Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 641–657
https://doi.org/10.1093/jrsssa/qnae105
Advance access publication 5 October 2024
Discussion Paper

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society
A

Efficient statistical inference methods for assessing changes in species' populations using citizen science data

Emily B. Dennis[1,2], Alex Diana[3], Eleni Matechou[2] and Byron J.T. Morgan[2]

[1]Butterfly Conservation, Manor Yard, East Lulworth, Dorset BH20 5QP, UK
[2]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, UK
[3]School of Mathematics, Statistics and Actuarial Science, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK

Address for correspondence: Byron J.T. Morgan, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, UK. Email: b.j.t.morgan@kent.ac.uk

Read before The Royal Statistical Society at the Discussion Meeting on the 'Analysis of citizen science data' held at the Society's 2024 annual conference in Brighton on Tuesday, 3 September 2024, the President, Dr Andrew Garrett, in the Chair.

Abstract

The global decline of biodiversity, driven by habitat degradation and climate breakdown, is a significant concern. Accurate measures of change are crucial to provide reliable evidence of species' population changes. Meanwhile citizen science data have witnessed a remarkable expansion in both quantity and sources and serve as the foundation for assessing species' status. The growing data reservoir presents opportunities for novel and improved inference but often comes with computational costs: computational efficiency is paramount, especially as regular analysis updates are necessary. Building upon recent research, we present illustrations of computationally efficient methods for fitting new models, applied to three major citizen science data sets for butterflies. We extend a method for modelling abundance changes of seasonal organisms, firstly to accommodate multiple years of count data efficiently, and secondly for application to counts from a snapshot mass-participation survey. We also present a variational inference approach for fitting occupancy models efficiently to opportunistic citizen science data. The continuous growth of citizen science data offers unprecedented opportunities to enhance our understanding of how species respond to anthropogenic pressures. Efficient techniques in fitting new models are vital for accurately assessing species' status, supporting policy-making, setting measurable targets, and enabling effective conservation efforts.

Keywords: biodiversity change, citizen science, concentrated likelihood, generalized abundance index, occupancy models, variational Bayes

1 Introduction

An existential crisis of our time is the alarming decrease of biodiversity, due to anthropogenic factors such as climate breakdown and loss of habitat. Producing robust measures of change is vital for evaluating species' status, understanding rates of change, and monitoring responses to pressures, as well as progress of conservation actions, such as towards biodiversity targets (Butchart et al., 2010). Both data and appropriate statistical models are critical for measuring biodiversity change. Citizen (or community) science (CS) data, where information is gathered by voluntary participants, are increasingly used for this purpose (Chandler et al., 2017; Silvertown, 2009).

CS data from systematic, designed surveys and monitoring schemes, collected by skilled and committed volunteers, have long been used to produce estimates of species' status, whereas

Received: November 23, 2023. **Revised:** June 25, 2024. **Accepted:** August 9, 2024

observations of species from less structured, opportunistic or mass-participation sampling, attracting contributions from wider society, are increasingly gathered (Pocock et al., 2017), broadening the scope to measure changes in populations from extensive geographic areas and for a variety of taxa. The many sources of CS data provide vast opportunities for biodiversity monitoring, but require suitable analytical approaches, for example to deal with sources of bias (Isaac & Pocock, 2015). Johnston et al. (2023) outlined a diversity of challenges for biodiversity monitoring using CS data, relating to dealing with observer behaviour, data structures, statistical models, and communication. This paper focuses upon one such key challenge, which is in the computational cost of analysing CS data, particularly with increasing data and complex models, leading to the 'necessity to identify and develop suitable modifications to improve computational efficiency and scalability, adapting traditional (and developing new) methods to big data' (McCrea et al., 2023).

In this paper, we present illustrations of computationally efficient methods for analysing CS data. Our work is motivated by applications to data for British butterflies, to model changes in both abundance and distribution, but the methods and overall need for efficiency apply to CS data for a range of taxa and locations. The global decline of insects has been widely reported, particularly in western Europe and North America (Wagner et al., 2021), yet there is an ongoing need for robust data and rigorous analysis methods (Didham et al., 2020; Thomas et al., 2019). Many taxa and geographic regions are lacking in sufficient data to appropriately assess trends, but there is a wealth of data gathered on butterflies in the UK. Butterflies are the most comprehensively monitored invertebrate taxon and their population status provides a valuable indicator for changes in biodiversity as they respond rapidly to environmental change.

We demonstrate efficient analysis methods for three sources of CS data for UK butterflies: (i) The UK Butterfly Monitoring Scheme (UKBMS) began in 1976 and is one of the longest-running insect monitoring schemes in the world. Counts of butterflies are made each year by recorders walking transects for six months from 1 April, according to a strict protocol (Pollard & Yates, 1993). (ii) The annual Big Butterfly Count (BBC) launched in 2010 and is a UK-wide mass-participation CS project in which members of the public record how many individual butterflies are seen for 15 min periods during 23–24 days in late July and early August. It is the largest CS project of its kind in the world, with almost 95,000 participants in 2023. (iii) The Butterflies for the New Millenium (BNM) data base collates opportunistic records of where butterflies are seen, and consists of over 14 million records (Fox et al., 2023). These data are unstandardized, and more than 80 similar recording schemes exist for various taxa in the UK (BRC, 2022; Pocock et al., 2015) and are typically used for describing changes in species' distributions.

Analyses of UKBMS and BNM data feed in to regular reporting such as the 'State of the UK's Butterflies' (Fox et al., 2023), Red List assessments of extinction risk (Fox et al., 2022), government biodiversity indicators (JNCC, 2022), and multitaxa outputs such as the State of Nature (Burns et al., 2023), all of which contribute to providing robust evidence for conservation, policy development, and the wider state of biodiversity.

During a time of biodiversity change, frequent analysis updates are essential for monitoring species' populations, and for understanding responses to both pressures and conservation action. Efficient computational methods are therefore vital, especially in CS surveys, which typically involve data sets that are large and continuously growing in size.

We are motivated by the need to measure population changes, by modelling abundance and distribution using the data sets mentioned above. In this paper, we summarize our recent research, describe new methods and applications, and suggest avenues for future development. We start with models for abundance, where we demonstrate how in certain circumstances it is possible to greatly reduce the dimensionality of the effective model parameter space, considerably reducing computation time. We then consider opportunistic data, where we show how variational inference (VI) can provide an efficient alternative to Markov chain Monte Carlo (MCMC) for Bayesian inference when fitting occupancy models.

2 Generalized abundance index

Models for abundance data, such as from the UKBMS, need to account for two key features: seasonality and missing data. Seasonality results in counts that vary throughout the season according to the emergence of one or more generations of adult insects over time in any year. Missing data arise in two ways. Firstly, some visits are missed by volunteers. UKBMS transects are sampled by committed and skilled volunteers, walking transects weekly across six months of the year, but inevitably some weeks are missed: Dennis et al. (2013) estimated that this occurs for roughly 8 of 26 weeks of the transect season. Secondly, there is turnover in transects sampled each year; for example transects may cease to be monitored, and new transects continue to be introduced to the scheme.

A variety of methods have produced analyses of these data. Dennis et al. (2013) presented a method based on using generalized additive models (GAMs) to model seasonality, followed by a generalized linear model (GLM) analysis to account for the annual variation in the transects sampled. This approach produced annual indices of abundance for each species, which can be used to estimate time trends in abundance. Alternatively, Matechou et al. (2014) proposed a stopover model approach, which enables estimation of within-season survival probabilities. Dennis et al. (2016) then proposed the generalized abundance index (GAI) approach, which provides a framework for both of these types of models, as well as a third alternative, which describes seasonality parametrically using an appropriate mixture of distributions, such as Normal (see Dennis et al., 2022, for examples).

In outline, for a given species, in any year we suppose that counts are obtained at S sites, each visited on at most V occasions. Each count, $y_{s,v}$, for site s and visit v is regarded as the realization of a random variable, such as Poisson, with expectation $\lambda_{s,v} = N_s a_{s,v}$, where the likelihood then takes the form

$$L(\mathbf{N}, \theta; \mathbf{y}) \propto \prod_{s=1}^{S} \prod_{v=1}^{V} \exp(-N_s a_{s,v})(N_s a_{s,v})^{y_{s,v}}. \tag{1}$$

Here the N_s are parameters describing site abundance, and $a_{s,v}$ denotes a function determined by parameters, $\boldsymbol{\theta}$, describing seasonal variation. Here and later the product over visits only includes terms corresponding to when visits are made. Background is provided by Dennis et al. (2016), who derive a concentrated-likelihood approach that substantially reduces the number of parameters to estimate via maximum likelihood. Briefly, using maximum likelihood, the site parameters can be estimated by

$$N_s = \frac{y_{s,.}}{a_{s,.}}, \tag{2}$$

where we use the dot notation to indicate summation corresponding to visited sites. The total observed count for each site is therefore rescaled to account for incomplete sampling within the season. Substitution of the estimates of equation (2) into equation (1) results in a Poisson likelihood which can be maximized with respect to only the parameters $\boldsymbol{\theta}$. Counts can be expected to be overdispersed relative to the Poisson and/or contain additional zeros, for example due to small counts at the ends of the season. Alternative discrete distributions such as negative binomial and zero-inflated Poisson, respectively, may then be appropriate, as described in Dennis et al. (2016).

3 Two adaptations of the GAI

We now provide two new adaptations of the GAI. Firstly, the model is extended to include appropriate site and year effects using an *annual model* integrated within the GAI, combined with the use of concentrated likelihood. Secondly, we adapt the GAI to improve the analysis of BBC data, a ‘snapshot’, mass-participation CS scheme.

3.1 An extended GAI with site and time effects

The basic GAI, described above, is a static model, where data for each year are analysed separately. However, to deal with the turnover of sites surveyed in the UKBMS each year, the Poisson GLM stage of Dennis et al. (2013), based on the *annual model* of ter Braak et al. (1994), is typically used. Here the abundance estimates $\{\hat{N}_{s,r}\}$, for sites s and now year r, are the dependent values for an additive model with year and site effects and the estimated year effects are then used to form abundance indices. Use of log-linear regression in this way is widely used in ecology, for example using TRIM (Trends and Indices for Monitoring Data, Bogaart et al., 2020; van Strien et al., 2004). This two-stage GAI approach is now a valuable tool for analysing seasonal count data in the UK (Fox et al., 2021, 2023; JNCC, 2022; UKBMS, 2023) and beyond (Schmucki et al., 2016; Van Swaay et al., 2020). In practice, the proportion of the species' flight period surveyed for a given site and year is typically included as a weighting in the GLM stage, such that better sampled sites have a higher contribution to the estimated species' abundance indices (Brereton et al., 2018).

A drawback of the two-stage GAI approach is the need to bootstrap to account for variance propagation between the two model stages. Bootstrapping is time consuming for large data sets, and can also be problematic when resampling small data sets, for example for rare species. We now describe an alternative approach which effectively incorporates the annual model within the GAI, extending the model to consider all years at once, and thus solving the issue of variance propagation.

In brief, we expand the Poisson likelihood of equation (1) to counts $y_{s,v,r}$ where r denotes one of Y successive years, with a corresponding expansion to $a_{s,v,r}$. We then incorporate the expression for the annual model $N_{s,r} = e^{\alpha_s+\beta_r}$ - see ter Braak et al. (1994) - which results in the following expression for the log-likelihood, ignoring an additive constant.

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \sum_r \sum_s \sum_v \{ - e^{(\alpha_s+\beta_r)} a_{s,v,r} + y_{s,v,r}(\alpha_s + \beta_r) + y_{s,v,r} \log(a_{s,v,r}) \}. \tag{3}$$

Here $\{\alpha_s\}$ and $\{\beta_r\}$ are respectively site and year effects to be estimated. As above, we use a concentrated likelihood approach to form maximum-likelihood parameter estimates efficiently, by concentrating out the parameters $\boldsymbol{\alpha}$, resulting in the log-likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \sum_r \sum_s \sum_v \left[-\frac{y_{s,.,.} e^{\beta_r} a_{s,v,r}}{\sum_j e^{\beta_j} a_{s,.,j}} + y_{s,v,r} \{\beta_r - \log(\sum_j e^{\beta_j} a_{s,.,j})\} \right], \tag{4}$$

which we can maximize efficiently with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Full details are given in Section S1 of the online supplementary material.

The aim here is to produce a new model which can be fitted efficiently and produce estimates of uncertainty. From the above modelling, estimates of error result from inverting the estimated Hessian at the maximum-likelihood estimates, which is far more efficient than bootstrapping. Profile confidence intervals are also easily obtained.

We illustrate this extended GAI approach with application to UKBMS data spanning 1976–2022 for two species: Chalk Hill Blue *Polyommatus coridon* and Gatekeeper *Pyronia tithonus*. Chalk Hill Blue is a species confined to calcareous grassland in southern England, with UKBMS counts of around 900,000 individuals from ~ 460 sites, whereas Gatekeeper is a widespread butterfly species across southern Britain, with UKBMS counts from more than 4,300 sites, counting over 3.3 million individuals.

Options for describing seasonal variation (parameters associated with $\boldsymbol{\theta}$) are flexible, as in the original GAI, but here we use a Normal distribution, since both species are univoltine (one generation per year), and thus parameters $\{\mu_r\}$ and $\{\sigma_r\}$ are estimated for each year, which describe the mean flight dates and length of the flight period, respectively (see Figure S1, online supplementary material where we see the tendency for earlier emergence over time). Flight periods are thus assumed to be fixed over sites within each year.

Figure 1 shows the estimates of $\{\beta_r\}$ from the extended GAI, with 95% confidence intervals (CI) produced from the estimated Hessian at the maximum-likelihood estimates. Error estimates were scaled to account for overdispersion in the Poisson GAI formulation.

In order to estimate a linear trend over time, the parameters $\boldsymbol{\beta}$ may also be expressed by a linear form, where we set $\beta_r = \gamma r, \forall\ r,$ in equation (4) and maximize the resulting log likelihood with respect to the parameters $\boldsymbol{\theta}$ and γ. The intercept parameter for a linear regression on year cannot be estimated due to confounding. This is also the case for the annual model (McCullagh & Nelder, 1989, p. 63, van Strien et al., 2004), where we therefore set $\beta_1 = 0$. Trend lines based on the extended GAI with the linear form are shown in Figure 1(i), and due to the lack of intercept, for ease of comparison we equate the ordinate at the middle year to that from a posthoc linear regression (also shown).

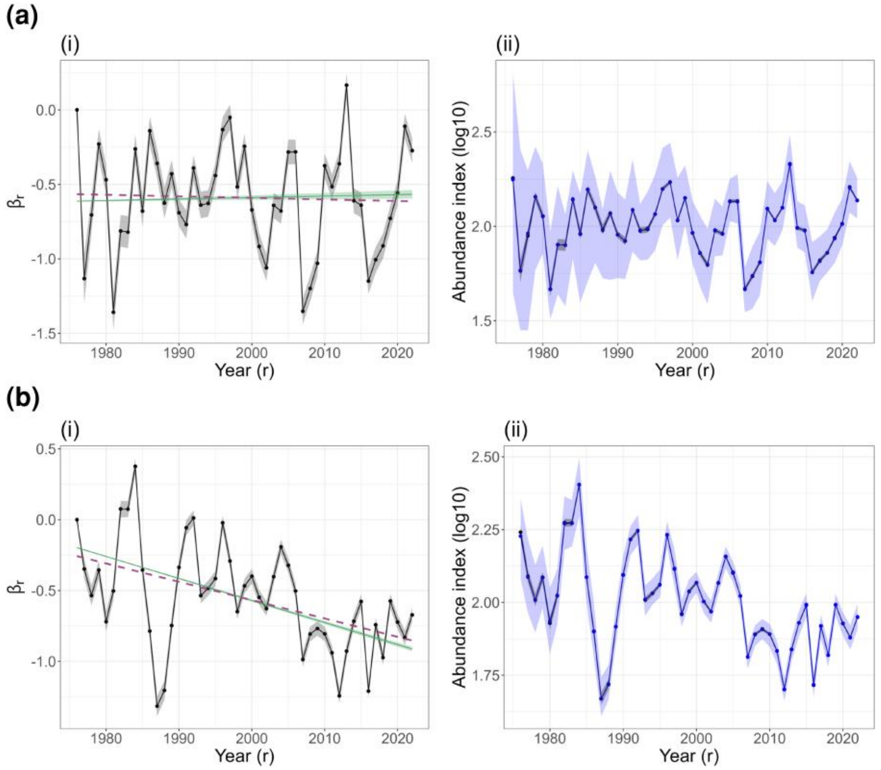

Figure 1. Results from applying the extended (generalized abundance index) GAI to two species: Chalk Hill Blue a) and Gatekeeper b). Plots (i) show the estimates of $\{\beta_r\}$ (black) with 95% confidence intervals (CI). Linear trend estimates for $\{\beta_r\}$ from the extended GAI are shown in green (with 95% CI). The trend line from a posthoc linear regression through the estimates of $\{\beta_r\}$ is also shown (purple dashed line). Plots (ii) compare abundance indices from the extended GAI (black) and two-stage GAI (blue), with 95% CI produced from the estimated Hessian at the maximum-likelihood estimates and from nonparametric bootstrapping respectively. Abundance indices correspond to $\{\beta_r\}$ converted to the $\log_{10}$ scale with a mean of 2, as is standard practice for UK Butterfly Monitoring Scheme indices.

Abundance indices produced from the extended GAI are virtually indistinguishable from indices produced from the two-stage GAI (Figure 1(ii)). The two-stage GAI was based on the implementation used for annual analyses of UKBMS data, where a GAM is used for the first stage. Despite the similarity in the indices, the 95% CI produced from bootstrapping the two-stage GAM are much wider than those estimated from the Hessian for the extended GAI. This difference is due to differences between the models, rather than due to differences in the methods of error estimation. For example the extended GAI does not make a distributional assumption (Poisson) about the site abundance estimates, **N**, as is the case in the GLM of two-stage GAI. The standard errors from the extended GAI may also be underestimated if the data are overdispersed relative to the Poisson distribution assumed, which could be explored in future work using alternative discrete distributions.

The implementation of the two-stage GAI includes a weighting in the GLM, accounting for the proportion of the flight period sampled, whereas the extended GAI does not. Interestingly, the close resemblance of the two indices in Figure 1(ii) implies that the new extended GAI effectively has this weighting built-in, by modelling the counts directly so that better-sampled sites have a greater contribution to the likelihood.

The extended model took 42 min to run for the Chalk Hill Blue, whereas the nonparametric bootstrap (with 1,000 replicates) for the two-stage GAI took 98 min. For the more widespread Gatekeeper, the extended GAI took 1.3 hr, whereas the two-stage GAI bootstrap took approximately 7 hr, but based on only 200 replicates. Analyses were parallelized across four CPUs on a computer equipped with Intel Core i7-8,700@3.2 GHz with 32 GB of RAM. Clearly, there are substantial time savings to be had for the 59 UK butterfly species, and potentially far more for more numerous taxa. We have seen that the advantages of the extended GAI are not confined to efficiency, but may also involve greater parameter precision provided the model and distributional assumptions are appropriate. In the Discussion, we outline the potential contributions of future work on the extended GAI.

3.2 Analysis of big butterfly count data

The Big Butterfly Count (BBC; https://bigbutterflycount.org) is an annual survey of common butterfly species which encourages wide participation particularly from members of the general public. The BBC has a high media profile and attracts a large number of participants, many of whom may have limited or no prior experience of biodiversity monitoring. The sampling protocol is minimal: participants simply count numbers of individuals seen of widespread butterfly (and day-flying moth) species for 15 min from any location.

The BBC generates a large amount of data for the UK's widespread butterfly species - more than 11 million butterflies have been counted since 2011- but to date few analyses have been undertaken and annual reporting of BBC results is based on simple comparisons with respect to the previous year only. Using data for 2011–2014, Dennis, Morgan, Brereton, et al. (2017) demonstrated that estimates of change in abundance from BBC were comparable to those estimated from standardized monitoring (UKBMS data), but that the short snapshot sampling period of three weeks results in bias caused by the inter-annual variation in species' flight periods.

This is demonstrated in Figure 2a which shows the estimated flight period for Marbled White *Melanargia galathea* for four example years. This species is a univoltine (single-generation), summer-flying species, which is therefore likely to be particularly susceptible to phenological bias (in the estimation of timing) with respect to the BBC sampling period. Flight periods were estimated from UKBMS data using the GAI with a spline formulation for $\{a_v\}$, which is fixed across sites within a given year. The timing of peak emergence varies year-to-year, and thus the proportion of the flight period sampled by the BBC varies annually (see Figure 2b). For example in 2011 the BBC only captures the tail end of the flight period, whereas in 2012 more than half of the flight period is captured.

Here we describe a modification of the GAI for producing abundance indices and trends from BBC data. The GAI was developed for standardized monitoring data, typically collected along transects, where sites are clearly defined and revisited many times within and across years. Mass-participation CS data such as from the BBC does not have such a structure; it consists of many locations, which often only have one count undertaken. Hence we define a BBC 'site' to be a 1km × 1km square and pool BBC data to this spatial scale. BBC sites may then have counts across multiple days within the BBC sampling period, and potentially multiple entries (15 min counts) on a given day.

To accommodate these multiple entries, the GAI likelihood for a given year now takes the form

$$L \propto \prod_{m=1}^{M}\prod_{v=1}^{V} \exp(-N_m a_{m,v}\kappa_{m,v})(N_m a_{m,v})^{y_{m,v,\cdot}}, \tag{5}$$

where $\kappa_{m,v}$ is the number of entries for site m and visit v, and $y_{m,v,\cdot} = \sum_{\kappa} y_{m,v,\kappa}$ is the sum of the counts over those entries.

The number of sites in a mass-participation CS dataset such as BBC is high, and we adopt a concentrated-likelihood approach to reduce the functional size of the parameter space, as in the previous section, where equation (2) becomes

$$N_m = \frac{y_{m,\cdot,\cdot}}{\sum_{v=1}^{V} a_{m,v}\kappa_{m,v}} \tag{6}$$

The BBC sampling period represents just a snapshot of most species' flight periods, therefore accurate flight period estimation from applying the GAI to BBC data is not possible for all species. Instead, for each species and year, we produce estimates of $\{N_m\}$ for each BBC site

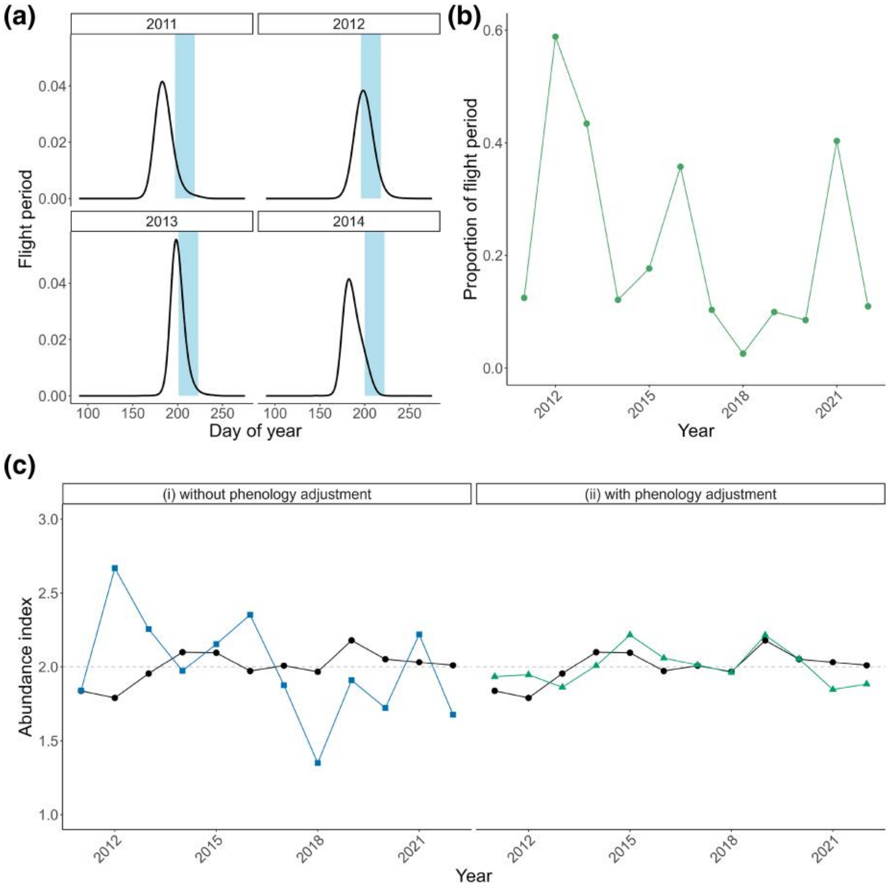

Figure 2. Demonstration of the phenology adjustment approach for the Marbled White butterfly: a) flight period curves estimated from UK Butterfly Monitoring Scheme (UKBMS) data for 4 yr. The blue shaded areas represents the Big Butterfly Count (BBC) sampling period each year. b) the proportion of the Marbled White flight period covered by the BBC sampling period each year. c) relative abundance indices produced from the generalized abundance index applied to UKBMS data (black), from BBC data without phenology adjustment (i, blue squares), and from BBC data with phenology adjustment (ii, green triangles). Indices are on the $\log_{10}$ scale with a mean value of 2 (indicated by the horizontal dashed lines).

(1km × 1km square) from equation (6) using daily flight period estimates of $\{a_v\}$ from the GAI applied to UKBMS data, where flight periods are assumed to be the same across sites. The average of $\{N_m\}$ provides an overall measure of BBC abundance per year and species (Dennis et al., 2016).

Figure 2c demonstrates the benefit of the phenology adjustment approach for the Marbled White butterfly. Fluctuations in the BBC abundance index produced without phenology adjustment (Figure 2c(i)) largely reflect year-to-year variation in the proportion of the flight period captured by the BBC each year, whereas adjusting for phenology produces an index that shows a pattern more similar to the UKBMS abundance index (Figure 2c(ii))). Applying the phenology adjustment approach is less influential for a multivoltine species such as the Comma *Polygonia c-album* (Figure S2, online supplementary material), for which the BBC sampling period covers a smaller, and less variable, proportion of the overall flight period each year.

Dennis et al. (2024) describe the new phenology adjustment approach for snapshot citizen science data in full, including the use of bootstrapping to estimate uncertainty. The approach is applied to BBC data for 17 species and explored further via simulation. The method enables data from snapshot CS schemes such as BBC to contribute to monitoring biodiversity. BBC results receive high-profile media coverage and are already beginning to reflect the advantages of the new analyses outlined here.

4 Efficient occupancy model fitting

Occupancy models (Altwegg & Nichols, 2019; MacKenzie et al., 2018) are widely used for inferring species distributions from presence/absence data at multiple sites, across a single or multiple seasons. They have been employed to extract meaningful species distribution trends from opportunistic citizen science data, addressing challenges associated with nonsystematic sampling and variable observation effort (Isaac et al., 2014; Kéry et al., 2010). However, as the size of the corresponding presence/absence data increases, occupancy models can be computationally demanding, especially in a Bayesian framework.

We start by defining the standard occupancy model that we fit in this section. We assume that there are n sampling units, where each sampling unit corresponds to a site in a specific year. In each sampling unit, we have a set of observations y_i, equal to 1/0 if the species was detected/not detected. The sampling unit to which observation i belongs is denoted by k_i. The hierarchical model representation is

$$\begin{cases} \text{logit}(\psi_j) = X_j^{\psi} \beta_{\psi} \\ z_j \sim \text{Be}(\psi_j) \\ \text{logit}(p_i) = X_i^{p} \beta_p \\ y_i \sim \begin{cases} \text{Be}(p_i) & \text{if } z_{k_i} = 1 \\ 0 & \text{if } z_{k_i} = 0 \end{cases} \end{cases} \tag{7}$$

where z_j is the occupancy state of sampling unit j; β_{ψ} are the covariate coefficients of the occupancy probability ψ; β_p are the covariate coefficients of the detection probability p; X_j^{ψ} and X_i^{p} are the available covariates for sampling unit j and observation i, respectively.

Similar versions of this model have been considered in Diana et al. (2023) and in Doser et al. (2023), who employed Gaussian processes to account for spatio-temporal autocorrelation in the occupancy probability. Bayesian inference for these occupancy models can be easily performed using Markov chain Monte Carlo (MCMC) (see Diana et al., 2023; Doser et al., 2023, who employed a Pólya-Gamma sampling scheme for logistic regression models Polson et al., 2013). In this MCMC framework, the z terms from equation (7), indicating species presence/absence at each site, are treated as latent variables and hence inferred and updated, typically at each MCMC iteration. Inferring the latent variables z allows us to easily write the complete data likelihood—see for example King (2014) and Newman et al. (2023)—for the observations in y, as shown in equation (7). However, when the number of sites is large, this leads to a computationally intensive MCMC, even when efficient model-fitting approaches, such as the Pólya-Gamma scheme, are employed for updating the model parameters. In addition when the complete data likelihood is used there may be correlations between estimators which can also slow down MCMC—see Newman et al. (2023) and Borowska and King (2022), where a subset of latent states are not treated as auxiliary variables but are integrated out numerically to reduce the correlation between latent variables. Therefore, MCMC-based inference can be prohibitively slow, which limits its application to large data sets such as CS data. For example, for the data sets considered in Diana et al. (2023), obtaining acceptable effective sample sizes from the posterior distributions of all parameters requires running times of around 19 hr (on an Intel Core i7-10610U@1.8 GHz). An obvious alternative is to use classical inference to fit occupancy models since it can be much faster (see for example the approach of Dennis, Morgan, Freeman, et al., 2017). In this case, the likelihood function is written by marginalizing over the z variables and hence the observed data likelihood is used for inference. Expressions for the complete and observed data likelihood for occupancy models are given in Section S2 of the online supplementary material. However, quantifying uncertainty around functions of parameters can sometimes be computationally intensive, relying on bootstrap methods when closed form expressions of variances are not available. Bayesian inference also offers the potential to more readily account for spatio-temporal autocorrelation in the occupancy probabilities.

Variational inference (VI) has been proposed as an alternative tool to overcome the computational issues of MCMC (Jordan et al., 1999). VI is traditionally faster than MCMC-based approaches because it transforms the problem from sampling (from a posterior distribution) to optimization. Therefore, VI combines the speed of classical inference, with the interpretability of Bayesian inference. However, while MCMC-based inference always recovers the true posterior (given enough MCMC iterations), in VI the true posterior distribution $p(\theta \mid y)$ is approximated using an appropriate flexible family of distributions, which is called the variational family. We denote the variational distributions by $q_\lambda(\theta)$, where λ is the set of variational parameters, the observed data likelihood by $p(y \mid \theta)$ and the prior distribution by $p(\theta)$. The parameters λ corresponding to the optimal variational distribution can be found by minimizing the Kullback–Leibler (KL) divergence between the true posterior distribution $p(\theta \mid y)$ and the variational distribution $q_\lambda(\theta)$. It can be proved that this is equivalent to finding the $q_\lambda(\theta)$ that minimizes the quantity

$$\mathbf{E}_{\theta \sim q_\lambda(\theta)}\left[\log p(y, \theta) - \log q_\lambda(\theta)\right] \tag{8}$$

which is known as the Evidence Lower BOund (ELBO), and forms the basis of VI inference. We note that $\log p(y, \theta) = \log \{p(y \mid \theta)p(\theta)\}$ and more details on VI can be found in Blei et al. (2017).

In VI, it is common to assume that parameters are a-posteriori independent, which is equivalent to assuming a variational family of the form $q(\theta) = \prod_j q_j(\theta_j)$ (the *mean-field assumption*) since this assumption considerably simplifies the inference. However, if the assumption is not valid, then posterior variance is underestimated (Wang & Titterington, 2005). In the case of occupancy models, occupancy and detection probability are independent conditionally on z, but not independent of z. Hence, assuming (according to the mean field assumption) that they are independent of z implies that ψ and p are independent of each other a-posteriori, which clearly does not hold. If the dependence structure of the model is ignored, then it leads to underestimation of the posterior variance of the occupancy and detection probability parameters, (β^ψ, β^p) (Clark et al., 2016).

However, if the observed data likelihood is used for inference, instead of the complete data likelihood, with the latter being common practice in a Bayesian framework, then we do not need to assume that (β^ψ, β^p) are independent of z, or of each other. In the observed data-likelihood case, we assume, as is standard in VI (Titsias & Lázaro-Gredilla, 2014), that $q_\lambda(\theta)$ is a multivariate normal distribution, that is, if $\theta = (\beta^\psi, \beta^p)$ are the model parameters, we assume $\theta \sim q_\lambda(\theta) = N(\mu, \Sigma)$, where the variational parameters are $\lambda = (\mu, C)$, with μ the mean of the variational distribution and C the Cholesky factor of the covariance matrix Σ.

We perform inference using stochastic gradient descent. Computing the gradient of equation (8) with respect to λ is complicated by the fact that λ itself appears in the expectation. To overcome this issue, we use the reparameterization trick (Rezende et al., 2014), which is a cornerstone of variational inference as it allows us to easily obtain these types of gradients. More details on the inference are presented in Section S2 of the online supplementary material.

To investigate the efficacy of our novel VI approach, we have performed a small preliminary simulation study to assess the coverage of Bayesian credible intervals generated using our procedure. Across the simulations, we varied the number of sites and the average occupancy and detection probability. We have chosen n, the number of sites, to be 500, 1,000, and 2,000 and ψ and p to be 0.25, 0.5, and 0.75. For each simulation, we ran 500 replications. The coverage was computed across 4 covariate coefficients for occupancy and 4 covariate coefficients for detection, with the coefficients randomly set to be either −1 or 1. Results are reported in Table 1 where it is seen that the 95% posterior credible intervals have the nominal coverage.

We also analysed the dataset of Ringlet butterflies collated through the Butterflies for the New Millennium (BNM) recording scheme run by Butterfly Conservation, using records collected between 1970 and 2014, which is also used in Diana et al. (2023). The data set consists of > 2 million records from ∼ 140,000 unique 1 km^2 (defined as sites), of which > 218,000 detections of Ringlet have been made, and nondetections were produced using observations of other butterfly species (Kéry et al., 2010). In this case, we do not account for spatial autocorrelation, but we model year as a factor variable and assume independence between sites, a point which we discuss in the next section. We also use relative list length, obtained by dividing the list length, which is the number of species recorded for a given site/date, divided by the maximum recorded list length

Table 1. Coverage of the 95% Bayesian credible intervals generated using variational Bayes, for varying n, the number of sites, and values of ψ and p

n / $\psi = p$	0.25	0.5	0.75
500	0.950	0.949	0.961
1,000	0.961	0.953	0.952
2,000	0.951	0.948	0.951

in a neighbouring square area of 50x50 km, as a covariate for detection probability and model the effect of the day of the year using a second-order polynomial of Julian date.

In this case, convergence was determined by assessing when parameter updates were smaller than a prespecified tolerance, which required approximately 1 hr of computation time (results were again obtained on an Intel Core i7-10610U@1.80 GHz). In Figure 3, we present the posterior distribution of the occupancy index, which is the year-specific occupancy probability, here assumed the same for all sites. The increase in precision over time reflects the growth of the underlying opportunistic CS data set, which has shown ongoing expansion, particularly since the mid-1990s (Fox et al., 2023).

The occupancy probabilities show similarities with results presented in Diana et al. (2023), in particular an increasing trend in recent years reflecting an expansion in the range of Ringlet in the UK and similar drops in the species' prevalence in mid-1970s and early and late-1990s are identified. However, in this case, the pattern is less smooth, as expected, since the effect of year is not constrained in any way. Furthermore, a direct comparison cannot be made, since Diana et al. (2023) fit a more complex model accounting for spatio-temporal correlation, and plot an occupancy index rather than the annual occupancy probabilities.

In Figure 4, we present the posterior distribution of the detection probability throughout the year which can be interpreted as an estimate of species' flying time. Unlike in Diana et al. (2023), in this example detection probability does not vary with year, and thus represents an average for 1970–2014. We also note here that we do not account for spatial effects on the probability of detection. However, the species' flight patterns are likely to vary in space, and hence the curve in Figure 4 represents a mixture of spatially varying curves. Additionally, we note that the peak of detection probability in this case is lower than that obtained by Diana et al. (2023), which is due to the different model structure for detection and occupancy probabilities, as discussed above. The 95% posterior credible interval of the coefficient of relative list length is (0.881, 0.883).

We have shown that using a VI framework to fit occupancy models efficiently to CS data shows promising results, with various avenues for future development-some of which we discuss in the next section.

5 Discussion

CS data have an important rôle in the future of biodiversity monitoring, but the ongoing growth of such data requires novel statistical models along with efficient inference methods and available computer code. In this paper, we have built upon recent work and demonstrated that concentrated likelihoods can greatly reduce the dimensionality of the effective model parameter space and that variational inference can substantially reduce computational time for large CS data sets. The approaches proposed in this paper address computational problems, but also lead to model developments for CS data.

The GAI is efficient due to the use of concentrated likelihood (Dennis et al., 2016). There is a broad analogy here with the efficiency that can result from adopting a hidden Markov model, when appropriate, and forming a forward algorithm for likelihood construction; see for example Cowen et al. (2017). We have shown concentrated likelihood to be useful in the new extension of the GAI to incorporate the annual model which conveniently allows for formal variance propagation, negating the need for time-consuming bootstrapping. A similar approach is given in Bravington et al. (2021), where the two stages for density surface models, one involving a detection probability and the other a GAM, are combined into one.

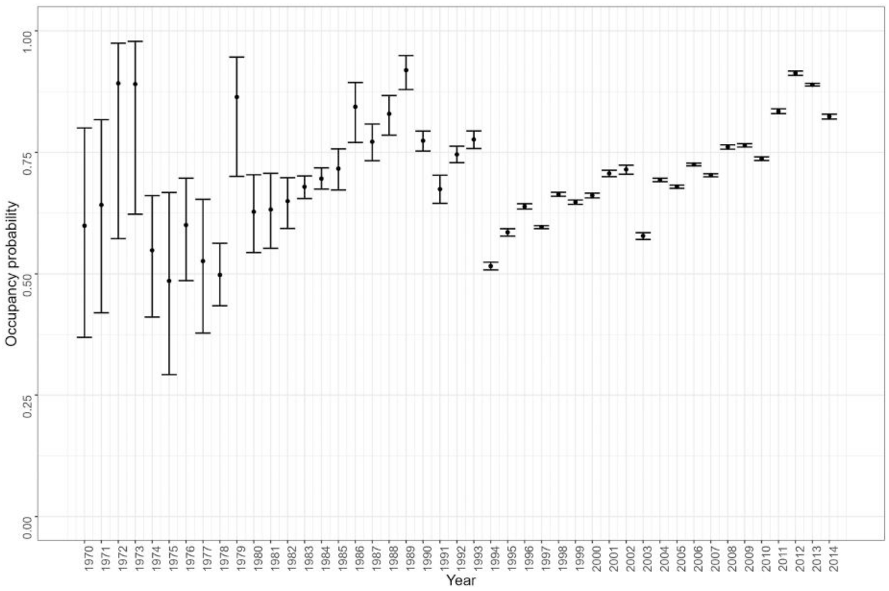

Figure 3. 95% posterior credible intervals of the occupancy probabilities of each year for Ringlet. The dots represent the posterior medians.

The extended GAI has been introduced in this paper, with application to two univoltine species, but future work will test wider application to more species, in particular with varying quantities of data and life histories, including species with two or more generations per year, requiring more complex functions to describe seasonal variation (Dennis et al., 2016). As mentioned, alternative distributions to the Poisson may also be explored, in particular to assess for effects of overdispersion and the associated impact on estimates of uncertainty.

In both adaptations of the GAI presented, species' flight periods were assumed to be the same across sites within each year, as is typical in the production of abundance trends for UK butterflies (UKBMS, 2023). However, greater flexibility in the seasonal variation function **a** can be readily accounted for through the inclusion of appropriate spatial covariates. For example this is done by Schmucki et al. (2016) and Dennis et al. (2022). This is similarly a future direction to explore for occupancy models, accounting for spatial variation in detection probabilities, for example due to spatial variation in species' phenologies, via spatial covariates or spatial effects. See also for example Clark and Altwegg (2019) and Dennis et al. (2019) for models with spatial variation in occupancy.

The extended GAI also provides a basis for further extensions within this computationally efficient approach, such as formal data integration by maximizing a joint likelihood; see Besbeas et al. (2002) and Schaub and Kéry (2021). For example, UKBMS (count data from standardized monitoring) and BBC (mass-participation CS data) could be used to produce new urban butterfly indicators. Integrated modelling approaches can optimize the use of available CS data sets, but present outstanding questions and challenges (Isaac et al., 2020; Johnston et al., 2023; Zipkin et al., 2021), including the need for computationally efficient approaches for data integration.

We have introduced the basic occupancy model within a VI framework and have discussed the use of the observed, instead of the complete, data likelihood for avoiding the standard issue of underestimated posterior variances when parameters are assumed to be independent. In this case, the dependence structure in the model parameters and latent variables was such that the use of the observed data likelihood allowed us to use VI without having to assume that parameters are a-posteriori independent. Intuitively, the same approach could be employed in other ecological models within a VI framework, although at the moment this is only an intuition and future work would need to explore the quality of inference for different types of data and corresponding models.

We believe that VI provides a powerful and versatile framework for efficiently fitting a wide range of ecological models. The advantage of VI is that it combines the speed of classical inference

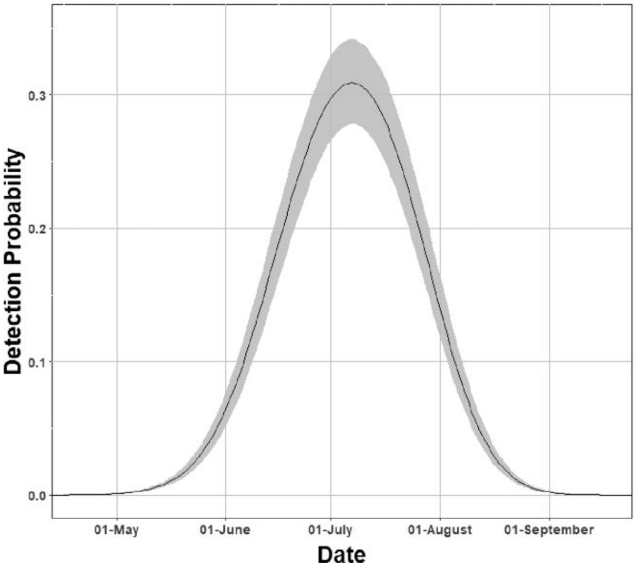

Figure 4. Posterior median and 95% posterior credible intervals of the detection probability in each week for Ringlet.

with the interpretability of Bayesian inference. VI relies on the ability to obtain the gradient of the likelihood function, which might seem like an obstacle. However, it is possible to take advantage of recent developments of deep learning methodologies such as the use of automatic differentiation, which can automatically compute gradients of this type (as long as the likelihood function is tractable), such as for example using the package TMB (Kristensen et al., 2016). We envisage that analyses currently based on hidden Markov models for likelihood computation—see for example Cowen et al. (2017), Besbeas and Morgan (2019, 2020) and McClintock et al. (2020)—can be a fruitful research avenue, since automatic differentiation engines would allow us to compute gradients by differentiating through the forward recursion used to compute the likelihood in this case. The use of automatic differentiation also enables us to consider highly nonlinear extensions, such as neural networks, to be introduced in the model. However, as discussed, particular attention needs to be paid to the chosen variational family, since that will determine the accuracy of the approximation.

We have considered a simple occupancy model, in comparison to models that have been employed previously for data of this type, as a means of discussing VI and the use of observed data likelihood within the VI framework. The flexible spatio-temporal models, for example based on Gaussian processes (Diana et al., 2023; Doser et al., 2022), which have been developed within an MCMC framework could be considered within this VI framework in future work. However, since this leads to a regression with as many covariates as the number of support of points, assuming a variational approximation with a full covariance matrix is computationally prohibitive, since the number of parameters of the Cholesky factor C scales quadratically with the number of covariates. One option to overcome this problem is to induce sparsity on the inverse of the covariance matrix by zeroing elements of the Cholesky factor C (Tan & Nott, 2018). For example, it is possible to assume a variational approximation where the covariate coefficients for the spatial approximation are independent in the posterior. Although this step can potentially reintroduce bias in the model, since it assumes a-posteriori independence of parameters, it leads to a substantial reduction in the number of parameters, making the model feasible to estimate. Finally, in cases where the observed data likelihood cannot be obtained, for example in complex data-generating processes or models with individual random effects, then the complete data likelihood and efficient MCMC inference (see for example King et al., 2023, who devise an importance sampling approach for ecological models with random effects) may provide the only viable alternative, at least at the moment.

