Strebel, N., Kéry, M., Guélat, J. & Sattler, T. (2022) Spatiotemporal modelling of abundance from multiple data sources in an integrated spatial distribution model. *Journal of Biogeography*, 49(3), 563–575.

Szabo, J.K., Vesk, P.A., Baxter, P.W.J. & Possingham, H.P. (2010) Regional avian species declines estimated from volunteer-collected long-term data using list length analysis. *Ecological Applications*, 20(8), 2157–2169.

van Strien, A.J., van Swaay, C.A. & Termaat, T. (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, 50(6), 1450–1458.

SUPPORTING INFORMATION

Web Appendices and Figures referenced in Sections 4 and 5 are available with this paper at the Biometrics website on Wiley Online Library. We have implemented our model in the package *FastOccupancy*, available on GitHub (Diana, 2022).

How to cite this article: Diana, A., Dennis, E.B., Matechou, E., & Morgan, B.J.T. (2023) Fast Bayesian inference for large occupancy datasets. *Biometrics*, 1–13. https://doi.org/10.1111/biom.13816