To some extent formal design considerations do not arise with the data that we have considered, however there is an issue of nonrandom sampling (Boyd, Powney, et al., 2023; Johnston et al., 2023) and the need to account for issues such as preferential sampling and spatial and temporal biases (Altwegg & Nichols, 2019; Boersch-Supan et al., 2019; Conn et al., 2017; Pati et al., 2011). See for example papers by King et al. (2023) and by Lahoz-Monfort et al. (2014), respectively on sampling the data, and on how to design studies when resources are limited.

We have demonstrated efficient analysis methods for sources of CS data for UK butterflies, but efficient statistical inference methods are needed for understanding population changes from CS data for a wide range of taxa and locations. For example the GAI approach, or related models, has been applied to moths, bees and beetles (Dennis et al., 2021; Fox et al., 2021; Matechou et al., 2018). Development of efficient inference for occupancy models is also vital given their application to various taxa (for example Boyd, August, et al., 2023; Burns et al., 2023; Outhwaite et al., 2019). The need for efficiency will continue to increase with the growth of data sets such as the Global Biodiversity Information Facility (GBIF), which has amassed more than 2.5 billions occurrences of more than one million species (GBIF.org, 2023).

CS data are increasingly 'big', not only in terms of volume, but also involving characteristics such as variety. Farley et al. (2018) and McCrea et al. (2023) discuss the 'Four Vs Framework' in which data may be characterized as 'big'. Analysing CS data for biodiversity monitoring presents various challenges (Johnston et al., 2023), but methods also need to be suitably scalable for these increasingly large data sets.

The examples in this paper are based upon analyses of data featuring observations of species submitted by citizen scientists, but computational challenges also arise from other data types, for example with the growth of data from technological advances such as automated interpretation of images submitted by citizen scientists (Terry et al., 2020; van Klink et al., 2022).

It may be argued that high performance computing (HPC) and cloud computing can be used to address the challenge of fitting computationally demanding models to CS data (Farley et al., 2018). For example, a supercomputer has enabled the production of occupancy trend estimates for thousands of UK species using Bayesian occupancy models (Boyd, August, et al., 2023; Outhwaite et al., 2019). However ultimately, as data sets continue to grow, and models become increasingly complex, we argue that a trade-off is needed, and that using more and more computing resources is not a simple solution. Fitting computationally demanding methods can make appropriate model validation and inference difficult: for example variable selection may become impractical (Johnston et al., 2023), as well as suitable goodness-of-fit assessment. Achieving model convergence for all parameters in MCMC may also become difficult (Boyd, August, et al., 2023; Outhwaite et al., 2019).

Furthermore, to maximize the use of CS data in biodiversity monitoring, there is a need for statistical approaches that are appropriately disseminated and accessible for use in practice. Johnston et al. (2023) suggest that 'accessible communication of novel methods could democratize analysis of these data and thus enable CS data to reach their broadest potential'. Statistical methods that depend upon HPC may be a barrier for analysis to those without easy or affordable access to such resources, thus hindering the potential of biodiversity monitoring with CS data globally (Pocock et al., 2018).

Efficient methods are also crucial for producing frequent analysis updates for reporting on the status of species, particularly as time lags in data availability continue to reduce. The need for accurate reporting is ever necessary in monitoring species' status, measuring against biodiversity targets, supporting policy-making and guiding effective conservation effort. This paper has presented examples for fitting computationally efficient models to CS data, but, as also suggested by Johnston et al. (2023), with the growth of such data and its importance for biodiversity monitoring (Callaghan et al., 2021; Pocock et al., 2018), there is an ongoing need to develop efficient statistical inference methods, with the potential to learn from developments in mainstream statistics.

Acknowledgments

We thank the Associate Editor and two referees for their positive and helpful reviews. The UKBMS is organized and funded by Butterfly Conservation, the British Trust for Ornithology (BTO), and the Joint Nature Conservation Committee (JNCC). The UKBMS is indebted to all volunteers who contribute data to the scheme. We are also 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, and also the citizen scientists who submitted Big Butterfly Count data.

Conflicts of interest: None declared.

Funding

B.M. was supported by a Leverhulme Emeritus Fellowship, EM-2017-061. This work was partly funded by NERC grant NE/T010045/1.

Data availability

Data and code associated with the analyses in this paper are available as follows: (i) example of the extended GAI approach applied to UKBMS data https://github.com/EBDennis/Extended_GAI_UKBMS_example (ii) example of the GAI applied to BBC data https://github.com/EBDennis/GAI_BBC_example (iii) example of VI applied to BNM data for Ringlet https://github.com/AlexDiana/VBOccupancy.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

Author contributions statement

All authors devised the paper. E.D. curated the data sets. E.D. did the computing for the new GAI modelling and A.D. did the computing for the VI analyses. All authors collaborated in the writing of the paper, and E.D. and A.D. produced the online supplementary material.

References

Altwegg R., & Nichols J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, *10*(1),8–21. https://doi.org/10.1111/mee3.2019.10.issue-1

Besbeas P., Freeman S. N., Morgan B. J. T., & Catchpole E. A. (2002). Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, *58*(3), 540–547. https://doi.org/10.1111/j.0006-341X.2002.00540.x

Besbeas P., & Morgan B. J. T. (2019). Exact inference for integrated population modelling. *Biometrics*, *75*(2), 475–484. https://doi.org/10.1111/biom.13045

Besbeas P., & Morgan B. J. T. (2020). A general framework for modelling population abundance data. *Biometrics*, *76*(1), 281–292. https://doi.org/10.1111/biom.v76.1

Blei D. M., Kucukelbir A., & McAuliffe J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Boersch-Supan P. H., Trask A. E., & Baillie S. R. (2019). Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. *Biological Conservation*, *240*, 108286. https://doi.org/10.1016/j.biocon.2019.108286

Bogaart P., van der Loo M., & Pannekoek J. (2020). *rtrim: Trends and Indices for Monitoring Data*. https://CRAN.R-project.org/package=rtrim. R package version 2.1.1.

Borowska A., & King R. (2022). Semi-complete data augmentation for efficient state space model fitting. *Journal of Computational and Graphical Statistics*, *32*(1), 19–35. https://doi.org/10.1080/10618600.2022.2077350

Boyd R. J., August T. A., Cooke R., Logie M., Mancini F., Powney G. D., Roy D. B., Turvey K., & Isaac N. J. (2023). An operational workflow for producing periodic estimates of species occupancy at national scales. *Biological Reviews*, *98*(5), 1492–1508. https://doi.org/10.1111/brv.v98.5

Boyd R. J., Powney G. D., & Pescott O. L. (2023). We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, *38*(6), 521–531. https://doi.org/10.1016/j.tree.2023.01.001

Bravington M. V., Miller D. L., & Hedley S. L. (2021). Variance propagation for density surface models. *Journal of Agricultural, Biological, and Environmental Statistics*, *26*(2), 306–323. https://doi.org/10.1007/s13253-021-00438-2

BRC (2022). Biological Records Centre Home Page. www.brc.ac.uk Date accessed November 13, 2023.

Brereton T. M., Botham M. S., Middlebrook I., Randle Z., Noble D., & Harris S. E. A. (2018). *United Kingdom butterfly monitoring scheme report for 2017*. (Technical Report). Centre for Ecology & Hydrology, Butterfly Conservation, British Trust for Ornithology and Joint Nature Conservation Committee.

Burns F., Mordue S., al Fulaij N., Boersch-Supan P. H., Boswell J., Boyd R. J., Bradfer-Lawrence T., de Ornellas P., de Palma A., de Zylva P., Dennis E. B., Foster S., Gilbert G., Halliwell L., Hawkins K., Haysom K. A., Holland M. M., Hughes J., Jackson A. C., …Gregory R. D. (2023). State of Nature 2023, the State of Nature partnership, www.stateofnature.org.uk.

Butchart S. H. M., Walpole M., Collen B., van Strien A., Scharlemann J. P. W., Almond R. E. A., Baillie J. E. M., Bomhard B., Brown C., Bruno J., Carpenter K. E., Carr G. M., Chanson J., Chenery A. M., Csirke J., Davidson N. C., Dentener F., Foster M., Galli A., …Watson R. (2010). Global biodiversity: Indicators of recent declines. *Science*, *328*(5982), 1164–1168. https://doi.org/10.1126/science.1187512

Callaghan C. T., Poore A. G., Mesaglio T., Moles A. T., Nakagawa S., Roberts C., Rowley J. J., VergÉs A., Wilshire J. H., & Cornwell W. K. (2021). Three frontiers for the future of biodiversity research using citizen science data. *BioScience*, *71*, 55–63. https://doi.org/10.1093/biosci/biaa131

Chandler M., See L., Copas K., Bonde A. M. Z., López B. C., Danielsen F., Legind J. K., Masinde S., Miller-Rushing A. J., Newman G., Rosemartin A., & Turak E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

Clark A. E., & Altwegg R. (2019). Efficient Bayesian analysis of occupancy models with logit link functions. *Ecology and Evolution*, *9*(2), 756–768. https://doi.org/10.1002/ece3.2019.9.issue-2

Clark A. E., Altwegg R., & Ormerod J. T. (2016). A variational Bayes approach to the analysis of occupancy models. *PloS One*, *11*(2), e0148966. https://doi.org/10.1371/journal.pone.0148966

Conn P. B., Thorson J. T., & Johnson D. S. (2017). Confronting preferential sampling in wildlife surveys: Diagnosis and model-based triage. *Methods in Ecology and Evolution*, *8*(11), 1535–1546. https://doi.org/10.1111/mee3.2017.8.issue-11

Cowen L., Besbeas P. T., Morgan B. J. T., & Schwarz C. (2017). Hidden Markov models for extended batch data. *Biometrics*, *73*(4), 1321–1331. https://doi.org/10.1111/biom.12701

Dennis E. B., Brereton T. M., Morgan B. J. T., Fox R., Shortall C. R., Prescott T., & Foster S. (2019). Trends and indicators for quantifying moth abundance and occupancy in Scotland. *Journal of Insect Conservation*, *23*(2), 369–380. https://doi.org/10.1007/s10841-019-00135-z

Dennis E. B., Fagard-Jenkin C., & Morgan B. J. T. (2022). rGAI: An R package for fitting the generalized abundance index to seasonal count data. *Ecology and Evolution*, *12*(8), e9200. https://doi.org/10.1002/ece3.9200

Dennis E. B., Freeman S. N., Brereton T., & Roy D. B. (2013). Indexing butterfly abundance whilst accounting for missing counts and variability in seasonal pattern. *Methods in Ecology and Evolution*, *4*(7), 637–645. https://doi.org/10.1111/mee3.2013.4.issue-7

Dennis E. B., Kéry M., Morgan B. J. T., Coray A., Schaub M., & Baur B. (2021). Integrated modelling of insect population dynamics at two temporal scales. *Ecological Modelling*, *441*, 109408. https://doi.org/10.1016/j.ecolmodel.2020.109408

Dennis E. B., Morgan B. J. T., Brereton T. M., Roy D. B., & Fox R. (2017). Using citizen science butterfly counts to predict species population trends. *Conservation Biology*, *31*(6), 1350–1361. https://doi.org/10.1111/cobi.2017.31.issue-6

Dennis E. B., Morgan B. J. T., Freeman S. N., Brereton T. M., & Roy D. B. (2016). A generalized abundance index for seasonal invertebrates. *Biometrics*, *72*(4), 1305–1314. https://doi.org/10.1111/biom.12506

Dennis E. B., Morgan B. J. T., Freeman S. N., Ridout M. S., Brereton T. M., Fox R., Powney G. D., & Roy D. B. (2017). Efficient occupancy model-fitting for extensive citizen-science data. *PLoS One*, *12*(3), e0174433. https://doi.org/10.1371/journal.pone.0174433

Dennis E. B., Morgan B. J. T., Harrower C. A., Bourn N. A. D., & Fox R. (2024). Incorporating phenology to estimate species' population trends from snapshot citizen-science data. *Journal of Agricultural, Biological, and Environmental Statistics, To appear*.

Diana A., Dennis E. B., Matechou E., & Morgan B. J. T. (2023). Fast Bayesian inference for large occupancy data sets, using the Pólya-Gamma scheme. *Biometrics*, *79*, 2503–2515. https://doi.org/10.1111/biom.13816

Didham R. K., Basset Y., Collins C. M., Leather S. R., Littlewood N. A., Menz M. H. M., Müller J., Packer L., Saunders M. E., Schönrogge K., Stewart A. J. A., Yanoviak S. P., & Hassall C. (2020). Interpreting insect declines: Seven challenges and a way forward. *Insect Conservation and Diversity*, *13*(2), 103–114. https://doi.org/10.1111/icad.v13.2

Doser J. W., Finley A. O., & Banerjee S. (2023). Joint species distribution models with imperfect detection for high-dimensional spatial data. *Ecology*, *104*, e4137. https://doi.org/10.1002/ecy.4137

Doser J. W., Finley A. O., Kéry M., & Zipkin E. F. (2022). spoccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, *13*, 1670–1678. https://doi.org/10.1111/mee3.v13.8

Farley S. S., Dawson A., Goring S. J., & Williams J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, *68*, 563–576. https://doi.org/10.1093/biosci/biy068

Fox R., Dennis E. B., Brown A. F., & Curson J. (2022). A revised red list of British butterflies. *Insect Conservation and Diversity*, *15*, 485–495. https://doi.org/10.1111/icad.v15.5

Fox R., Dennis E. B., Harrower C. A., Blumgart D., Bell J. R., Cook P., Davis A. M., Evans-Hill L. J., Haynes F., Hill D., Isaac N. J. B., Parsons M. S., Pocock M. J. O., Prescott T., Randle Z., Shortall C. R., Tordoff G. M., Tuson D., & Bourn N. A. D. (2021). The State of Britain's Larger Moths 2021. Butterfly Conservation, Rothamsted Research and UK Centre for Ecology & Hydrology, Wareham, Dorset, UK.

Fox R., Dennis E. B., Purdy K. M., Middlebrook I., Roy D. B., Noble D. G., Botham M. S., & Bourn N. A. D. (2023). The State of the UK's Butterflies 2022. Butterfly Conservation, Wareham, UK.

GBIF.org (2023). GBIF Home Page. https://www.gbif.org Date accessed November 13, 2023.

Isaac N. J. B., Jarzyna M. A., Keil P., Dambly L. I., Boersch-Supan P. H., Browning E., Freeman S. N., Golding N., Guillera-Arroita G., Henrys P. A., Jarvis S., Lahoz-Monfort J., Pagel J., Pescott O. L., Schmucki R., Simmonds E. G., & O'Hara R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

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. https://doi.org/10.1111/bij.12532

Isaac N. J. B., van Strien A. J., August T. A., de Zeeuw M. P., & Roy D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*, 1052–1060. https://doi.org/10.1111/mee3.2014.5.issue-10

JNCC (2022). UK Biodiversity Indicators 2022 - C6 Insects of the wider countryside (butterflies). https://jncc.gov.uk/our-work/ukbi-c6-insects-of-the-wider-countryside/ Date accessed November 8, 2023.

Johnston A., Matechou E., & Dennis E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*, 103–116. https://doi.org/10.1111/mee3.v14.1

Jordan M. I., Ghahramani Z., Jaakkola T. S., & Saul L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233. https://doi.org/10.1023/A:1007665907178

Kéry M., Gardner B., & Monnerat C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, *37*, 1851–1862. https://doi.org/10.1111/j.1365-2699.2010.02345.x

King R. (2014). Statistical ecology. *Annual Review of Statistics and Its Application*, *1*, 401–426. https://doi.org/10.1146/annurev-statistics-022513-115633

King R., Sarzo B., & Elvira V. (2023). When ecological individual heterogeneity models and large data collide: An importance sampling approach. *The Annals of Applied Statistics*, *17*, 3112–3132. https://doi.org/10.1214/23-AOAS1753

Kristensen K., Nielsen A., Berg C. W., Skaug H., & Bell B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21. https://www.jstatsoft.org/index.php/jss/article/view/v070i05. https://doi.org/10.18637/jss.v070.i05

Lahoz-Monfort J. J., Harris M. P., Morgan B. J. T., Freeman S. N., & Wanless S. (2014). Exploring the consequences of reducing survey effort for detecting individual and temporal variability in survival. *Journal of Applied Ecology*, *51*, 534–543. https://doi.org/10.1111/1365-2664.12214

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 ed.). Academic Press.

Matechou E., Dennis E. B., Freeman S. N., & Brereton T. (2014). Monitoring abundance and phenology in (multivoltine) butterfly species: A novel mixture model. *Journal of Applied Ecology*, *51*, 766–775. https://doi.org/10.1111/jpe.2014.51.issue-3

Matechou E., Freeman S. N., & Comont R. (2018). Caste-specific demography and phenology in bumblebees: Modelling Beewalk data. *Journal of Agricultural, Biological and Environmental Statistics*, *23*, 427–445. https://doi.org/10.1007/s13253-018-0332-y

McClintock B. T., Langrock R., Gimenez O., Cam E., Borchers D. L., Glennie R., & Patterson T. A. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, *23*, 1878–1903. https://doi.org/10.1111/ele.v23.12

McCrea R., King R., Graham L., & Börger L. (2023). Realising the promise of large data and complex models. *Methods in Ecology and Evolution*, *14*, 4–11. https://doi.org/10.1111/mee3.v14.1

McCullagh P., & Nelder J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

Newman K., King R., Elvira V., de Valpine P., McCrea R. S., & Morgan B. J. T. (2023). State-space models for ecological time series: Practical model-fitting. *Methods in Ecology and Evolution*, *14*, 26–42. https://doi.org/10.1111/mee3.v14.1

Outhwaite C. L., Powney G. D., August T. A., Chandler R. E., Rorke S., Pescott O. L., Harvey M., Roy H. E., Fox R., Roy D. B., Alexander K., Ball S., Bantock T., Barber T., Beckmann B. C., Cook T., Flanagan J., Fowles A., Hammond P., …Isaac N. J. B. (2019). Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970–2015. *Scientific Data*, *6*, 1–12. https://doi.org/10.1038/s41597-019-0269-1

Pati D., Reich B. J., & Dunson D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, *98*, 35–48. https://doi.org/10.1093/biomet/asq067

Pocock M. J. O., Chandler M., Bonney R., Thornhill I., Albin A., August T., Bachman S., Brown P. M. J., Cunha D. G. F., Grez A., Jackson C., Peters M., Rabarijaon N. R., Roy H. E., & Danielsen F., 2018). A vision for global biodiversity monitoring with citizen science. In *Advances in Ecological Research* (Vol. *59*, pp. 169–223). Elsevier.

Pocock M. J. O., 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*, 475–493. https://doi.org/10.1111/bij.12548

Pocock M. J. O., Tweddle J. C., Savage J., Robinson L. D., & Roy H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *PloS One*, *12*, e0172579. https://doi.org/10.1371/journal.pone.0172579

Pollard E., & Yates T. J. (1993). *Monitoring butterflies for ecology and conservation: The British butterfly monitoring scheme*. Chapman & Hall.

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*, 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Rezende D. J., Mohamed S., & Wierstra D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning* (pp. 1278–1286). PMLR.

Schaub M., & Kéry M. (2021). *Integrated population models: Theory and ecological applications with R and JAGS*. Elsevier.

Schmucki R., Pe'er G., Roy D. B., Stefanescu C., Van Swaay C. A. M., Oliver T. H., Kuussaari M., Van Strien A. J., Ries L., Settele J., Musche M., Carnicer J., Schweiger O., Brereton T. M., Harpke A., Heliölä J., Kühn E., & Julliard R. (2016). A regionally informed abundance index for supporting integrative analyses across butterfly monitoring schemes. *Journal of Applied Ecology*, *53*, 501–510. https://doi.org/10.1111/jpe.2016.53.issue-2

Silvertown J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*, 467–471. https://doi.org/10.1016/j.tree.2009.03.017

Tan L. S. L., & Nott D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, *28*(2), 259–275. https://doi.org/10.1007/s11222-017-9729-7

ter Braak C. J. F., van Strien A. J., Meijer R., & Verstrael T. J. (1994). Analysis of monitoring data with many missing values: which method? In E. J. M. Hagemeijer, & T. J. Verstrael (Eds.), *Bird Numbers 1992. Distribution, Monitoring and Ecological Aspects: Proceedings of the 12th International Conference of IBCC and EOAC, Noordwijkerhout, The Netherlands* (pp. 663–673). Statistics Netherlands, Voorburg/Heerlen & SOVON, Beek-Ubbergen.

Terry J. C. D., Roy H. E., & August T. A. (2020). Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, *11*(2), 303–315. https://doi.org/10.1111/mee3.v11.2

Thomas C., Jones T. H., & Hartley S. E. (2019). 'Insectageddon': A call for more robust data and rigorous analyses. *Global Change Biology*, *25*(6), 1891–1892. https://doi.org/10.1111/gcb.2019.25.issue-6

Titsias M., & Lázaro-Gredilla M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International conference on machine learning* (pp. 1971–1979). PMLR.

UKBMS (2023). UKBMS official statistics. https://ukbms.org/official-statistics Date accessed November 8, 2023.

van Klink R., August T., Bas Y., Bodesheim P., Bonn A., Fossøy F., Høye T. T., Jongejans E., Menz M. H. M., Miraldo A., Roslin T., Roy H. E., Ruczyński I., Schigel D., Schäffler L., Sheard J. K., Svenningsen C., Tschan G. F., Wäldchen J., …Bowler D. E. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*, *37*, 872–885. https://doi.org/10.1016/j.tree.2022.06.001

van Strien A., Pannekoek J., Hagemeijer W., & Verstrael T. (2004). A loglinear Poisson regression method to analyse bird monitoring data. In A. Anselin (Ed.), *Bird Numbers 1995, Proceedings of the International Conference and 13th Meeting of the European Bird Census Council, Pärnu, Estonia*. Bird Census News 13 (2000) (pp. 33–39).

Van Swaay C. A. M., Dennis E. B., Schmucki R., Sevillega C. G., Aghababyan K., Åström S., Balalaikins M., Bonelli S., Botham M., Bourn N., Brereton T., Cancela J. P., Carlisle B., Collins S., Dopagne C., Dziekanska I., Escobés R., Faltynek Fric Z., Feldmann R., …Roy D. B. (2020). Assessing Butterflies in Europe - Butterfly Indicators 1990-2018 Technical report. Butterfly Conservation Europe & ABLE/eBMS (www.butterfly-monitoring.net).

Wagner D. L., Grames E. M., Forister M. L., Berenbaum M. R., & Stopak D. (2021). Insect decline in the anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, *118*(2), e2023989118. https://doi.org/10.1073/pnas.2023989118

Wang B., & Titterington D. M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *International Workshop on Artificial Intelligence and Statistics* (pp. 373–380). PMLR.

Zipkin E. F., Zylstra E. R., Wright A. D., Saunders S. P., Finley A. O., Dietze M. C., Itter M. S., & Tingley M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*(1), 30–38. https://doi.org/10.1002/fee.v19.1

Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 658–673
https://doi.org/10.1093/jrsssa/qnae096
Advance access publication 13 September 2024
Discussion Paper

Frequentist prediction sets for species abundance using indirect information

Elizabeth Bersson and Peter D. Hoff

Department of Statistical Science, Duke University, Durham, NC 27701, USA

Address for correspondence: Elizabeth Bersson, Department of Statistical Science, Duke University, Durham, NC 27701, USA. Email: elizabeth.bersson@duke.edu

[Read before The Royal Statistical Society at the Discussion Meeting on the 'Analysis of citizen science data' held at the Society's 2024 annual conference in Brighton on Tuesday, 3 September 2024, the President, Dr Andrew Garrett, in the Chair.]

Abstract

Citizen science databases that consist of volunteer-led sampling efforts of species communities are relied on as essential sources of data in ecology. Summarizing such data across counties with frequentist-valid prediction sets for each county provides an interpretable comparison across counties of varying size or composition. As citizen science data often feature unequal sampling efforts across a spatial domain, prediction sets constructed with indirect methods that share information across counties may be used to improve precision. In this article, we present a nonparametric framework to obtain precise prediction sets for a multinomial random sample based on indirect information that maintain frequentist coverage for each county. We detail a simple algorithm to obtain prediction sets for each county using indirect information where the computation time does not depend on the sample size and scales nicely with the number of species considered. The indirect information may be estimated by a proposed empirical Bayes procedure based on information from auxiliary data. Our approach makes inference for under-sampled counties more precise, while maintaining area-specific frequentist validity for each county. Our method is used to provide a useful description of avian species abundance in North Carolina, USA based on citizen science data from the eBird database.

Keywords: categorical data, conformal prediction, empirical Bayes, exchangeability, frequentist coverage, nonparametric

1 Introduction

Understanding species abundance across heterogeneous spatial areas is an important task in ecology. Citizen science databases that consist of observations of species counts gathered by volunteers are increasingly regarded as one of the richest sources of data for such a task. One of the largest such data sources is the eBird database in which citizen scientists throughout the world input counts of bird sightings (B. Sullivan et al., 2009). In addition to its use for describing avian species abundance, eBird is a principal resource for understanding global biodiversity and is widely used in constructing and implementing conservation action plans (B. L. Sullivan et al., 2017).

More generally, analyses from such databases may be used for informing policy, conservation efforts, habitat preservation, and more, for which understanding species prevalence for non-overlapping geographic areas, such as counties across a state or country, is important. In practice, species abundance from citizen science data are commonly summarized within areas such as counties by empirical proportions from a sample, as in, e.g. Arnold et al. (2021) and Camerini and Groppali (2014). Such proportions can be used to construct a prediction set for each county that provides a description of species prevalence for that county with guaranteed frequentist coverage.

Received: November 24, 2023. **Revised:** June 24, 2024. **Accepted:** August 9, 2024

Given the impact on policy design, corresponding uncertainty quantification is of particular import (Lele, 2020), and so it is desirable that precise prediction sets maintain a target coverage rate regardless of the county's size or composition. This is challenging as a common feature of citizen science data is unequal sampling efforts that results in some counties with large amounts of data information and others with very little. Using direct procedures that only make use of within-county information, a prediction set may be imprecise in these counties with low sampling efforts. This suggests using indirect information such as data from neighbouring counties to improve prediction set precision for a given county.

In this article, we describe species abundance across sampling areas such as counties with frequentist-valid prediction sets that are constructed to contain an unobserved bird with $1 - \alpha$ probability. That is, a valid prediction set for a given county is a set of avian species such that an unobserved bird will belong to one of those species with $1 - \alpha$ probability in a frequentist sense. We develop a valid nonparametric prediction method that allows for information to be shared across counties. Specifically, our approach results in prediction sets with guaranteed frequentist coverage for each county that are constructed with the incorporation of indirect or prior information. We detail and provide code for an empirical Bayes procedure to estimate such prior information from auxiliary data such as neighbouring counties. If this indirect information used to construct the prediction sets is accurate, the prediction sets will be smaller than direct sets that only make use of within-county information.

Identifying avian species that are present across a spatial domain is of general interest in ecology (Lebrun et al., 2012; Shanahan & Possingham, 2009; Twedt et al., 2010). To this aim, in Section 4, we detail the usefulness of the proposed approach in summarizing the eBird citizen science data. For this task, frequentist-valid prediction sets provide useful summaries of the data that may be used to compare information across subregions and better inform policy. In particular, a frequentist-valid prediction set consists of species that are likely to be present in a given area and can be used to draw statistically robust conclusions regarding future observations. Such a prediction set is constructed based on observed species counts, and, by maintaining a specified frequentist coverage rate, reflects the uncertainty in the data. Commonly reported summaries of observed species abundance data, such as a list of observed proportions or counts do not reflect the uncertainty in the data. Moreover, the approach we propose constructs frequentist-valid prediction sets with the incorporation of auxiliary information. This is desirable as sharing information across counties generally results in smaller prediction sets as compared to direct prediction approaches, particularly so in counties with low sampling efforts. In contrast, indirect approaches constructed with auxiliary information such as a Bayesian prediction set may reflect uncertainty in the data, but are not calibrated.

2 Methodology

2.1 Background and notation

For county $j \in \{1, \ldots, J\}$, let $\boldsymbol{X}_j$ be a vector of length K where $X_{j,i} = x_{j,i}$ is the observed count of species i over some set sampling period that may vary across counties. We model $\boldsymbol{X}_j$ with a K-dimensional multinomial distribution with $N_j = \sum_{i=1}^{K} x_{j,i}$ trials and population proportions vector $\boldsymbol{\theta}_j$,

$$\boldsymbol{X}_j \sim MN_K(\boldsymbol{\theta}_j, N_j). \tag{1}$$

We construct a prediction set $A_\alpha(\boldsymbol{X}_j)$ for an observation of a new bird arising from the same distribution, $\boldsymbol{Y}_j \sim MN_K(\boldsymbol{\theta}_j, 1)$, where $\boldsymbol{Y}_j \in \mathcal{Y}$ for $\mathcal{Y} = \{(y_1, \ldots, y_K) : \sum_{i=1}^{K} y_i = 1, y_i \in \{0, 1\}$ $(i = 1, \ldots, K)\}$. Let $\boldsymbol{y}_j^{(k)} \in \mathcal{Y}$ denote a prediction of category k, that is, let $\boldsymbol{y}_j^{(k)}$ be a vector of length K with a one at index k and zeros elsewhere. In particular, we are interested in a prediction set for $\boldsymbol{Y}_j$ that maintains frequentist validity for some error rate α. Formally, we refer to this as an α-valid prediction set:

Definition (α-Valid Prediction Set). An α-valid prediction set for a predictand $\boldsymbol{Y}_j \in \mathcal{Y}$ is any subset A_α of the sample space $\mathcal{Y}$ that contains $\boldsymbol{Y}_j$ with probability greater than or equal to $1-\alpha$,

$$P_\theta\big(\boldsymbol{Y}_j \in A_\alpha(\boldsymbol{X}_j)\big) \geq 1-\alpha, \quad \forall\, \theta, \tag{2}$$

where the probability is taken with respect to $\boldsymbol{Y}_j$ and $\boldsymbol{X}_j$.

Additionally, small or precise α-valid prediction sets are of particular interest, where prediction set size is measured by expected cardinality, that is, expected number of the K categories in the sample space included in the prediction set.

2.2 Order-based prediction for a single area

A standard approach to construct α-valid prediction sets for each county or area is with a direct method that only makes use of within-area information. As such, we first consider construction of a prediction set for a single area j, using only data from area j. For ease of notation, we drop the area-identifying subscript in this subsection.

For multinomial data in general, if the event probability vector $\boldsymbol{\theta}$ is known, an α-valid prediction set is any combination of categories such that their event probabilities cumulatively sum to be greater than or equal to $1-\alpha$. Equivalently stated, an α-valid prediction set may be constructed by excluding categories such that the cumulative sum of the excluded categories' event probabilities is less than α. Such a prediction set may be constructed by admitting categories in some pre-specified order into the prediction set until the cumulative sum of their event probabilities is at least $1-\alpha$. The resulting prediction set will have $1-\alpha$ coverage regardless of the ordering used to admit categories. In fact, the class of all α-valid prediction sets may be constructed by following this procedure for non-strict total orderings of categories.

Perhaps intuitively, constructing such a prediction set by including categories with the largest event probabilities will result in the smallest α-valid prediction set. In the terminology of ordering, this corresponds to constructing a prediction set based on an ordering of categories that matches the ordering of the elements in $\boldsymbol{\theta}$. We refer to this optimal ordering as the oracle ordering:

Theorem 1 (Oracle order-based prediction). Let $\boldsymbol{Y} \sim MN_K(\boldsymbol{\theta}, 1)$ for $\boldsymbol{\theta}$ known. Then,

1. the class of all α-valid prediction sets for a given $\boldsymbol{\theta}$ consists of prediction sets of the form,

$$A_\alpha^{\theta,o} = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^{K} \mathbb{1}(o_k \geq o_l)\theta_l \right] > \alpha \right\}, \tag{3}$$

for some vector $\boldsymbol{o} \in \mathbb{R}^K$, and

2. the **oracle ordering** is that which corresponds to the increasing order statistics of $\boldsymbol{\theta}$,

$$\boldsymbol{o}^\theta = \{\boldsymbol{o} : \theta_m < \theta_n \Rightarrow o_m < o_n \quad \forall\, m, n \in \{1, \ldots, K\}, m \neq n\},$$

and $A_\alpha^{\theta,o^\theta}$ has the smallest cardinality among all orderings.

In practice, $\boldsymbol{\theta}$ is unknown, but a prediction set may be constructed based on an observed sample $\boldsymbol{X} = \boldsymbol{x}$. It turns out, in fact, that any conditional α-valid prediction set can be written similarly to the previous construction (equation (3)), where the cumulative sum is computed with respect to the empirical proportions given by $\boldsymbol{x}$ and $\boldsymbol{y}$. This is a generalization of the conformal prediction framework, a popular machine learning approach to construct prediction regions based on measuring conformity (or non-conformity) of a predictand to an observed sample (Vovk et al., 2005).

Theorem 2 (α-valid order-based prediction). Let $\boldsymbol{X} \sim MN_K(\boldsymbol{\theta}, N)$, $\boldsymbol{Y} \sim MN_K(\boldsymbol{\theta}, 1)$. Then, every conformal α-valid prediction set based on observed data $\boldsymbol{x}$ can be written

$$A_\alpha(\boldsymbol{x}) = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^{K} \mathbb{1}\left(o_k \geq o_l\right) \frac{x_l + y_l^{(k)}}{N+1} \right] > \alpha \right\}, \tag{4}$$

for some vector $\boldsymbol{o} \in \mathbb{R}^K$.

Note that the prediction set depends on the vector $\boldsymbol{o}$ only through the order of its elements. Such a vector may be defined based on the data, or otherwise. We elaborate in the remainder of this section.

For any ordering of the K categories, constructing a prediction set following Theorem 2 results in a prediction set with guaranteed finite-sample $1 - \alpha$ frequentist coverage. The choice of ordering, however, will impact prediction set precision, that is, the set's cardinality. For inference for a single area, a natural approach is to order the categories with respect to their empirical proportions. The empirical proportions are unbiased for population proportions, so, if the area has a large sample size, an ordering based on the empirical proportions will approximate the oracle ordering well. It turns out this approach is well-motivated by classical prediction approaches. Specifically, a standard direct prediction method constructs a prediction set separately for an area based on an area-specific conditional pivotal quantity (Faulkenberry, 1973; Tian et al., 2022). For a multinomial population, $\boldsymbol{Y} \,|\, \boldsymbol{X} + \boldsymbol{Y}$ is such a quantity that follows a multivariate hypergeometric distribution which does not depend on the event probability vector. See Thatcher (1964) for work on prediction sets of this type for binomial data. A prediction set constructed to contain species belonging to a highest mass region of this pivotal distribution is obtained by including species with the largest empirical counts until their cumulative proportion sum exceeds $1 - \alpha$,

$$A_\alpha^D(\boldsymbol{x}) = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^{K} \mathbb{1}\left(\left(x_k + y_k^{(k)} \right) \geq \left(x_l + y_l^{(k)} \right) \right) \frac{x_l + y_l^{(k)}}{N+1} \right] > \alpha \right\}. \tag{5}$$

This direct prediction set based on an ordering of the empirical proportions is appealing as it is easy to interpret and has finite-sample guaranteed $1 - \alpha$ frequentist coverage. For an area with low sampling effort, though, the empirical proportions will not precisely estimate the true proportions. As a result, a prediction set may have prohibitively large cardinality such that it is not practically useful. For such an area, incorporating indirect information from neighbouring counties can improve the estimates of the county proportions and thereby increase the precision of a prediction set.

2.3 Order-based prediction for multiple areas

In general, in analysing small area data, that is, areal data featuring small within-area sample sizes in some areas, it is common to utilize indirect methods that share information across areas (Rao & Molina, 2015). The eBird database is a rich data source, and inference in any given county may be improved upon by taking advantage of auxiliary data using an indirect method. In this subsection, we detail how information from neighbouring counties may be used in estimating an ordering of categories to improve prediction set precision.

As opposed to a direct prediction set based on an ordering corresponding to within-county empirical proportions, an indirect prediction set can be constructed similarly whereby species are admitted into the prediction set based on an ordering corresponding to empirical posterior proportions estimated from a hierarchical model. Such an estimate may be obtained based on a conjugate Dirichlet prior distribution parameterized with a common concentration hyperparameter for the J areas,

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J \sim \text{Dirichlet}_K(\boldsymbol{\gamma}). \tag{6}$$

Given a hyperparameter $\boldsymbol{\gamma} \in \mathbb{R}^K$, the posterior expectation of the proportions $\boldsymbol{\theta}_j$ in county j is $\tilde{\boldsymbol{x}}_j/(N_j + \sum_{i=1}^{K} \gamma_i)$, where $\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j + \boldsymbol{\gamma}$. In this way, $\tilde{\boldsymbol{x}}_j$ may be interpreted as a posterior vector of counts for county j. Then, an α-valid prediction set based on $\tilde{\boldsymbol{x}}_j$ is,

$$A^I_\alpha(\boldsymbol{x}_j) = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^{K} \mathbb{1}\left(\left(\tilde{x}_{j,k} + y^{(k)}_k\right) \geq \left(\tilde{x}_{j,l} + y^{(k)}_l\right)\right) \frac{x_{j,l} + y^{(k)}_l}{N_j + 1} \right] > \alpha \right\}. \tag{7}$$

By Theorem 2, $A^I_\alpha(\boldsymbol{x}_j)$ is an α-valid procedure, and it is constructed based on prior information. Specifically, it differs from the direct set given in equation (5) in that categories are admitted into the prediction set based on an ordering determined by posterior counts that incorporate indirect information $\boldsymbol{\gamma}$, as opposed to an ordering based on the observed sample. Moreover, it has been shown that if the indirect information used is accurate, $A^I_\alpha(\boldsymbol{x}_j)$ may be more precise than a direct prediction set with the same coverage rate (Bersson & Hoff, 2024; Hoff, 2023).

In total, $A^D_\alpha(\boldsymbol{x}_j)$ and $A^I_\alpha(\boldsymbol{x}_j)$ are both α-valid prediction procedures. They differ in the order in which species are admitted into the prediction sets, as species are admitted into the direct set in terms of decreasing empirical proportions and into the indirect set in terms of decreasing posterior counts. As a result, for an area with a small sample size, incorporating accurate prior information can result in an ordering used to construct a prediction set that more accurately approximates the oracle ordering as the empirical proportions might be too unstable. Of note, these two approaches are equivalent for a uniform prior $\boldsymbol{\gamma} = c\mathbf{1}$, for any constant c. This includes, for example, a standard noninformative prior $c = 1$, a standard objective Bayes Jeffrey's prior $c = 1/2$, and an improper prior $c = 0$.

2.4 Empirical Bayes estimation of indirect information

To obtain an α-valid indirect prediction set for county $j \in \{1, \ldots, J\}$, all that is required is an estimate of the prior concentration parameter $\boldsymbol{\gamma}$. We propose an empirical Bayesian approach whereby values of $\boldsymbol{\gamma}$ to be used for county j are estimated from data collected in neighbouring counties. Specifically, we use the maximum likelihood estimate of the marginal likelihood based on the conjugate hierarchical model given by equations (1) and (6),

$$\begin{aligned} \boldsymbol{\gamma}_j &= \arg\max_{\boldsymbol{\gamma}} \log p\left(\bigcup_{l \in L_j} \boldsymbol{X}_l \middle| \boldsymbol{\gamma} \right) \\ &= \arg\max_{\boldsymbol{\gamma}} \log \prod_{l \in L_j} \left[\frac{\Gamma(\sum_{i=1}^{K} \gamma_i)}{\Gamma(\sum_{i=1}^{K} x_{l,i} + \gamma_i)} \times \prod_{i=1}^{K} \frac{\Gamma(x_{l,i} + \gamma_i)}{\Gamma(\gamma_i)} \right], \end{aligned} \tag{8}$$

where $L_j \subseteq \{1, \ldots, J\} \backslash \{j\}$ is a non-empty set containing the indices of counties neighbouring county j. Information is shared across neighbouring counties to inform an estimate of the prior for county j, and, when estimated in this way, the prior concentration represents an across-county pooled prior concentration. This optimization problem can be solved numerically with a Newton–Raphson algorithm. See Appendix B for details and derivation of such an algorithm. Code to implement this procedure in the R Statistical Programming language is available online, see Section 5.

When $\boldsymbol{\gamma}_j$ is estimated using data independent of area j and used to construct $A^I_\alpha(\boldsymbol{x}_j)$, the finite-sample coverage guarantee of $A^I_\alpha(\boldsymbol{x}_j)$ holds regardless of the accuracy of the estimated prior hyperparameter. If the estimated vector $\boldsymbol{\gamma}_j$ is accurate, then $A^I_\alpha(\boldsymbol{x}_j)$ may also be more precise than direct prediction approaches.

3 Simulation study

To illustrate how the incorporation of indirect information can affect precision of prediction sets, we compare expected set cardinality obtained from the indirect and direct prediction methods for a single simulated area. In contrast to the eBird data, for example, the analysis of this section corresponds to that of one county. Because citizen science data such as these often feature unequal

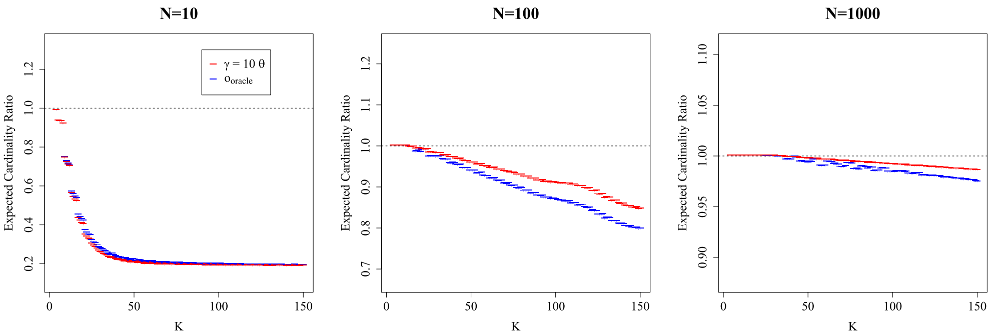

Figure 1. Monte Carlo approximations (±1 standard deviation) of the expected cardinality ratio of indirect method based on an accurate prior with moderate precision to direct method, and expected cardinality ratio of α-valid prediction set given the oracle ordering to direct method.

sampling efforts across counties, we are particularly interested in demonstrating the difference in cardinality between these two approaches for a range of sample sizes $N = 10, 100, 1{,}000$. Moreover, we compare results for varying number of categories K. Throughout, we consider a low entropy regime in which $\lceil K/4 \rceil$ categories unequally split nearly all of the probability mass, and the rest of the categories have nearly probability 0. While we do not necessarily expect real populations in practice to have such a distribution, it is chosen to clearly demonstrate the benefit of including indirect information in the construction of prediction sets that maintain frequentist coverage.

In one construction of indirect prediction sets, we consider a prior based on full information with moderate prior precision $\boldsymbol{\gamma} = \boldsymbol{\theta} \times 10$. We compare with direct prediction sets given by equation (5), or, equivalently, indirect prediction sets constructed with a uniform prior $\boldsymbol{\gamma} = \mathbf{1}$. Finally, we compare the approaches to α-valid order-based prediction sets obtained based on an oracle ordering. Results comparing Monte Carlo approximations of the expected prediction set cardinality ratios between the various approaches obtained from 25,000 replications are displayed in Figure 1.

As all methods considered are α-valid procedures, the crucial difference between them is the incorporation of indirect information. Utilizing accurate prior information in the construction of prediction sets generally results in prediction sets distinctly smaller than direct sets, particularly so if there are a large number of categories relative to the sample size. This is evidenced by the red dashes in Figure 1 showing the expected cardinality ratios of the indirect to direct prediction sets are always at or below a value of 1. An accurate prior may be one that approximates the true probability mass vector well with large precision relative to sample size, as seen in the left plot of Figure 1 for sample size $N = 10$. More generally, though, all that is needed is a prior that results in posterior counts that accurately approximate the oracle ordering of categories. We discuss the three sample size regimes in detail below.

For a small sample size of $N = 10$, the prior $\boldsymbol{\gamma}$ used to construct the indirect prediction sets is an informative prior with strong precision in that the scale used is equal to the sample size in this case. As a result, the posterior distributions contain notably more information than what is in each simulated dataset. As a result, the ordering of categories induced by the posterior counts, used to construct the indirect prediction sets, are accurately approximating the oracle ordering of categories. This is evidenced by the nearly identical behaviour of the two cardinality ratios explored. In conjunction with the instability of the direct method in the presence of such a small sample size, this results in notably smaller cardinality of the indirect set as compared to the direct set, even for relatively small total numbers of categories. At its best, the indirect prediction set is about 80% smaller than the direct set.

For a moderate sample size of $N = 100$, the prior precision used to construct the indirect prediction sets is not overwhelming as compared to the sample size, and hence the posterior counts do not approximate the oracle ordering as well as in the regime with a smaller sample size. This is evidenced by the divergence of the red and blue dashes in the middle plot of Figure 1. Still, particularly as the number of categories increases for fixed N, the benefit of utilizing prior information of this type is highlighted by the decline of the cardinality ratio of the indirect to direct prediction sets (red lines). For example, in the case of $N = 100$ and $K = 150$, the indirect prediction set constructed with $\boldsymbol{\gamma}$ is about 15% smaller than the direct prediction set.

A similar but less pronounced pattern is seen in the presence of a larger sample size of $N = 1{,}000$. For this sample size with $K \le 150$, all methods considered perform relatively similarly. However, as the number of categories increases, there is a distinct gain in prediction set precision given the input of indirect information in prediction set construction.

3.1 Comparison with Bayesian prediction sets

For multi-group count data such as we are considering, a standard Bayesian prediction set may be constructed from the posterior predictive distribution (see, for example, Gelman et al., 2014 Section 1.3) based on the hierarchical model given by equations (1) and (6). Specifically, the posterior predictive distribution for group j is a multinomial distribution for a single trial with proportions vector $\tilde{\boldsymbol{x}}_j/(N_j + \sum_{i=1}^{K} \gamma_i)$. Then, a Bayesian prediction set A_α^B may be taken to be the categories corresponding to the highest mass region of this predictive distribution such that the categories' cumulative predictive probability sum exceeds $1 - \alpha$,

$$A_\alpha^B(\boldsymbol{x}_j) = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^{K} \mathbb{1}(\tilde{x}_{j,k} \ge \tilde{x}_{j,l}) \frac{\tilde{x}_{j,l}}{N_j + \sum_{i=1}^{K} \gamma_i} \right] > \alpha \right\}. \tag{9}$$

Such a prediction set may have smaller cardinality than the indirect or direct approach, but it will not be α-valid. In fact, the frequentist coverage may arbitrarily be nearly zero under a misspecified prior parameter with a large scale.

To illustrate the relationship between cardinality and coverage, we compare the Bayesian prediction method to the proposed indirect approach (7) in simulation. We follow the same simulation set-up as before and, for both methods, consider a prior based on full information with moderate prior precision $\boldsymbol{\gamma} = \boldsymbol{\theta} \times 10$. Results for a moderate sample size of $N = 100$ are plotted in Figure 2. By construction, the indirect approach is α-valid and maintains frequentist coverage rates at or above the nominal level. In contrast, the Bayesian prediction sets may be narrower than the indirect sets, but, correspondingly, the frequentist coverage rate of the Bayesian sets may fall below the nominal rate.

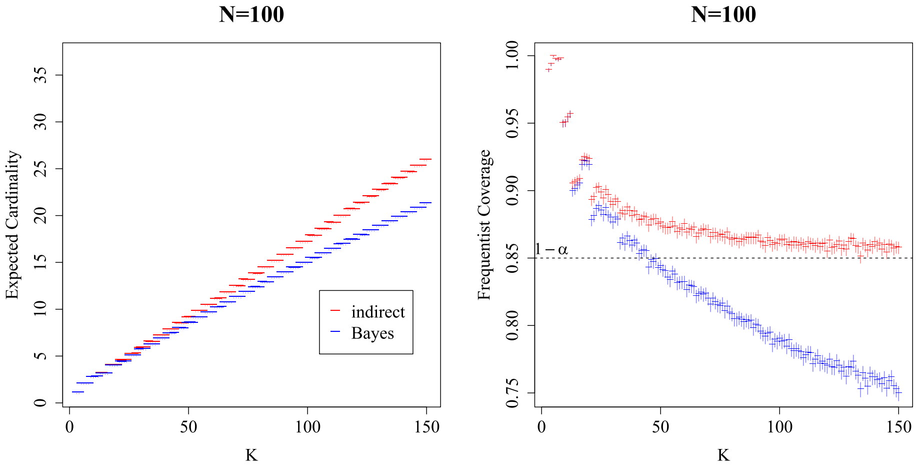

Figure 2. (Left) Monte Carlo approximations (±1 standard deviation) of the expected cardinality of the indirect α-valid prediction set (indirect) and Bayesian posterior predictive set (Bayes), both constructed with prior concentration $\boldsymbol{\gamma} = \boldsymbol{\theta} \times 10$. (Right) Monte Carlo approximations (with 95% Clopper–Pearson intervals) of the frequentist coverage rate of each method, both constructed for an error rate of $\alpha = 0.15$.

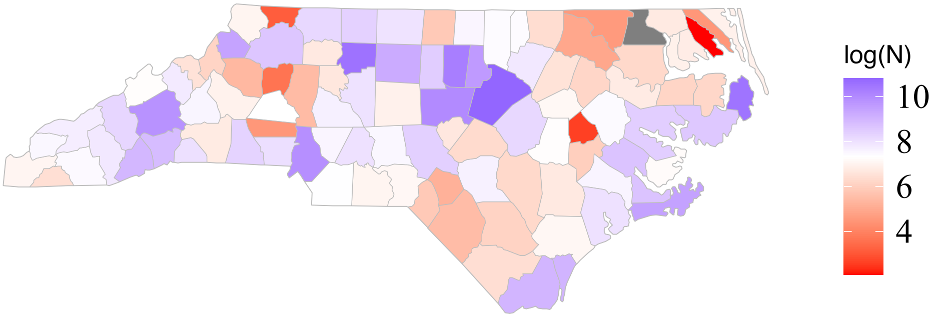

Figure 3. Within-county log sample size.

4 Summarizing eBird species abundance data

In this section, we describe avian species abundance in North Carolina, USA from eBird data obtained from citizen-uploaded complete checklists of species observations in the first week of May 2023. Across the 99 counties, 393 unique species were identified. Some species such as the Northern Cardinal, Carolina Wren, and American Robin were identified frequently. Many others like the Northern Saw-whet Owl and the Solitary Sandpiper were rarely seen; in fact, 50% of species were seen fewer than 100 times each across the entire state. Moreover, within-county sample sizes vary drastically (Figure 3) from approximately 50,000 individual birds identified in Wake County, one of the most populous counties in NC that contains the state's capital, to only 8 in Pasquotank County, a small coastal county consisting of about 1/30th of the human population of Wake County.

As motivated in Section 1, describing such data with α-valid prediction sets for each county provides a useful summary with unambiguous statistical interpretation. That is, with at least probability $1 - \alpha$, an unobserved bird in a given county will belong to a species contained in the specified prediction set, where the probability is taken with respect to the random sample and the predictand. Here, we demonstrate the usefulness of this approach in gaining better understanding of species abundance. Moreover, we elaborate on the benefit of utilizing indirect information in the construction of practically useful sets that are precise, particularly for counties with small within-county sample sizes.

For each county in NC, we construct an indirect prediction set based on a prior hyperparameter estimated from data in the five nearest neighbouring counties, following the procedure described in Section 2.4. The eBird data consist of independent samples collected across the state, so samples are independent across counties. As a result of this independence, finite-sample coverage of the indirect prediction approach is guaranteed. We compare the cardinality of these indirect prediction sets to that of direct prediction sets, both of which maintain at least 95% coverage for each county. The cardinality ratios of the indirect to direct prediction sets across the counties in NC are plotted in Figure 4. To highlight the impact of within-county sample size, the lower quantile sample sizes are overlaid on their respective county.

In general, the incorporation of indirect information in the construction of prediction sets results in notably smaller cardinality of the indirect prediction sets as compared with that of the direct prediction sets. Of the 99 counties in NC, indirect sets have smaller cardinality in 65, and the two approaches result in the same cardinality in 20 counties. The improvement in cardinality is particularly conspicuous in counties with small to moderate sample sizes, as evidenced by the sample sizes of counties corresponding to the smallest cardinality ratios, seen in Figure 4. Moreover, 10 counties have trivial direct sets consisting of all K species, while only two counties with the smallest within-county sample sizes, 8 and 14, have trivial indirect prediction sets. For the county with the third smallest sample size (24), the indirect prediction set only includes 80 species, or about 20% of all possible species, while the direct prediction set is the trivial set.

Overall, even in counties with larger sample sizes, it is most common for the indirect and direct prediction sets to contain a different set of species. In fact, the indirect and direct prediction sets

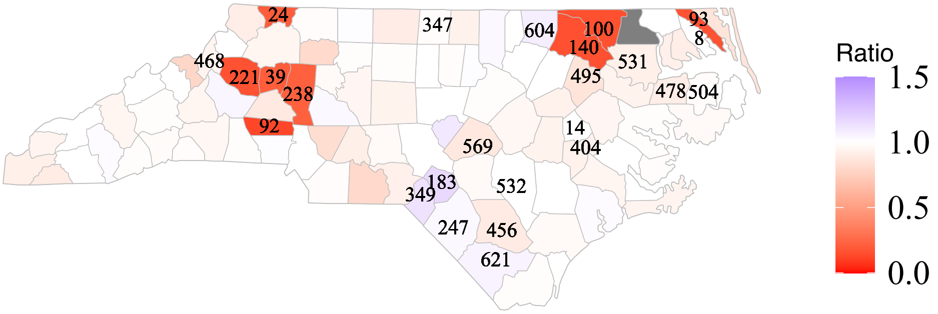

Figure 4. Cardinality ratio of indirect to direct prediction sets. Prior hyperparameters estimated with an empirical Bayesian procedure based on five nearest neighbours for each county. The lowest quantile sample sizes are overlaid on their respective counties.

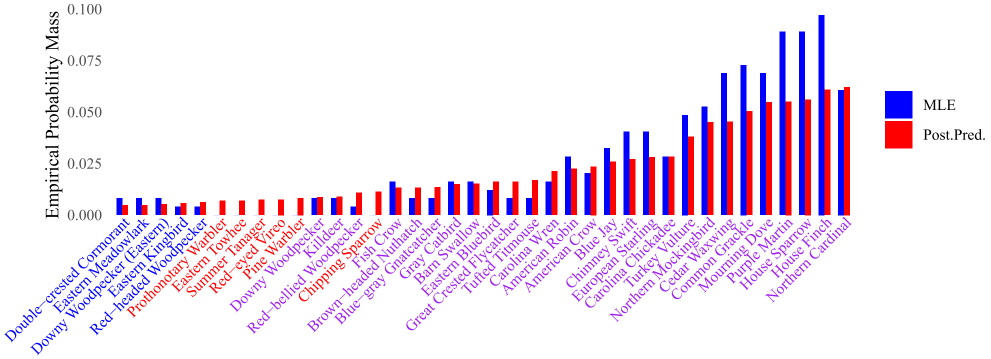

Figure 5. Empirical probability masses of species included in only the indirect (Prothonotary Warbler, Eastern Towhee, Summer Tanager, Red-eyed Vireo, Pine Warbler, and Chipping Sparrow), only the direct (Double-crested Cormorant, Eastern Meadowlark, Downy Woodpecker (Eastern), Eastern Kingbird, and Red-headed Woodpecker), or both (all remaining species) prediction sets, sorted by the posterior proportion (Post.Pred). The empirical proportion (MLE) and the posterior proportion based on a prior estimated from data in nearest five neighbouring counties are plotted.

disagree for nearly every county in NC. They are equivalent for only six counties where they are not both trivial sets. Commonly, this discrepancy corresponds with smaller indirect sets, and hence highlights the benefit of inclusion of indirect information in the construction of prediction sets.

4.1 Order-based prediction in Robeson County

To further compare the two approaches and elucidate the role of the ordering of the species, we elaborate on the construction of indirect and direct prediction sets for Robeson County. Robeson is located near the southeastern border of NC and features a moderately small within-county sample size of 247 birds observed, with species-specific observation counts ranging from zero to ten. The two prediction sets have nearly the same cardinality but contain differences in species inclusion. Specifically, the indirect prediction set contains 33 species, and the direct set contains 32, with an overlap of 27 species.

To illustrate the role of the ordering used in the construction of α-valid prediction sets, the empirical proportions based on the observed sample [maximum likelihood estimate (MLE)] and posterior proportions (Post.Pred) are plotted in Figure 5 for the union of included species in the two sets. In the figure, the species are sorted by increasing posterior proportions. The indirect and direct sets include species based on the posterior and empirical distributions, respectively. Discrepancies between the indirect and direct sets occur when these two distributions disagree.

Table 1. Percentage of all birds observed within each respective county, for select species included in either the indirect set (Pine Warbler and C. Sparrow) or the direct set (D. Cormorant and E. Kingbird)

	D. Cormorant	E. Kingbird	Pine Warbler	C. Sparrow
Robeson	0.81%	0.4%	0.00%	0.00%
NC-017	0.00%	2.85%	1.97%	2.63%
NC-047	0.00%	0.00%	1.29%	0.64%
NC-051	0.00%	0.68%	2.62%	5.59%
NC-093	0.00%	0.00%	1.09%	0.55%
NC-165	0.00%	0.86%	2.58%	5.16%
γ	0.00	1.33	3.42	4.69

Note. Estimated prior hyperparameter γ for Robeson County recorded in the last row.

From Figure 5, it is easy to see the indirect prediction set consists of the species with the 33 largest posterior predictive proportions. In contrast, the direct set consists of species with the largest sample probability mass. Naturally, the ordering of these two estimates agree for species common to the region, and, as such, there is a fair amount of overlap of species inclusion.

As a result of our estimation procedure for the prior hyperparameter γ for Robeson County, the disparity between inclusion or exclusion of a species among the two prediction set methods is further elucidated by examining species presence in neighbouring counties. In short, species with more frequent occurrence in neighbouring counties will have a larger estimated prior count than those seen rarely in neighbouring counties. Species occurrences in neighbouring counties are displayed in Table 1 for a select few species along with the estimated γ for Robeson County, obtained by solving equation (8) using data in these neighbouring counties.

Intuitively, species that are seen in neighbouring counties with some relative frequency, such as the Chipping Sparrow or Pine Warbler, are probably also present in Robeson County, and hence should be included in a prediction set. In practice, these species have a comparatively high estimated prior of about 5 and 4, respectively, and hence are included in the indirect prediction set even though they were not recorded as being observed in Robeson County in the dataset. Alternatively, consideration of indirect information yields the conclusion that species like the Eastern Kingbird and Cormorant may be rare in the area in general, as reflected by small γ values, and thus these species are not included in the indirect prediction set.

4.2 Inference among species with tied observed counts in Haywood County

In species abundance data, particularly for areas or counties with small sample sizes, it is common for multiple species to have the same observed count. A feature of the construction of the direct order-based prediction approach as presented is that species with the same observed counts will either be jointly included or excluded from the prediction set. As a result, a direct prediction set constructed from a sample with tied species counts may have increased cardinality over an indirect prediction set that does not necessarily jointly admit all species with tied observed counts. If the direct set has increased cardinality for this reason, the direct set will also have increased coverage over the indirect set.

When constructing a prediction set based on the empirical proportions without consideration of indirect information, as in the construction of the direct set, this may commonly occur, and there is no clear approach to choose among the species with tied counts without further information than what is provided in the sample in that county. One could randomly choose to include one of the species from the set of species with tied counts, for example, but a more principled manner is to utilize indirect information to determine which species should be included. This is the mechanism used by the indirect prediction approach when the prior hyperparameter is a real-valued vector estimated from indirect information. As such, a more nuanced benefit of utilizing indirect information in the construction of a prediction set is the capacity to include a select few categories with tied empirical proportions.

Table 2. Percentage of all birds observed within each respective county for species included in either both prediction sets (Bobolink) or only the direct set (L. Flycatcher, R. Hawk, C. Yellowthroat, and E. Kingbird)

	L. Flycatcher	R. Hawk	C. Yellowthroat	E. Kingbird	Bobolink
Haywood	0.21%	0.24%	0.21%	0.24%	0.24%
NC-021	0.06%	0.29%	0.18%	0.43%	0.14%
NC-099	0.00%	0.08%	0.16%	0.00%	0.00%
NC-115	0.07%	0.14%	0.21%	0.28%	0.00%
NC-173	0.18%	0.18%	0.66%	0.09%	1.41%
NC-175	0.00%	0.30%	1.00%	0.54%	0.84%
γ	0.4	1.29	2.43	1.49	2.15

Note. Estimated prior hyperparameter γ for Haywood County recorded in the last row. Species are sorted by posterior proportion.

To demonstrate, we elaborate on species inclusion in the indirect and direct prediction sets in Haywood County. Haywood is popular destination in the Blue Ridge Mountains, located near the western border of North Carolina. It features a moderately large within-county sample size of roughly 4,000 birds observed. In Haywood County, the indirect prediction set contains 70 species, and the larger direct set contains 74. In the construction of these prediction sets, the ordering of species with regards to the posterior proportions and the empirical proportions agree for most species. As a result, all 70 species included in the indirect set are also included in the direct set. The disparity in species inclusion occurs primarily as a result of tied counts of species occurrence in the sample.

Empirical proportions in Haywood and neighbouring counties are reported in Table 2 for the five species included in Haywood County's prediction sets with the smallest posterior proportions. The species with the four smallest posterior proportions are included only in the direct set, and the other species, the Bobolink, is included in both the indirect and direct sets. The Bobolink was observed nine times in the sample from Haywood County, or about 0.24% of the Haywood sample. For an ordering determined by either the empirical counts or the posterior counts, this species is required to be included in the order-based prediction set to guarantee $1 - \alpha$ coverage. Two of the other species, the Red-shouldered Hawk and Eastern Kingbird, were each also observed nine times in the sample from Haywood, and, by construction of the order-based prediction approach, must also be included in the direct set. When admitting the species into a prediction set by posterior counts based on the real-valued prior hyperparameter γ estimated from data in neighbouring counties, as in the indirect approach considered, the 'tie' among these three species is broken, and only one, the Bobolink, is included in the indirect prediction set.

4.3 Comparison with Bayesian prediction sets

The α-valid prediction approaches may also be compared to the Bayesian prediction method detailed in equation (9). As the Bayesian method and the indirect method are constructed based on the same hierarchical working model, the Dirichlet prior parameter in (9) may be estimated following the same empirical Bayes approach used for the indirect prediction method. While the Bayesian sets may be smaller than the direct or indirect sets, the Bayesian sets are not α-valid, and thus the frequentist coverage of a set for any given county may fall below the nominal level.

As conveyed in the cardinality comparison between the indirect and direct prediction sets, incorporating auxiliary information in the construction of the prediction sets results in improved cardinality overall. Specifically, the Bayesian prediction sets are smaller than the direct sets in 59 of the 99 counties in North Carolina. In comparing the two methods that utilize auxiliary information, the Bayesian sets are smaller than indirect sets in 33 counties, and they have the same cardinality in 15 counties. Overall, the indirect method outperforms both the Bayesian method and the direct method in terms of minimizing prediction set cardinality. Moreover, the Bayesian approach is not α-valid, and thus the frequentist coverage rates of the Bayesian prediction sets may fall below the nominal level.

5 Discussion

Species abundance data collected across heterogeneous areas is increasingly important in understanding biodiversity. Some of the largest sources of such data are citizen science databases for which volunteers spearhead the data collection. As a result of the civilian-led scientific effort, such data often feature unequal sampling across a spatial domain where some areas have large within-area sample sizes and others have much smaller within-area sample sizes.

In this article, we propose summarizing species abundance data of this type with valid prediction sets that are constructed by sharing information across areas. Utilizing indirect information may result in smaller prediction sets than otherwise achievable with direct methods. Meanwhile, maintaining validity of the prediction sets for each area allows for an accessible interpretation that enables a straightforward comparison across areas. In particular, maintaining interpretable statistical guarantees on a descriptor of such data is important as analyses from such data often have far reaching policy implications. Smaller prediction sets may be attainable based on Bayesian inference of a spatial hierarchical model such as that presented in Tang et al. (2023), for example, but these approaches introduce bias and a resulting prediction set would not retain the nominal frequentist coverage rate guarantee for each county.

The usefulness of our approach for summarizing citizen science data is motivated in part to combat the common problem of varying sampling efforts across areas. We detail how α-valid prediction sets can be constructed with the incorporation of indirect information to improve within-county prediction set precision and propose an empirical Bayes procedure to do so. Incorporation of accurate indirect information results in a narrower prediction set for a given county than a direct prediction set by exploiting data in nearest neighbouring counties. The proposed empirical Bayes procedure is based on a standard hierarchical model that is straightforward to understand, and the authors provide code for implementation.

There may, however, be a benefit to utilizing a more structured prior that incorporates indirect information in a more complex manner such as a prior that weights data from different parts of the state differently. For example, a model based on a learned intrinsic distance between counties was shown in Christensen and Hoff (2022) to fit a subset of the eBird data better than standard methods based on geographic adjacency structure. In the sample analysed in Section 4, we found an indirect prediction set constructed with a hyperparameter estimated from five nearest neighbours results in overall narrower prediction sets than a direct approach, but it would be valuable to explore if this can be further improved upon with a more detailed prior. More broadly, different applications may warrant an alternative information sharing prior if, for example, there is no notion of spatial distance across the different areas. For example, it may be of interest to compare species abundance variation across different time frames for a given county.

All replication codes for this article, including functions to implement the empirical Bayes estimation procedure for the prior hyperparameter, are available at https://github.com/betsybersson/FreqPredSets_Indirect.

Conflicts of interest: No competing interest is declared.

Funding

None to declare.

Data availability

The eBird project is managed by the Cornell Lab of Ornithology, and the raw data are publicly available online at https://science.ebird.org/en/use-ebird-data. The sample analysed in Section 4 is available at https://github.com/betsybersson/FreqPredSets_Indirect/tree/main/data.

Author contributions statement

E.B. and P.H. conceived the experiments, E.B. conducted the experiments, E.B. and P.H. analysed the results, E.B. and P.H. wrote and reviewed the manuscript.

Appendix A: Proofs

Remark 1 (Concerning Theorem 1). We first elaborate on the construction of an order-based prediction set following equation (3). To test if an element $\boldsymbol{y}^{(k)}$ in the sample space $\mathcal{Y}$ is included in a prediction set for a given vector $\boldsymbol{o}$ and known event probability vector $\boldsymbol{\theta}$, the cumulative sum of event probabilities of the categories corresponding to the minimum element of $\boldsymbol{o}$ up to element k, following the ordering of $\boldsymbol{o}$, is computed. If this cumulative sum is greater than the error rate α, then element k is included in the prediction set. As a result, all elements with such cumulative sums greater than α are included in the prediction set. The elements with such cumulative sums less than or equal to α are not included. Therefore, by construction, $P(\boldsymbol{Y} \notin A_\alpha^{\theta,o} \mid \boldsymbol{\theta}) \leq \alpha$. Consequently,

$$P(\boldsymbol{Y} \in A_\alpha^{\theta,o} \mid \boldsymbol{\theta}) = 1 - P(\boldsymbol{Y} \notin A_\alpha^{\theta,o} \mid \boldsymbol{\theta}) \geq 1 - \alpha,$$

so $A_\alpha^{\theta,o}$ is α-valid.

Proof of Theorem 1. (1): Note first that $\boldsymbol{o}$ enters equation (3) only through the ordering of its elements. As such, without loss of generality, consider vectors of the form $\boldsymbol{o} \in \{0, 1\}^K$. When constructing a prediction set following equation (3) based on such a vector $\boldsymbol{o}$, the category space is effectively divided into two disjoint subsets,

$$\mathcal{Y}_0 = \{\boldsymbol{y}^{(k)} \in \mathcal{Y} : o_k = 0\}$$
$$\mathcal{Y}_1 = \{\boldsymbol{y}^{(k)} \in \mathcal{Y} : o_k = 1\},$$

such that, by construction,

$$A_\alpha^{\theta,o} = \begin{cases} \mathcal{Y} & \text{if } \sum_{\{j : \boldsymbol{y}^{(j)} \in \mathcal{Y}_0\}} \theta_j > \alpha \\ \mathcal{Y}_1 & \text{else} \end{cases}.$$

For a given error rate α, clearly any α-valid prediction set may be constructed under considerations of permutations of the vector $\boldsymbol{o}$ of the form $\boldsymbol{o} \in \{0, 1\}^K$.

(2): We wish to show no other ordering results in an α-valid prediction set with strictly smaller cardinality than $\boldsymbol{o}^\theta$. In words, consider switching the ordering of one element at a time. This will always result in a prediction set with the same cardinality or greater cardinality than that under $\boldsymbol{o}^\theta$. More formally, let $\tilde{\boldsymbol{o}} = \{o_1^\theta, o_2^\theta, \ldots, o_{k^*}^\theta, o_{k^*-1}^\theta, o_{k^*+1}^\theta, \ldots, o_K^\theta\}$, that is, equivalent to $\boldsymbol{o}^\theta$ with the (k^*)th and $(k^* - 1)$th ordering flipped. But, by construction, $\theta_{k^*} \geq \theta_{k^*-1}$, so $|A_\alpha^{\theta,o^\theta}| \leq |A_\alpha^{\theta,\tilde{o}}|$. □

Proof of Theorem 2. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{N+1} \sim \text{i.i.d.} MN_K(\boldsymbol{\theta}, 1)$. Then, we wish to construct a conformal prediction set for $\boldsymbol{Y}_{N+1}$ based on an observation of $\boldsymbol{X} = \sum_{i=1}^N \boldsymbol{Y}_i$, and some conformity measure C.

To determine if a candidate category $k \in \{1, \ldots, K\}$ is included in a $1 - \alpha$ conformal prediction set, the conformal algorithm proceeds as follows, see Section 2.1 of Bersson and Hoff (2024) for more details:

1. Set $\boldsymbol{y}_{N+1} = \boldsymbol{y}^{(k)}$ where $\boldsymbol{y}^{(k)}$ is a vector of length K with a 1 in the kth index and 0s elsewhere.
2. For $j = 1, \ldots, N+1$, compute conformity scores $c_j = C(\boldsymbol{x} - \boldsymbol{y}_j + \boldsymbol{y}^{(k)}, \boldsymbol{y}_j)$.
3. Set

$$p_k = \frac{\{\#j \in \{1, \ldots, N+1\} : c_{N+1} \geq c_j\}}{N+1}$$

More compactly, and by symmetry in the problem, this conformal p-value may be equivalently written as:

$$p_k = \sum_{l=1}^{K} \mathbb{1}(o_k^* \geq o_l^*) \frac{x_l + y_l^{(k)}}{N+1},$$

where $o_k^* = c_{N+1}$ and $o_l^* = c_j$ for a $j \in \{1, \ldots, N\}$ such that $\boldsymbol{y}_j = \boldsymbol{y}^{(l)}$. Then, for prediction mis-coverage rate α, the category k is included in the prediction set if $p_k > \alpha$. A prediction set constructed from this procedure may be concisely written as follows:

$$A_\alpha(\boldsymbol{x}) = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^{K} \mathbb{1}\left(o_k \geq o_l\right) \frac{x_l + y_l^{(k)}}{N+1} \right] > \alpha \right\},$$

for some $\boldsymbol{o} \in \mathbb{R}^K$. □

Appendix B: Maximization of the Marginal Multinomial-Dirichlet Likelihood

In this section, we detail a Newton–Raphson algorithm to maximize the log marginal likelihood of a conjugate multinomial-Dirichlet model, sometimes referred to as the Dirichlet-multinomial compound distribution:

$$\begin{aligned} \boldsymbol{X}_j &\sim MN_K(\boldsymbol{\theta}_j, N_j), \text{ independently for } j = 1, \ldots, J \\ \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J &\sim \text{Dirichlet}_K(\boldsymbol{\gamma}). \end{aligned}$$

The log likelihood of the marginal likelihood is as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}) \propto \sum_{j=1}^{J} \Bigg[&\log \Gamma\left(\sum_{i=1}^{K} \gamma_i \right) - \log \Gamma\left(N_j + \sum_{i=1}^{K} \gamma_i \right) \\ &+ \sum_{i=1}^{K} \log \Gamma(x_{j,i} + \gamma_i) - \sum_{i=1}^{K} \log \Gamma(\gamma_i) \Bigg]. \end{aligned}$$

Define

$$\Psi(s) = \frac{d}{ds}\log\Gamma(s) = -\xi + \sum_{n=0}^{\infty}\left[\frac{1}{n+1} - \frac{1}{n+s}\right],$$

where ξ is the Euler–Mascheroni constant. Then, it is straightforward to obtain the first and second derivatives of the marginal log likelihood,

$$\frac{d}{d\gamma_k} = \sum_{j=1}^{J}\left[\Psi\left(\sum_{i=1}^{K}\gamma_i\right) - \Psi\left(N_j + \sum_{i=1}^{K}\gamma_k\right) + \Psi(x_{j,k} + \gamma_k) - \Psi(\gamma_k)\right]$$

$$\frac{d}{d\gamma_k^2} = \sum_{j=1}^{J}\left[\Psi'\left(\sum_{i=1}^{K}\gamma_i\right) - \Psi'\left(N_j + \sum_{i}\gamma_i\right) + \Psi'(x_{j,k} + \gamma_k) - \Psi'(\gamma_k)\right]$$

$$\frac{d}{d\gamma_k d\gamma_{k'}} = \sum_{j=1}^{J}\left[\Psi'\left(\sum_{i=1}^{K}\gamma_i\right) - \Psi'(N_j + \sum_{i=1}^{K}\gamma_i)\right],$$

where Ψ' is the trigamma function. Let $\boldsymbol{g}$ be the gradient vector of length K and $\boldsymbol{H}$ the Hessian matrix. Finally, Newton's method updates $\boldsymbol{\gamma}$ as follows:

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} - \boldsymbol{H}^{-1}(\boldsymbol{\gamma}^{(t)})\boldsymbol{g}(\boldsymbol{\gamma}^{(t)}),$$

where the algorithm is iterated until convergence.

Solving for the MLE of the unknown parameter in the Dirichlet-multinomial compound distribution is covered in literature dating back to at least (Mosimann, 1962), and it is stated in Wallach (2008) that the likelihood is concave. For a detailed overview of the Dirichlet-multinomial compound distribution, see Ng et al. (2011). For derivations and comparisons of various algorithms to solve the maximization problem, including computational expense and initialization sensitivity, we refer the reader to Minka (2000) and Wallach (2008).

References

Arnold Z. J., Wenger S. J., & Hall R. J. (2021). Not just trash birds: Quantifying avian diversity at landfills using community science data. *PLoS One*, *16*(9), 1–14. https://doi.org/10.1371/journal.pone.0255391

Bersson E., & Hoff P. D. (2024). Optimal conformal prediction for small areas. *Journal of Survey Statistics and Methodology*, *n/a*(2325-0992), smae010. https://doi.org/10.1093/jssam/smae010

Camerini G., & Groppali R. (2014). Landfill restoration and biodiversity: A case of study in Northern Italy. *Waste Management and Research*, *32*(8), 782–790. https://doi.org/10.1177/0734242X14545372

Christensen M. F., & Hoff P. D. (2022). A flexible and interpretable spatial covariance model for data on graphs. *Environmetrics*, *n/a*(n/a), e2879. https://doi.org/10.1002/env.2879

Faulkenberry G. D. (1973). A method of obtaining prediction intervals. *Journal of the American Statistical Association*, *68*(342), 433–435. https://doi.org/10.1080/01621459.1973.10482450

Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., & Rubin D. B. (2014). *Bayesian data analysis* (3rd ed.). CRC Press, Taylor & Francis Group.

Hoff P. (2023, May). Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, *29*(2), 901–928. https://doi.org/10.3150/22-BEJ1484

Lebrun J. J., Thogmartin W. E., & Miller J. R. (2012, August). Evaluating the ability of regional models to predict local avian abundance. *Journal of Wildlife Management*, *76*(6), 1177–1187. https://doi.org/10.1002/jwmg.374

Lele S. R. (2020). How should we quantify uncertainty in statistical inference? *Frontiers in Ecology and Evolution*, *8*, 1–18. https://doi.org/10.3389/fevo.2020.00035

Minka T. P. (2000). *Estimating a Dirichlet distribution* (Technical report).

Mosimann J. E. (1962). On the compound multinomial distribution, the multivariate beta-distribution, and correlations among proportions. *Biometrika*, *49*(2), 65. https://doi.org/10.2307/2333468

Ng K. W., Tian G.-L., & Tang M.-L. (2011, May). Dirichlet–multinomial distribution. In *Dirichlet and related distributions : Theory, methods and applications* (1st ed., Vol. 895, chapter 6, pp. 199–225). John Wiley & Sons, Incorporated.

Rao J. N. K., & Molina I. (2015). *Small area estimation* (2nd ed.). John Wiley and Sons, Inc.

Shanahan D. F., & Possingham H. P. (2009, October). Predicting avian patch occupancy in a fragmented landscape: Do we know more than we think? *Journal of Applied Ecology*, *46*(5), 1026–1035. https://doi.org/10.1111/j.1365-2664.2009.01694.x

Sullivan B., Wood C., Iliff M., Bonney R., FInk D., & Kelling S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Sullivan B. L., Phillips T., Dayer A. A., Wood C. L., Farnsworth A., Iliff M. J., Davies I. J., Wiggins A., Fink D., Hochachka W. M., Rodewald A. D., Rosenberg K. V., Bonney R., & Kelling S. (2017, April). Using open access observational data for conservation action: A case study for birds. *Biological Conservation*, *208*, 5–14. https://doi.org/10.1016/j.biocon.2016.04.031

Tang B., Clark J. S., Marra P. P., & Gelfand A. E. (2023, March). Modeling community dynamics through environmental effects, species interactions and movement. *Journal of Agricultural, Biological, and Environmental Statistics*, *28*(1), 178–195. https://doi.org/10.1007/s13253-022-00520-3

Thatcher A. R. (1964). Relationships between Bayesian and confidence limits for predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 176–210. https://doi.org/10.1111/j.2517-6161.1964.tb00551.x

Tian Q., Nordman D. J., & Meeker W. Q. (2022, November). Methods to compute prediction intervals: A review and new results. *Statistical Science*, *37*(4), 580–597. https://doi.org/10.1214/21-STS842

Twedt D. J., Tirpak J. M., Jones-Farrand D. T., Thompson F. R., Uihlein W. B., & Fitzgerald J. A. (2010, August). Change in avian abundance predicted from regional forest inventory data. *Forest Ecology and Management*, *260*(7), 1241–1250. https://doi.org/10.1016/j.foreco.2010.07.027

Vovk V., Gammerman A., & Shafer G. (2005). *Algorithmic learning in a random world*. Springer.

Wallach H. M. (2008). *Structured topic models for language*. University of Cambridge.

Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 674–699
https://doi.org/10.1093/jrsssa/qnae108
Advance access publication 21 October 2024
Discussion Paper

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society
A

Extreme-value modelling of migratory bird arrival dates: insights from citizen science data

Jonathan Koh[1] and **Thomas Opitz**[2]

[1]Institute of Mathematical Statistics and Actuarial Science, Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[2]Biostatistics and Spatial Processes (UR546), INRAE, Avignon, France

Address for correspondence: Jonathan Koh, Institute of Mathematical Statistics and Actuarial Science, Oeschger Centre for Climate Change Research, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland. Email: jonathan.koh@stat.math.ethz.ch

[To be read before The Royal Statistical Society at the Discussion Meeting on the Analysis of citizen science data’ held at the Societys 2024 annual conference in Brighton on Tuesday, 3 September 2024, the President, Dr Andrew Garrett, in the Chair]

Abstract

Citizen science mobilizes many observers and gathers huge datasets but often without strict sampling protocols, resulting in observation biases due to heterogeneous sampling effort, which can lead to biased predictions. We develop a spatio-temporal Bayesian hierarchical model for bias-corrected estimation of arrival dates of the first migratory bird individuals at their breeding sites. Higher sampling effort could be correlated with earlier observed dates. We implement data fusion of two citizen-science datasets with fundamentally different protocols (Breeding Bird Survey, eBird) and obtain posterior distributions of the latent process, which contains four spatial components endowed with Gaussian process priors: species niche; sampling effort; position and scale parameters of annual first arrival date. The data layer consists of four response variables: counts of observed eBird locations (Poisson); presence–absence at observed eBird locations (Binomial); BBS occurrence counts (Poisson); first arrival dates (generalized extreme-value). We devise a Markov chain Monte Carlo scheme and check by simulation that the latent process components are identifiable. We apply our model to several migratory bird species in the northeastern US for 2001–2021 and find that the sampling effort significantly modulates the observed first arrival dates. We exploit this relationship to effectively bias-correct predictions of the true first arrivals.

Keywords: Bayesian hierarchical model, bias correction, bird phenology, opportunistic data, sampling effort, species distribution

1 Introduction

1.1 The rise of citizen science

Defining citizen science and its boundaries is difficult (Haklay et al., 2021). Broadly, the field involves a collaborative approach to scientific inquiry that engages volunteers and nonprofessionals in the collection, analysis, and interpretation of data. This participatory model is not new. Friedrich Wilhelm Herschel, a musician by training, discovered the planet Uranus in 1781 using his telescopes. Charles Darwin conducted crowd-sourcing projects in the 19th century by recruiting acquaintances and travellers to write to him about their observations (Browne, 1996). The longest-running citizen science initiative is the annual Christmas Bird Count (Bock & Root, 1981), a census administered by the National Audubon Society of wintering birds in the Western Hemisphere by volunteer birdwatchers since 1900.

Received: November 24, 2023. **Revised:** July 15, 2024. **Accepted:** August 9, 2024

Citizen science has gained significant traction over the past two decades, spurred on by growing world literacy levels and the advent of technological devices and infrastructures for gathering, reporting, sharing, and storing data (Newman et al., 2012; Wynn, 2017). These datasets are often collected at relatively low cost and sometimes with unconventional funding sources (Silvertown, 2009). Zooniverse,[1] the world's largest platform for volunteer-based research, has seen a surge to over 1.9 Million registered volunteers since its inception in 2009. Climate*prediction*.net runs climate modelling experiments using the home computers of thousands of volunteers. There is now a peer-reviewed journal (Bonney, Cooper, et al., 2016) dedicated to disseminating research on citizen science, and the 2019 Citizen Science Association biannual conference attracted 818 registered delegates from 28 countries.

Citizen science is expected to grow in importance. Technological progress continually improves how machines and citizen science work together; for example, volunteer-classified training sets have already been used to improve the performance of machine learning approaches in astronomy (Marshall et al., 2015). Fraisl et al. (2020) note that traditional data sources are not sufficient for measuring the United Nations Sustainable Development Goals (UN, 2015), and sources from citizen science are required to better inform policies and actions. Citizen-science initiatives can also have community-level impacts on the participants (Jordan et al., 2012), such as by empowering them, giving them a voice in local environmental decision-making (Bonney, Phillips, et al., 2016), and increasing public awareness of the scientific process. Though citizen science's global economic value has been estimated to exceed 2.5 billion USD annually, the majority of data collected through citizen-science initiatives has not yet reached analysis in peer-reviewed literature (Theobald et al., 2015). There is a need to develop and disseminate statistical methods that facilitate wider scientific use of these datasets.

1.2 Analysis of data for species monitoring

Much of the recent growth in citizen science has occurred in the ecological and environmental sciences (Fraisl et al., 2022; McKinley et al., 2015; Pocock et al., 2018). A prominent example in ecology is the eBird project (Cornell Lab of Ornithology, 2022), launched in 2002, which has led to a database of over 200 Gigabytes providing information on over 600 million bird observations. Another important bird monitoring program within the scope of citizen science is the North American Breeding Bird Survey (BBS). It was launched in 1966 for official monitoring purposes and engages large numbers of trained volunteers in collecting standardized data on bird populations according to a rather strict sampling protocol along several selected routes and during each year. BBS data can be considered of high quality but sampling is relatively sparse in space and occurs only during fixed periods in the year, whereas the sampling of eBird data is much more heterogeneous but provides a generally denser spatio-temporal coverage, as we will detail in Section 2. We exploit these two data sources in this work. Regarding nonavian species, another recent example, among many others, is the deployment of camera traps for mammal monitoring (Hsing et al., 2022).

Many observation data in ecology are *opportunistic*, i.e. they were collected incidentally or without a predefined research question in mind, so information on the criteria applied by the observer for sampling and reporting observations is limited. However, opportunistic data often allow for a spatio-temporal coverage of species monitoring that would not be attainable with protocol-based data collected by professionals at relatively sparse and preselected sampling locations. Citizen-science data are particularly valuable for statistical inference on low-probability events (e.g. the presence of rare species individuals, or unusual or extreme phenological events) and on occurrence times of events, such as the arrival of migratory birds at their breeding site in spring. The general data quality from citizen-science initiatives is high (Kosmala et al., 2016), although the need for statistical methods to account for different data biases prevails. Isaac et al. (2014) discuss bias-correction approaches for ecological trend estimates, and conclude that opportunistic data would be further enhanced if information on the sampling effort at the data collection points could be captured. They further identify four key dimensions of biases induced by volunteer sampling: uneven sampling in time and space, uneven effort per visit, and uneven detectability of species individuals, which can vary by observer, species, and land cover. For example, detectability could be higher in open landscapes than in dense forests. Moreover, the expertise of individual observers for detecting and identifying species may increase over time due to

[1] https://www.zooniverse.org/

increasing training, experience, and knowledge-sharing with other observers, thus reducing observation biases (Johnston et al., 2018; Kelling et al., 2015).

Using eBird, statistical modelling of sampling effort expressed through various criteria was discussed by Tang et al. (2021). More generally, a large body of literature has emerged to characterize heterogeneous sampling effort and quantify it from available data (e.g. Fink et al., 2020; Gelfand & Shirota, 2019; Johnston et al., 2021, 2023).

Another problem arises when only the presence of species is reported in a nonexhaustive way but not their absence (*presence-only data*). Then, areas and times without any reported occurrences could correspond either to the true absence of the species or to the presence of the species but the absence of any observer. The BBS and eBird datasets do not explicitly report absences, but both rely on observation protocols requiring that all species detected at the sampled location and identified by the observer be reported, leading to exhaustive sampling for detected and identified bird individuals. In eBird, data entries satisfying this protocol are known as *checklist data*, with a unique space–time coordinate associated with each checklist; these account for the vast majority of eBird entries. Checklist data in eBird are quality-controlled by experts, and nonexpert observers can train in birding with resources offered by eBird, such as tutorials, online courses and bird identification apps. Observer skill for detecting and correctly identifying various bird species still varies across observers, even for experts. Due to the large number of observers contributing to the eBird dataset, our model will capture an observation effort that represents the average observation skill.

We adopt the common approach of considering species not reported within a checklist as being absent from the space–time location of the checklist. By systematically adding such pseudo-absence information to the dataset, we obtain a binary observation for each combination of checklist and species, with 1 corresponding to the presence and 0 corresponding to (pseudo-) absence of the species. We can then estimate the species presence probability using available predictors with a binomial likelihood.

1.3 Extreme-value analysis for migratory bird arrivals

Extreme-value theory (EVT, Coles, 2001) is a branch of probability theory and statistics that originated in the first half of the 20th century. It provides a theoretically justified framework to model extreme events, i.e. the tails of data distributions. It has been extensively used in fields such as finance and environmental sciences, especially for climate data. Applications in ecology have so far mostly focused on extremes of abiotic environmental processes (Gaines & Denny, 1993; Katz et al., 2005) which can have a strong influence on biotic processes, for example when very low winter temperatures contribute to limiting the extent of outbreaks of forest pests (Thibaud et al., 2016). However, direct applications of EVT to biotic variables in ecology are scarce, due perhaps to the strongly discretized nature of species observation data, often collected in the form of presence–absence information or relatively small occurrence counts. By contrast, standard EVT is more focused on variables measured on a continuous scale. Regarding human biology, some extreme-value studies exist and have considered extreme sports performances (e.g. for professional swimming competitions, Spearing et al., 2021) and longevity (Belzile et al., 2022). Recent ecological studies leveraging EVT aim to model species accumulation curves (Borda-de Água et al., 2021), extremes of species movements (Wijeyakulasuriya et al., 2019) or the first arrival dates of migratory birds (Wijeyakulasuriya et al., 2024). However, the aforementioned approaches do not explicitly consider the specific observation biases inherent to ecological datasets, which is the focus of our work.

The first arrivals of migratory birds at their destination for a given year are an example of phenological events, i.e. of recurring life history events of species, such as migration or breeding. Bird migration is a complex phenomenon taking place over large geographic scales (Somveille et al., 2015), with possible temporal trends in migration patterns resulting from changes in climate, land cover and other ecological and evolutionary factors (Conklin et al., 2021; Cotton, 2003). The study of migration patterns, and particularly of arrival times at the breeding site, has attracted strong interest (e.g. Linden, 2011; Youngflesh et al., 2021). Approximately 20% of all bird species are concerned by migration, which allows birds to adapt to seasonal cycles in climatic stresses and in the availability of resources such as food, especially during their breeding period.

Previous studies often used rather complex approaches to define and estimate a date that can be viewed as representative of the arrival of birds at their breeding site. Throughout this paper, we

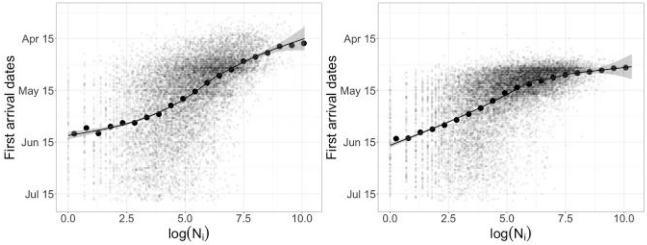

Figure 1. Scatterplots of the logarithm of positive checklist counts (x-axes) and first arrivals (y-axes, with later dates corresponding to lower values) observed at each pixel-year combination for *Chimney Swift* (left) and *Chestnut-sided Warbler* (right). The larger dots show 20 binned estimates, the black line a smooth fitted curve, and the shaded region the 95% pointwise confidence intervals of the curve.

will simply write 'first arrival' to refer to the calendar date of first arrival. Youngflesh et al. (2021) based the representative date for a given area and year on both the first arrival of an individual bird and the first local maximum of the species' detection probabilities estimated from eBird data. In our approach, we focus only on the first arrival for each pixel of a spatial mesh covering the study area, i.e. we model the minimum of all dates of occurrence of the species during the year. We use the generalized extreme-value (GEV) distribution motivated by EVT to model this sample extreme of all observed occurrence times in the year, similar to Wijeyakulasuriya et al. (2024). We treat dates as a continuous variable, which is sensible since the distribution of observed first arrivals spans over a long enough period so that the discretization effects inherent to dates have a weak influence; Figure 1 suggests that observed dates span over several months. Although distributions for discrete extremes have been proposed in the literature (e.g. Hitz et al., 2024; Prieto et al., 2014; Ranjbar et al., 2022), they usually come with overhead for numerical computation and modelling. Moreover, asymptotic theory is limited to heavy-tailed variables, by contrast with our setting where tails have a natural finite bound and are therefore light-tailed.

In the northeastern US, birds arrive during spring from more southern regions where they have spent overwintering. Our goal is to analyse how the probability distribution of the first arrivals varies across bird species and space in this region, and in response to factors related to climate and land cover. The mapping of the arrival at breeding sites should be restricted to locations where birds are present during the year, i.e. locations that provide a suitable environment for birds to breed, and so the considered locations must be part of the spatial area forming the ecological niche occupied by the bird species for breeding. We thus design our statistical model to predict both the niche and the first arrivals. At locations where the species presence probability is very low according to the niche model, we will not map arrival times.

To our knowledge, the modelling approach developed here is the first that aims to appropriately capture the interplay of the niche, the sampling effort and the observed first arrivals to provide bias-corrected predictions of the true first arrivals. Wijeyakulasuriya et al. (2024) used only eBird data and focused on modelling the spatial dependence among the first arrivals for each year using the class of max-infinitely divisible models, so that first arrivals at locations without observations during certain years can be predicted. We extend their approach and consider the niche and the sampling effort when predicting first arrivals, and we focus on revealing spatial patterns that remain stable across the whole study period. The approach of Youngflesh et al. (2021) is based on a different metric for the arrival dates, with a less direct and intuitive interpretation, and it does not account for sampling effort except for choosing a study area with a relatively high overall sampling effort in eBird.

1.4 Bayesian hierarchical modelling of complex ecological data

Observation data in ecology often do not directly measure the latent (i.e. not directly observed) processes of interest, such as the ecological niche, the sampling effort and the timing of a phenological event in our case. This is due to observation biases and complex interactions among such processes. Moreover, different data sources (e.g. eBird and BBS) can contribute complementary information about the same latent process (e.g. the ecological niche). The data we use here provide a concrete illustration. The eBird checklists are available for spatio-temporal locations chosen by the observers and provide generally good spatio-temporal coverage of the study area, especially for the most recent years in the study period. By contrast, the BBS observations are only available along predefined routes around 40 km long, with up to 50 stops separated by around 800 m at which observers can report occurrence numbers of detected bird species. Time intervals for observation are also imposed by the study protocol. Therefore, observation always takes place at prespecified spatio-temporal locations in BBS, with the land-cover type at those locations marked by the presence of a usually relatively large road. As a consequence, BBS data may be less representative of all possible land-cover types in comparison to eBird, and also provide no direct information about events taking place outside the prescribed observation time interval, such as first arrivals. On the other hand, within BBS many routes have been sampled during each year since the 1970s, so the temporal coverage of BBS over the full study period is more homogeneous and complete than eBird. Combining information from both datasets offers the possibility of improved inferences for properties of the niche and of phenological events of bird species.

With Bayesian hierarchical models (BHMs, Banerjee et al., 2003), data are assumed to be generated conditional on latent processes, which in turn are conditioned on hyperparameters (e.g. the variance or the spatial correlation range). This makes it possible to account for spatial patterns, complex relationships, and uncertainties. Using Bayes' Theorem, the combination of prior information based on expert knowledge and the data likelihood results in posterior distributions for latent processes and hyperparameters that reflect updated beliefs about the processes of interest given the observed data (van de Schoot et al., 2021).

An interesting application of BHMs in ecology is the fusion of presence-only data, available through large datasets but with strongly heterogeneous and unknown sampling effort, with presence–absence data, available through smaller datasets but with known sampling effort, to infer species distribution maps (e.g. Gelfand & Shirota, 2019) using so-called integrated species distribution models. An often important element of such approaches is the inclusion of a latent Gaussian field that represents spatial heterogeneity in preferential sampling, similar to the foundational work of Diggle et al. (2010).

To infer the complex BHM, we design for mapping the niche, observation effort, and first arrivals, we devise a Markov chain Monte Carlo (MCMC) scheme with Vecchia likelihood approximations (Katzfuss & Guinness, 2021; Vecchia, 1988) and Metropolis-adjusted Langevin algorithm (MALA, Roberts & Rosenthal, 2001) updates for the latent Gaussian fields. Our approach marks the first use of the Vecchia approximation in Bayesian hierarchical modelling for spatial extremes, extending other recent approaches in EVT (Huser et al., 2023; Majumder et al., 2024).

1.5 Outline of the paper

Section 2 presents the datasets and extensive preprocessing steps. In Section 3, we recall the basics of EVT and introduce the BHM, we use to identify the spatial variability of three aspects (sampling effort; niche of a given species; date of arrival of the first individuals during spring migration for a given species). Details of latent Gaussian process models and MCMC inference are presented in Section 4, together with a simulation study that confirms the statistical identifiability of the model components. Results for estimated first arrivals and other model components are presented for several species in Section 5. A discussion and an outlook towards future research related to our approach and to citizen-science data more generally concludes the paper in Section 6.

2 Datasets and preprocessing steps

2.1 Bird observation data

We extracted bird observation data from eBird and BBS databases for the study period 2001–2021. Both databases provide separate data files for each state in the selected study area in the northeastern US composed of the nine following states: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. We merged the extracted data into a single dataset. We used the full observation record of the eBird Basic Dataset, where each entry reports the observation of one or several individuals of a species, and retained the following attributes: name of species, count of observed individuals, longitude, latitude, date, duration of observation (i.e. what we consider as temporal sampling effort), and a flag indicating if the observation is part of a so-called checklist. Recall that a checklist is an observation event where observers report all species they observe, provided they succeed in identifying the species. For BBS, we extracted information about observed species at the level of the survey route and year.

The first arrivals of migratory birds in the study area are known to take place at the earliest towards the end of March. We identify migratory species as having no observations during the winter months of December and January in eBird. For the migratory species characterized by this property, we remove a few February occurrences that are understood as carryovers from the preceding autumn season. Only migratory species that had at least 200 occurrences in eBird for 2021 and were also observed in BBS are kept. This selection procedure led us to consider around 50 species as migratory, and we apply our modelling approach to the 10 species with the largest number of reported presences in eBird: Red-eyed Vireo, Eastern Wood-Pewee, Chimney Swift, Great Crested Flycatcher, Warbling Vireo, Veery, Chestnut-sided Warbler, Magnolia Warbler, Purple Martin, Blackburnian Warbler, in decreasing order of occurrence entries.

The following two data tables were generated from eBird:

- *eBird checklists:* each row corresponds to one checklist in eBird, with attributes longitude, latitude, year, and duration of observation;
- *eBird species occurrence:* for each migratory species, the table contains the same number of rows as there are checklists, and each row contains the attributes of the checklist, the name of the species, and the binary presence–absence flag to indicate if the species was observed. Species observations flagged as ‘flyover’ in eBird (i.e. where birds did not nest or breed near the observation location) were declared as absences.

We define four response variables for the components of our regression model based on BBS species counts and eBird data aggregated to a regular pixel grid of 20 km width. This pixel size is similar to the spatial mapping unit used for most official eBird communications and provides a good compromise between relatively noisy data behaviour at very small scales and the intended modelling of spatial variability at relatively small scales. Our response variables are:

- *BBS species counts:* the number of observed birds, available for each combination of migratory species, route, and year, where only routes surveyed during the specific year are considered; we use the Poisson distribution for this variable;
- *eBird checklist count:* the number of available checklists, calculated for each configuration of year and pixel; we use the Poisson distribution for this variable;
- *eBird occurrence count:* the number of checklists (among all available checklists) for which the species was reported present, calculated for each configuration of species, year, and pixel; we use the binomial distribution for this variable;
- *eBird first arrivals:* the minimum date of observation, expressed as the number of days since 1st January, calculated for each configuration of species, year, and pixel where at least one observation of the species occurred; we use the GEV distribution for the negated variable, which therefore represents a maximum.

Some of the above datasets and their components are illustrated in Figure 3.

2.2 Effect of observation effort on first arrivals

Figure 1 shows observed first arrivals plotted against positive checklist counts, for each pixel-year combination and two species. Observed dates tend to occur later during the year when fewer checklists are available, i.e. when sampling effort was lower in space. This gives strong motivation for modelling the influence of sampling effort on observed dates, and then using the estimated relationship to bias-correct predictions of true dates by setting the sampling effort to a very high, saturating value during the prediction step.

Figure 2 further explores whether the duration of observation during a checklist helps explain the residuals obtained after fitting a local regression curve in Figure 1. We detect a slight effect of

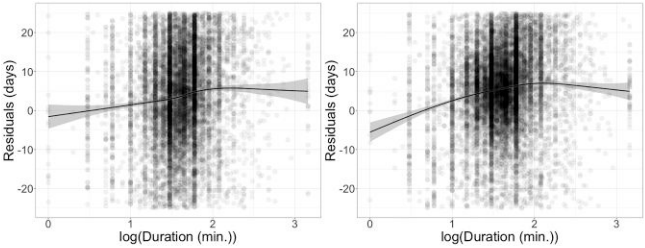

Figure 2. Scatterplots of the residuals (in days) of the smooth local regression fitted to first arrivals in Figure 1 (with later dates corresponding to lower values) as a function of the logarithm (with base 10) of duration (in minutes) of observation (*x*-axis), for *Chimney Swift* (left) and *Chestnut-sided Warbler* (right). To improve readability, the range of values shown for the residuals does not include all points.

duration for the two example species in Figure 2, with relatively earlier arrivals arising for longer durations. This suggests that both the number of checklists and the duration of observation per checklist are relevant components of the overall sampling effort modulating the distribution of the observed first arrival of birds at their breeding sites, with higher effort associated with earlier observation of the first arrival.

To explain the remaining variability in the residuals in Figures 1 and 2, we conjecture that other factors related to spatial variation in bird phenology and in species presence are relevant. We construct our model to estimate spatial effects for these properties and then use them to correct predicted first arrivals.

2.3 Climate and land-cover data

The North Atlantic Oscillation (NAO) index describes changes in the strength of two recurring pressure patterns in the atmosphere over the North Atlantic: a low near Iceland, and a high near the Azores Islands. Positive NAO indicates that these features are strong, creating a big pressure difference between them. Strongly positive values are linked to warm conditions across the Eastern US and Northern Europe, and cold conditions across Southern Europe. Negative NAO indicates that these features are relatively weak, and the pressure difference between them is smaller. Strongly negative NAO is linked to cold conditions in the Eastern US and Northern Europe, and warm conditions in Southern Europe. As a covariate at annual temporal resolution, we use the average NAO for March, which was already identified as a relevant climate predictor for bird migration timing by Wijeyakulasuriya et al. (2024).

We use the 2021 National Land Cover Database (NLCD) of the US to extract land-cover information. To facilitate model construction and identifiability, we will not use land-cover variables as covariates within our BHM. Instead, we will compare posterior maps of the niche, observation effort and components of the first arrival distribution with land cover maps to detect and interpret significant correlations between these model components and land cover. Land cover proportions for the pixel grid are reported in the Appendix (Figure A1) for four land-cover types, defined by merging original NLCD categories: Developed areas (including all areas with buildings and infrastructure such as roads); Forest; Vegetation (excluding forest but including planted and cultivated land); Water (including wetlands).

Figure 3 shows a schematic summarizing the different data sources and preprocessing steps of our approach, and the overall workflow of the model. The plot in Panel 1 illustrates a strong increase in eBird checklists and in spatial coverage between the beginning and the end of the study period; the plot in Panel 2 highlights that the median duration per checklist varies substantially in both space and time. In 2001, we see relatively longer durations in specific spatial regions, while in 2021 the durations are more homogeneous across larger areas and also generally longer. The display in Panel 3 shows the occurrence dates of observations of a species pooled together for a pixel-year, and we extract the minimum day of the year, which will serve as an observation for the GEV response variable in our BHM. Finally, the figure in Panel 4 shows the number of BBS stops in 2001 and 2021.

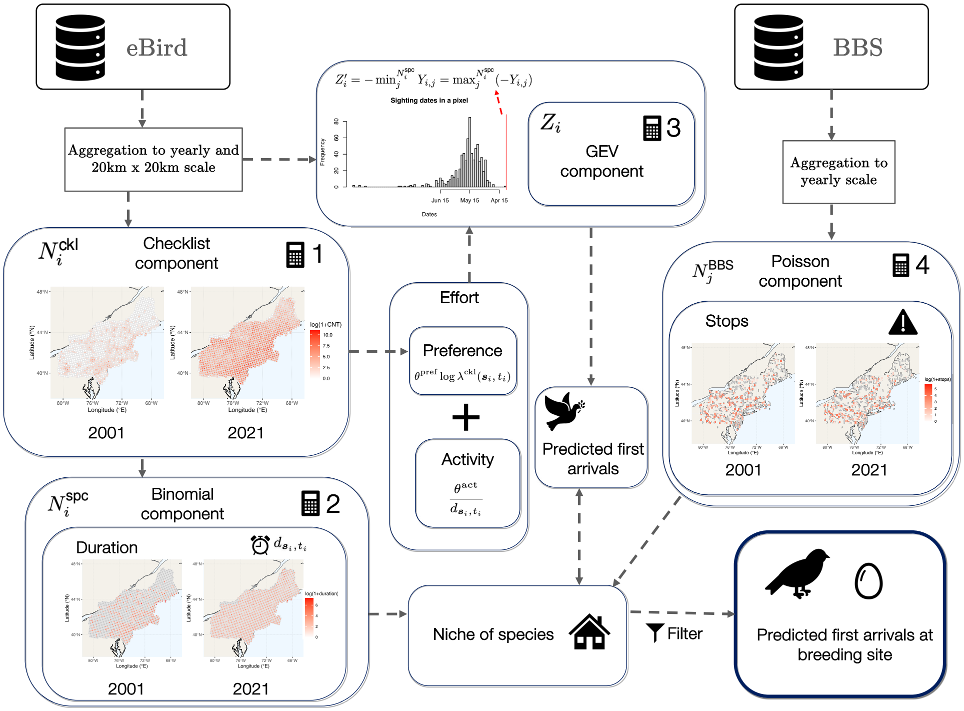

Figure 3. Schematic summarizing the data sources, the data preprocessing steps and the four components in the data layer of the BHM (panels labelled with numbers 1–4). The dashed lines illustrate our modelling choices and how model components interact, but they need not show the chronology of how these components are estimated, since the whole model is estimated jointly using a fully Bayesian approach. The plots in the panels numbered '1', '2', and '4' show the number of available eBird checklists, the median duration spent per checklist in each pixel, and the number of stops along a route from the BBS data, respectively, in 2001 (left display in each panel) and 2021 (right display).

3 Bayesian hierarchical model

3.1 General setting and notations

Figure 3 highlights the four components in the data layer of the BHM we propose, and we here detail its notations and structure.

We write $\mathcal{S} \times \mathcal{T}$ for the space–time study domain, and denote by A_i, $i = 1, \ldots, D \times T$, the set of nonintersecting space–time cells based on a division of the study area $\mathcal{S}$ into $D = 1{,}268$ pixels, replicated over $T = 21$ years. We identify each cell A_i with a representative location $(\boldsymbol{s}_i, t_i)$, such as its barycenter, where $\boldsymbol{s}_i \in \mathcal{S}$ and $t_i = 1, \ldots, T$.

Let N_i^{ckl} be the number of eBird checklists in A_i. If $N_i^{\text{ckl}} > 0$, then $N_i^{\text{spc}} \in \{0, 1, \ldots, N_i^{\text{ckl}}\}$ denotes the sum of the binary indicators of detecting a given species from each of the N_i^{ckl} checklists. Binary indicators equal 1 if the species was detected and 0 otherwise. Given $N_i^{\text{spc}} > 0$, we write $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,N_i^{\text{spc}}}) \in [1, 365]^{N_i^{\text{spc}}}$ to denote the corresponding first arrivals in $\boldsymbol{s}_i$ within the year t_i from the N_i^{spc} observed species occurrences. Based on these arrival dates, we calculate $Z'_i = -\min_j^{N_i^{\text{spc}}} Y_{i,j} = \max_j^{N_i^{\text{spc}}} (-Y_{i,j})$, the first arrival in year t_i within $\boldsymbol{s}_i$. We set $Z_i = -\log(-Z'_i/366)$ to reflect that Z'_i has a natural lower bound at -366, and model $Z_i > 0$ instead.

3.2 Extreme-value distribution for first arrivals

Provided that the partitioning of $\mathcal{S}$ provides enough 'effectively independent' arrival times within each pixel-year A_i, EVT (Coles, 2001) motivates using a GEV to model the transformed first arrivals in A_i. Invoking EVT, we assume convergence in distribution of each Z_i to a limiting GEV as $N_i^{\mathrm{spc}} \to \infty$.

Let M_n denote the maxima of independent and identically distributed random variables $X_1, \ldots, X_n$. If M_n converges in distribution to a nondegenerate distribution after linear renormalization, then this distribution must be a GEV; this theoretical result was further extended and shown to hold for maxima over dependent data, although under technical assumptions that are difficult to formally validate in applications. In practice, for fixed n large enough (where $n = N_i^{\mathrm{spc}}$ in our setting), one uses the approximation

$$\Pr(M_n \le z) \approx \begin{cases} \exp\left\{-\exp\left(-\dfrac{z-\mu}{\sigma}\right)\right\}, & \xi = 0, \\ \exp\left\{-\left[1+\xi\left(\dfrac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & \xi \neq 0, \end{cases} \tag{1}$$

where the right-hand side of (1) is the distribution function of the GEV, defined on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$ with location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and shape $\xi \in \mathbb{R}$. The shape parameter determines the tail behaviour of the GEV: for $\xi > 0$, the GEV has support with a finite lower endpoint and an upper heavy tail of power-law form; for $\xi < 0$, the GEV has a finite upper endpoint but is unbounded below; for $\xi = 0$, the support of the GEV has no finite bounds, and the upper tail exhibits an exponential decay rate. In our application, some sample sizes N_i^{spc} can be very small and dates $Y_{i,j}$, $j = 1, \ldots, N_i^{\mathrm{spc}}$ can be dependent, but the three-parameter GEV remains a flexible model to accommodate relevant distributional properties of the maxima Z_i, even in cases where the theoretical asymptotics are not reached. The transformation from Z_i' to Z_i is very close to linear in a large neighbourhood of $Z_i' = 100$ (corresponding to approximately mid-April), so it does not substantially modify the shape ξ but the parameters μ and σ. We incorporate linear predictors into the GEV regression equation used to model Z_i in our BHM: one for the location μ and another for the scale σ.

3.3 Data layer of the BHM

We construct a system of four regression equations with latent Gaussian processes used in the linear predictors. We denote the response variables, as described in Section 2.1, as follows: N_i^{ckl}—eBird checklist count per pixel-year i; N_j^{BBS}—BBS counts of the species of interest per year-route j; N_i^{spc}—eBird occurrence count for the species of interest per pixel-year i; Z_i—transformed negated first arrivals for the species of interest in eBird per pixel-year i.

For the response variable N_i^{ckl} reporting the checklist counts, we weight the Poisson intensities according to the surface areas $\mathcal{A}_i$ of the pixels (i.e. we use an offset of $\log \mathcal{A}_i$ in the linear predictor), since some pixels at the boundary of the study region have smaller area. For the response variable N_j^{BBS} reporting the BBS occurrence counts, observations are made along a subset of 50 equidistant stops along each of the routes, and we weight the Poisson intensities accordingly with the number of stops (between 1 and 50) that were visited. To alleviate notations in the predictor formulas below, we do not explicitly write down these weights related to pixel surface areas and road lengths. Moreover, routes can span across several pixels, so we construct the Poisson intensity parameters as a weighted mean of the parameters for the involved pixels, with weights w_k defined as the percentage length inside each of the cells A_k.

The hierarchical system of the four regression equations is

$$\begin{aligned} N_j^{\mathrm{BBS}} \mid \lambda^{\mathrm{BBS}}, \boldsymbol{\theta}_{\mathrm{bbs}} &\sim \mathrm{Pois}\left\{\sum_{k \in \mathrm{route}_j} \omega_k \lambda^{\mathrm{BBS}}(\boldsymbol{s}_k; \boldsymbol{\theta}_{\mathrm{bbs}})\right\}, \\ N_i^{\mathrm{ckl}} \mid \lambda^{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{ckl}} &\sim \mathrm{Pois}\left\{\lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{ckl}})\right\}, \\ N_i^{\mathrm{spc}} \mid N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}, \boldsymbol{\theta}_{\mathrm{spc}} &\sim \mathrm{Bin}\{N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{spc}})\}, \\ Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma &\sim \mathrm{GEV}\{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_\sigma), \xi\}, \end{aligned}$$

where

$$\boldsymbol{\theta}_{\text{bbs}}, \boldsymbol{\theta}_{\text{ckl}}, \boldsymbol{\theta}_{\text{spc}}, \boldsymbol{\theta}_{\mu}, \boldsymbol{\theta}_{\sigma} \sim \text{Hyperpriors}$$

are hyperparameters for the random predictors λ^{BBS}, λ^{ckl}, p^{spc}, μ, and σ that govern the different model components; we discuss their specifics next. Checklist occurrences form a point pattern, and due to the structure of the N_i^{ckl}-related model component, we model them as a spatio-temporal log-Gaussian Cox point process (Møller et al., 1998), discretized according to the cells A_i.

3.4 Structure of latent Gaussian processes

We write $X(\bullet)$ to denote any latent Gaussian effect, here either indexed by pixels $s \in \mathcal{S}$ or years $t \in \mathcal{T}$. We use Gaussian process priors for all latent effects. We further use the exponential covariance function to parametrize the spatial effects, given as follows for two locations s_1 and s_2:

$$\text{Cov}\{X(s_1), X(s_2)\} = \sigma^2 \exp(-\|s_1 - s_2\|/\kappa), \quad \sigma, \kappa > 0,$$

with Euclidean distance $\|\bullet\|$, standard deviation $\sigma > 0$ and range parameter $\kappa > 0$.

The four spatial Gaussian process priors included in our model are

$$\begin{array}{ll} X^{\text{pref}}(\bullet) \sim \mathcal{GP}(\boldsymbol{\omega}_1), & X^{\text{niche}}(\bullet) \sim \mathcal{GP}(\boldsymbol{\omega}_2), \\ X^{\text{GEV}-\mu}(\bullet) \sim \mathcal{GP}(\boldsymbol{\omega}_3), & X^{\text{GEV}-\sigma}(\bullet) \sim \mathcal{GP}(\boldsymbol{\omega}_4), \end{array} \tag{2}$$

where $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3$, and $\boldsymbol{\omega}_4$ contain individual range and standard deviation parameters that control the spatial dependence of each process. These hyperparameters are assigned identical joint penalized complexity priors (Fuglstad et al., 2018). We also impose sum-to-zero constraints on all spatial effects to aid with identifiability. Spatial patterns of the niche of the species are captured by X^{niche}, whereas spatial patterns of the sampling effort are modelled through X^{pref}. Moreover, we include a temporal latent effect X^{year} to capture the variability in checklist counts across years, and set the prior $X^{\text{year}}(\bullet) \sim \mathcal{GP}(\boldsymbol{\omega}_5)$, where $\boldsymbol{\omega}_5$ contains the range and standard deviation parameters of a temporal exponential covariance function.

The structure of link functions and linear predictors, explained in detail subsequently, is:

$$\begin{aligned} \log \lambda^{\text{BBS}}(s_i) &= \beta_0^{\text{BBS}} + X^{\text{niche}}(s_i), \\ \log \lambda^{\text{ckl}}(s_i, t_i) &= \beta_0^{\text{ckl}} + X^{\text{year}}(t_i) + X^{\text{pref}}(s_i), \\ \text{cloglog}\{p^{\text{spc}}(s_i, t_i)\} &= \beta_0^{\text{spc}} + X^{\text{niche}}(s_i) + \frac{\beta^{\text{act}}}{d_{s_i,t_i}}, \\ \mu(s_i, t_i) &= g\{\beta_0^{\text{GEV}-\mu} + X^{\text{GEV}-\mu}(s_i) + \beta_1^{\text{GEV}-\mu}\text{NAO}_{t_i} \\ &\quad + \theta^{\text{niche}-\text{GEV}} X^{\text{niche}}(s_i), x_{\text{effort}}(s_i, t_i)\}, \\ \log \sigma(s_i) &= \beta_0^{\text{GEV}-\sigma} + X^{\text{GEV}-\sigma}(s_i). \end{aligned}$$

The climate covariable, i.e. the NAO value for year t_i, is denoted as NAO_{t_i}, and we include it inside the link function g (detailed below) as a fixed effect modulating the upper bound of the GEV location parameter. The scaling parameter $\theta^{\text{niche}-\text{GEV}}$ controls the sharing of the niche spatial effect. All scaling hyperparameters and intercepts in the model, along with ξ, are given flat normal prior distributions with mean 0 and variance 100. The complementary log-log link function $\text{cloglog}(p) = \log(-\log(1-p))$ for binomial probabilities allows interpreting them as the probabilities of observing at least one count in a Poisson model, which is equal to $1 - \exp(-\lambda)$, with Poisson intensity $\lambda = \exp(x)$. This ensures consistency with the Poisson model for the BBS counts.

We differentiate between two dimensions of sampling effort that we combine into x_{effort}. The first one is modelled through the space–time intensity λ^{ckl} of checklists, which we call the *preference*. We hypothesize that it influences the first arrivals Z_i, where a higher *preference* likely leads to more and thus earlier sightings in the year. The second dimension of sampling effort that we consider here is the median duration of observation per checklist in a pixel-year d_{s_i,t_i}, which we call the *activity*. This effect could influence both the binomial probability of observing the species of interest in a checklist, and again the first arrivals. A higher *activity* could increase the probability of observing the species in a checklist, which could also induce an earlier observed first arrival. We combine the two effects *preference* and *activity* into

$$x_{\text{effort}}(s_i, t_i) = \theta^{\text{eff}} + \theta^{\text{pref}} \log \lambda^{\text{ckl}}(s_i, t_i) + \frac{\theta^{\text{act}}}{d_{s_i,t_i}} \tag{3}$$

with real-valued scaling hyperparameters θ^{pref} (typically expected to be positive) and θ^{act} (typically expected to be negative), and an intercept θ^{eff} representing a baseline sampling effort that is modulated according to the checklist abundance and observer activity. With this structure, setting an infinite duration d_{s_i,t_i} causes the activity term to vanish.

The next subsection explains the choice we make for the nonlinear link function g of the GEV location parameter.

3.5 Saturated sharing towards the GEV predictor

Sharing latent processes across carefully chosen regression equations, as implemented in Section 3.4, enables us to leverage data from several response variables to identify the four latent processes. Specifically, we want to identify the effect of spatio-temporally heterogeneous sampling effort and correct it in predictions, such as those of the true first arrivals.

The most often used approach of sharing latent effects between regression equations is via a linear structure where the linear predictor of one regression equation has an additive component $X(\bullet)$, while another has an additive component $\beta \times X(\bullet)$, with a scaling factor $\beta \in \mathbb{R}$. This yields models akin to the preferential-sampling framework of Diggle et al. (2010) and Pati et al. (2011), and to Koh et al. (2023) and Yadav et al. (2023) in the context of joint modelling of occurrences and sizes of wildfires or landslides. However, such linear sharing is not flexible enough for sharing the sampling effort towards the linear predictor of the μ parameter of the GEV for first arrival dates. Letting the effort tend to infinity with linear sharing would lead to first arrivals tending to 1st January, the smallest possible value. As illustrated by the exploratory data analysis in Figure 1, the observation effort effect on first arrivals saturates at very high levels. Therefore, we share x_{effort} with the μ component of the GEV through the following upper-bounded function:

$$g(x_{\text{bound}}, x_{\text{effort}}) = \frac{\exp(x_{\text{bound}})}{1 + \exp(-x_{\text{effort}})}, \tag{4}$$

where the first argument determines the upper bound in the GEV location parameter, which is modulated by the sampling effort in the second argument and can be attained with unlimited (i.e. infinite) effort. The g-function is always positive, which is appropriate here when considering as response the transformed dates $Z_i > 0$. Figure 4 illustrates that the class of functions (4) is bounded and monotonically increasing, therefore allowing the effect of the sampling effort on the first arrivals to saturate with high effort.

The structure we put into (4) is also inspired by species accumulation curves in ecology (Colwell et al., 1994), but here we replace the species abundance with the first observed arrival dates. To our knowledge, such an approach has not yet been considered in the literature. It establishes a representation that is both parsimonious and ‘physically meaningful’, a feature that is desirable in statistical modelling in various contexts such as landslide prediction (Opitz et al., 2022) or temperature extremes (see the 2024 preprint "Integration of physical bound constraints to alleviate shortcomings of statistical models for extreme temperatures" from Noyelle et al.).

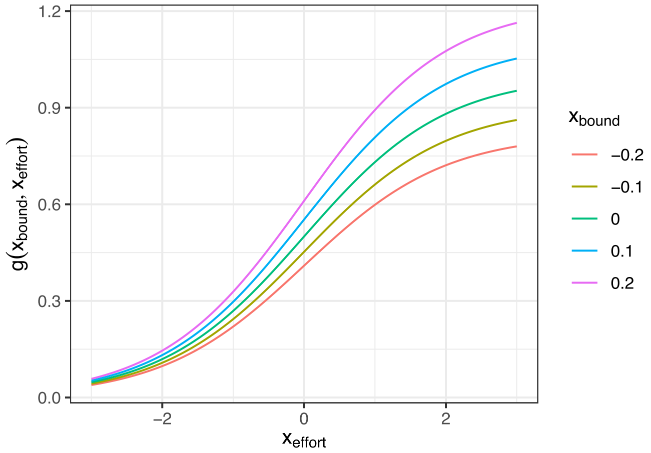

Figure 4. Plots of the nonlinear sharing function g in (4) for varying x_{effort} (x-axis) and x_{bound} (colours).

4 Simulation-based Bayesian inference

Fast and relatively accurate off-the-shelf implementations of Bayesian hierarchical inference exist, such as the INLA framework (Rue et al., 2017). However, due to certain specifics of our model, such as the nonlinear sharing of latent spatial effects and the inclusion of linear predictors in several parameters of the GEV, it is not feasible to use them here. Instead, we develop an MCMC sampler for simulation-based inference.

4.1 Vecchia approximation for Gaussian components

Due to the large number of spatial cells in our case study ($D = 1{,}268$), keeping spatial latent Gaussian effects and their conditional distributions numerically tractable is challenging. Various solutions have been proposed to tackle this fundamental problem in spatial statistics with large-dimensional data, including Gauss–Markov random fields used to approximate certain types of covariance functions based on stochastic partial differential equations (SPDE approach, Lindgren et al., 2011), or nearest-neighbour Gaussian processes (NNGPs, Datta et al., 2016). Here, we use another flexible and popular option to construct approximate Gauss–Markov representations of any Gaussian covariance function through the Vecchia (1988) approximation, which leads to a sparse Cholesky factor of the precision matrix. We write $\{x_1, \ldots, x_D\}$ for the set of observations of a Gaussian field X evaluated at locations $s_1, \ldots, s_D \in \mathcal{S}$, and we consider a permutation $m: \{1, \ldots, D\} \to \{1, \ldots, D\}$ defining a reordering of the observations. Based on this reordering, we can write $H(i; m) = \{j \in \{1, \ldots, D\} : m(j) < m(i)\}$ for the 'history' of the ith index based on the permutation m, with $\boldsymbol{x}_{H(i;m)}$ denoting the corresponding subvector of observations. For a chosen permutation m, the exact joint density of the observations can be written as the product of conditional densities, i.e.

$$f(x_1, \ldots, x_D) = f(x_{m(1)}) \prod_{i=2}^{D} f(x_{m(i)} \mid \boldsymbol{x}_{H(i;m)}).$$

The Vecchia approximation still represents a valid Gaussian process, but assumes that

$$f(x_1, \ldots, x_D) \approx \hat{f}(x_1, \ldots, x_D) = f(x_{m(1)}) \prod_{i=2}^{D} f(x_{m(i)} \mid \boldsymbol{x}_{S(i;m)}), \tag{5}$$

where $S(i; m) \subseteq H(i; m)$ and $|S(i; m)| = k$. The approximation reduces the conditioning history in the conditional densities, which decreases the computational complexity of evaluating the density from $\mathcal{O}(D^3)$ to $\mathcal{O}(Dk^3)$. This reduction is considerable if $k \ll D$, as in our setting with $D > 1{,}000$. An active research area (Katzfuss & Guinness, 2021) is to understand how one should choose the permutation m and order the conditioning variables in (5) to obtain approximations that satisfy certain criteria of optimality. Vecchia (1988) suggested to order the conditioning variables lexicographically based on their spatial coordinates, but this approach has been shown to be inefficient (Guinness, 2018). Instead, we follow the ordering proposed by the latter study, which ensures that points are ordered in a quasi-random fashion. We select $k = 5$ conditioning locations in our case study to make inference feasible. The simulation study detailed in Section 4.3 confirms that estimation works well.

4.2 Proposal scheme for MCMC

We draw posterior inference on the parameters using a general MCMC sampling scheme with customized Metropolis–Hastings (MH) updates. The hyperparameters are updated using Gibbs sampling. The most computationally intensive part of our scheme involves sampling from the latent Gaussian components, which are updated jointly through a MALA (Roberts & Rosenthal, 2001) designed to speed up mixing of the chains. We write $\boldsymbol{x}$ for the current state of the latent Gaussian component of interest, $\boldsymbol{x}^\star$ for a proposal for that component, and $\boldsymbol{y}$ for the data vector. A discrete approximation of the Langevin diffusion implies that MALA is an MH algorithm with proposal distribution

$$(\boldsymbol{x}^\star \mid \boldsymbol{x}) \sim N_p\left(\boldsymbol{x} + \frac{\delta^2}{2}\nabla_\theta \log \pi(\boldsymbol{x} \mid \boldsymbol{y}),\, \delta^2\, \mathbf{W}_\mathbf{p}\right), \tag{6}$$

where tuning parameters are the prewhitening matrix $\mathbf{W}_\mathbf{p}$, used to account for the correlation between parameters within the component, and the stepsize δ. For example, if we set $\boldsymbol{x} = \{X^{\text{niche}}(s_1),\, \dots,\, X^{\text{niche}}(s_D)\}$, the gradient in (6), when updating this component, is

$$\begin{aligned}\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x} \mid \boldsymbol{y}) \approx\, & \nabla_{\boldsymbol{x}} \log \pi_{\text{BBS}}(\boldsymbol{N}^{\text{BBS}} \mid \boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log \pi_{\text{spc}}(\boldsymbol{N}^{\text{spc}} \mid \boldsymbol{x}) \\ & + \nabla_{\boldsymbol{x}} \log \pi_{\text{GEV}}(\boldsymbol{Z} \mid \boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log \pi_{\text{Vecchia}}(\boldsymbol{x}),\end{aligned}$$

where the first three terms represent the gradients of log-likelihoods for the model components that X^{niche} is included in, and

$$\nabla_{\boldsymbol{x}} \log \pi_{\text{Vecchia}}(\boldsymbol{x}) = -\tilde{\Sigma}_{\boldsymbol{\omega}_2}^{-1}\boldsymbol{x},$$

where the $D \times D$ precision matrix $\tilde{\Sigma}_{\boldsymbol{\omega}_2}^{-1}$ is the Vecchia-approximated the precision of the Gaussian field prior assigned to X^{niche}, as detailed in §4.1 and implied by (5). The precision matrix is parametrized by the vector $\boldsymbol{\omega}_2 = (\sigma, \kappa)$, see §3.4, to which we assign a hyperprior.

For our data application, we parallelize and run separate chains for each species. We generate 80, 000 posterior samples and discard the first 60, 000 as the burn-in period. We then perform a thinning operation of the Markov chains and keep one of four consecutive samples, thus obtaining 5, 000 samples overall to perform posterior inference. We divide the number of posterior samples and the burn-in period by two for our simulation study for computational reasons. We monitor the convergence and mixing of the chains through trace plots and by assessing the effective sample size. Code implemented for the MCMC procedure is available at https://github.com/kohrrelation/mcmc_birds. To obtain a shorter burn-in period and to avoid having initial values too far from the region where posterior mass concentrates, we ran test chains that model each model component separately first (where we can estimate all parameters except the sharing parameters). When possible, we then used the last values of those chains as starting values when fitting the full model. Calculations were performed on UBELIX,[2] the HPC cluster at the University of Bern, and the computational time of our algorithm there is roughly 6 s per iteration.

[2] https://www.id.unibe.ch/hpc

4.3 Simulation study

We check by simulation that the model parameters are identifiable and can be estimated appropriately through posterior mean estimates obtained from the above Bayesian inference procedure. We especially wish to check whether the spatio-temporal field of first arrivals and the niche of the species are identifiable. This verification is important for the proposed models owing to their high structural complexity with a relatively large number of parameters. For the simulation study, we mimic our data application and choose the same spatial domain and temporal replicates, i.e. we simulate datasets at the $D = 1{,}268$ pixels over the $T = 21$ years from 2001 to 2021 with the same covariates and data structures, including the climate covariate (NAO), the number of BBS stops and the median duration in each pixel-year. The values we set for latent Gaussian processes and hyperparameters are similar to those obtained when having fitted this model to the species *Chimney Swift*. To test model robustness, we allow for potential model misspecification and overdispersion in the checklist counts. We thus add another layer of randomness to the count component, giving

$$N_i^{\text{ckl}} \mid \Lambda_i^{\text{ckl}}, \boldsymbol{\theta}_{\text{ckl}} \sim \text{Pois}(\Lambda_i^{\text{ckl}}),$$
$$\Lambda_i^{\text{ckl}} \sim \Gamma\{r, r\lambda^{\text{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\text{ckl}})\},$$

where $\Gamma(a, b)$ is the Gamma distribution with shape $a > 0$ and scale $b > 0$. As a result, N_i^{ckl} has a negative binomial distribution with mean $\lambda^{\text{ckl}}(\boldsymbol{s}_i, t_i)$ and an inflated variance $\lambda^{\text{ckl}}(\boldsymbol{s}_i, t_i) + \lambda^{\text{ckl}}(\boldsymbol{s}_i, t_i)^2/r$. We set $r = 10$ and estimate all the parameters for 100 independent replicates of such data.

The different model components are generally well identified and reproduced by our estimation procedure. The results in the Appendix (Figure A2) suggest that the posterior means of the median first arrivals estimated with this procedure are indeed essentially unbiased, and the truth is mostly well recovered. Importantly, the temporal evolution and spatial patterns are also captured satisfactorily. For example, in our simulation setting, we have spatially more homogeneous and later first arrivals for the earlier years such as 2003, and the estimated posterior means reproduce this property. The boxplots in Figure A2 suggest that first arrivals for certain pixels are harder to estimate than others, but the truth falls within the predicted interquartile ranges in most cases. We obtained similar simulation results in the setting where there is no model misspecification (not shown), i.e. $r = \infty$.

5 Results

We apply the model constructed and checked in Sections 3 and 4 to the bird data described in Section 2, where we run separate models for the 10 species with largest occurrence numbers in eBird. We report a selection of results for four species: *Chimney Swift*, *Great Crested Flycatcher*, *Chestnut-sided Warbler,* and *Purple Martin*. The MCMC trace plots in the Appendix (Figure A3) for one of the aforementioned species indicate that the chains are mixing relatively well, although some parameters like θ^{act} and $\beta_0^{\text{GEV}-\mu}$ still display autocorrelation after the burn-in period and thinning.

To compare our approach with a simpler one where only the first arrivals are considered but without modelling the niche or sampling effort, we also report results for a model using only the Z_i-variables but not N_i^{BBS}, N_i^{ckl}, and N_i^{spc}. We fit this model separately and call it the *GEV-only* model, for which $\theta^{\text{niche}-\text{GEV}}$ is null and x_{effort} set to a constant.

Figure 5 illustrates that there have been more checklists in recent years, and the model captures these temporal differences and also the spatial pattern in observed checklist numbers.

The posterior means of the sharing parameters are all positive for θ^{pref} and mostly negative (or positive with 95% posterior credible interval including zero) for θ^{act} (see Table 1), which corroborates our exploratory findings from Section 2 that higher sampling effort leads to earlier observed dates of first arrival.

Table 1 indicates that the sharing parameter $\theta^{\text{niche}-\text{GEV}}$ often has posterior credible intervals that do not contain zero, with most posterior means in the range [0, 0.15]. This means that birds tend to be observed earlier in areas where they are relatively abundant. For $\beta_1^{\text{GEV}-\mu}$, posterior means are generally hovering around [0.01, 0.03] but often not containing 0 in its credible intervals, indicating that first arrivals can take place slightly earlier in years with high NAO. Higher NAO and temperature anomaly correspond to warmer weather, which could explain earlier arrivals. However,

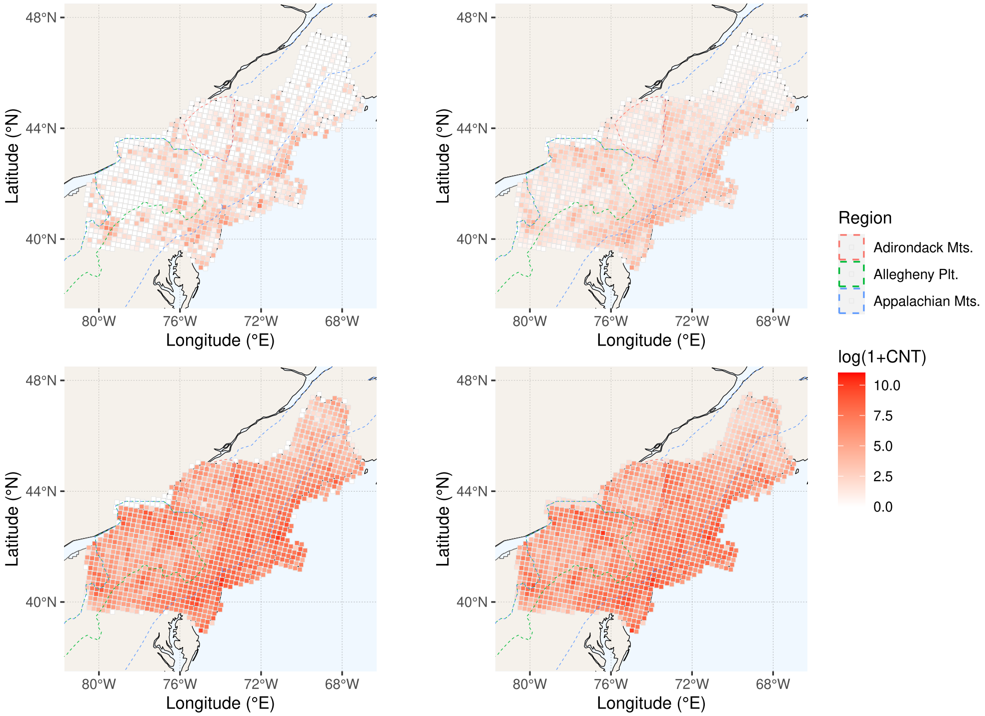

Figure 5. Observed numbers of checklists, N_i^{ckl} (left) against model-based posterior mean estimates of the Poisson mean number of checklists, $\lambda^{\text{ckl}}(\boldsymbol{s}_i, t_i)$ (right), using the model fitted to *Chimney Swift* in 2001 (top) and 2021 (bottom).

for the majority of the species, we investigate these slightly positive NAO coefficients appear not to be significant, as their credible intervals include zero. We conclude that the slight NAO effect on first arrivals could be present for some species but there is no strong signal of a significant NAO effect across all species; Wijeyakulasuriya et al. (2024) obtained a similar result when applying their analyses to the *Magnolia Warbler*.

The estimated GEV shape parameters are negative for all species and have posterior means around −0.6 to −0.9, indicating that the estimated first arrival dates have a fixed lower bound.

5.1 Excursion sets of latent spatial fields

We study the spatial patterns in the posterior estimation of the four latent spatial fields in (2) for the sampling effort, the species niche, and the μ and σ parameters of the distribution of the first arrivals.

To identify areas with very high or low values of the spatial random effects, we utilize credible sets for excursion regions (Bolin & Lindgren, 2015). We evaluate where the fields exceed or fall below certain thresholds. To visualize excursion sets simultaneously for all values of the probability level $\alpha \in (0, 1)$, Bolin and Lindgren (2015) introduced the positive and negative excursion functions $F_u^+(s) = 1 - \inf\{\alpha \mid s \in \mathrm{E}_{u,\alpha}^+\} \in [0, 1]$ and $F_u^-(s) = 1 - \inf\{\alpha \mid s \in \mathrm{E}_{u,\alpha}^-\} \in [0, 1]$, where $\mathrm{E}_{u,\alpha}^+$ is the subregion with maximal surface area in which the spatial field fully exceeds the threshold u with probability α (with an analogous definition of $\mathrm{E}_{u,\alpha}^-$ for threshold deficits below u). Figure 6 highlights these excursion functions for the species *Purple Martin*. Sampling effort is high in the southeastern areas near the ocean and strongly decreases in the further northwestern areas. The niche of the species tends to concentrate in southwestern areas. The fields related to first arrivals show more fragmented spatial patterns for this species.

5.2 Correlation of latent spatial fields with land cover

To further explain the patterns in the four latent spatial fields, we calculate the correlation of their posterior means with the land-cover maps described in Section 2 and shown in the Appendix

Table 1. Examples of predictions related to first arrivals for two pixel-years in 2022

Species	Chimney Swift	Great Crested Flycatcher	Chestnut-sided Warbler	Purple Martin
$\hat{\theta}^{\text{pref}}$	0.191 (0.184,0.202)	0.204 (0.199,0.21)	0.187 (0.183,0.191)	0.2 (0.178,0.217)
$\hat{\theta}^{\text{act}}$	−0.15 (−0.217,−0.061)	−0.818 (−0.911,−0.696)	−0.548 (−0.619,−0.454)	−0.03 (−0.269,0.236)
$\hat{\theta}^{\text{niche}-\text{GEV}}$ $(\times 10^{-2})$	4.9 (4.664,5.134)	4 (3.894,4.133)	0.2 (0.17,0.278)	6 (5.541,6.443)
$\beta_1^{\text{GEV}-\mu}$ $(\times 10^{-1})$	0.1 (−0.083,0.204)	0.2 (0.125,0.331)	0.3 (0.295,0.409)	0 (−0.384,0.526)
Observed	NA	NA	NA	NA
Predicted (GEV only)	28/05	27/05	01/07	18/06
Predicted	09/05	03/05	21/05	~~07/06~~
Debiased	03/04	13/04	03/05	~~28/03~~
Observed	01/05	04/05	04/05	29/06
Predicted (GEV only)	25/05	20/05	15/05	19/05
Predicted	09/05	15/05	12/05	~~12/05~~
Debiased	22/04	05/05	03/05	~~07/04~~

Note. The first five rows report the species and the posterior mean estimates of selected parameters, with 95% credible intervals in brackets. The next rows give the observed, predicted (with the GEV-only model, and with the full model), and debiased (with the full model) median first arrivals (in day/month format) at the two pixels indicated by the two triangles in Figure 8); the first four rows here correspond to the pixel in the southeastern area, the last four rows to the pixel located in the middle of the study region. An 'NA' value in the Observed row indicates that there were no observations within that pixel in 2022. The strikethroughs over the dates indicate that our full model estimates a species presence probability of less than 1% for that pixel-year.

(Figure A1). This can improve our understanding of the land-cover types that drive the four latent processes for a given species.

Figure 7 shows these correlations for four species. The correlations between the estimated latent field for sampling effort and land-cover proportions are very similar across species, which is expected since the checklist count data for the corresponding regression equation are the same in the four models. Sampling effort has strong positive correlation with Developed (built) areas, weaker correlation with Water-dominated or Vegetation-dominated areas, and negative correlation with Forest. Regarding the latent field for niche, it shows different patterns for different species, for example with positive correlations with Forest for species that are known to breed in forests (e.g. *Chestnut-sided Warbler*), and with Developed land cover for species known to breed in Developed areas (e.g. *Chimney Swift*). The latent fields describing the spatial variability in first arrivals show generally weaker correlations with specific land-cover types. For example, *Chimney Swift* tends to settle earlier in Developed and Vegetation areas than in Water- and Forest-dominated areas.

5.3 Predicting true first arrivals

To illustrate the utility of our model, we predict the true first arrivals in all pixels. Firstly, Figure 8 shows that the model reproduces generally well the spatial variability in the probability of observing species presence in eBird, with some differences in empirical and model-predicted probabilities due to the additional information provided by BBS data. We use this component of the model to avoid predicting first arrivals at locations where the presence of the species is uncertain in terms of the estimated niche. Therefore, we do not map predictions of first arrival dates at pixel-years where the estimated presence probability $p^{\text{spc}}\{(s_i, t_i)\}$ is smaller than 1%. This feature of our model avoids making predictions in areas where data give no clear signal of a species' breeding activity.

For the prediction of the true first arrivals, we set $\lambda^{\text{ckl}}(s_i, t_i)$ and d_{s_i,t_i} to be close to infinity to mimic infinite sampling effort, so we expect to retrieve the true and unbiased dates of the first arrival. This approach is illustrated in Figure 9 for the *Great Crested Flycatcher*, where we also show that the model captures the spatial pattern in the observed first arrivals.

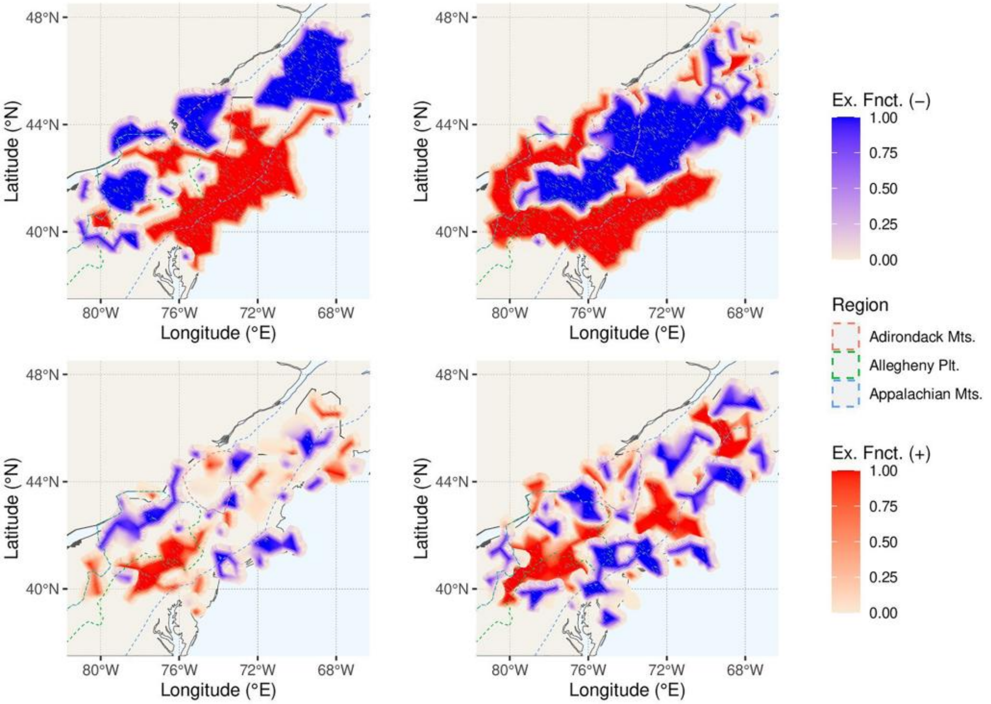

Figure 6. Excursion set analysis for the four latent spatial fields X^{effort} (top left), X^{niche} (top right), $X^{\text{GEV}-\mu}$ (bottom left), and $X^{\text{GEV}-\sigma}$ (bottom right) estimated by the model fitted to the *Purple Martin*. Thresholds are fixed at 2, 1, 0.1, and 0.1 for the four fields, using the same order.

To expand on our prediction procedure, we focus on two pixels, marked with triangles in Figure 8. The first pixel is chosen near the southwestern boundary of our study region and has 83% Forest cover but only 6% Developed land cover. It is located in a rather secluded region with generally low sampling effort. In 2022, there were no reported sightings of the four species of interest in this pixel. Table 1 highlights that the model can extrapolate and draw posterior predictive samples for that pixel-year. The second pixel is chosen in the central area of the study region. For example, a first arrival was observed in 2022 for the *Purple Martin* on the 29th of June, but Table 1 (last two rows) shows that our model flags this pixel as being uncertain to be located in the species' niche; it is thus possible that in general no breeding activity takes place in this pixel. This illustrates the capability of our model to prevent inaccurate extrapolation to spatial regions with very low species presence probability. Table 1 also shows that the GEV-only model generally estimates a later first arrival than that from the full model after correcting biases. The former predictions usually fall outside of what is expected of the four species' phenology in our study region, i.e. late spring/early summer rather than early spring.

6 Discussion and outlook

6.1 Statistical learning from participatory data

Participatory data and citizen-science programs advance and share scientific knowledge, thanks particularly to the data's large volume, diversity, and spatio-temporal coverage. However, sampling protocols for these data sources are usually lenient, with different levels of observer expertise. Statistical techniques are required to correct for heterogeneous sampling effort and observation biases.

In this work, we developed a BHM in which the sampling effort and its influence on the observation bias of a phenological event can be quantified and then corrected during prediction. Our results show that the required corrections are often substantial. We could also estimate spatial probabilities of species presence, and then filter out implausible regions when predicting the phenological event, since it might never occur.

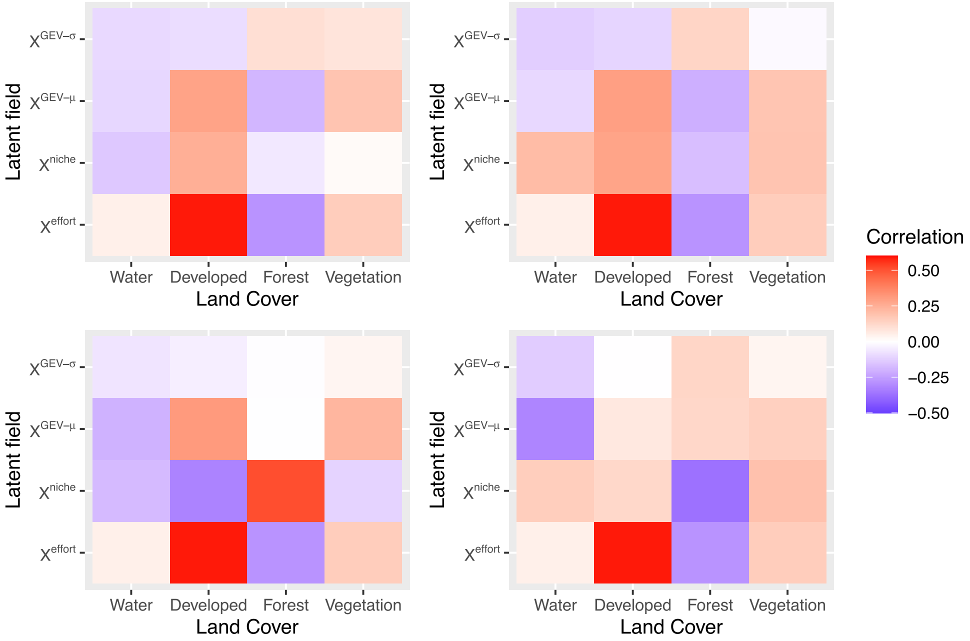

Figure 7. Correlations of land-cover proportions (*x*-axis) with posterior means of the latent fields X^{effort}, X^{niche}, $X^{\text{GEV}-\mu}$, and $X^{\text{GEV}-\sigma}$ (*y*-axis), for four species: *Chimney Swift* (top left), *Great Crested Flycatcher* (top right), *Chestnut-sided Warbler* (bottom left), and *Purple Martin* (bottom right).

6.2 Ecological data fusion using latent processes

We carefully constructed statistical models to infer and interpret latent components based on incomplete and biased observations of the true ecological processes. BHMs have many advantages: they allow us to identify various latent processes and their interactions, to separate the process describing how data are observed from the latent ecological processes we seek to reveal, to keep track of uncertainties and assess them through posterior sampling, and to incorporate expert knowledge. The Bayesian framework is also relatively robust to noisy data and overdispersion in count data thanks to the inclusion of appropriately designed random effects. Data fusion in Integrated Species Distribution Models (e.g. Miller et al., 2019, and the 2023 preprint "The Point Process Framework for Integrated Modelling of Biodiversity Data" of Adjei et al.) combining opportunistic and protocol-based data, is facilitated by shared latent components. Further extensions to better consider temporal dynamics are very actively explored in the current literature but are challenging since very complex ecological and observational dynamics must be disentangled.

Machine learning and deep learning tools are used in ecology for their skill in predicting observed data, though they may be less effective at capturing complex latent processes. Further developments are needed to allow such algorithms to predict the processes of interest when these are not directly observed, to extrapolate beyond the range of observed data (as we have done by setting the observation effort to a very high level in predictions), and to avoid propagating or even reinforcing biases from the input data when generating predictive outputs (Dunson, 2018). Future research could devise mechanisms similar to the sharing of random effects in BHMs but implemented within the architectures of general machine learning and deep learning approaches.

6.3 Modelling sampling effort

The eBird sampling effort has various dimensions. We have here considered the annual number of visits to an area (using checklists) and the median duration of each of those visits. Understanding how the combination of these two aspects influences the number of detected species individuals is a complex task (e.g. see Tang et al., 2021), and would merit exploration beyond the additive

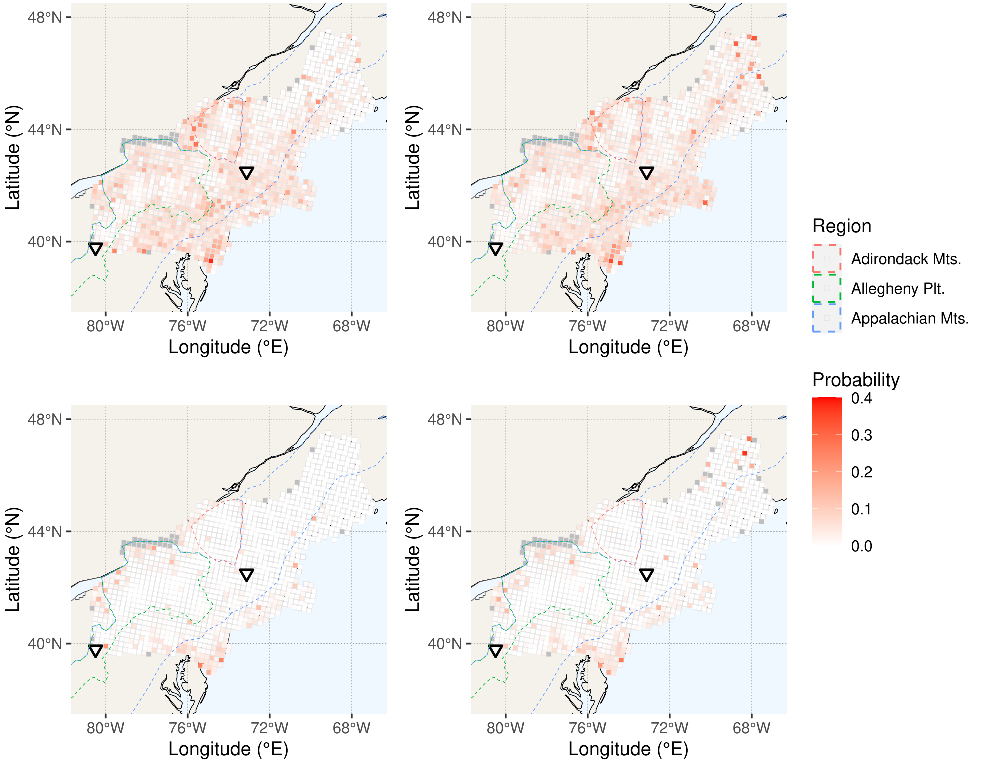

Figure 8. Empirical mean probability (left displays) of observing the species *Great Crested Flycatcher* (top) and *Purple Martin* (bottom) in checklists aggregated over all years; posterior mean probability (right displays) of observing the species from the presence model by using the mean checklist duration aggregated over all years as covariate. The two triangles are the two pixels used in the predictions of first arrivals detailed in Section 5.3.

structure of x_{effort} that we posed in (3). Other relevant attributes that we did not consider here are partially available for eBird checklists, such as the surface of the area where observations took place, or the distance over which the observers moved during the checklist event. Such information could be further studied for a more complete characterization of the sampling effort and its impact on observation biases for species distributions. Identifying sampling effort through BHMs could also be of interest in other domains, for example for reports of hail events and damages, and more generally for citizen-science data in climate science.

In cases where the performance of data fusion of opportunistic data with protocol-cased data could be improved by collecting additional exhaustive field data, it would be instrumental to conduct simulation studies using BHMs to explore how new sampling locations can be placed in an optimal and cost-efficient way.

6.4 Detectability of species

Observation biases related to the detectability of species remain a major challenge since they are difficult to identify with available data, i.e. it is possible that individuals of a species were present but not observed. This risk of nondetection can vary with land cover, species behaviour, observer experience, and other factors. In our model, detectability can be viewed as being included in the niche process, which could more precisely be called the *detectable niche*. It is difficult to devise a general statistical approach to further disentangle the niche and the detectability. Expert knowledge about the conditions of detectability could provide further insights and improve models, but including it was beyond the scope of this work. It would also require designing models adapted to species-specific properties, and could hinder simple intercomparison of models for different species.

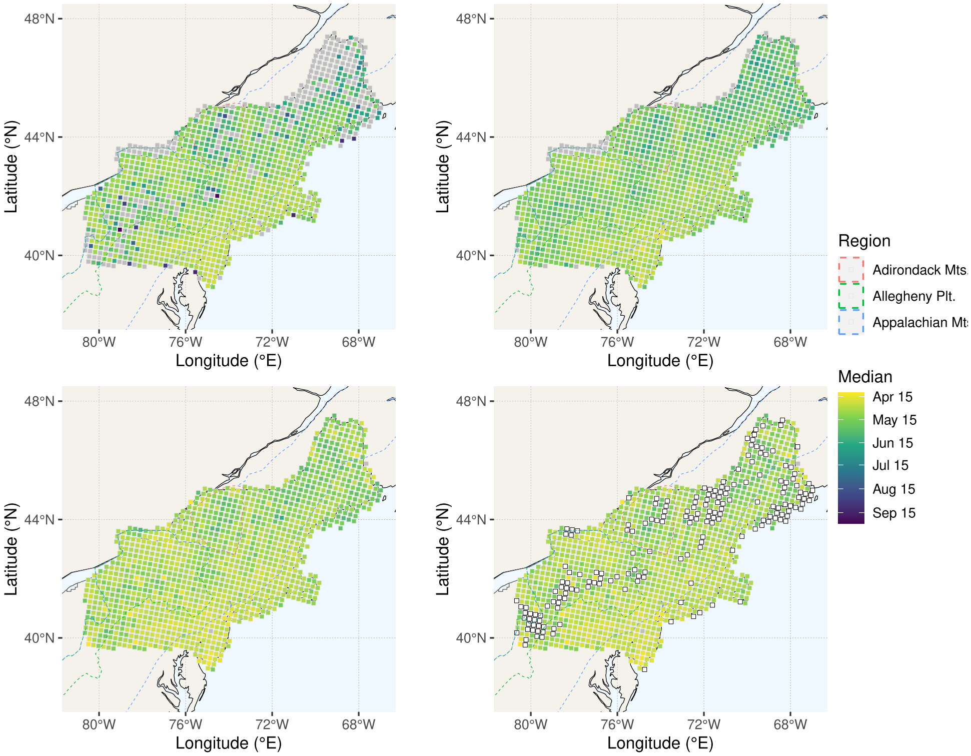

Figure 9. Maps for the species *Great Crested Flycatcher*. Observed first arrivals of *Great Crested Flycatcher* in 2022 (top left), posterior mean of the GEV median first arrivals in 2022 (top right), bias-corrected mean of the GEV median first arrivals in 2022 (bottom left), and the same plot as before though with indications (in white rectangles over the corresponding pixels) for where the model estimates a presence probability of less than 1% for the species (bottom right). The grey pixels in the top left display indicate where no species occurrences were recorded in 2022, or where no checklists and therefore no median durations were available in that pixel.

6.5 Validation of results

Validation of modelling results is inherently challenging with data that are obtained through heterogeneous sampling efforts and ridden with unknown biases. We can only validate directly whether the model appropriately reproduces data as they have been observed. For example, we performed such checks for checklist counts and observed first arrivals in Section 5. However, the latent processes of interest, as well as sampling effort and observation biases, can be entangled in complex ways. Direct validation of inferences on the latent processes of interest is thus difficult. Careful construction of interpretable models and appropriate inference algorithms is paramount.

External protocol-based data sources not suffering from these drawbacks can be used, but they often have much smaller spatial, temporal, and species coverage. The fusion of data sources, such as in our study with BBS and eBird data, can not only improve the estimation of certain model components, but could also be used for validation purposes at space–time locations covered by both datasets.

Data collection projects are increasingly initiated with advanced sensing technologies that require no human observers, therefore providing opportunities to collect large amounts of data obeying strict sampling protocols. For instance, one could track bird arrivals with radar data (e.g. Nussbaumer et al., 2021, or the vogelwarte.ch project). Camera traps combined with artificial intelligence for species identification can provide near-continuous temporal coverage, although with rather limited spatial coverage.

6.6 Possible extensions of our approach

We focused on identifying past and present spatio-temporal trends in ecological processes. In species distribution modelling, important research concerns the potential impact of climate change and its interplay with other dynamic processes (e.g. land use change). Our modelling approach could be refined to establish more detailed relationships between ecological dynamics and climate variables, to then be used for projecting first arrivals under future climate simulations, similar to Wijeyakulasuriya et al. (2024).

We have fixed the spatial pixel mesh using a size similar to that routinely used in modelling results published by the eBird project. One could explore the influence of the mesh size on the properties of the observed first arrivals. For example, a larger mesh size leads to larger numbers of observations within each pixel. This could reduce the observation bias. On the other hand, it would also decrease the spatial resolution of predictive maps and hamper the identification of relationships between ecological processes and land cover. Further research could investigate the role of the mesh choice, or how to optimally combine results from different mesh sizes.

Regarding temporal trends, the seasonal variation of sampling effort and of the species presence probability, along with how these properties interact with first arrivals, could also be modelled more precisely.

If some spatio-temporal domains are visited at high frequency, the corresponding sampling effort could be regarded as being exhaustive, and the structure of the model, e.g. the g-function in (4), could be modified accordingly to improve the identifiability of model components.

Another extension consists of studying interactions among different species in terms of their niche and phenological events, e.g. to infer which species groups show similar spring migration patterns. One could implement joint species distribution models with each having a set of regression equations and with the possibility to share certain spatial random effects between several species; this could be feasible with our current approach for a small number of species. Alternatively, one could fit models separately for each species as we have done here, and then apply classification approaches on the posterior estimates of latent model components to group species that show similar model behaviour.

6.7 Ecological extreme-value analysis

EVT could play a more prominent role in ecological science. An important domain of application is the analysis and probabilistic prediction of extreme climatic and environmental events, which can strongly influence population dynamics and biodiversity. Here, the tools from EVT could pave the way towards ecological analyses focusing not only on long-term climate averages but also on the influence of specific extreme weather events. A second domain of application, of which the present work is part, concerns the analysis of extreme phenological events using EVT. Although analyses of this type are still in their early infancy, we hope that our work will motivate more widespread applications of extreme-value techniques.

Acknowledgements

The authors thank the Editor and two anonymous referees for their numerous valuable comments and suggestions to improve the quality of the manuscript.

Conflicts of interest: None declared.

Funding

No funding sources are declared.

Data availability

The data in this article are shared publicly on https://github.com/kohrrelation/mcmc_birds.

Appendix

A.1 Land-Cover Plots

Figure A1 shows plots of the four land-cover covariates defined as proportions of certain land-cover categories in the pixels of the study area.

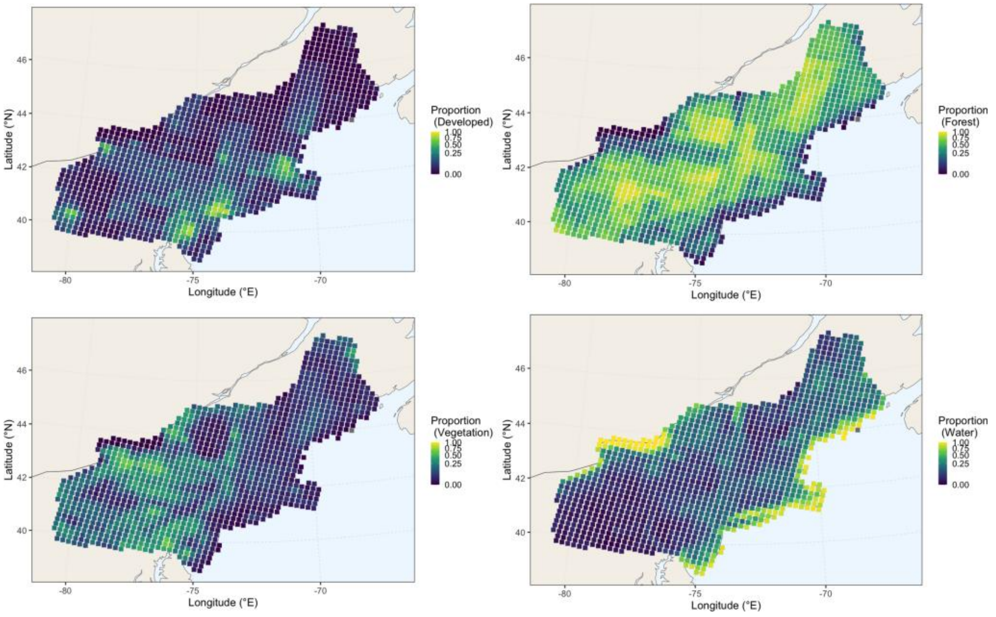

Figure A1. Land-cover proportions in the study area (NLCD, 2021) according to four types described in Section 2.

A.2 Plots from the Simulation Study

Figure A2 shows plots assessing the identifiability of the observed first arrivals in our simulation study.

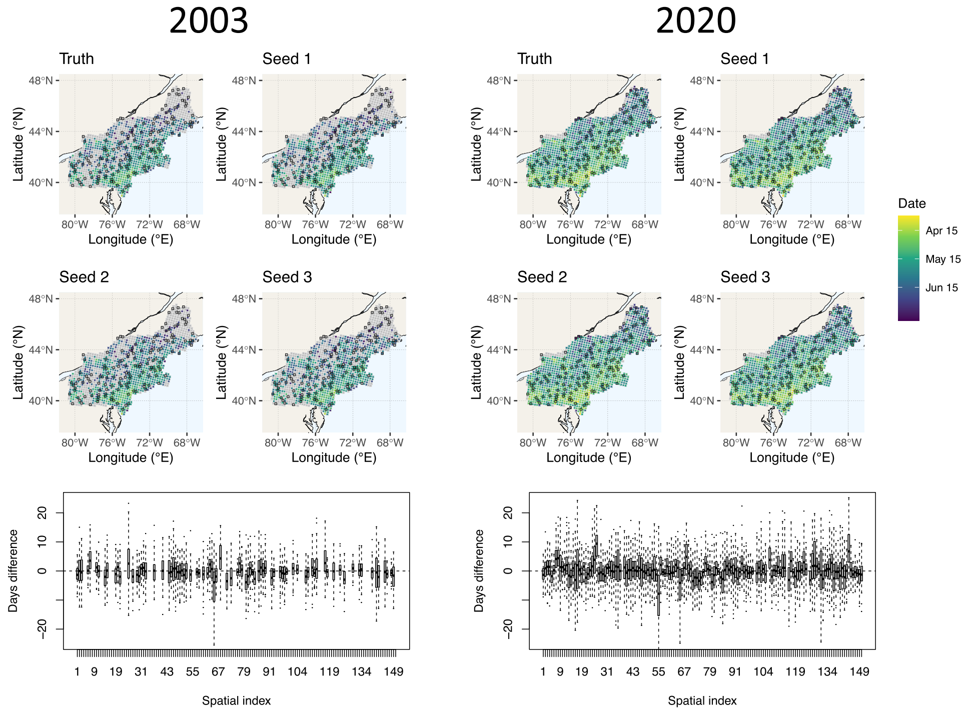

Figure A2. Simulation study: Truth vs. posterior means for the median first arrivals in 2003 (left) and 2020 (right), shown with spatial maps (for three different seeds in the simulation study), and with boxplots of the differences in days between true and predicted arrival dates, based on 100 simulations. The boxplot shows results for 150 randomly chosen pixels, marked with black boxes in the spatial plots.

A.3 Further Plots of MCMC Results

Figure A3 shows trace plots from the MCMC run of the model on the species *Great Crested Flycatcher*. Figure A4 shows spatial plots of the four latent spatial fields from the model estimated for the species *Chimney Swift*.

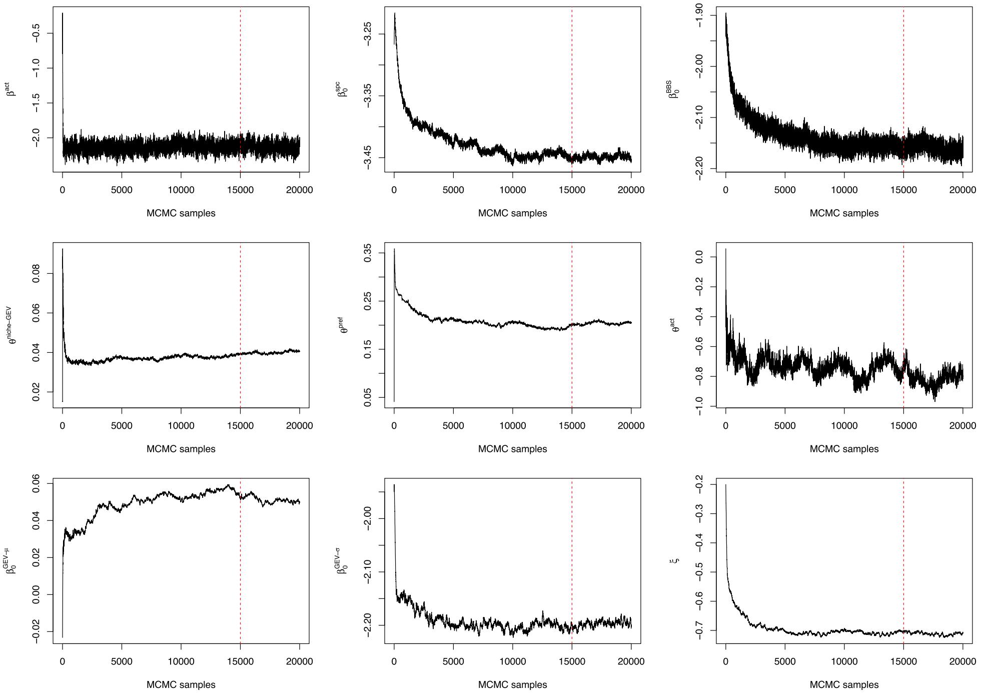

Figure A3. MCMC trace plots for some fixed effect and sharing parameters (top left to bottom right): β^{act}, β_0^{spc}, β_0^{BBS}, $\theta^{niche-GEV}$, θ^{pref}, θ^{act}, $\beta_0^{GEV-\mu}$, $\beta_0^{GEV-\sigma}$, and ξ from our model for the species *Great Crested Flycatcher*. The dashed line indicates the burn-in period.

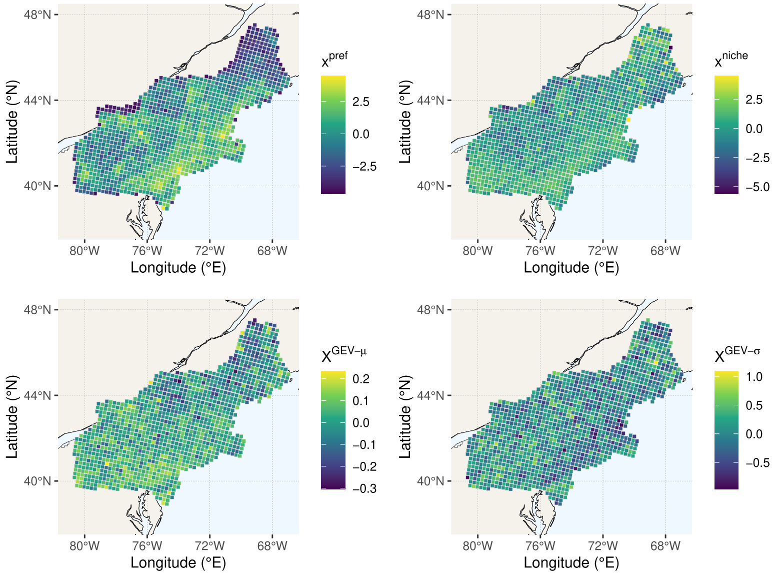

Figure A4. Posterior means plots of the four spatial random effects in (2) for the model fitted to the species *Chimney Swift*.

References

Banerjee S., Carlin B. P., & Gelfand A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Belzile L. R., Davison A. C., Gampe J., Rootzén H., & Zholud D. (2022). Is there a cap on longevity? A statistical review. *Annual Review of Statistics and Its Application*, *9*(1), 21–45. https://doi.org/10.1146/statistics.2022.9.issue-1

Bock C. E., & Root T. L. (1981). The Christmas bird count and avian ecology. *Studies in Avian Biology*, *6*, 17–23. http://hdl.handle.net/1969.3/24710

Bolin D., & Lindgren F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(1), 85–106. https://doi.org/10.1111/rssb.12055

Bonney R., Cooper C., & Ballard H. (2016). The theory and practice of citizen science: Launching a new journal. *Citizen Science: Theory and Practice*, *1*(1). https://doi.org/10.5334/cstp.56

Bonney R., Phillips T. B., Ballard H. L., & Enck J. W. (2016). Can citizen science enhance public understanding of science? *Public Understanding of Science*, *25*(1), 2–16. https://doi.org/10.1177/0963662515607406

Borda-de Água L., Alirezazadeh S., Neves M., Hubbell S. P., Borges P. A. V., Cardoso P., Dionísio F., & Pereira H. M. (2021). *Species accumulation curves and extreme value theory*, Ecology, biodiversity and conservation (pp. 211–226). Cambridge University Press.

Browne J. (1996). Charles Darwin: A biography, vol. i, voyaging. *Journal of the History of Biology*, *29*(2), 314–316. https://doi.org/10.1126/science.268.5214.1196

Coles S. (2001). *An introduction to statistical modeling of extreme values*. Springer.

Colwell R. K., Coddington J. A., & Hawksworth D. L. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *345*(1311), 101–118. https://doi.org/10.1098/rstb.1994.0091

Conklin J. R., Lisovski S., & Battley P. F. (2021). Advancement in long-distance bird migration through individual plasticity in departure. *Nature Communications*, *12*(1), 4780. https://doi.org/10.1038/s41467-021-25022-7

Cornell Lab of Ornithology (2022). eBird Basic Dataset, Version: EBD relNov-2022.

Cotton P. A. (2003). Avian migration phenology and global climate change. *Proceedings of the National Academy of Sciences*, *100*(21), 12219–12222. https://doi.org/10.1073/pnas.1930548100

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. https://doi.org/10.1080/01621459.2015.1044091

Diggle P. J., Menezes R., & Su T.-L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(2), 191–232. https://doi.org/10.1111/j.1467-9876.2009.00701.x

Dunson D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters*, *136*, 4–9. https://doi.org/10.1016/j.spl.2018.02.028

Fink D., Auer T., Johnston A., Ruiz-Gutierrez V., Hochachka W. M., & Kelling S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, *30*(3), e02056. https://doi.org/10.1002/eap.2056

Fraisl D., Campbell J., See L., Wehn U., Wardlaw J., Gold M., Moorthy I., Arias R., Piera J., Oliver J. L., Masó J., Penker M., & Fritz S. (2020). Mapping citizen science contributions to the un sustainable development goals. *Sustainability Science*, *15*(6), 1735–1751. https://doi.org/10.1007/s11625-020-00833-7

Fraisl D., Hager G., Bedessem B., Gold M., Hsing P.-Y., Danielsen F., Hitchcock C. B., Hulbert J. M., Piera J., Spiers H., Thiel M., & Haklay M. (2022). Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, *2*(1), 64. https://doi.org/10.1038/s43586-022-00144-4

Fuglstad G.-A., Simpson D., Lindgren F., & Rue H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*(525), 445–452. https://doi.org/10.1080/01621459.2017.1415907

Gaines S. D., & Denny M. W. (1993). The largest, smallest, highest, lowest, longest, and shortest: Extremes in ecology. *Ecology*, *74*(6), 1677–1692. https://doi.org/10.2307/1939926

Gelfand A. E., & Shirota S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, *89*(3), e01372. https://doi.org/10.1002/ecm.1372

Guinness J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics*, *60*(4), 415–429. PMID: 31447491. https://doi.org/10.1080/00401706.2018.1437476

Haklay M., Fraisl D., Tzovaras B. G., Hecker S., Gold M., Hager G., Ceccaroni L., Kieslinger B., Wehn U., Woods S., Nold C., Balázs B., Mazzonetto M., Ruefenacht S., Shanley L. A., Wagenknecht K., Motion A., Sforzi A., Riemenschneider D., … Vohland K. (2021). Contours of citizen science: A vignette study. *Royal Society Open Science*, *8*(8), Article 202108. https://doi.org/10.1098/rsos.202108

Hitz A. S., Davis R. A., & Samorodnitsky G. (2024). Discrete Extremes. *Journal of Data Science*, 524–536. https://doi.org/10.6339/24-JDS1120

Hsing P.-Y., Hill R. A., Smith G. C., Bradley S., Green S. E., Kent V. T., Mason S. S., Rees J., Whittingham M. J., Cokill J., Scientists M. C., & Stephens P. A. (2022). Large-scale mammal monitoring: The potential of a citizen science camera-trapping project in the United Kingdom. *Ecological Solutions and Evidence*, *3*(4), e12180. https://doi.org/10.1002/2688-8319.12180

Huser R., Stein M. L., & Zhong P. (2023). Vecchia likelihood approximation for accurate and fast inference with intractable spatial max-stable models. *Journal of Computational and Graphical Statistics*, *33*(3), 978–990. https://doi.org/10.1080/10618600.2023.2285332

Isaac N. J. B., van Strien A. J., August T. A., de Zeeuw M. P., & Roy D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*(10), 1052–1060. https://doi.org/10.1111/mee3.2014.5.issue-10

Johnston A., Fink D., Hochachka W. M., & Kelling S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*(1), 88–97. https://doi.org/10.1111/mee3.2018.9.issue-1

Johnston A., Hochachka W. M., Strimas-Mackey M. E., Gutierrez V. R., Robinson O. J., Miller E. T., Auer T., Kelling S. T., & Fink D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265–1277. https://doi.org/10.1111/ddi.v27.7

Johnston A., Matechou E., & Dennis E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103–116. https://doi.org/10.1111/mee3.v14.1

Jordan R. C., Ballard H. L., & Phillips T. B. (2012). Key issues and new approaches for evaluating citizen-science learning outcomes. *Frontiers in Ecology and the Environment*, *10*(6), 307–309. https://doi.org/10.1890/110280

Katz R. W., Brush G. S., & Parlange M. B. (2005). Statistics of extremes: Modeling ecological disturbances. *Ecology*, *86*(5), 1124–1134. https://doi.org/10.1890/04-0606

Katzfuss M., & Guinness J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, *36*(1), 124–141. https://doi.org/10.1214/19-STS755

Kelling S., Johnston A., Hochachka W. M., Iliff M., Fink D., Gerbracht J., Lagoze C., La Sorte F. A., Moore T., Wiggins A., & Wong W. K. (2015). Can observation skills of citizen scientists be estimated using species accumulation curves? *PLoS One*, *10*(10), e0139600. https://doi.org/10.1371/journal.pone.0139600

Koh J., Pimont F., Dupuy J.-L., & Opitz T. (2023). Spatiotemporal wildfire modelling through point processes with moderate and extreme marks. *The Annals of Applied Statistics*, *17*(1), 560–582. https://doi.org/10.1214/22-AOAS1642

Kosmala M., Wiggins A., Swanson A., & Simmons B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/10.1002/fee.2016.14.issue-10

Linden A. (2011). Using first arrival dates to infer bird migration phenology. *Boreal Environment Research*, *16*, 49–60. http://hdl.handle.net/10138/232775

Lindgren F., Rue H., & Lindström J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Majumder R., Reich B. J., & Shaby B. A. (2024). Modeling extremal streamflow using deep learning approximations and a flexible spatial process. *The Annals of Applied Statistics*, *18*(2), 1519–1542. https://doi.org/10.1214/23-AOAS1847

Marshall P. J., Lintott C. J., & Fletcher L. N. (2015). Ideas for citizen science in astronomy. *Annual Review of Astronomy and Astrophysics*, *53*(1), 247–278. https://doi.org/10.1146/astro.2015.53.issue-1

McKinley D. C., Miller-Rushing A. J., Ballard H. L., Bonney R., Brown H., Evans D. M., French R. A., Parrish J. K., Phillips T. B., Ryan S. F., Shanley L. A., Shirk J. L., Stepenuck K. F., Weltzin J. F., Wiggins A., Boyle O. D., Briggs R. D., Chapin III S. F., Hewitt D. A., Preuss P. W., & Soukup M. A. (2015). Investing in citizen science can improve natural resource management and environmental protection. *19*.

Miller D. A., Pacifici K., Sanderlin J. S., & Reich B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/mee3.2019.10.issue-1

Møller J., Syversveen A. R., & Waagepetersen R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, *25*(3), 451–482. https://doi.org/10.1111/1467-9469.00115

Newman G., Wiggins A., Crall A., Graham E., Newman S., & Crowston K. (2012). The future of citizen science: Emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298–304. https://doi.org/10.1890/110294

Nussbaumer R., Bauer S., Benoit L., Mariethoz G., Liechti F., & Schmid B. (2021). Quantifying year-round nocturnal bird migration with a fluid dynamics model. *Journal of The Royal Society Interface*, *18*(179), Article 20210194. https://doi.org/10.1098/rsif.2021.0194

Opitz T., Bakka H., Huser R., & Lombardo L. (2022). High-resolution Bayesian mapping of landslide hazard with unobserved trigger event. *The Annals of Applied Statistics*, *16*(3), 1653–1675. https://doi.org/10.1214/21-AOAS1561

Pati D., Reich B. J., & Dunson D. B. (2011, March). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, *98*(1), 35–48. https://doi.org/10.1093/biomet/asq067

Pocock M. J., Chandler M., Bonney R., Thornhill I., Albin A., August T., Bachman S., Brown P. M., Cunha D. G. F., Grez A., Jackson C., Peters M., Rabarijaon N. R., Roy H. E., Zaviezo T., & Danielsen F. (2018). Chapter six - a vision for global biodiversity monitoring with citizen science. In D. A. Bohan, A. J. Dumbrell, G. Woodward, & M. Jackson (Eds.), *Next generation biomonitoring: Part 2*, Volume 59 of *advances in ecological research* (pp. 169–223). Academic Press.

Prieto F., Gómez-Déniz E., & Sarabia J. M. (2014). Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention*, *71*, 38–49. https://doi.org/10.1016/j.aap.2014.05.005

Ranjbar S., Cantoni E., Chavez-Demoulin V., Marra G., Radice R., & Jaton K. (2022). Modelling the extremes of seasonal viruses and hospital congestion: The example of flu in a Swiss hospital. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *71*(4), 884–905. https://doi.org/10.1111/rssc.12559

Roberts G. O., & Rosenthal J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*(4), 351–367. https://doi.org/10.1214/ss/1015346320

Rue H., Riebler A., Sørbye S. H., Illian J. B., Simpson D. P., & Lindgren F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, *4*(1), 395–421. https://doi.org/10.1146/statistics.2017.4.issue-1

Silvertown J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*(9), 467–471. https://doi.org/10.1016/j.tree.2009.03.017

Somveille M., Rodrigues A. S., & Manica A. (2015). Why do birds migrate? A macroecological perspective. *Global Ecology and Biogeography*, *24*(6), 664–674. https://doi.org/10.1111/geb.2015.24.issue-6

Spearing H., Tawn J., Irons D., Paulden T., & Bennett G. (2021). Ranking, and other properties, of elite swimmers using extreme value theory. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(1), 368–395. https://doi.org/10.1111/rssa.12628

Tang B., Clark J. S., & Gelfand A. E. (2021). Modeling spatially biased citizen science effort through the eBird database. *Environmental and Ecological Statistics*, *28*(3), 609–630. https://doi.org/10.1007/s10651-021-00508-1

Theobald E., Ettinger A., Burgess H., DeBey L., Schmidt N., Froehlich H., Wagner C., HilleRisLambers J., Tewksbury J., Harsch M., & Parrish J. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Thibaud E., Aalto J., Cooley D. S., Davison A. C., & Heikkinen J. (2016). Bayesian inference for the Brown–Resnick process, with an application to extreme low temperatures. *Annals of Applied Statistics*, *10*(4), 2303–2324. https://doi.org/10.1214/16-AOAS980

UN (2015). A/res/70/1 un general assembly transforming our world: The 2030 agenda for sustainable development. *Seventieth session of the General Assembly on 25 Sept 2015*.

van de Schoot R., Depaoli S., King R., Kramer B., Märtens K., Tadesse M. G., Vannucci M., Gelman A., Veen D., Willemsen J., & Yau C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1. https://doi.org/10.1038/s43586-020-00001-2

Vecchia A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(2), 297–312. https://doi.org/10.1111/j.2517-6161.1988.tb01729.x

Wijeyakulasuriya D. A., Hanks E. M., & Shaby B. A. (2024). Modeling first arrival of migratory birds using a hierarchical max-infinitely divisible process. *Journal of Agricultural, Biological and Environmental Statistics*, https://doi.org/10.1007/s13253-024-00624-y

Wijeyakulasuriya D. A., Hanks E. M., Shaby B. A., & Cross P. C. (2019). Extreme value-based methods for modeling elk yearly movements. *Journal of Agricultural, Biological and Environmental Statistics*, *24*(1), 73–91. https://doi.org/10.1007/s13253-018-00342-2

Wynn J. (2017). *Citizen science in the digital age : Rhetoric, science, and public engagement*. Rhetoric, culture, and social critique. The University of Alabama Press Tuscaloosa.

Yadav R., Huser R., Opitz T., & Lombardo L. (2023). Joint modelling of landslide counts and sizes using spatial marked point processes with sub-asymptotic mark distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *72*(5), 1139–1161. https://doi.org/10.1093/jrsssc/qlad077

Youngflesh C., Socolar J., Amaral B. R., Arab A., Guralnick R. P., Hurlbert A. H., LaFrance R., Mayor S. J., Miller D. A., & Tingley M. W. (2021). Migratory strategy drives species-level variation in bird sensitivity to vegetation green-up. *Nature Ecology & Evolution*, *5*(7), 987–994. https://doi.org/10.1038/s41559-021-01442-y

Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 700–740
Discussion Paper Contribution

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Proposer of the vote of thanks and contribution to the Discussion of the 'Discussion Meeting on the Analysis of citizen science data'

Ben Swallow[1,2]

[1]School of Mathematics and Statistics, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK
[2]Centre for Research into Ecological and Environmental Modelling, University of St Andrews, Buchanan Gardens, St Andrews KY16 9LZ, UK

Address for correspondence: Ben Swallow, School of Mathematics and Statistics, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK; Centre for Research into Ecological and Environmental Modelling, University of St Andrews, Buchanan Gardens, St Andrews KY16 9LZ, UK. Email: bts3@st-andrews.ac.uk

1 Introduction

Data are a vital part of statistical inference, grounding and calibrating statistical models, and supporting prediction and forecasting. Whilst data are routinely collected for analyses, they can be very expensive and/or resource intensive to collect. This can lead to prohibitively small-scale sampling, making it hard to draw conclusions at a population level. Citizen science (CS) data, however, can provide an alternative that has the potential to provide data at spatial and/or temporal resolutions that would not otherwise be feasible. These data utilize a network of highly skilled amateurs who routinely observe and make record of personal sightings.

Due to the often nonstructured nature of data collection, it is particularly important with CS data to model the observation process in which these data are observed and recorded to help account for the lack of randomized sampling and potential structural biases. This is particularly important given that they often feed into important national and international monitoring schemes, underpinning legal obligations of conservation policy, for example, Newson et al. (2017).

A particular challenge with CS data is the potential for a large range of biases in them, including from oversampling (that is rare or noteworthy events being more likely to be reported, spatial bias in high-population regions, visiting 'good' locations with a higher chance of successful observations); miss-identification; partial observations or checklists being submitted; and multiple observations of the same individuals. In order to alleviate these potential biases, rich metadata needs to be collected to help explain the process in which the data were collected, recorded, and submitted.

Other challenges include high data resolution leading to 'big data' computational challenges, including hardware demands, a need for efficient algorithms and multiple heterogeneities across the population and data sources. In addition, communication and interaction with domain experts or nonstatisticians are common due to the nature of both data collectors and the end users. A lack of statistical involvement in the design and implementation of data collection can make many of these issues further challenging to overcome.

It is, therefore, particularly welcome to comment on the three novel approaches presented here.

Received: September 6, 2024. **Accepted:** December 16, 2024

2 Efficient statistical inference methods for assessing changes in species' populations using citizen science data

The two main developments of this paper are, firstly, to incorporate time and site effects through an annual model integrated within the generalized abundance index. A concentrated likelihood framework is proposed to significantly reduce the dimension of the parameter space. The approach is fitted to both data from the UK Butterfly Monitoring Scheme and the Big Butterfly Count. Secondly, the authors propose the use of variational inference to speed up the efficiency of fitting the occupancy model framework. Using the observed data likelihood, required assumptions of independence between model parameters are avoided and a reparameterization provides an approach for efficient stochastic gradient descent. A small simulation study is implemented to study coverage probabilities of credible intervals for the parameters, with the intervals having nominal coverage. The approach is also applied to a common species of butterfly from the Butterflies for the New Millennium recording scheme.

2.1 Comments and questions

It seems the new development of a two-stage model offers differing estimates of uncertainties, however, which of these is correct is uncertain. Some simulations would be advantageous to determine which of these is more realistic. Wald-type standard errors from the Hessian will likely be unreliable in rarer species and/or small studies.

An advantage of the approach is that some very simple adjustments to the model, such as changing the discrete family distribution, can be made to determine impacts on uncertainty propagation. For example, a negative binomial or quasi-Poisson distribution could be compared using simple (quasi-)information criteria to select between these. There is an advantage in allowing different levels of variation across different levels of heterogeneity in the population. Global dispersion parameters from these alternative families can determine species-level variation across spatial and temporal regions, with additional variation allocated to the random effects in a hierarchical manner.

Furthermore, species populations and associated distribution are inherently spatial phenomena. This paper, amongst others, paves the way for accounting for the spatial correlation beyond independent site effects, combined with efficient inference for binomial models (e.g. Belmont et al., 2024; Held & Holmes, 2006). An alternative without quite as much computational overhead would be to incorporate spatial covariates that may account for much of the correlation.

Finally, as there are many species, some very rare, some discussion of multispecies approaches to learn across similar species would be fruitful, perhaps in conjunction with integrated population models that are already discussed in the paper. A common species with a similar ecological niche could be highly informative of the distributions of an undersampled species.

3 Frequentist prediction sets for species abundance using indirect information

The authors develop a frequentist-valid prediction sets framework for eBird data in North Carolina counties, assuming a nonparametric framework to obtain sets for a multinomial random sample. The approach uses data from more sampled counties to better inform undersampled counties through an integrated approach. An empirical Bayes framework for estimating prediction sets uses the maximum marginal likelihood calculated using the Newton–Raphson method to determine parameter values. The approach utilizes neighbouring counties to indirectly inform estimates, where more data diminishes the impact of neighbouring counties and data-scarce counties benefit from updated posteriors from informed neighbours. This has benefits when sampling is comparable and unbiased by reducing the size of the prediction set in most of the species tested.

3.1 Comments and questions

The authors show that assuming independence of observations across county, frequentist coverage probabilities are theoretically guaranteed. As discussed above, the assumption is perhaps not valid with CS data due to preferential sampling of observers, incomplete lists and observer heterogeneity. The coverage probabilities are above expected for moderately large K (see Figure 2), so some discussion of this in relation to potential over-confident standard errors would be welcome. Further attempts to account for the sampling process could alleviate this (e.g. Bird et al., 2014). Further work could develop a simulation incorporating this heterogeneity to determine the impact of preferential sampling on the prediction sets and determine sensitivity to this assumption or how to account for it if not realistic.

4 Extreme-value modelling of migratory bird arrival dates: insights from citizen science data

The authors study the first arrival dates of migrating birds from the global eBird CS project, and the more structured Breeding Bird Survey. Focusing on the northeastern US states, they treat the first arrival date as an extreme event, and develop an extreme value theory approach to predicting first arrival dates, accounting for survey effort and environmental niche. A complex hierarchical Bayesian model incorporating four latent Gaussian process priors to explain the spatial covariance in arrival dates across regions, sampling effort, and environmental niches is proposed. The data from two sources are fused together with common model structures to alleviate sampling bias in the CS data.

4.1 Comments and questions

There is evidently difficulty with validating latent variable models such as these. Variation in goodness-of-fit tests and information criteria for latent variable models are made more difficult in the complex observation process of CS data. More work is definitely needed here, particularly in estimating effective numbers of parameters. The suggestion of combining CS with more structured data definitely shows promise (e.g. Kelling et al., 2019; Robinson et al., 2020) and needs further work to determine when and how this should be done and how valuable it is in a specific scenario.

Multivariate extreme approaches are in their infancy but determining marginal dependence, e.g. through copulas, multivariate Generalized Pareto Distributions or tensor products of Gaussian Processes, would be a worthwhile extension to this approach. This may allow for sharing of information across species, where early arrivals of one species may be able to suggest similar species may also arrive early.

The Markov chain Monte Carlo (MCMC) chain mixing shown in Figure 12 shows there is a strong correlation between model components that may not be fully alleviated by the Metropolis-adjusted Langevin algorithm approach used. The authors highlight that a Gauss Markov random field may not be feasible due to nonlinearity of the model, but perhaps some components could be embedded within the INLA-SPDE approach with others estimated by MCMC (Gómez-Rubio & Rue, 2018)?

The accuracy of the Vecchia approximation used will also depend on the validity of the sparsity assumption, which is linked to the smoothness of the surface. This may work well for some species but less well for others with more patchy distributions.

Finally, all species chosen had almost identical preference parameter estimates for $\hat{\theta}^{pref}$ (Table 1). It would be interesting to choose some species that clearly show a very strong or weak observer preference, to show these are estimable and not prior-dependent.

5 Summary and looking forward

The three papers presented here show a wide computational variability of statistical approaches for tackling applied ecological questions using CS data. These include novelties covering a mix of computational, methodological, and philosophical approaches to model these complex data. Harnessing the power of information contained within CS data, they aim to account for (some) of the potential biases or sub-optimality of sampling inherent in their chosen schemes.

Ecological sciences have made CS data a cornerstone of the domain, giving much opportunity for methodological development; however, there remains many other areas of applied sciences where the power of these data could be better realized (e.g. epidemiology/infectious disease dynamics; environmental surveillance; public health). Further computational demands (including methods, hardware integration, and software) are needed to analyse the vast quantities of data that can be obtained from these schemes. As shown in these papers, the combination of differing data sources and their associated heterogeneities can provide significant benefits over single sources of data. With those benefits comes a greater need for improved methods for relative and absolute goodness-of-fit for these data and models, however, to allow for differing resolutions and collection procedures. The three contributions here show the potential for further development in analysing CS data and highlight that further work is still needed from statisticians working in this area.

It therefore gives me great pleasure to propose the vote of thanks.

Conflicts of interest: None declared.

References

Belmont J., Martino S., Illian J., & Rue H. (2024). 'Spatio-temporal occupancy models with INLA.' *Methods in Ecology and Evolution*, *15*(11), 2087–2100. https://doi.org/10.1111/2041-210X.14422.

Bird T. J., Bates A. E., Lefcheck J. S., Hill N. A., Thomson R. J., Edgar G. J., Stuart-Smith R. D., Wotherspoon S., Krkosek M., Stuart-Smith J. F., Pecl G. T., Barrett N., & Frusher S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*(6), 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Gómez-Rubio V., & Rue H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, *28*(5), 1033–1051. https://doi.org/10.1007/s11222-017-9778-y

Held L., & Holmes C. C. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, *1*(1), 145–168. https://doi.org/10.1214/06-BA105

Kelling S., Johnston A., Bonn A., Fink D., Ruiz-Gutierrez V., Bonney R., Fernandez M., Hochachka W. M., Julliard R., Kraemer R., & Guralnick R. (2019). Using semistructured surveys to improve citizen science data for monitoring biodiversity. *Bioscience*, *69*(3), 170–179. https://doi.org/10.1093/biosci/biz010

Newson S. E., Evans H. E., Gillings S., Jarrett D., Raynor R., & Wilson M. W. (2017). Large-scale citizen science improves assessment of risk posed by wind farms to bats in southern Scotland. *Biological Conservation*, *215*(1), 61–71. https://doi.org/10.1016/j.biocon.2017.09.004

Robinson O. J., Ruiz-Gutierrez V., Reynolds M. D., Golet G. H., Strimas-Mackey M., & Fink D. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity & Distributions*, *26*(8), 976–986. https://doi.org/10.1111/ddi.v26.8

https://doi.org/10.1093/jrsssa/qnaf009
Advance access publication 12 February 2025

Seconder of the vote of thanks and contribution to the Discussion of 'the Discussion Meeting on the Analysis of citizen science data'

Kerrie Mengersen

QUT Centre for Data Science, Queensland University of Technology (QUT), Brisbane, Australia

Address for correspondence: Queensland University of Technology (QUT), Centre for Data Science.
Email: k.mengersen@qut.edu.au

Thank you for the opportunity to second the vote of thanks at this meeting. I want to start by saying that after reading the three papers presented in this session, it's such a wonderful time to be a statistician! There are three reasons for saying this.

The first is that, as we all know, we are in the best of professions. Statistics is currently one of the most in-demand professions globally, in no small part due to the need to 'do something' with the wealth of data that is now available. This creates an ideal space for us to flourish as researchers and practitioners. The high profile of statistics as a trusted 'data-whisperer' is in no small part due to the great efforts of our professional societies, and the RSS is one of the global leaders in this endeavour, so thank you.

The second reason is that we not only have a fantastic new range of tools thanks to the rise of statistical machine learning, but we also have the compute power to implement our methods and shift the horizon. Models and algorithms that were 'nice but not practical' a decade ago are now common tools and established research directions. Emerging algorithmic enablers—such as foundation models and quantum computing—promise to open even more doors, creating again a new round of challenges and opportunities for statisticians as we formulate, validate and integrate solutions.

The third reason is that we now have a huge range of new data sources available to us, to wrangle, model and understand. Data from digital sensors, cameras, satellites, social media, and—yes—citizen science provide rich information in their own right and as a complement to more traditional sources of data. As statisticians, we have an obligation to embrace these new data sources and be the trusted voice for researchers, practitioners, managers, and the public regarding their benefits and deficits. If the data are rubbish, then we can—and should—say so. However, if there is merit, we have an opportunity—indeed a responsibility—to contribute to research, tools, and applications that enhance the utility of these data, extract information in a rigorous manner and merge the resultant insights with other sources of knowledge, thereby creating a true data-informed learning and decision-making environment.

Which brings me to the three papers under discussion at this meeting. In accord with the theme of the meeting, the common ambition of these papers is to enhance the utility of citizen science data. Although all three papers focus on applications in ecology, the methodological innovations are generalizable to a wide range of other contexts.

Within this broad remit, the three papers take slightly different directions. Two papers focus primarily on improved models for estimation and inference which correct for sampling bias. Elizabeth and Peter articulate a two-stage approach, with neighbourhood information entering as 'indirect information' through the concentration parameter of the Bayesian prior in one of the stages. In contrast, Jonathan and Thomas propose a strictly Bayesian hierarchical spatio-temporal model that similarly adjusts for heterogeneous sampling effort. Both papers comment on computational efficiency. Emily and coauthors focus primarily on this aspect and propose variational inference as a solution.

I have so many comments and questions about each of these papers, but in this forum I would like to make just three points.

First, all three authors deal with what I would call 'well-behaved' citizen science data. That is, the data are sourced from large, public studies with an underlying design, albeit not to the standard of a randomized controlled study. This is great, and the contributions are very worthwhile. However, there is more to be done: there are other biases in these data that need to be addressed to increase their trust and utility, and there are citizen science data from more 'wild' studies that also have great value. So keep going!

Second, the papers advance the practice of data fusion. Elizabeth and Peter focus on one study—the eBird survey—and fuse information across the geographic sites; Jonathan and Thomas 'see this and raise it' to the merger of two separate studies with different protocols, the eBird study and the North American Breeding Bird Survey. Emily and coauthors raise the ante again to the merger of three butterfly datasets. This work touches on a large literature that rectifies biases and gaps in this way—I call out earlier work of Sylvia Richardson in epidemiology, and the United Nations Global Platform for enabling the use of different types of remotely sensed data to monitor the Sustainable Development Goals, as examples.

The papers also advance the practice of model fusion. In particular, Elizabeth and Peter suggest a merger of frequentist and Bayesian methods to achieve the desired improvement in the size and precision of the area-specific species prediction sets. This 'smooshing' of models and philosophies is now commonplace and potentially very effective, but again as statisticians we need to ensure that this is undertaken in a rigorous manner, that all of the properties of the resultant approach are understood, that comparisons with other approaches are thorough and fair, and that 'edge cases' are carefully considered. Furthermore, we should be actively stepping into the space of fusing machine learning and statistical models, particularly for the analysis of citizen science data. I think that there is much more work to be done in this area.

My third point is that all of the papers work hard to 'respect' the citizen science data. That is, their new tools aim to address the characteristics of the data rather than 'make the shoe fit' with existing tools. This is important, and pushes us to consider the relative benefits and drawbacks of employing models at all in this context. This is a longer discussion for another time.

It would be remiss of me to not mention the efforts of colleagues in my part of the world to capitalize on the enormous potential of citizen science data. A special call-out to early work by Mark Burgman, Hugh Possingham, and Sama Low Choy in eliciting expert information to enhance ecological modelling, and more recently Edgar Santos-Fernandez and Julie Vercelloni for their leadership in using citizen-contributed classifications of coral cover in underwater images to improve monitoring of the health of the Great Barrier Reef. Through Julie, this effort has expanded to the Pacific Islands, which are particularly vulnerable to socio-industrial development and climate change.

In summary, I applaud all of the authors in this session for their very welcome, substantive contributions to the three areas that I touched on in my introduction: to raising the profile of statisticians as 'data-whisperers' in the development and application of trusted methods, to extending our computational capability, and to advancing the analysis of new data sources—in particular, citizen science data. It is with great pleasure that I second the vote of thanks for these stimulating, cutting-edge papers.

Conflicts of interest: none declared.

https://doi.org/10.1093/jrsssa/qnaf013
Advance access publication 11 February 2025

Andrej Srakar's contribution to the Discussion of 'the Discussion Meeting on the Analysis of citizen science data'

Andrej Srakar

Institute for Economic Research and University of Ljubljana, Ljubljana, Slovenia

Address for correspondence: Andrej Srakar, Institute for Economic Research, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia. Email: srakara@ier.si

This contribution discusses the papers of Dennis et al., Bersson and Hoff, and Koh and Opitz.

Paper by Dennis et al. (2024) mentions reducing the functional model parameter space. As a mathematician, I wonder which mathematical spaces (such as metric spaces of Banach or Hilbert space variety) to use to study citizen science (CS) data. Could inclusion of tensor data approaches and tensorial 'spaces' be of help? Namely studying CS data leads to high-dimensional problems which gives this question a special importance. As the paper uses concentrated likelihood functions, I wonder whether using approaches which rely on intractable likelihoods (such as indirect inference and in particular approximate Bayesian computation) would provide responses to many open problems addressing analysis of CS data. In general,

This is a discussion contribution to the Discussion Meeting on the 'Analysis of citizen science data' at the 2024 Royal Statistical Society conference in Brighton, United Kingdom.

do you see any special properties of CS data which would give priority to Bayesian over frequentist approaches?

Paper by Bersson and Hoff (2024) uses borrowing of information, and approach related to empirical Bayes procedures. Interesting could be to refer to and use recent thread of research from Ascolani, Prünster, Lijoi, and coauthors (e.g. Ascolani, Franzolini et al. 2024) on nonparametric priors with full-range borrowing of information. Also, addressing geographic adjacency structure seems an extremely interesting aspect of work with CS data. This immediately leads to spatiotemporal but also random matrix perspectives, and it would be great hear more about future possibilities here.

Paper by Koh and Opitz (2024) mentions development of bird and ecology related applications of CS data. Could they be found in economics, in medicine and healthcare, in robotics or in law? Could they be fruitfully applied and found in any scientific and sector field? Usage of Gaussian latent random fields bring questions how to study asymptotics of approaches studying CS data—is there any aspect to be careful about? Using MCMC and MALA algorithms questions how to approach computational efficiency for modelling CS data in the future? One also wonders whether modelling species behaviour using microsimulation or even agent-based simulation approaches be used combined with CS data and how.

Papers providing mathematical and statistical approaches to study CS data are highly important based on the present spread of such initiatives as well as the fact that leading journals reject paper using CS data on the ground of their questionable validity. It would be great for all three papers to explain how their findings can be used to mitigate such problems in the future.

Conflicts of interest: None declared.

References

Bersson E., & Hoff P. D. (2025). Frequentist prediction sets for species abundance using indirect information. *Journal of the Royal Statistical Society: Series A*, *188*(3), 658–673. https://doi.org/10.1093/jrsssa/qnae096

Dennis E. B., Diana A., Matechou E., & Morgan B. J. T. (2025). Efficient statistical inference methods for assessing changes in species' populations using citizen science data. *Journal of the Royal Statistical Society: Series A*, *188*(3), 641–657. https://doi.org/10.1093/jrsssa/qnae105

Koh J., & Opitz T. (2025). Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data. *Journal of the Royal Statistical Society: Series A*, *188*(3), 674–699. https://doi.org/10.1093/jrsssa/qnae108

https://doi.org/10.1093/jrsssa/qnaf017
Advance access publication 8 February 2025

Benjamin R. Baer's contribution to the Discussion of 'the Discussion Meeting on Analysis of citizen science data'

Benjamin R. Baer

Centre for Research into Ecological and Environmental Modelling, School of Mathematics and Statistics, University of St Andrews, St Andrews KY16 9SS, Scotland

Address for correspondence: Benjamin R. Baer, Centre for Research into Ecological and Environmental Modelling, School of Mathematics and Statistics, University of St Andrews, St Andrews KY16 9SS, Scotland. Email: benjamin.baer@st-andrews.ac.uk

We congratulate Dennis, Diana, Matechou, and Morgan, the authors of 'Efficient statistical inference methods for assessing species' populations using citizen science data', on an interesting contribution to the literature on computationally efficient algorithms for statistical problems involving citizen science data and thank them for an interesting conversation about their work.

We start by briefly reviewing the problem set up of the first example. Consider counts Y_{svr} of butterflies where $s = 1, \ldots, 4{,}328$ denotes sites, $r = 1, \ldots, 47$ denotes years, and $v = 1, \ldots, 26$ denotes weeks/visits. Dennis et al. (2024) define the model

$$\begin{aligned} Y_{svr} &\sim \mathrm{Pois}(n_{sr} a_{vr}), \\ \log n_{sr} &= \alpha_s + \beta_r, \\ a_{vr} &= \varphi(v; \mu_r, \sigma_r), \end{aligned}$$

where φ denotes the normal density. With independent observations, the log likelihood up to an additive constant is

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) := \sum_{s,v,r} -e^{\alpha_s + \beta_r} \varphi(v; \mu_r, \sigma_r) + Y_{svr}\{\alpha_s + \beta_r + \log \varphi(v; \mu_r, \sigma_r)\}.$$

Let $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}})$ be maximizers of ℓ. Throughout a '$\bullet$' in a subscript denotes aggregation over that index. Applying the first order condition, Dennis et al. (2024, sec. S1) show that $e^{\hat{\alpha}_s} = Y_{s\bullet\bullet} / \sum_{v,r} e^{\hat{\beta}_r} \varphi(v; \hat{\mu}_r, \hat{\sigma}_r)$ for all s. Thus the log profile likelihood $\ell_p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) := \ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ admits $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}})$ as maximizers. In their data example with Gatekeepers, the sample size is $n = 527{,}588$, and the (profiled) parameter dimension is $p = 141 = 3 \cdot 47$ so that $n \gg p$. Without the profiling step, there would be an additional 4,328 parameters, so profiling results in a large reduction.

Now, we consider a further computational speedup. The parameter $\boldsymbol{\alpha}$ is the only one that varies with site s: once it is profiled out to create ℓ_p, the remaining parameters do not depend on site. The following calculation shows that the observations may, therefore, be aggregated over site.

Proposition 1 Up to an additive constant,

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{v,r} Y_{\bullet vr}\{\beta_r + \log \varphi(v; \mu_r, \sigma_r)\} - Y_{\bullet\bullet\bullet} \log\left\{\sum_{v,r} e^{\beta_r} \varphi(v; \mu_r, \sigma_r)\right\}.$$

Proof. Denote '$\doteq$' as equality up to additive constants. The log profile likelihood satisfies

$$\begin{aligned}
&\sum_{s,v,r} -e^{\hat{\alpha}_s + \beta_r} \varphi(v; \mu_r, \sigma_r) + Y_{svr}\{\hat{\alpha}_s + \boldsymbol{\beta}_r + \log \varphi(v; \mu_r, \sigma_r)\} \\
&= \sum_{s,v,r} -e^{\beta_r} \frac{Y_{s\bullet\bullet}}{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})} \varphi(v; \mu_r, \sigma_r) \;\; + Y_{svr}\left[\log\left\{\frac{Y_{s\bullet\bullet}}{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})}\right\} + \beta_r + \log \varphi(v; \mu_r, \sigma_r)\right] \\
&\doteq \sum_{s,v,r} -Y_{s\bullet\bullet} \frac{e^{\beta_r} \varphi(v; \mu_r, \sigma_r)}{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})} \;\; + Y_{svr}\{\beta_r + \log \varphi(v; \mu_r, \sigma_r)\} - Y_{svr} \log\left\{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})\right\} \\
&= \sum_{v,r} -Y_{\bullet\bullet\bullet} \frac{e^{\beta_r} \varphi(v; \mu_r, \sigma_r)}{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})} \;\; + \sum_{v,r} Y_{\bullet vr}\{\beta_r + \log \varphi(v; \mu_r, \sigma_r)\} - Y_{\bullet\bullet\bullet} \log\left\{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})\right\} \\
&\doteq \sum_{v,r} Y_{\bullet vr}\{\beta_r + \log \varphi(v; \mu_r, \sigma_r)\} - Y_{\bullet\bullet\bullet} \log\left\{\sum_{v',r'} e^{\beta_{r'}} \varphi(v'; \mu_{r'}, \sigma_{r'})\right\}. \qquad \square
\end{aligned}$$

The form of the profile likelihood involves a sum over only visits (denoted by v) and years (denoted by r). This reduces the size of the problem from $n = 527{,}588$ to $1{,}222 = 26 \cdot 47$, which reduces the run time to evaluate the profile likelihood and, hence, the run time of an algorithm to compute associated maximum likelihood estimators.

This note illustrates that inspecting the form of objective functions can lead to simplifications and computational speedups. Separately, if the simplification had not occurred and a sum over $n = 527{,}588$ observations remained in the objective function, then an algorithm that scales well with large n, such as stochastic gradient descent, may be preferred to limited-memory BFGS, the algorithm employed by Dennis et al. (2024).

Conflicts of interest: The author has no conflicts of interest.

Reference

Dennis E. B., Diana A., Matechou E., & Morgan B. J. (2024). Efficient statistical inference methods for assessing changes in species' populations using citizen science data. *Journal of the Royal Statistical Society: Series A*, *188*(3), 641–657. http://dx.doi.org/10.1093/jrsssa/qnae105

https://doi.org/10.1093/jrsssa/qnaf011
Advance access publication 11 February 2025

Jafet Belmont, Sara Martino, Janine Illian, and Håvard Rue's contribution to the Discussion of the 'Discussion meeting on the analysis of citizen science data'

Jafet Belmont[1], Sara Martino[2], Janine Illian[1] and Håvard Rue[3]

[1]School of Mathematics and Statistics, University of Glasgow, 132 University Pl, Glasgow G12 8TA, Glasgow, G12 8QQ, UK
[2]Department of Mathematics, Norwegian University of Science and Technology, Trondheim, Norway
[3]CEMSE Division, King Abdullah University of Science and Technology, KAUST, Thuwal, Saudi Arabia

Address for correspondence: Jafet Belmont, School of Mathematics and Statistics, University of Glasgow, 132 University Pl, Glasgow G12 8TA, Glasgow, G12 8QQ, UK. Email: Jafet.BelmontOsuna@glasgow.ac.uk

The authors highlight the prominent role of occupancy modelling in the context of citizen science (CS) data and the need to account for imperfect detection. They comment on the computational cost of fitting these models and, consequently, the importance of computational efficiency, which they address with a variational inference (VI) approach.

We suggest considering integrated-nested Laplace approximation (INLA) as an alternative, more efficient fitting approach that can be used in the current context since occupancy models can now be fitted using the library `R-INLA` (Belmont et al., 2024), provided detection depends on fixed effects only. To illustrate this, we have used INLA to fit the model discussed in the paper to the Ringlet butterfly dataset (R code accessible through https://github.com/Ecol-Stats/CS_data_analysis). This took about 30 s on a standard Apple M2 Pro machine, i.e. about 120 times faster than the VI approach with very similar estimates, see Figure 1a,b.

As an additional benefit, it is straightforward to include other complex model structures, such as smooth terms or, specifically, spatio-temporal effects as these are readily available in `R-INLA`. The latter are particularly relevant in the context of ecological data as spatial random effects can account for spatial dependencies not captured by covariates avoiding wrong inference (Wright

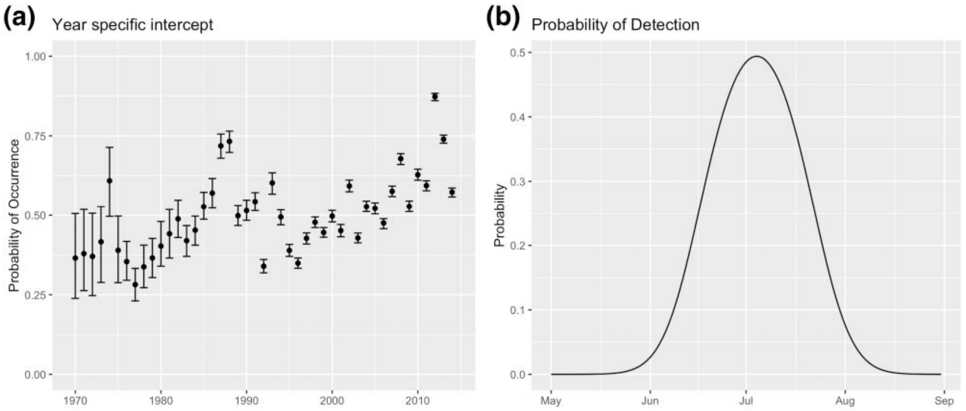

Figure 1. (a) Posterior median and 96% credible interval for yearly probability of occurrence. (b) Posterior median for the probability of detection. Both results refer to the model in Dennis et al. (2024).

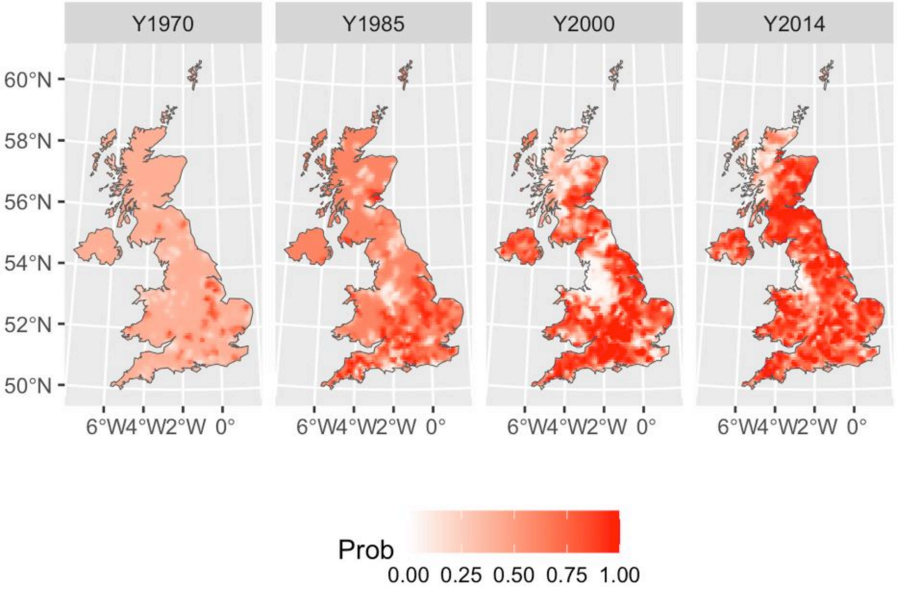

Figure 2. Posterior median of the probability of occurrence of the Ringlet butterfly (*Aphantopus hyperantus*) in space and across time, for the model including a spatial term.

et al., 2019). In an extended model, we have included a spatially structured random effect using the SPDE approach (Lindgren et al., 2011) as implemented in R-INLA, similar to the model in Diana et al. (2023), but omitting yearly random effects for the detection probabilities (computation time approximately 5 min). Figure 2 shows the estimated probability of occurrence over the whole domain for four different years; the results resemble those in Diana et al. (2023).

We agree with the authors on the need for developing accurate, computationally efficient—yet still practically relevant and accessible—methods to address the computational challenges resulting from large CS data. The INLA framework, along with the R-INLA package, addresses exactly this. Therefore, we believe that the growing community of R-INLA users can greatly benefit from its relevance to CS data as well as its general versatility. For instance, the package provides goodness of fit and prediction accuracy measures (Adin et al., 2024; Van Niekerk et al., 2023), and a straight forward way of combining different data structures into a single global model (Illian et al., 2013; Martino et al., 2021; Panunzi et al., 2024).

However, we also agree that for some complex data structures, e.g. where detection may have to depend on random effects, Markov chain Monte Carlo as a less computationally efficient approach remains the method of choice for Bayesian inference. Striking the right balance here requires a continuous dialogue between practitioners and method developers to adequately analyse CS data and to encourage innovation in tackling complex ecological questions (Illian & Burslem, 2017).

Conflicts of interest: All the authors listed made substantial contributions to the manuscript and qualify for authorship, and no authors have been omitted. We warrant that none of the authors has any conflict of interest in regard to this manuscript.

References

Adin A., Krainski E. T., Lenzi A., Liu Z., Martínez-Minaya J., & Rue H. (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics*, *62*(1), 100843. https://doi.org/10.1016/j.spasta.2024.100843

Belmont J., Martino S., Illian J., & Rue H. (2024). Spatio-temporal occupancy models with INLA. *Methods in Ecology and Evolution*, *15*(11), 2087–2100. http://dx.doi.org/10.1111/2041-210X.14422

Dennis E. B., Diana A., Matechou E., & Morgan B. J. (2025). Efficient statistical inference methods for assessing changes in species' populations using citizen science data. *Journal of the Royal Statistical Society: Series A*, *188*(3), 641–657. https://doi.org/http://dx.doi.org/10.1093/jrsssa/qnae105

Diana A., Dennis E. B., Matechou E., & Morgan B. J. T. (2023). Fast Bayesian inference for large occupancy datasets. *Biometrics*, *79*(3), 2503–2515. https://doi.org/10.1111/biom.13816

Illian J. B., & Burslem D. F. (2017). Improving the usability of spatial point process methodology: An interdisciplinary dialogue between statistics and ecology. *Advances in Statistical Analysis: AStA: A Journal of the German Statistical Society*, *101*(4), 495–520. https://doi.org/10.1007/s10182-017-0301-8

Illian J. B., Martino S., Sørbye S. H., Gallego-Fernández J. B., Zunzunegui M., Esquivias M. P., & Travis J. M. (2013). Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods in Ecology and Evolution*, *4*(4), 305–315. https://doi.org/10.1111/mee3.2013.4.issue-4

Lindgren F., Rue H., & Lindström J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Martino S., Pace D. S., Moro S., Casoli E., Ventura D., Frachea A., Silvestri M., Arcangeli A., Giacomini G., Ardizzone G., & Jona Lasinio G. (2021). Integration of presence-only data from several sources: A case study on dolphins' spatial distribution. *Ecography*, *44*(10), 1533–1543. https://doi.org/10.1111/ecog.2021.v44.i10

Panunzi G., Moro, S., Marques I., Martino S., Colloca F., Ferretti F., & Jona Lasinio G. (2024). Estimating the spatial distribution of the white shark in the Mediterranean Sea via an integrated species distribution model accounting for physical barriers. *Environmetrics*, *36*(1), e2876. http://dx.doi.org/10.1002/env.2876

Van Niekerk J., Krainski E., Rustand D., & Rue H. (2023). A new avenue for Bayesian inference with INLA. *Computational Statistics & Data Analysis*, *181*(1), 107692. https://doi.org/10.1016/j.csda.2023.107692

Wright W. J., Irvine K. M., & Higgs M. D. (2019). Identifying occupancy model inadequacies: Can residuals separately assess detection and presence? *Ecology*, *100*(6), e02703. https://doi.org/10.1002/ecy.2019.100.issue-6

https://doi.org/10.1093/jrsssa/qnaf022
Advance access publication 28 February 2025

Léo R. Belzile and Rishikesh Yadav's contribution to the Discussion of 'the Discussion Meeting on the Analysis of citizen science data'

Léo R. Belzile and Rishikesh Yadav

Department of Decision Sciences, HEC Montréal, Montréal, Canada

Address for correspondence: Léo R. Belzile, Department of Decision Sciences, HEC Montréal, 3000, ch. Côte-Sainte-Catherine, Montréal (Québec), Canada H3T 2A7. Email: leo.belzile@hec.ca

We congratulate Koh and Opitz for this stimulating piece of work. The complex framework adopted by the authors offers crucial insights into migratory patterns and species behaviours that simpler models may overlook. Much effort has gone into the model specification to account for sampling biases and to ensure interpretability, and it shows.

Hierarchical Bayesian models are well suited for this type of modelling exercise, but the inference is complicated and preprocessing tedious. Assuming separability of space and time effects simplifies the problem, and offers the possibility of using leave-one-year-out cross-validation, although this is computationally intensive. The sharing of random effects allows one to borrow strength across data sources, but may lead to model misspecification without great care. Adding covariates, such as land cover, could reduce the residual variability, but we acknowledge that their effect may be nonlinear, and suitable smooths would add multiple fixed effect parameters.

One concern is the slow convergence and poor mixing observed in Figure 12, even after thinning. We wonder what the effective sample size is after burn-in. Although standard methods, such as adaptive MCMC (Andrieu & Thoms, 2008; Rosenthal, 2011), could help, hyperparameters may be strongly correlated due to shared components, and joint updates may be necessary to increase the efficiency of the sampler. The performance of Metropolis adjusted Langevin algorithm (MALA) is highly sensitive to the global tuning parameter or prewhitening matrix; locally adaptive schemes could fare better (e.g. Girolami & Calderhead, 2011; Rue & Held, 2005, Section 4.4.1).

Although the Markov chains seem to stabilize eventually, running multiple chains could be used to check whether they reach a unique stationary distribution. Model predictions at the data level (e.g. Figure 9) look sensible, but it is unclear whether individual model components are identifiable. For example, consider the function relating the generalized extreme value distribution location parameter μ with the sampling effort, $g(x_{\text{bound}}, x_{\text{effort}}) = \exp(x_{\text{bound}})/\{1 + \exp(-x_{\text{effort}})\}$. The functional form of eq. (4) implies that we cannot distinguish between parameters for x_{bound} when x_{effort} is low. Increases in g (and thus in μ) lead to an earlier minimum arrival rate.

We believe data fusion of related databases observed at different locations or resolutions has great potential in the field of spatial extreme value analysis as there is limited information available and pooling can help partly alleviate this.

Conflicts of interest: The authors have no conflict of interest to declare.

References

Andrieu C., & Thoms J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, *18*(4), 343–373. https://doi.org/10.1007/s11222-008-9110-y

Girolami M., & Calderhead B. (2011, March). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *73*(2), 123–214. https://doi.org/10.1111/j.1467-9868.2010.00765.x. ISSN 1369-7412

Rosenthal J. S. (2011). Optimal proposal distributions and adaptive MCMC. In S. Brooks, A. Gelman, G. Jones, & X. Meng, (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 93–111). CRC Press. https://doi.org/10.1201/b10905-5

Rue H., & Held L. (2005). *Gaussian Markov random fields: Theory and applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press. https://doi.org/10.1201/9780203492024. ISBN 9780203492024.

https://doi.org/10.1093/jrsssa/qnaf016
Advance access publication 8 February 2025

Abdelaati Daouia and Gilles Stupfler's contribution to the Discussion of the 'Discussion Meeting on the Analysis of citizen science data'

Abdelaati Daouia[1] and Gilles Stupfler[2]

[1]Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France
[2]Department of Mathematics & CNRS, University of Angers, Angers, France

Address for correspondence: Gilles Stupfler, Laboratoire Angevin de Recherche en Mathématiques, 2 Boulevard de Lavoisier, Angers Cedex 49045, France. Email: gilles.stupfler@univ-angers.fr

We congratulate Jonathan Koh and Thomas Opitz for developing a sophisticated but fully operational spatiotemporal Bayesian hierarchical model for an explicit bias-corrected estimation of arrival dates from combined heterogeneous datasets. An econometrician may have thought, however, that this problem also belongs to the area of (frequentist) frontier analysis, which provides a rather different, and, we hope, useful perspective. More precisely, for a given species of birds and each pixel of a spatial mesh covering the study area, let Y represent the arrival date reported by an observer and T be the corresponding year of observation. As highlighted in the paper, the conditional distribution of Y given T obviously has a tail bounded to the left. The finite lower extremity of the support of Y given $T = t$ defines the frontier point that corresponds to the population left-endpoint of all dates of occurrence of the species during the year t. This frontier function of t can be estimated by a global envelopment spline smoother under/without shape (e.g. concavity) constraints with automatic selection information criteria of the number and location of knots (Daouia et al., 2016), or by a local polynomial approach based on local extreme value statistics with adaptive selection of the tuning parameters (Jirak et al., 2014). Both approaches are based on the characterization of the yearly first arrival dates as a regression function in a nonparametric regression model with one-sided errors. The resulting spline or local polynomial boundary would then describe the yearly evolution of the desired first arrival date at the local zone considered. This avoids setting up a complex Bayesian hierarchical structure, as well as the Generalized Extreme Value approach whose goodness-of-fit to the data is not always easy to assess. Furthermore, instead of restricting the analysis to each pixel, thus resulting in a discrete geographical differentiation of the first arrival date, we suggest using all the data covering the study area and incorporating the observation location (i.e. longitude and latitude), as well as exogenous climate, land cover and other ecological factors into the estimation procedure as covariates (Wang et al., 2020), which would provide spatially smoother predictions of the true first arrivals. Proceeding in this way would also implicitly integrate the sampling effort, reflected by both preference and activity dimensions, as this would naturally show up in the confidence intervals, whether produced by asymptotics or bootstrapping. We think that this would be a useful complement to the very interesting approach constructed here.

Conflicts of interest: None declared.

Funding

This research was supported by the French National Research Agency under the grants ANR-19-CE40-0013 (ExtremReg project), ANR-23-CE40-0009 (EXSTA project), ANR-17-EURE-0010 (EUR CHESS) and ANR-11-LABX-0020-01 (Centre Henri Lebesgue). A. Daouia and G. Stupfler acknowledge financial support from the TSE-HEC ACPR Chair "Regulation and systemic risks". G. Stupfler acknowledges further support from the Chair Stress Test, RISK Management and Financial Steering of the Foundation Ecole Polytechnique.

References

Daouia A., Noh H., & Park B. U. (2016). Data envelope fitting with constrained polynomial splines. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *78*(1), 3–30. https://doi.org/10.1111/rssb.12098

Jirak M., Meister A., & Reiss M. (2014). Adaptive function estimation in nonparametric regression with one-sided errors. *Annals of Statistics*, *42*(5), 1970–2002. https://doi.org/10.1214/14-AOS1248

Wang L., Xue L., & Yang L. (2020). Estimation of additive frontier functions with shape constraints. *Journal of Nonparametric Statistics*, *32*(2), 262–293. https://doi.org/10.1080/10485252.2020.1721494

https://doi.org/10.1093/jrsssa/qnaf018
Advance access publication 10 February 2025

Dani Gamerman's contribution to the Discussion of 'the Discussion Meeting on the Analysis of citizen science data'

Dani Gamerman Sr.

Department of Statistical Methods, Institute of Mathematics, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Address for correspondence: Dani Gamerman, Department of Statistical Methods, Institute of Mathematics, Universidade Federal do Rio de Janeiro, Av. Athos da Silveira Ramos, 149 Cidade Universitária, Rio de Janeiro, RJ 21941-909, Brazil. Email: dani@im.ufrj.br

This contribution discusses the paper by Dennis et al. (2025). The RSS must be congratulated for this timely discussion session, with three great papers.

The model-based approach adopted in this paper is welcomed. It was particularly pleasing to see (observed and unobserved) temporal components presented and discussed in this paper. This paper will certainly be a useful reference for the analysis of archaeologically dated ancient remains we are currently contemplating. My contribution to the discussion addresses the possible extension of the ideas of this paper to temporal sequences of infinite-dimensional observational processes.

Locations of occurrences are as relevant as their counts for point pattern analyses. Inhomogeneous Poisson processes (IPP) are frequently used as their observational distribution. Their intensity functions are based on two components, as in the paper: one for explaining occurrences and one for explaining detection of occurrences. This was implemented for presence-only analyses in Moreira and Gamerman (2022), also using Bayes, but with no model approximation. Prediction of the locations and the total number of unobserved occurrences are easily provided. This prediction was obtained in the multidisciplinary study of the impact of pre-Columbian occupation of Amazonia (Peripato et al., 2023), and it attracted a sizeable media attention (see Gamerman, 2024). Could the approach of this paper be adapted to such scenario or to its temporal generalization when marks (e.g. dates) are associated with sequences of point patterns (e.g. archaeological remains)?

I will not dwell much over computational aspects other than sharing the concern of the authors about possible errors associated with approximations to the likelihood and models. I appreciate the efforts reported in Table 1. Could the results be expanded to include comparisons against the respective true posteriors (e.g. using Markov Chain Monte Carlo as a gold standard), other parameters, and their (co)variances? The analyses of Amazonia reported results about the true posterior despite the IPP likelihood intractability and despite having to work with databanks with millions of rows.

There are many nice features in the paper by Dennis et al. (2025) and I congratulate the authors for all their comprehensive and hard work.

Conflicts of interest: none declared.

References

Dennis, E. B., Diana, A., Matechou, E. & Morgan, B. J. T. (2025). Efficient statistical inference methods for assessing changes in species' populations using citizen science data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *188*(3), 641 –657. https://doi.org/10.1093/jrsssa/qnae105

Gamerman, D. (2024). Uncovering Amazonia: How statistics dismantled a myth. *Significance*, *21*(5): 6–10. https://doi.org/10.1093/jrssig/qmae071

Moreira, G. A., & Gamerman, D. (2022). Analysis of presence-only data via exact Bayes, with model and effects identification. *The Annals of Applied Statistics*, *16*(3), 1848–1867. https://doi.org/10.1214/21-AOAS1569

Peripato, V., Levis, C., Moreira, G. A., Gamerman, D., Aragao, L. E. O. C., *et al.* (2023). More than 10,000 pre-Columbian earthworks are still hidden throughout Amazonia. *Science*, *382*(6666), 103–109. https://doi.org/10.1126/science.ade2541

https://doi.org/10.1093/jrsssa/qnaf010
Advance access publication 11 February 2025

Raphaël Huser and Andrew Zammit-Mangion's contribution to the Discussion of the 'Discussion Meeting on the Analysis of citizen science data'

Raphaël Huser[1] and Andrew Zammit-Mangion[2]

[1]Statistics Program Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[2]School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia

Address for correspondence: Raphaël Huser, Statistics Program Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. Email: raphael.huser@kaust.edu.sa

We are grateful for the opportunity to discuss the paper entitled '*Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data*' (Koh & Opitz, 2025). We congratulate the authors for conducting such an elaborate data analysis of bird arrivals showing how multiple citizen-science data products of varying quality, completeness, and space-time resolutions, can be fused together and modelled jointly using a complex Bayesian hierarchical model. The model implements different likelihood functions over a shared latent structure, and accounts for the sampling effort to remove systematic biases.

Our comment focuses on the methods for making inference, and how alternative approaches can be used to substantially reduce the computational cost. The authors perform Bayesian inference using a customized Metropolis–Hastings algorithm, where hyperparameters are updated using Gibbs sampling, and where the latent Gaussian components are updated using block proposals based on the Metropolis-adjusted Langevin algorithm. They report that each iteration takes approximately 6 sec., so a single model fit obtained by drawing 80 000 posterior samples takes about 5.5 days. While such an inference time may be acceptable if the model needs to be fitted only a few times, it is clearly too expensive if repeated fitting is necessary, e.g., to conduct an extensive cross-validation study, to analyse different bird species in different study regions, or to incorporate new data, as is often required in citizen science. Recent years have seen the emergence of neural networks being used for amortized inference (see Zammit-Mangion et al., 2025, for a recent review), both in the context of point estimation (e.g. Sainsbury-Dale et al., 2023, 2024) and full posterior inference (e.g., Radev et al., 2022; Radev, Schmitt, Pratz, et al., 2023; Radev, Schmitt, Schumacher, et al., 2023) for models with intractable or unavailable likelihood functions. With such neural 'amortized' approaches, inference with new data can be performed repeatedly at a fraction of the time needed using standard Markov chain Monte Carlo methods, after an initial computational cost is incurred to train a neural network using training data simulated from the model of interest. Amortized methods for complex spatio-temporal Bayesian hierarchical models are still in their infancy, but see Zammit-Mangion and Wikle (2020) for an early example. Although neural inference methods can be used to estimate model parameters, they will likely struggle to make inference on latent variables, which are often numerous and highly correlated a posteriori. One solution is to adopt an empirical Bayes approach, where hyperparameters are estimated in a first step (e.g., using neural Bayes estimators; see Sainsbury-Dale et al., 2024), and latent variables are then inferred in a second step conditional on the estimated hyperparameters, perhaps by using a learning network such as that proposed by Liu and Liu (2020). This strategy remains to be tested, especially in cases where the model has a relatively large number of hyperparameters (i.e., more than 20, say) and when the trained neural networks need to be adaptable to small model changes (e.g., to different covariates values). In principle, however, this strategy would allow for fast inference for a wide range of Bayesian hierarchical models, including models for data fusion, thus enabling the analysis of citizen-science data at unprecedented scale.

Conflicts of interest: None declared.

References

Koh J., & Opitz T. (2025). Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data. *Journal of the Royal Statistical Society: Series A*, *188*(3), 674–699. https://doi.org/10.1093/jrsssa/qnaf012

Liu L., & Liu L. (2020). Localizing and amortizing: Efficient inference for Gaussian processes. In S. J. Pan & M. Sugiyama (Eds.), *Proceedings of The 12th Asian Conference on Machine Learning* (Vol. *129*, pp. 823–836). PMLR.

Radev S. T., Mertens U. K., Voss A., Ardizzone L., & Köthe U. (2022). BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(4), 1452–1466. https://doi.org/10.1109/TNNLS.2020.3042395

Radev S. T., Schmitt M., Pratz V., Picchini U., Köethe U., & Buerkner P.-C. (2023). JANA: Jointly Amortized Neural Approximation of complex Bayesian models. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023)* (Vol. 216, pp. 1695–1706). PMLR.

Radev S. T., Schmitt M., Schumacher L., Elsemüller L., Pratz V., Schälte Y., Köthe U., & Bürkner P. -C. (2023). BayesFlow: Amortized Bayesian workflows with neural networks. *Journal of Open Source Software*, *8*(89), 5702. https://doi.org/10.21105/joss

Sainsbury-Dale M., Richards J., Zammit-Mangion A., & Huser R. (2024). Neural Bayes estimators for irregular spatial data using graph neural networks. *Journal of Computational and Graphical Statistics*.

Sainsbury-Dale M., Zammit-Mangion A., & Huser R. (2024). Likelihood-free parameter estimation with neural Bayes estimators. *The American Statistician*, *78*(1), 1–14. https://doi.org/10.1080/00031305.2023.2249522

Zammit-Mangion A., Sainsbury-Dale M., & Huser R. (2025). Neural methods for amortized inference. *Annual Reviews of Statistics and Its Application*, *12*. https://doi.org/10.1146/annurev-statistics-112723-034123

Zammit-Mangion A., & Wikle C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, *37*, 100408. https://doi.org/10.1016/j.spasta.2020.100408

https://doi.org/10.1093/jrsssa/qnaf012
Advance access publication 11 February 2025

Kuldeep Kumar's contribution to the Discussion of 'the Discussion Meeting on the Analysis of citizen science data'

Kuldeep Kumar

Bond University, Australia

Address for correspondence: Kuldeep Kumar, Bond University, Australia. Email: kkumar@bond.edu.au

This contribution discusses the papers of 'Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data' by Jonathan Koh and Thomas Opitz.

First of all, I would like to congratulate the authors for modelling such a complex issue as migratory bird arrival using citizen science data. As mentioned by the authors, citizen science data often suffers from inconsistent quality, including a lack of standardization. The modelling becomes more complex because the response variables include Generalized Extreme Value distributions, as well as values distributed according to Binomial and Poisson distributions. Migratory bird arrival dates can be influenced by the interaction of various factors, reflecting the complex interplay between environmental conditions, biological processes, and ecological interactions. I am unsure how many variables the authors have considered; naturally, the model will become more complex if additional variables are included. There are several cutting-edge machine learning models, such as Artificial Neural Networks and Stochastic Gradient Boosting, which could be applied to extreme value modelling. The validity of the data in these models can be tested by dividing the data into training and testing sets, which is particularly crucial in the context of citizen science data. I am curious whether the authors have compared the validity of their predictions using Bayesian hierarchical models with these machine learning models.

Conflicts of interest: none declared.

https://doi.org/10.1093/jrsssa/qnaf014
Advance access publication 11 February 2025

Subhash Lele's contribution to the Discussion of 'the Discussion Meeting on the Analysis of citizen science data'

Subhash Lele

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2R3

Address for correspondence: Subhash Lele, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2R3. Email: slele@ualberta.ca

I applaud Drs. Bersson and Hoff for the paper 'Frequentist Prediction Sets for Species Abundance using Indirect Information' for bringing to attention data collected by citizen scientists. These data are particularly important because they increase the spatial and temporal extent for many bird monitoring surveys conducted by trained researchers. I have a couple of concerns and a suggestion.

1. Detection error occurs when a bird species that is present is not detected because of various factors such as season, time of the day, rainy day, and such. Correcting for detection error is a major issue in ecological studies. How do authors take into account detection error?
2. Structural zeros occur when a species may not be present at a location due to specific habitat characteristics. How do authors take into account such structural zeros? In their presentation, the number of species K is assumed to be the same in every location. This assumption does not seem realistic.
3. I would like to draw attention to a hierarchical model approach described in Lele and Allen (2006) that calibrates expert opinion or soft data such as collected by E-bird with hard observations collected by trained researchers. Florio et al. (2004) combined satellite data with large spatial extent and ground truth data with smaller spatial extent using a similar approach. In Lele and Allen (2006) formulation, the weight of the soft data in the analysis is automatically determined. As was illustrated, sometimes soft data can, in fact, deteriorate the quality of inference. The moral was that not all soft data are informative and hence should not be used without checking its usefulness and relevance. Could one conduct similar analysis to get prediction sets that combine not just the citizen science data from neighbouring counties but also data systematically collected by bird researchers?

Conflicts of interest: none declared.

References

Florio, E. N., Lele, S. R., Chi Chang, Y., Sterner, R., & Glass, G. E., (2004). Integrating AVHRR satellite data and NOAA ground observations to predict surface air temperature: A statistical approach. *International Journal of Remote Sensing*, *25*(15), 2979–2994. https://doi.org/10.1080/01431160310001624593

Lele, S. R., & Allen, K. L. (2006). On using expert opinion in ecological analyses: A frequentist approach. *Environmetrics*, *17*(7), 683–704. https://doi.org/10.1002/env.786

https://doi.org/10.1093/jrsssa/qnaf015
Advance access publication 18 February 2025

Allan Reese's contribution to the Discussion of the 'Discussion Meeting on the Analysis of citizen science data'

R. Allan Reese

Forston, Dorset, UK

Address for correspondence: R. Allan Reese, Forston, Dorset, UK. Email: allan@eurolyme.uk

Koh and Opitz present an impressive piece of modelling that covers a far larger geographical area than our own reporting of first-arrival dates (Reese & Tucker, 2019), which by comparison equates to some 9 by 9 20 km pixels. Our data, however, covered more than 130 years (1880s–2010s). We speculated on sources of bias, some of which are corroborated in the new work, and how these may have changed over time. Koh and Opitz's work would usefully be made more accessible to ornithologists, with the mathematics simplified and more biological interpretation.

One minor point is the coding of dates. The crude day number within a year introduces a small anomaly: each date after February is larger by one in leap years. This effect is easily removed by coding March 1 as Day 1, using negative values for January and February—which do not anyway occur in these data. The authors could also check for 'weekend bias', which we found in more recent (post-1950) reports.

Most studies, including our own, consider first arrival dates in a strict sense: 'the first Cuckoo of spring'. The definition here, 'first arrivals of migratory birds *at their destination*' is more obviously a latent variable. Many migrants may stop in a suitable habitat and, for example, sing before continuing their journey. The second definition, 'a date that can be viewed as representative of the arrival of birds at their breeding site', smacks more of a population average than the earliest pioneer. The span of dates in Figure 1 may reflect sampling variation, but also the large geographical spread. Surely one should be modelling the contours of the northward progress? The pixels appear to be coded with a single index ($i = 1,\ldots,1,268$) rather than in rows successively northward.

Our paper looked briefly at 'effort' based on the estimated number of observers, with equivocal results. For me, Figure 1 needed some decoding, and I would advise revising the labels and the caption. Values of $\log(N)$—I assume base e—do not easily convey a magnitude: the value of 10 equates to 22,000 observations in a pixel-year. I zoomed in on the PDF, and the observations appear to extend to at least 9 (8,000). If these were spread randomly across the year (e.g. not just wildfowl counts in winter), there is an extremely small Poisson probability of a day without data. Labelling with natural numbers, including the minimum and maximum number of observations, would be more understandable. Nor is it obvious why the Y-axis is inverted. Plotting the day number ascendant leads to the more natural description: more observations in an area lead to earlier dates with smaller variance. I would try a curve with a single inflexion and an asymptote.

A final point is that the interest in first dates is generally in relation to change. If the sources of bias and the direction of effect are largely unchanged, then changes in arrival dates have importance and need interpretation.

Conflicts of interest: none declared.

Reference

Reese R. A., & Tucker J. J. (2019). Descriptive models for first arrival dates of migrants in an inland UK county. *Ringing & Migration*, *34*(1), 63–69. https://doi.org/10.1080/03078698.2019.1759896

https://doi.org/10.1093/jrsssa/qnaf019
Advance access publication 11 February 2025

Stefano Rizzelli's contribution to the Discussion of the 'Discussion Meeting on the Analysis of citizen science data'

Stefano Rizzelli

Department of Statistics, University of Padova, Padova, Italy

Address for correspondence: Stefano Rizzelli, Department of Statistics, University of Padova, Padova, Italy.
Email: stefano.rizzelli@unipd.it

This contribution discusses the paper 'Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data' by Koh and Opitz, whom I congratulate on this work. Of particular interest is their use of Generalized Extreme Value distributions for modelling maxima Z_i of (transformed) birds arrival times, performing parameter inference, and point prediction through a Bayesian approach. One of its advantages is the possibility of incorporating field knowledge through prior distributions. Since the variables Z_i plausibly have an upper bound, it would be interesting to incorporate this information by assigning ξ a prior supported on negative values, and to assess the sensitivity of the proposed method in moving from a vague to an informative prior specification. Another advantage is the possibility of issuing forecasts of maxima through predictive distributions, as remarked in Coles and Pericchi (2003) and Padoan and Rizzelli (2022, 2024). Expanding the analysis in §5.3 to include predictive median (or mode), checking how much predictive distributions concentrate around it and whether predictive intervals cover truly observed arrival times, would also be interesting.

As the authors remark, fitting models grounded on extreme value theory (EVT) to extremes of biotic processes is a rare practice, which may seem surprising given the relevance of extreme events in life sciences. Besides EVT primary focus on continuous variables, this may be due to a relative inflexibility of the more established extreme value models, arising from EVT fundamental theorems (de Haan & Ferreira, 2006). While they allow to successfully account e.g. for spatial dependence (Davison et al., 2012), there are as yet no commonly agreed-upon assumptions to derive extreme value models in the presence of covariates. The need for extensions of classical EVT in this sense is evidenced by the plethora of applications for which traditional extreme value models have been expanded in different ways, according to problem-specific desiderata, without, however, dwelling with probabilistic foundations of such extensions and mathematical guarantees on the proposed methodologies, which is far from trivial but crucial to establish how reliably they can be used in other areas. Some efforts in this direction have been made by Bobbia et al. (2021), Chernozhukov (2005), Daouia et al. (2022), and Dombry et al. (2023), among others. Rigorous attempts to also account for time heterogeneity can be found in Bücher Jennessen (2024) and Einmahl et al. (2016, 2022).

Recent progress is remarkable, but there is still a long way to go to develop and consolidate rich, interpretable and mathematically sound models. Bridging the gap between theory and practice remains a key challenge for the advancement of extreme value analysis, making its toolbox increasingly usable and trustworthy for practitioners among different fields.

Conflicts of interest: The author declares to have no conflict of interest.

References

Bobbia B., Dombry C., & Varron D. (2021). Extreme quantile regression in a proportional tail framework. *Transactions of A. Razmadze Mathematical Institute*, *175*, 13–32. https://rmi.tsu.ge/transactions/TRMI-volumes/175-1/v175(1)-2.pdf

Bücher A., & Jennessen T. (2024). Statistics for heteroscedastic time series extremes. *Bernoulli*, *30*(1), 46–71. https://doi.org/10.3150/22-BEJ1560

Chernozhukov V. (2005). Extremal quantile regression. *Annals of Statistics*, *33*(2), 806–839. https://doi.org/10.1214/009053604000001165

Coles S., & Pericchi L. (2003). Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *52*(4), 405–416. https://doi.org/10.1111/1467-9876.00413

Daouia A., Gijbels I., & Stupfler G. (2022). Extremile regression. *Journal of the American Statistical Association*, *117*(539), 1579–1586. https://doi.org/10.1080/01621459.2021.1875837

Davison A. C., Padoan S. A., & Ribatet M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, *27*(2), 161–186. https://doi.org/10.1214/11-STS376

de Haan L., & Ferreira A. (2006). *Extreme value theory: An introduction*. Springer.

Dombry C., Padoan S. A., & Rizzelli S. (2023). *Asymptotic theory for Bayesian inference and prediction: From the ordinary to a conditional peaks-over-threshold method*. arXiv eprint 2310.06720, available at: https://arxiv.org/abs/2310.06720.

Einmahl J. H. J., de Haan L., & Zhou C. (2016). Statistics of heteroscedastic extremes. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *78*(1), 31–51. https://doi.org/10.1111/rssb.12099

Einmahl J. H. J., Ferreira A., de Haan L., Neves C., & Zhou C. (2022). Spatial dependence and space–time trend in extreme events. *Annals of Statistics*, *50*(1), 30–52. https://doi.org/10.1214/21-AOS2067

Padoan S. A., & Rizzelli S. (2022). Consistency of Bayesian inference for multivariate max-stable distributions. *Annals of Statistics*, *50*(3), 1490–1518. https://doi.org/10.1214/21-AOS2160

Padoan S. A., & Rizzelli S. (2024). Empirical Bayes inference for the block maxima method. *Bernoulli*, *30*(3), 2154–2184. https://doi.org/10.3150/23-BEJ1668

https://doi.org/10.1093/jrsssa/qnaf020
Advance access publication 11 February 2025

Maozai Tian’s contribution to the Discussion of the ‘Discussion meeting on the Analysis of citizen science data’

Maozai Tian

Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China

Address for correspondence: Maozai Tian, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China. Email: mztian@ruc.edu.cn

1 Comments and suggestions

I enthusiastically congratulate the authors on a stimulating article that provides a cutting-edge technology for implementing data fusion of heterogenous datasets and may inspire many follow-up works.

The paper develops a spatiotemporal Bayesian hierarchical model for bias-corrected estimation of arrival dates of the first migratory bird individuals at a breeding site, devises an MCMC scheme, and checks by simulation that the latent process components are identifiable. The proposed model is used to several migratory bird species and resulted in an interesting finding that the sampling effort significantly modulates the observed first arrival date. The relationship is effective in making bias-correct predictions of the true first arrivals.

Here are three comments:

First, the article develops a spatiotemporal Bayesian hierarchical model with two levels. The level-1 model is with the hierarchical system of the four regression equations in Section 3.3 on pages 7 and 8,

$$\begin{cases} N_j^{\mathrm{BBS}} \mid \lambda^{\mathrm{BBS}}, \boldsymbol{\theta}_{\mathrm{bbs}} \sim \mathrm{Pois}\left\{ \sum\limits_{k \in \mathrm{route}_j} \omega_k \lambda^{\mathrm{BBS}}(\boldsymbol{s}_k; \boldsymbol{\theta}_{\mathrm{bbs}}) \right\} \\ N_i^{\mathrm{ckl}} \mid \lambda^{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{ckl}} \sim \mathrm{Pois}\left\{ \lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{ckl}}) \right\} \\ N_i^{\mathrm{spc}} \mid N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}, \boldsymbol{\theta}_{\mathrm{spc}} \sim \mathrm{Bin}\{N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{spc}})\} \\ Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \mathrm{GEV}\{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_\sigma), \xi\}. \end{cases} \tag{1}$$

At level-2 model, the level-1 hyperparameters λ^{BBS}, λ^{ckl}, p^{spc}, μ, and σ become outcomes:

$$\begin{cases} \log \lambda^{\mathrm{BBS}}(\boldsymbol{s}_i) = \beta_0^{\mathrm{BBS}} + X^{\mathrm{niche}}(\boldsymbol{s}_i) \\ \log \lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i) = \beta_0^{\mathrm{ckl}} + X^{\mathrm{year}}(t_i) + X^{\mathrm{pref}}(\boldsymbol{s}_i) \\ \mathrm{cloglog}\{p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i)\} = \beta_0^{\mathrm{spc}} + X^{\mathrm{niche}}(\boldsymbol{s}_i) + \frac{\beta^{\mathrm{act}}}{d_{\boldsymbol{s}_i, t_i}} \\ \mu(\boldsymbol{s}_i, t_i) = g\Big\{\beta_0^{\mathrm{GEV}-\mu} + X^{\mathrm{GEV}-\mu}(\boldsymbol{s}_i) + \beta_1^{\mathrm{GEV}-\mu}\mathrm{NAO}_{t_i} \\ \qquad\qquad + \theta^{\mathrm{niche-GEV}} X^{\mathrm{niche}}(\boldsymbol{s}_i), x_{\mathrm{effort}}(\boldsymbol{s}_i, t_i)\Big\} \\ \log \sigma(\boldsymbol{s}_i) = \beta_0^{\mathrm{GEV}-\sigma} + X^{\mathrm{GEV}-\sigma}(\boldsymbol{s}_i). \end{cases} \tag{2}$$

In practice, the existence of such citizen science data with hierarchies is neither accidental nor ignorable, it is a common phenomenon. To ignore this, hierarchical data structure risks overlooking the importance of group effects and may also render invalid many of the traditional statistical analysis techniques used for studying data relationships. On the other hand, hierarchical models take hierarchical data structure into account and have also many applications in statistics. However, the aforementioned level-2 model fails in taking into account the inherent association among the four regression equations at level-1. So, it needs to impose, for example, some strict constraints on the joint distribution of the latent effect vector $(X^{\mathrm{niche}}(\boldsymbol{s}_i), X^{\mathrm{year}}(t_i), X^{\mathrm{pref}}(\boldsymbol{s}_i), X^{\mathrm{niche}}(\boldsymbol{s}_i), X^{\mathrm{GEV}-\mu}(\boldsymbol{s}_i), X^{\mathrm{GEV}-\sigma}(\boldsymbol{s}_i))$. Obviously, it is not suitable to assume that these latent effects be assigned independent and identical distribution, see, for example, Amini et al. (2021), Li et al. (2014), Ma and Tian (2024), Rotejanaprasert et al. (2023), Yu et al. (2022), and Tian and Yu (2024).

Second, what are the conditions for the identifiability of the model? There are seven hyperparameters $\beta_0^{\mathrm{BBS}}, \beta_0^{\mathrm{ckl}}, \beta_0^{\mathrm{spc}}, \beta^{\mathrm{act}}, \beta_0^{\mathrm{GEV}-\mu}, \beta_1^{\mathrm{GEV}-\mu}$, and $\beta_0^{\mathrm{GEV}-\sigma}$ as the fixed effects to estimate. Whether so many hyperparameter estimators will lead to the superposition of errors and the failure of the method. Refer to Yu et al. (2022) and Tian and Yu (2024).

Third, how to monitor the convergence of Markov Chain Monte Carlo algorithm for the proposed spatiotemporal Bayesian hierarchical model at each step? See, for example, Gnecco et al. (2024).

Acknowledgments

The authors thank the RSS Journals Manager Judith Shorten for her kind invitation to make presentation with discussion of the paper. Prof. Maozai Tian's work was supported by the Natural Science Foundation of Beijing (No. 1242005).

Conflicts of interest: None declared.

References

Amini P., Moghimbeigi A., Zayeri F., Tapak L., Maroufizadeh S., & Verbeke G. (2021). Longitudinal joint modelling of ordinal and overdispersed count outcomes: A bridge distribution for the ordinal random intercept. *Computational and Mathematical Methods in Medicine*, *2021*(1), 5521881. https://doi.org/10.1155/2021/5521881

Gnecco N., Terefe E. M., & Engelke S. (2024). Extremal random forests. *Journal of the American Statistical Association*, *119*(548), 3059–3072. https://doi.org/10.1080/01621459.2023.2300522

Li Q., Pan J., & Belcher J. (2014). Bayesian inference for joint modelling of longitudinal continuous, binary and ordinal events. *Statistical Methods in Medical Research*, *25*(6), 2521–2540. https://doi.org/10.1177/0962280214526199

Ma S. P., & Tian M. Z. (2024). A censored quantile transformation model for Alzheimer's disease data with multiple functional covariates. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *188*(2), 515–538. https://doi.org/10.1093/jrsssa/qnae061

Rotejanaprasert C., Lawpoolsri S., & Sa-angchai P. (2023). Projecting malaria elimination in Thailand using Bayesian hierarchical spatiotemporal models. *Scientific Reports*, *13*(1), 7799. https://doi.org/10.1038/s41598-023-35007-9

Tian M. Z., & Yu K. M. (2024). Maozai Tian and Keming Yu's contributions to the Discussion of "Safe Testing" by Peter Grüunwald, Rianne de Heide, and Wouter Koolen. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *86*(5), 1160. https://doi.org/10.1093/jrsssb/qkae076

Yu Z., Yu K. M., Härdle W. K, Zhang X. L., Wang K., & Tian M. Z. (2022). Bayesian spatio-temporal modeling for inpatient hospital costs of alcohol-related disorder. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(2), 644–667. https://doi.org/10.1111/rssa.12963

https://doi.org/10.1093/jrsssa/qnaf021
Advance access publication 10 February 2025

Authors' reply to the Discussion of 'Efficient statistical inference methods for assessing changes in species' populations using citizen science data' at the 'Discussion meeting on the analysis of citizen science data'

Emily B. Dennis[1,2], **Alex Diana**[3], **Eleni Matechou**[2] **and Byron J.T. Morgan**[2]

[1]Butterfly Conservation, Manor Yard, East Lulworth, BH20 5QP Dorset, UK
[2]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF Kent, UK
[3]School of Mathematics, Statistics and Actuarial Science, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ Essex, UK

Address for correspondence: Byron J.T. Morgan, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF Kent, UK. Email: B.J.T.Morgan@kent.ac.uk

Emily Dennis is unable to comment on the issues raised by the Discussion due to maternity leave.

We are grateful for the wide-ranging comments on our paper, which have given us encouragement and provide pointers to future research.

We have focused on providing analyses of data provided by *Butterfly Conservation* in the UK. For both abundance and distribution data, analyses are repeated regularly on sets of 59 species of butterflies and a subset of the approximately 900 species of larger moths. What we need therefore are efficient, flexible methods, where citizen-science records typically contain minimal covariate information.

Ben Swallow

We thank Ben Swallow for his wide-ranging comments covering citizen-science data, and for the particular comments relating to our paper.

He mentions the importance of modelling the observation process, which resonates with the paper by Moreira and Gamerman (2022), mentioned below. We note here one way in which this is not possible in comparison with planned surveys conducted by known individuals. In the latter case, there is often information available regarding the observers themselves, for instance regarding age and experience, and these are typically not available with citizen-science data.

As we say, the extended Generalised Abundance Index (GAI), incorporating the annual model, is at a preliminary stage of development. However, the examples that we have shown are promising, and we are in the process of undertaking a wider study. The annual model is much used in ecology and we believe that our new approach should have wide-ranging applications. This ties in also with the comments of Ben Baer and our responses to these given below. It is not uncommon for fitted models in ecology to result in parameter estimates which are then analysed further, for example using Generalised Linear Models (GLMs). In order then to account for error propagation some form of timeconsuming bootstrapping is used. Differences in the estimation of uncertainty are mentioned, requiring more research, however incorporating an annual model within the GAI avoids the arbitrariness of a post hoc analysis which might, for example, depend upon the use of a particular GLM.

The attraction of the GAI family of models is that it allows for a range of different discrete distributions for the analysis of count data, and for comparisons to be made using information criteria—see Dennis et al. (2016). In addition relevant covariates, such as northing, easting and relating to land cover, are easily included. Northing and easting can reveal spatial features; an illustration is given in Dennis et al. (2022), showing how the Common Blue butterfly, *Polyommatus icarus*, moves from having two broods in the south of Britain to having only one in the north, where the colder climate only allows one brood to be completed each year. More complex inclusion of spatial covariates results in the occupancy maps in Diana et al. (2023), which are reproduced in the Discussion contribution by Jafet and coauthors using Integrated nested Laplace approximations (INLA), which also incorporate spatial correlation.

The possibility of multi-species modelling is a good one, and applications are mentioned in McCrea et al. (2023) and Moreira and Gamerman (2022). The benchmarking used in the paper to produce presence–absence data uses observations on species other than target species to induce absences, and so is a particular type of multi-species analysis. UK butterflies are generally classified as habitat specialists, migrants, and wider countryside; specialists are only found where the habitat is favourable to them. This point is relevant to the utility of multi-species analyses, related to the construction of indicators which describe how groups of individual species behave. Indicators are widely used to summarize complex, multi-species analyses. However, indicators are only useful if they describe relatively homogeneous groups of species. Methods of cluster analysis and ordination are used to check this, and illustrations using functional principal component analysis are in Dennis et al. (2019), which illustrate in a two-dimensional plot the relationships between butterfly occupancy indices. It is an interesting point that modelling of data on specialists might benefit from appropriate analyses of data from wider countryside species with overlapping ranges.

Kerrie Mengersen

We welcome Kerrie's mention of relevant work in other areas and also the citizen-science research taking place in the antipodes. We very much endorse the need for rigor in model fusion, emphasized also in McCrea et al. (2023). As a comment on integrated modelling, in our paper, we analyse both count and distribution data on British butterflies, but separately, using different models in each case as abundance and distribution are inherently different. Typically data in the first case relate to individuals while in the second case they refer to species. As we mention in the Discussion to our paper, we are developing a new urban butterfly indicator based on an integrated analysis of UKBMS and BBC data, building on the material of Section 3 of the paper.

Citizen-science data are increasing in size, spatio-temporal coverage and species monitored, especially with the emergence of new technologies, such as cameras, acoustic, and DNA-based approaches. Therefore, they provide considerable opportunities for species' monitoring and timely intervention. At the same time, the unstructured, often opportunistic and biased process by which such data are collected presents modelling challenges, many of which remain open. For the data sets considered in our paper, reporting preferences, observer differences and spatial bias

(Johnston et al., 2023) are just some of the known and unavoidable biases in the data generating process. To maximize the benefit of citizen-science data for reliable and timely monitoring, these biases have to be addressed within a modelling framework, and we agree with Kerrie that more work is needed in the area.

Machine learning methods, especially deep models, are some of the most flexible and scalable tools currently available to us, and therefore it is indeed beneficial to incorporate them within statistical models of the type we present in our paper. For example, deep models can be used in place of logistic regression models to increase the flexibility of our framework. Variational inference approaches, such as the one we developed, provide a natural way forward for this task, thanks to the now widely available automatic differentiation tools.

Andrej Srakar

We thank Andrej for the interesting and thought-provoking ideas.

Our use of concentrated likelihood is a simple way of reducing the effective size of parameter spaces for particular survey data. More generally, the approach has been found to be useful in time-series analysis for example; see Durbin and Koopman (1997).

As we have shown, complex models in ecology may well have many parameters, and one way to deal with this is to use appropriate random effects. Dennis (2015) compared this approach with using concentrated likelihood and found the latter to be preferable for our models. In his contribution to the paper Discussion, Ben Baer both simplified the expression for a likelihood and computed a dimension reduction in one application from 527588 to 1222; further discussion with him has shown that additional concentration of the likelihood in this case allows us to go even further. Thus for the GAI-based models in particular, and the data being analysed, we can probably achieve the efficiency that we need at present. In terms of efficiency, Ben also suggests using stochastic gradient descent (which is used in the paper), and we need to look into the value of that for the GAI-based models. The potential of using Banach, Hilbert, or tensor spaces for our work is as yet unknown. However data and models are becoming more extensive and complex—McCrea et al. (2023)—and these suggestions may well gain traction.

In some of our work, likelihoods are not intractable, when methods of indirect inference may not be the best approach. Approximate Bayesian Computation, ABC, now has a wide literature—see for example Beaumont (2019) and Gutmann and Corander (2016) as an illustration of efficiency in computing. Each application requires its own simulators and data-matching functions, in both cases with the potential for time-consuming computation. One of the drivers for our work has been the need for methods to analyse data arising from established surveys, which can be used by ecologists with limited computational power. The main issues with ABC are that it relies on identifying summary statistics, which can be challenging with large spatio-temporal data and models with many parameters, of the type we consider here. ABC also does not scale well as the number of parameters increases (the curse of dimensionality, Cranmer et al., 2020). Instead, we think that when the likelihood is tractable, as is the case for the data and corresponding models of this paper then likelihood-based inference and efficient computational tools, such as variational inference, hold considerable promise. However, we expect that Andrej's suggestions are highly relevant for the new data that are regularly appearing, as with the use of autonomous machines for data collection for example.

In our paper, we have used both classical inference and variational inference. The two classical inference illustrations are easily performed and described. In general, the value of Bayesian methods for certain data analysis is evident from examples: for instance the variational inference illustration and also Diana et al. (2023), combining a Pólya-Gamma scheme and spatio-temporal random effects using Gaussian processes, the exact Bayes method of Moreira and Gamerman (2022), the extension of INLA to occupancy problems by Belmont et al. (2024), and the sampling approach of King et al. (2023) combined with importance sampling.

Dani Gamerman

The approach proposed in Moreira and Gamerman (2022) is particularly useful for presence-only data. However, the data and model considered in our paper occupy an intermediate position between presence-only and presence–absence data. While it is true that observations are initially recorded as presence-only data, observers document all species they encounter. This allows absences to be inferred whenever an observer records a different species but not the one under consideration. This process, known as benchmarking, enables the identification of absences in addition to presences, distinguishing the data from typical presence-only data. We note that if observers were to report on only one species, the data would indeed be presence-only. We do recognize, however, that the data are not pure presence/absence data either, since that would require observers to explicitly document absences whenever no species is recorded, which is not the case here. We believe that in this case it is preferable to infer non-detections for each species using records of other species (assuming that all species share the same reporting probability, i.e. their detections are equally likely to be reported) than to ignore that information all together and treat the data as presence-only.

We note with interest the approach of Moreira and Gamerman (2022), of assuming a latent process of unobserved occurrences as a way to mitigate preferential sampling. Our approach does not account for potential preferential sampling, which is a known issue in citizen-science surveys since observers might tend to visit sites that are more likely to be occupied, which leads to the sampled sites being a biased sample of all sites. Preferential sampling is a well-known but open problem when modelling citizen-science data; see Section 3.1 in Johnston et al. (2023), who comment that '… (preferential sampling is) particularly a challenge if the drivers of observer site selection are aligned with the ecological process of interest, and if there are no variables that describe this preference in the model.' To the best of our knowledge, preferential sampling has not been considered in models of abundance or occupancy of the type we develop and use in this paper—though see Conn et al. (2017)—so this is certainly an exciting future research direction.

The approach of Moreira and Gamerman (2022) relies on the presence of relevant covariates, and may not apply to various types of citizen-science data. For example in the UK, a recorder using iRecord is required to add a location to a submitted presence entry. We do not think that in general there is information on other covariates. As also observed in Moreira and Gamerman (2022), observer information in citizen science data can be difficult to obtain. Another feature might relate to sample sizes, as data can be very large; see for example Diana et al. (2023).

Finally, on the question regarding (co)variances, simulation results in our paper show that the marginal posterior distributions of parameters from our variational inference approach and Markov chain Monte Carlo (MCMC), are very similar. In response to the question about the posterior covariances, we have run a new set of simulations to compare the posterior covariances of the parameters $(\beta_1^{\psi}, \ldots, \beta_{p_1}^{\psi}, \beta_1^{p}, \ldots, \beta_{p_2}^{p})$ obtained by variational inference against those ones obtained by MCMC (which should be the true ones). We have used 2 covariates for both detection and occupancy (including the intercept). Across 50 simulations and the 6 posterior covariances, the average relative error was 38%. It is likely that the bias is still due to the low number of replications or to the lack of convergence of the variational inference procedure, since the estimator of the posterior covariance matrix tends to exhibit slower convergence for the off-diagonal elements.

Benjamin R. Baer

The extended GAI presents a way of including the main-effects, annual, model within the standard GAI framework. This simple idea allows the GAI to be applied in a true dynamic fashion, for analysis of count data over time. It is preferable to the current approach in which the static GAI is applied separately to each year, followed by a Poisson GLM in order to obtain time trends, as explained. Having a dynamic model avoids the need for time-consuming bootstrap analyses in order to deal with the variance propagation resulting from the GLM applying to estimated values. It is also desirable to remove reliance on a particular GLM. This new approach is preliminary, and we plan extensive checking on simulated data to investigate performance and properties, building on the initial good results presented for real data. We thank Ben for his useful development of the model. We focused on using concentrated likelihood to estimate the α parameters to reduce the dimensionality of the parameter space for maximum likelihood. Ben has taken this work further, for a particular submodel in which the phenological function $a(., ., .)$ is not a function of site. This restriction is in fact one that is often used in practice—see Dennis et al. (2013). He then proceeds to derive an analytical form for the likelihood which will also enhance computational efficiency. We have discussed this work with Ben which has resulted in an additional simplification due to the concentrated likelihood for the β parameters also resulting in explicit parameter estimates. This in turn provides a further improvement in efficiency which we look forward to examining in detail in the near future. The point about stochastic steepest descent is an interesting one, as it has in fact been used in Section 4 of the paper. We shall certainly explore its wider relevance.

Jafet Belmont, Sara Martino, Janine Illian, and Håvard Rue

The flexibility and speed of R-INLA provides considerable modelling opportunities for spatio-temporal data, such as those considered in our paper. It is interesting to see how observation error, in the form of the probability of detection given presence in occupancy models—accounting for false negative observation error—can now be accounted for within R-INLA, and undoubtedly this will be widely used by practitioners to model their data in the future.

We agree with the reviewers that in the specific case of an occupancy model, using INLA is more efficient than the variational inference approach. However, we believe that the advantage of variational inference is its greater flexibility in considering potentially more complex sampling scenarios that arise in citizen-science data, since the approach only requires us to be able to express the likelihood in closed analytic form (potentially having to integrate over discrete variables if these are present). Moreover, using variational inference, it is possible to take advantage of the developments on automatic differentiation and potentially model occupancy and detection effects using deep models, as we discuss in response to Kerrie's comments. Although in the specific case study considered there are more efficient approaches, as shown by the reviewers, we think that in the complex observational processes that give rise to different types of citizen-science data there would be considerable scope to perform Bayesian (variational) inference.

Conflicts of interest: None declared.

Funding

The UKBMS is organized and funded by Butterfly Conservation, the British Trust for Ornithology (BTO), and the Joint Nature Conservation Committee (JNCC). The UKBMS is indebted to all volunteers who contribute data to the scheme. We are also 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, and also the citizen scientists who submitted Big Butterfly Count data. B.M. was supported by a Leverhulme Emeritus Fellowship. This work was partly funded by NERC grant NE/T010045/1.

References

Beaumont M. A. (2019). Approximate Bayesian computation. *Annual Review of Statistics and Its Application*, *6*, 379–403. https://doi.org/10.1146/annurev-statistics-030718-105212

Belmont J., Martino S., Illian J., & Rue H. (2024). Spatio-temporal occupancy models with INLA. *Methods in Ecology and Evolution*, *15*(11), 1923–2167. https://doi.org/10.1111/mee3.v15.11

Conn P. B., Thorson J. T., & Johnson D. S. (2017). Confronting preferential sampling in wildlife surveys: Diagnosis and model-based triage. *Methods in Ecology and Evolution*, *8*(11), 1535–1546. https://doi.org/10.1111/mee3.2017.8.issue-11

Cranmer K., Brehmer J., & Louppe G. (2020). The frontier of simulation-based inference. *PNAS*, *117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117

Dennis E. B. (2015). *Development of statistical methods for monitoring insect abundance* [Ph.D. thesis]. University of Kent. https://kar.kent.ac/49079/

Dennis E. B., Fagard-Jenkin C., & Morgan B. J. T. (2022). rGAI: An R package for fitting the generalized abundance index to seasonal count data. *Ecology and Evolution*, *12*(8), e9200. https://doi.org/10.1002/ece3.v12.8

Dennis E. B., Freeman S. N., Brereton T., & Roy D. B. (2013). Indexing butterfly abundance whilst accounting for missing counts and variability in seasonal pattern. *Methods in Ecology and Evolution*, *4*, 637–645. https://doi.org/10.1111/2041-210X.12053

Dennis E. B., Morgan B. J. T., Freeman S. N., Brereton T. M., & Roy D. B. (2016). A generalized abundance index for seasonal invertebrates. *Biometrics*, *72*(4), 1305–1314. https://doi.org/10.1111/biom.12506

Dennis E. B., Morgan B. J. T., Roy D. B., & Brereton T. M. (2019). Functional data analysis of multi-species abundance and occupancy data sets. *Ecological Indicators*, *104*(1–2), 156–165. https://doi.org/10.1016/j.ecolind.2019.04.070

Diana A., Dennis E. B., Matechou E., & Morgan B. J. T. (2023). Fast Bayesian inference for large occupancy data sets, using the Pólya-Gamma scheme. *Biometrics*, *79*(3), 2503–2515. https://doi.org/10.1111/biom.13816

Durbin J., & Koopman S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, *84*(3), 669–684. https://doi.org/10.1093/biomet/84.3.669

Gutmann M., & Corander J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, *17*, 1–47. //jmlr.csail.mit.edu/papers/volume17/15-017/15-017.pdf

Johnston A., Matechou E., & Dennis E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103–116. https://doi.org/10.1111/mee3.v14.1

King R., Sarzo B., & Elvira V. (2023). When ecological individual heterogeneity models and large data collide: An importance sampling approach. *The Annals of Applied Statistics*, *17*(4), 3112–3132. https://doi.org/10.1214/23-AOAS1753

McCrea R., King R., Graham L., & Börger L. (2023). Realising the promise of large data and complex models. *Methods in Ecology and Evolution*, *14*(1), 4–11. https://doi.org/10.1111/mee3.v14.1

Moreira G. A., & Gamerman D. (2022). Analysis of presence-only data via exact Bayes, with model and effects identification. *The Annals of Applied Statistics*, *16*(3), 1848–1867. https://doi.org/10.1214/21-AOAS1569

https://doi.org/10.1093/jrsssa/qnaf041
Advance access publication 11 April 2025

Authors' reply to the Discussion of 'Frequentist prediction sets for species abundance using indirect information' at the 'Discussion meeting on the analysis of citizen science data'

Elizabeth Bersson[1] and **Peter D. Hoff**[2]

[1]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Department of Statistical Science, Duke University, Durham, NC 27701, USA

Address for correspondence: Elizabeth Bersson, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email: ebersson@mit.edu

To begin, we would like to thank all discussants for their insightful comments on our article, and congratulate Dennis, Diana, Matechou, Morgan, Koh, and Opitz on their contributions to the discussion meeting that enabled such an enriching conversation on analysis of citizen science data. Much of the discussion comments pertain to how our proposed framework coincides with unique concerns of citizen science data. In what follows, we group common themes that arose in the discussion and respond.

1 Assumptions and applicability to citizen science data

Our work presents methodology to construct a prediction set based on avian species abundance count data from a given county. Moreover, we present an empirical Bayesian framework to incorporate indirect information to improve prediction set precision using auxillary data from neighbouring counties. The prediction set for a given county will have guaranteed frequentist coverage if two primary assumptions are upheld: The data are exchangeable within the county and are independent from the auxiliary data. To this end, Swallow, Mengersen, Srakar, and Lele each commented on some unique aspects of citizen science data that require particular attention in any statistical methodology, which opens a discussion on the practicality of the assumptions required for our approach.

1.1 Within-county exchangeability

For the frequentist coverage guarantee to hold, our proposed method relies on an assumption of exchangeability of samples. Swallow points out that this may be a concern as sampling bias is a warranted concern with citizen science data. Some concerns of biases with citizen science-collected data may be alleviated by pre-processing data such that only entries collected with some sufficient level of effort are used in an analysis. In the eBird database, data entries may be filtered using some of the ample meta-data available regarding the data collection session. For example, utilizing data from complete checklists (eBird, 2020) and from checklists uploaded by frequent contributors can ensure only data with a sufficient level of skill and effort are being utilized in an analysis (Horns et al., 2018).

More generally, for the coverage guarantee of our approach to hold, the sum of the count vectors obtained from each birding session must follow a multinomial distribution when conditioning on the total number of observed birds. This assumption will be violated, for example, if repeated samples are uploaded from individuals in a group who collected data together. These entries will be correlated, and the sum of the corresponding count vectors will not follow a multinomial distribution. In the eBird database, there is meta-data identifying if separate checklists were uploaded from a group birding session, and duplicate lists may be excluded from an analysis using this information. There is recent work that relaxes the exchangeability assumption in conformal prediction (Barber et al., 2023), and extending our method in this direction may be useful for citizen science datasets with more prevalent biases.

1.2 Between-county independence

In this work, we are motivated by utilizing auxiliary data from neighbouring counties to improve prediction set precision for a given county. This is accomplished via a hyperparameter that may be estimated from data that is independent from the sample in the county of interest.

As Srakar highlights, geographic adjacency structure is an inherently interesting aspect of much citizen science data. In our approach, we propose an estimation procedure that estimates the hyperparameter with auxiliary data from the nearest neighboring counties. In this way, we allow for information to be shared among regions that are geographically near one another.

Instead, we may aim to share information among counties that exhibit similar ecological patterns in a more nuanced manner, perhaps allowing for greater geographical distance. For example, in some contexts, it may be more useful to share information across counties with similar county-level environmental covariates. Accounting for more complex dependencies between the counties in this way can be integrated into a more heuristic-based estimation procedure for the indirect information. As an example, the distribution of the transformation of counts of observations of a single species can be approximately modelled with a normal distribution with a spatial dependence structure, such as a simultaneous auto-regressive (SAR) model (Singh et al., 2005) that includes a regression with county-level covariates. Then, the estimated hyperparameter may be taken to be an empirical Bayes estimate of the mean from this distribution. This perspective also pertains to Lele's suggestion of incorporation of alternative expert data.

1.3 Additional data concerns

In this work, we analyze eBird data from the state of North Carolina in the United States. For inference on each county, we take the categorical support to include any species seen in the entire state during the time frame we analyze. Lele notes this all-encompassing approach may result in structural zeros such that a species may have near zero probability of occurrence in a given county. While the support is the same in each county, the species' occurrence probability vector varies across state, so this is theoretically coherent.

An additional concern with citizen science data, raised by Lele, is detection error that arises when species are not detected due to various extenuating circumstances that affect data collection, including inclement weather, time of day, and others. Some of these concerns may be alleviated by aggregation, however, this solution is dependent upon the inferential question of interest.

2 Auxillary information

Lele discusses the nuance around incorporating auxiliary information in an analysis. Although in some settings it may be undesirable to incorporate indirect information that is uninformative, our approach is adaptable to such a scenario. For example, if many species with small observed counts have large hyperparameter values, then this reflects a general uncertainty of knowledge, and, consequently, will result in a large prediction set. Alternatively, if the indirect information is aligned with the sample, then a smaller prediction set will be obtained.

3 Coverage probability

Swallow points out that the coverage probabilities can be above the nominal rate, as seen in Figure 2. Indeed, our method guarantees a lower bound on the coverage rate of a prediction set (Eqn 2). The realized coverage rate of the indirect approach depends on, among others, the sample size N, as described in Bersson and Hoff (2024). Exact coverage may be achieved by utilizing a randomized confidence set algorithm, see Vovk et al. (2022) §2.2.6. In contrast to the α-validity guaranteed by the indirect approach, a Bayesian or empirical Bayesian prediction set can have arbitrarily poor coverage, depending on the prior hyperparameter.

Conflicts of interest: None declared.

References

Barber R. F., Candès E. J., Ramdas A., & Tibshirani R. J. (2023). Conformal prediction beyond exchangeability. *Annals of Statistics*, *51*(2), 816–845. ISSN 21688966. https://doi.org/10.1214/23-AOS2276

Bersson E., & Hoff P. D. (2024). Optimal conformal prediction for small areas. *Journal of Survey Statistics and Methodology*, *12*(5), 1464–1488. https://doi.org/10.1093/jssam/smae010

eBird (2020). *Birding as your 'Primary Purpose' and complete checklists* (Technical Report). https://support.ebird.org/en/support/solutions/articles/48000967748-birding-as-your-primary-purpose-and-complete-checklists.

Horns J. J., Adler F. R., & Şekercioğlu Ç H. (2018). Using opportunistic citizen science data to estimate avian population trends. *Biological Conservation*, *221*(5), 151–159. ISSN 00063207. https://doi.org/10.1016/j.biocon.2018.02.027

Singh B. B., Shukla G. K., & Kundu D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, *31*(2), 183–195. https://www.researchgate.net/publication/237110155

Vovk V., Gammerman A., & Shafer G. (2022). *Algorithmic learning in a random world* (2nd ed.). Springer US.

https://doi.org/10.1093/jrsssa/qnaf040
Advance access publication 4 April 2025

Authors' reply to the Discussion of 'Extreme-value modelling of migratory bird arrival dates: insights from citizen-science data'

Jonathan Koh[1,2] and Thomas Opitz[3]

[1]Institute of Mathematical Statistics and Actuarial Science, Oeschger Centre for Climate Change Research, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland
[2]Seminar for Statistics, ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland
[3]Biostatistics and Spatial Processes (UR546), INRAE, 228 route de l'Aérodrome, 84914 Avignon, France

Address for correspondence: Jonathan Koh, Seminar for Statistics, ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland. Email: jonathan.koh@stat.math.ethz.ch

Abstract

We respond to the discussion comments of the Proposer and Seconder of the vote of thanks, and to the eight other contributions discussing our work.

We warmly thank the Proposer and Seconder of the vote of thanks, and all discussants. We appreciate their numerous observations, suggestions, and questions on statistics for citizen science (CS) data in general, and their insightful and encouraging comments on our work in particular.

The Proposer raises perspicacious points on the challenges of working with CS data, highlighting the inherent structural biases and lack of appropriately randomized sampling involved with data collection. These challenges resonate strongly with the difficulties faced in our work, and have guided our considerations when building our model. Besides fostering stronger interactions with data collectors and end users of model output, better communication between statisticians, domain experts, and citizen scientists should continuously be stimulated, not just to ensure its continuity by increasing public awareness of the scientific process (Bonney et al., 2016) but also to improve data and model quality. As observed by Reese, '*Koh and Opitz's work would usefully be made more accessible to ornithologists, with the mathematics simplified and more biological interpretation*'.

We are inspired by the positive comments of the Seconder. We agree with the sentiment that statisticians are increasingly being trusted as 'data whisperers', and this comes at an exciting time with new combinations of models, algorithms, and computing power. Though we acknowledge the need to propose methods appropriate for other 'wilder' datasets in this domain, we hope that our work on a 'well-behaved' CS dataset will trigger new extensions and collaborations, and can be a reference point for further discussion on this important topic.

Several key aspects are fundamental for statistical methods to extract useful insights from the wealth of data sources and powerful AI tools that should guide new methodological developments. We have to explain the phenomena and achieve good predictive behaviour by separating the ecological signal from the noise and biases, thus allowing for generalization and extrapolation of results. Statistical models representing bias and noise explicitly with latent variables (e.g. through spatio-temporal random effects) provide interesting paths to this goal. Moreover, we should provide decision support through statistical inference (e.g. selecting variables and finding significant effects), and provide probabilistic statements to assess remaining uncertainties. We believe that Bayesian hierarchical models (BHMs) provide an appropriate framework to jointly address these aspects.

Regarding the discussion contributions on our work, we have thematically regrouped their comments along with our responses in the five sections below. Extreme-value aspects are addressed in Section 1; Section 2 discusses validating complex BHMs, especially when weakly structured CS data are involved. Considerations on model identifiability, efficient estimation algorithms, and the potential of machine learning (ML) techniques follow in Section 3. Possible new applications of CS statistics beyond ecology are summarized in Section 4. Section 5 concludes with a discussion on the opportunities of CS data and model fusion.

1 Extreme-value distributions and their priors

A key focus of our work is the analysis of extreme phenological events (i.e. life-history events of individuals of a species) using probability distributions motivated by extreme-value theory (EVT). We iterate our hope that this work will spur broader applications of these techniques in ecology and other fields concerned with biotic processes where they are still scarce, and we are appreciative of the insightful comments by the discussants on this theme.

Reese scrutinizes the definition of first arrival dates that we model with the generalized extreme value (GEV). It is indeed difficult for an observer to know whether a spotted bird has arrived at its *destination* (i.e. the breeding site) or has just stopped before continuing its journey. Nevertheless, we were able to implement some basic data filtering operations using the ‘flyover’ flag in eBird, which observers can toggle when they are unsure whether the bird has already arrived at its destination. We also believe that there might be a ‘weekend bias’, though this bias could be evened out across pixel-years in our spatio-temporal model. Reese’s idea to model dates consistently over leap years, for example by modelling the first date as 1st March, is also interesting; this is easy to implement and would only lead to a location shift for the fitted GEV. We agree that there should be more discussion between ecologists and statisticians to better differentiate and estimate both individual and populational behaviour of migratory bird arrivals.

To address Reese’s comments on Figure 1 of our main paper, our reasoning to invert the y-axis there is based on the EVT convention of modelling maxima instead of minima; extreme values are typically associated with the right tail. In our case, the higher our transformed variable (which we later use for modelling), the earlier the date. We also chose to have $\log(N_i)$, with N_i the number of checklists in pixel-year i, in the x-axes of that figure since we later model the checklist count with a log-link function. We appreciate Reese’s suggestion to estimate contours of northward migration of birds and will note this for follow-up work. Relatedly, Figure 1 shows a strong dependence of the first arrivals on latitude, a feature that has also been detected by our model; see Figure 9 of the main paper.

The Proposer suggests extending our approach to multivariate modelling by using copulas and especially multivariate extreme-value distributions, to share information across species and to enable joint inference of the first arrivals of multiple species. We agree that multivariate modelling would be a natural and worthwhile extension and could provide further insights into bird community structure. Regarding our study, we argue that it would pertain to a different modelling philosophy. It may require sacrificing important benefits of mono-specific modelling and of the conceptually and computationally useful conditional independence assumption in BHMs. Using multivariate probability distributions for the response variable would entail building our model components differently, and care would have to be taken not to lose computational tractability or omit important model components. A major drawback of the joint modelling approach is the inability to simply parallelize computations across tens to thousands of species as we do in our work. Choosing the right multivariate distribution also raises questions about choosing the copula that can capture the appropriate dependence class for the first arrivals of multiple species, and this is an avenue yet to be explored. An alternative approach to share information between species, more in line with our current model, would be to keep the conditional independence assumption inherent in our BHM but to additionally incorporate cross-species dependencies into the latent processes, for instance using tensor products of Gaussian processes as suggested by the Proposer. Multivariate Gaussian processes are already routinely used in joint species distribution models, which have their foundations on Chib and Greenberg (1998), with many recent developments (e.g. Chiquet et al., 2021; Ovaskainen & Abrego, 2020; Tikhonov et al., 2020). We concur that using computationally efficient representations (e.g. Gauss Markov random fields) could indeed make certain multivariate extensions feasible, even for complex models combining multiple response variables for each species.

Rizzelli advocates for more informative prior distributions for the shape parameter ξ of the GEV, which is expected to be negative in our case study due to the finite upper bound of the variable of

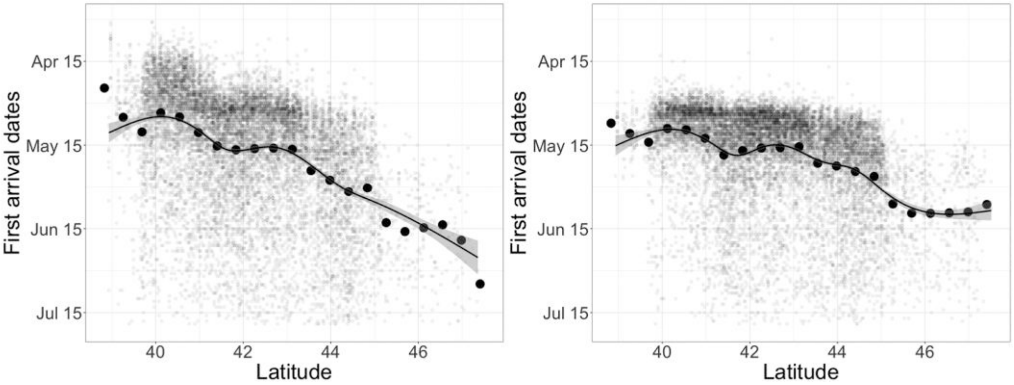

Figure 1. Scatterplots of the latitude (*x*-axes) and first arrivals (*y*-axes, with later dates corresponding to lower values) observed at each pixel-year combination for *Chimney Swift* (left) and *Chestnut-sided Warbler* (right). The larger dots show 21 binned estimates, the black line a smooth fitted curve, and the shaded region the 95% pointwise confidence intervals of the curve.

transformed first arrival days. This is a valid point, and a strategy in this direction is to use penalized complexity priors (Simpson et al., 2017), though the choice of the reference distribution when building this prior remains less clear in our application. Choosing the reference distribution as the GEV with $\xi = 0$, as advocated in Opitz et al. (2018), penalizes relatively heavy or short tails in the GEV, and we could truncate this prior such that only negatives values of ξ are possible. However, this would problematically push the posterior towards earlier arrivals in our application, especially when observations are scarce. A different approach is to model $\tilde{\xi} = \log(-\xi)$ instead, forcing ξ to be negative.

Modelling $\tilde{\xi}$ instead of ξ falls under the general strategy of reparametrization, which could also be motivated by parameter identifiability and computational reasons; we discuss this more in Section 3. One could also directly model the finite upper bound of the GEV as a parameter $B = \mu - \sigma/\xi$ with the usual parameter notation, and in our case incorporate the g function on this bound. Alternatively, one could avoid estimating this upper bound B but rather fix it to a known value, e.g. the value corresponding to 1st January or 1st March, or based on other physical constraints; see Noyelle et al. (2024) who develop this approach in the context of temperature extremes. In a reparametrized model, one of the three GEV parameters could be equal to the upper bound.

Rizzelli also discusses the general inflexibility of extreme-value models, especially for modelling contexts with biotic processes, and relatedly discusses EVT in the presence of covariates. When working with Peaks-Over-Threshold models and the resulting asymptotic Generalized Pareto distribution, common practice in the covariate setting is to assume conditional independence of the response given the covariates, such that the usual asymptotics from the classical 'i.i.d.' setting hold after the conditioning (e.g. see Koh et al., 2025). By contrast, deriving theoretical asymptotics for maxima over non-identically distributed variables is more involved; for example, see the results derived for extreme-value limits of upper-bounded distributions with seasonal patterns in Stein (2017), and for semiparametric tail estimation (i.e. without a limit probability distribution such as the GEV) in Einmahl and He (2023). We agree that rigorous attempts to provide mathematical guarantees in the presence of covariates, as done for example in the temporal heterogeneity setting, should be further explored. Though asymptotically justified distributions confer more robustness in data-scarce contexts, asymptotic properties remain difficult to verify formally in many data applications.

Daouia and Stupfler propose an interesting approach from econometrics based on frequentist frontier analysis (Daouia et al., 2016), with the frontier defined as the lower bound of arrivals, i.e. the first arrival date. They suggest using all data covering the study region and incorporating as covariates the spatial coordinates, exogenous climate, land cover, and other ecological factors into the estimation procedure in a nonparametric regression framework with one-sided errors. This could be applied in continuous space without needing a spatial pixel discretization. Instead, one would use a kernel with a data-driven choice of its bandwidth to pool nearby observations to estimate local parameters. However, as far as we understand, this approach would not attempt to explicitly debias the boundary estimate, which is the direction taken in our work and others in ecological statistics that explicitly capture the observation process (Gelfand & Shirota, 2019). We acknowledge that the sampling effort from the preference dimension would be captured indirectly by the width of confidence intervals of the estimated frontier, though it is less clear to us how one should systematically incorporate the activity component (i.e. the 'temporal effort' of an observer during an observation event). Pooling data locally around each prediction location could certainly provide spatially smoother predictions of the first arrivals. However, we believe that it would also be quite sensitive to the spatio-temporally heterogeneous distribution of sampled locations in our application, and it might be ill-adapted to the sharp spatial changes in land cover configurations and the corresponding changes in ecological processes. Nevertheless, we would be delighted to see future work combining our model with the one proposed.

2 Validation of inferences and debiased predictions

A central question when modelling CS data is how models, and the bias corrections for varying and unknown observational effort they provide, can be validated. Attenuating the possible biases in model-based predictions requires either explicitly identifying the effort or predicting only quantities that are not affected by varying effort under realistic assumptions. This is a challenging and fundamental topic of research in CS statistics, and not surprisingly appears as a common theme among many of the discussants.

The Proposer poses questions about validating latent variable models in the presence of complex observational processes. We reiterate that using structured data (i.e. collected by following a known predefined sampling design) to define validation datasets, or constructing integrated models combining structured and unstructured data sources, remain the best options to reliably validate inferences on observation biases. The Proposer also questions how one should estimate the effective number of parameters in our setting to assess model complexity and calculate information criteria that allow comparing different models. This appears challenging, and a reasonable approach could focus on the parameters describing the ecological process by integrating out the observational process before estimating the model's remaining degrees of freedom.

Our model's debiasing mechanism is based on appropriately identifying the observation process jointly with the other model components. An approach to better validate this mechanism is to randomly remove observation days to artificially reduce the observational effort, and then verify the observed first arrivals, calculated including the hold-out data, with the model-debiased ones estimated under lower observational effort. Similarly, one could assume that regions with high observational effort should already be at the saturation point of our g function, and verify whether the observed first arrivals in these regions are not significantly different from the model-debiased ones under high observational effort.

Rizzelli highlights the advantages of the Bayesian approach, such as the possibility of using predictive distributions for forecast evaluation, including using summary statistics like the predictive mode and median. We agree that validation schemes holding out recent data to mimic prediction with relatively short lead times, along with coverage checks on predictive distributions, are relevant when models are used to forecast into the near future. These approaches have been considered in previous studies with BHMs (e.g. in wildfire modelling contexts by Koh et al., 2023).

Tian emphasizes building hierarchical frameworks with identifiable components. Though we verify this for our model with a simulation study, we resonate with his call to explore the conditions that ensure identifiability of parameters in complex BHMs more generally.

We also agree with Belzile and Yadav that a possible model assessment approach is to use k-fold cross-validation designs appropriately stratified for spatial and temporal dependence. This is generally good practice, but it would be computationally demanding in our setting.

3 Conducting computations and inference

We here respond to the numerous comments on how model building, Bayesian computation and inference can be conducted for complex models such as ours.

Achieving computational efficiency while staying flexible with respect to model structure and inferrable quantities remains challenging despite continual progress in computing power and algorithmic efficiency. A recommended practice for BHMs is to structure their latent variables and hyperparameters to maximally facilitate their identification, and to keep auto-correlations low among the posterior samples from hyperparameters and latent variables.

3.1 Identifiability of latent spatio-temporal fields

Belzile and Yadav remark that some components of our latent model may be difficult to identify from data. Specifically, they notice that the field $x_{\text{bound}}(s, t)$, which is in the GEV location parameter $\mu(s, t) = \exp\{x_{\text{bound}}(s, t)\}/\{1 - \exp(-x_{\text{effort}}(s, t)\}$, could be hard to identify at space-time locations (s, t) where the field $x_{\text{effort}}(s, t)$ representing observational effort is low. We consider the way our model works in this situation rather as a feature than a problem: if the effort is low at (s, t), then the GEV location parameter will tend to be smaller, which implies that observed first arrivals are expected to occur relatively late in the year t, or not at all if observations are missing; this feature is coherent. Moreover, if the effort is low at (s, t), then the posterior of the field $x_{\text{bound}}(s, t)$ will be strongly influenced by the behaviour at surrounding locations and its intercept term, i.e. the model will borrow information from nearby locations where more data are available and from global behaviour. We do not see an identifiability issue here.

Regarding the identifiability of our latent fields, one could explore alternative parameterizations of the GEV that were shown to yield weaker correlations between the posterior samples of these three parameters, as outlined in Section 1. Moins et al. (2023) propose a reparameterization that reduces the dependence between the posterior samples of the three GEV parameters and improves identifiability. Using this alternative parametrization in our BHM could facilitate identification of the latent variables and parameters we incorporated into them. Another idea to aid identifiability is to impose additional modelling assumptions, e.g. as in Section 2, one could assume that data from high observational effort regions should already be the g function's saturation point, and impose this as a constraint in the model.

Tian questions whether additional constraints or prior dependencies should be imposed to further link the different Gaussian random effects in our model and ensure identifiability. We underline that these effects are of quite different nature, characterizing, for example, spatial niche variability, interannual variability, spatial preferential sampling, and spatial variation of true first arrivals. We do not see strong a priori links between these different effects and their sources of variability. Therefore, we believe that independent Gaussian process priors are more appropriate for our model.

3.2 MCMC-based inference and alternatives

Several discussants, including the Proposer, have commented on the relative slow convergence and mixing of some of the posterior chains in our model, and have asked (e.g. Tian) how one could efficiently monitor convergence of MCMC algorithms for complex hierarchical space-time model. Suggestions for algorithmic improvement from Belzile and Yadav include closer monitoring of effective sample size (ESS), using adaptive MCMC schemes instead of the global Metropolis-adjusted Langevin algorithm (MALA) we implemented, performing joint hyperparameter updates when posteriors of hyperparameters are correlated, and running multiple chains. We agree that these techniques could help improve our MCMC algorithm and its diagnostics. Another interesting option we see would be to use Hamiltonian MCMC instead of MALA to better explore high-density regions of the posterior distribution through the Hamiltonian dynamics (Neal, 2011). Moreover, restructuring the model form to improve identifiability, as discussed in the Section 3.1, could also aid convergence.

As the Proposer recalls, Gauss–Markov random field (GRMF) structures facilitate fast calculations with latent fields in spatial Bayesian frameworks. The INLA-SPDE approach is a well-known example (Lindgren et al., 2011; Rue et al., 2009); it uses GMRF approximations of spatial Matérn covariance functions based on the stochastic partial differential equation (SPDE) approach, combined with fast and accurate deterministic posterior approximations with the INLA method. The Proposer suggested that simulation-based MCMC inference, as implemented in our work, could be combined with the INLA-SPDE approach depending on the structure of the components of the

BHM. This could leverage INLA for model components where linear combinations of latent variables in GMRF form arise, whereas slower MCMC inference can be used only for parts with strongly nonlinear structures. We agree that such extensions bear promise. Combined inference of this type has been developed by Gómez-Rubio and Rue (2018), where INLA is run within an MCMC algorithm by conditioning on those other variables that would have impeded full INLA-based inference. This approach could indeed lead to performance gains for our model, although its implementation appears complex, and further research into such combined methods would be welcome in the setting of complex spatio-temporal Bayesian inference. Our main algorithmic difficulty preventing us from using faster inference tools comes from the nonstandard, nonlinear and parametric transformation of GMRFs that achieves a saturating effect when the observational effort increases to infinity, as also recognized by the Proposer. In this respect, the `inlabru` package (Bachl et al., 2019) with its extensions of the classical INLA method could offer help through implementation of methods that handle nonstandard nonlinear transformations of GMRFs, but to our knowledge, it does not yet cover the parametric transformations we would need. In response to Tian, who suspected possible issues with the relatively large number of fixed-effect coefficients (e.g. regression constants) in our model, we reiterate that they are endowed with Gaussian priors and are part of the linear predictors in the model (i.e. of multivariate Gaussian vectors), such that their estimation does not pose specific challenges.

3.3 Vecchia approximation

The Proposer questions whether the Vecchia approximation provides a sufficiently accurate representation of the underlying covariance function. Indeed, differences could arise between the approximation and the exact covariance function. However, we emphasize that the principal role of the Vecchia-approximated Gaussian field in our model is to serve as a rich prior with spatial dependence—what matters is the flexibility it provides for posterior inferences. We do not necessarily want to exactly reproduce the underlying covariance function, but rather obtain a prior that has similar and interpretable properties in terms of variance and correlation range. We do not use the Vecchia approximation to simulate data directly from the prior model, where stronger deviations could arise between the exact and approximated simulations.

3.4 Likelihood-free amortized inference

Huser and Zammit-Mangion outline the advantages of likelihood-free inference using neural networks (NNs) and recent developments in this area. Neural Bayes estimators are NNs trained on large amounts of simulated data to estimate (i.e. *predict* through the NN) the parameters of a stochastic process. They have emerged as powerful alternatives to likelihood-based estimation (Zammit-Mangion et al., 2025). The approach is still limited to models with a moderate number of parameters, since the parameter configurations used for simulating training data should cover the possible range of parameters to be estimated from real data. While estimating the large number of latent variables in a BHM through this approach appears unattainable, the discussants question whether hyperparameters, usually of moderate number in a BHM, could be estimated with the NN approach. Other parameters, especially the latent variables representing fixed effects and spatio-temporal effects, would be estimated separately. Intercepts could be considered as hyperparameters (i.e. estimated by the NN) or as latent variables (i.e. estimated separately). Given the proven benefits of neural Bayes estimators for estimating parameters of various spatial stochastic process classes, we recognize that they could also facilitate inference for spatial latent-process models. They would be especially beneficial in situations where parameters of the same model have to be estimated for a large number of different datasets (e.g. for different regions of the globe).

We comment on possible challenges for developing neural Bayes estimators in the setting of latent models for discrete responses with covariates. Three estimation steps could be necessary: (i) covariate coefficients (including intercepts) using a pseudo-likelihood regression technique; (ii) hyperparameters using a neural Bayes estimator; (iii) latent variables using a chosen technique (e.g. MCMC or INLA with fixed covariate coefficients and hyperparameters, or new NN-based approaches). Clearly, one would lose the benefits of performing joint inference on all model components, but could in turn gain in robustness and speed.

Neural Bayes estimators may be more difficult to implement when there are covariates. In latent-variable models for discrete response data (e.g. binary or count data), it is not possible to define classical standardized residuals, and marginal distributions and the dependence structure cannot be fully separated as in copula techniques for continuous margins. Therefore, it is more intricate to separate fixed covariate effects acting on marginal distributions from the random spatio-temporal effects governed by hyperparameters. For example, with binary and count data, the residuals take values outside the discrete support of the response. Recall that this is unlike continuous response distributions, such as the normal $\mathcal{N}(\mu_i, \sigma_i^2)$, where we could estimate marginal parameters μ_i and σ_i^2 in a first step, then generate standardized residuals by normalizing observations y_i to $(y_i - \mu_i)/\sigma_i$, and finally infer dependence hyperparameters from standardized residuals, using methods such as the NN estimator for the last step. This is not possible with discrete data, which also explains the popularity of fully Bayesian inference (i.e. with joint estimation of the posterior distributions of hyperparameters and latent variables) when modelling this data, which are very common in fields such as ecology.

To make the NN approach for estimating hyperparameters feasible in the presence of covariates, one could estimate the fixed-effect part with covariates in a preliminary step and then feed it into a NN as one of its inputs used for predicting the hyperparameters. However, there are essentially no constraints on the values of the fixed effect field, which presents a very high-dimensional input to the NN. This could make it very challenging to simulate a sufficiently large number of NN training instances that ensure good generalization behaviour.

We would be very interested to see how research on this topic could lead to NN estimators for latent-variable models that cope well with the aforementioned challenges.

3.5 Simplified estimation of shared latent fields

Instead of estimating certain hyperparameters differently, or separate from the other parameters, as suggested by Huser and Zammit-Mangion, one could also simplify the model by separately estimating some of the latent variable fields, or some of the regression equations.

We emphasize that the motivation for our BHM consists of jointly estimating all model components to allow for optimal information flow between components and to precisely assess posterior uncertainties. However, to alleviate computational cost and improve identifiability, one could instead separately estimate some of the shared fields by using only a subset of the response variables for which they appear in the predictor. The species niche field $X^{\text{niche}}(s)$ could be estimated using only the BBS data (with its $\lambda^{\text{BBS}}(s)$ parameter) and the binary presence–absence indicator for each checklist (with its $p^{\text{spc}}(s, t)$ parameter). We would not use the GEV response of first arrival dates to estimate the species niche field, for which it appears in the GEV location parameter $\mu(s, t)$. Now, assume we have estimated the species niche field separately from the GEV model in a first step. Then, in a second step, posterior means of the separately estimated fields could be included as a simple covariate (i.e. as a fixed effect) in the GEV model. We could proceed similarly for other spatial fields (e.g. the observational effort $\lambda^{\text{ckl}}(s, t)$) that we share towards the GEV response by estimating them separately in a preliminary step. Implementing this simplified model structure should drastically reduce computational cost while maintaining identifiability and reliable predictions.

3.6 Using ML algorithms

The main proposed utility of our model is the separation of the ecological signal from heterogeneous observational effort. This is made possible by identifying how the latter modifies outcomes when we consider several data sources with different observation design. Disentangling the roles of these different latent processes cannot be easily achieved with standard ML algorithms, even more so if they have black-box like behaviour. Nevertheless, we believe that ML models can still be useful if the predictor and response variables are adapted appropriately, but we are not aware of any existing general framework proposed to this end in the literature.

We provide a specific example where standard ML algorithms could be leveraged. Assume that only the presences of the individuals of various species were reported, with non exhaustive and varying observational effort, i.e. we have opportunistic presence-only data. Using a point-process framework, we can construct a response variable that can be estimated with essentially any classification algorithm as follows: The points of the point process are defined as the reported presences of a focal species for which we want to make predictions. A set of background points representing pseudo-absences of this species are constructed from the reported presences of other species for which similar observational effort can be reasonably assumed. Therefore, the background points represent a proxy for the general observational effort. Any classification algorithm trained to classify the points into 'species' vs. 'background' can then be used to predict the probability of species presence conditional on environmental predictors and location. The predicted probabilities can be interpreted as the relative abundances of the focal species with respect to the group of background species, which can provide valuable ecological insights. This so-called *target-group background* approach can be deployed even for purely opportunistic data without the need for strong assumptions on sampling design (Botella et al., 2020, 2021, Phillips et al., 2009).

In the above example, and certainly in many other CS models, ML approaches could be highly efficient, such as the use of NNs or stochastic gradient boosting techniques advocated by Kumar. Crucially, the model and the input data must be appropriately structured, which requires solid assumptions on how data were collected and how ecological and observational processes interact.

Kumar also encourages using ML models in extreme-value modelling. We concur with this sentiment, though this approach is less common when modelling count data; for an example, Koh (2023) uses gradient boosting to predict the parameters of discrete generalized Pareto distributions.

4 Pushing CS beyond ecology

CS data are commonly exploited in statistical ecology, and especially for species distribution modelling. Consequently, this research community has become a major driver for advancing statistical methods for CS purpose. As the Proposer and Srakar suggest, CS-based approaches are promising in various other areas of applied science, including epidemiology and public health, environmental surveillance, and possibly even in fields such as medicine, robotics, and law. We absolutely share this view. Due to its potential for collecting large amounts of diverse data, CS can provide valuable contributions to integrated modelling, for instance of entire ecosystems in the *One Health* perspective (Destoumieux-Garzón et al., 2018). CS also appears highly relevant for filling data gaps through the collection of data using mobile devices, especially in countries with low coverage of standardized datasets (e.g. in the Global South), and can help monitor and assess the sustainable development goals formulated by the United Nations.

It is challenging to establish statistical guarantees for reliable CS-based inferences similar to the asymptotic results in classical statistics, and no general theory has emerged so far. This is certainly due to the wide variety of application-specific data collection processes and the lack of well-known sampling designs, especially with opportunistic data. As Srakar highlights, more work should be invested into theoretical guarantees for CS-based approaches. We second this opinion and would like to incite statisticians to propose appropriate sampling design frameworks that preserve the simplicity and low cost of CS-based data collection while ensuring reliable inferences and predictions. Checklist-based approaches are an interesting example in species monitoring: observers must comply to exhaustively report relevant observations. Alternatively, the software interface used to report the species occurrences could explicitly ask the observers to confirm the nonobservation of certain other species not reported. Datasets collected with this protocol allows modellers to discern the absence of an observer (i.e. missing data) from the absence, or non detection, of any species individuals. One can then define pseudo-absences of individuals, which is crucial for incorporating observational effort and reducing bias in species distribution mapping, as highlighted in our case study.

However, ecological datasets covering past periods often contain only opportunistic presence-only data: the absence of a species is not explicitly reported, and the reporting is not exhaustive, i.e. it is possible that the observer detects an individual of a species but does not report it, so we are not in the convenient checklist setting mentioned before. Then, a strategy is needed to generate useful pseudo-absences, for example by using the target-group background approach outlined in Section 3.6. Given a focal species that one wants to model, the space-time locations where other species (the *target group*) were reported could be declared as pseudo-absences for the focal species. This requires carefully choosing the target group of species and the subsample scheme for its occurrences (Botella et al., 2020, 2021, Phillips et al., 2009; Van Strien et al., 2013). Knowledge of ecological experts is indispensable for this step, and it highlights the strong need for interdisciplinary work between data scientists and domain experts when developing statistical methods for CS data.

5 Opportunities of data and model fusion

By developing sampling approaches and models that integrate unstructured CS data with structured ones, we could benefit from both the often dense spatio-temporal coverage of CS data and the precisely known observation design and effort in structured data. The fundamental question was asked by the Proposer: When and how should we do this? Regarding spatial species distribution modelling, recent studies (e.g. Simmonds et al., 2020) highlight considerable added value from data integration, provided that models are appropriately structured using spatial random effects. Ideally, one would need covariates that strongly correlate with observational effort, but not with the species presence or abundance. However, such covariate data are typically unavailable in practice, except in very specific sampling situations (e.g. distance sampling). Instead, one usually needs both dataset-specific and shared spatial random effects to disentangle the observational effort from ecological effects. The flip side of implementing multiple spatial random effects and regression equations in these approaches is that estimation becomes computationally very intensive, and setting up models and algorithms may require considerable expertise and experience.

The Seconder further encourages statisticians to compare and hybridize statistical and ML approaches to tackle such complex statistical questions. We are curious how ML techniques could be adapted to mimic the estimation of spatial effects; for example, one could try to achieve this by feeding the predictor data, spatial coordinates, and response data into the prediction algorithm in a specific way. The example we outline in Section 3.6, where the modelling question is transformed into a classification problem for ML algorithms, is an example where combining expertise on statistical modelling with ML leverages the simple use of ML.

We also suggest to investigate how existing large CS datasets with limited knowledge about observational effort could be efficiently complemented by the additional collection of relatively small datasets in controlled surveys. The sampling design for newly collected standardized data would have to ensure reliable inferences and identification of observational effort through integrated models combining CS and survey data, while the cost of sampling in terms of required human and material resources must remain manageable. Modern technology for automated data collection (e.g. camera traps and other remote sensing devices) could be used for collecting data. Principles and techniques of *optimal design* can be applied (Müller, 2005; Pukelsheim, 2006) to develop variants of stratified or adaptive sampling, for example in a model-based perspective. BHMs, such as modifications of our model, could be useful for Bayesian optimal design experiments (Ryan et al., 2016) adapted to geostatistical applications (Diggle & Lophaven, 2006).

Once we have general standardized statistical methods to obtain trustworthy CS-based insights, routine publication of scientific results from CS would be facilitated. We are aware of the long road ahead, but are optimistic about the future.

Acknowledgments

The authors thank the Royal Statistical Society for facilitating the discussion meeting.

Conflict of interests: The authors declare no conflict of interests.

References

Bachl F. E., Lindgren F., Borchers D. L., & Illian J. B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, *10*(6), 760–766. https://doi.org/10.1111/mee3.2019.10.issue-6

Bonney R., Phillips T. B., Ballard H. L., & Enck J. W. (2016). Can citizen science enhance public understanding of science? *Public Understanding of Science*, *25*(1), 2–16. https://doi.org/10.1177/0963662515607406

Botella C., Joly A., Bonnet P., Munoz F., & Monestiez P. (2021). Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, *12*(5), 933–945. https://doi.org/10.1111/mee3.v12.5

Botella C., Joly A., Monestiez P., Bonnet P., & Munoz F. (2020). Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *PLOS One*, *15*(5), e0232078. https://doi.org/10.1371/journal.pone.0232078

Chib S., & Greenberg E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361. https://doi.org/10.1093/biomet/85.2.347

Chiquet J., Mariadassou M., & Robin S. (2021). The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, *9*, 588292. https://doi.org/10.3389/fevo.2021.588292

Daouia A., Noh H., & Park B. U. (2016). Data envelope fitting with constrained polynomial splines. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *78*(1), 3–30. https://doi.org/10.1111/rssb.12098

Destoumieux-Garzón D., Mavingui P., Boetsch G., Boissier J., Darriet F., Duboz P., Fritsch C., Giraudoux P., Le Roux F., & Morand S. (2018). The one health concept: 10 years old and a long road ahead. *Frontiers in Veterinary Science*, *5*, 14. https://doi.org/10.3389/fvets.2018.00014

Diggle P., & Lophaven S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, *33*(1), 53–64. https://doi.org/10.1111/sjos.2006.33.issue-1

Einmahl J. H., & He Y. (2023). Extreme value inference for heterogeneous power law data. *Annals of Statistics*, *51*(3), 1331–1356. https://doi.org/10.1214/23-AOS2294

Gelfand A. E., & Shirota S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, *89*(3), e01372. https://doi.org/10.1002/ecm.2019.89.issue-3

Gómez-Rubio V., & Rue H. (2018). Markov Chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, *28*(5), 1033–1051. https://doi.org/10.1007/s11222-017-9778-y

Koh J. (2023). Gradient boosting with extreme-value theory for wildfire prediction. *Extremes*, *26*(2), 273–299. https://doi.org/10.1007/s10687-022-00454-6

Koh J., Pimont F., Dupuy J.-L., & Opitz T. (2023). Spatiotemporal wildfire modelling through point processes with moderate and extreme marks. *The Annals of Applied Statistics*, *17*(1), 560–582. https://doi.org/10.1214/22-AOAS1642

Koh J., Steinfeld D., & Martius O. (2025). Using spatial extreme-value theory with machine learning to model and understand spatially compounding weather extremes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *481*, 20240763. https://doi.org/10.1098/rspa.2024.0763

Lindgren F., Rue H., & Lindström J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Moins T., Arbel J., Girard S., & Dutfoy A. (2023). Reparameterization of extreme value framework for improved Bayesian workflow. *Computational Statistics & Data Analysis*, *187*, 107807. https://doi.org/10.1016/j.csda.2023.107807

Müller P. (2005). Simulation based optimal design. *Handbook of Statistics*, *25*, 509–518. https://doi.org/10.1016/S0169-7161(05)25017-4

Neal R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (Vol:2(11), p. 2).

Noyelle R., Robin Y., Naveau P., Yiou P., & Faranda D. (2024). Integration of physical bound constraints to alleviate shortcomings of statistical models for extreme temperatures. HAL-04479249.

Opitz T., Huser R., Bakka H., & Rue H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes*, *21*(3), 441–462. https://doi.org/10.1007/s10687-018-0324-x

Ovaskainen O., & Abrego N. (2020). *Joint species distribution modelling: With applications in R*. Cambridge University Press.

Phillips S. J., Dudík M., Elith J., Graham C. H., Lehmann A., Leathwick J., & Ferrier S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Pukelsheim F. (2006). *Optimal design of experiments*. SIAM.

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*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Ryan E. G., Drovandi C. C., McGree J. M., & Pettitt A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*(1), 128–154. https://doi.org/10.1111/insr.v84.1

Simmonds E. G., Jarvis S. G., Henrys P. A., Isaac N. J., & O'Hara R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422. https://doi.org/10.1111/ecog.2020.v43.i10

Simpson D., Rue H., Riebler A., Martins T. G., & Sørbye S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28.https://doi.org/10.1214/16-STS576

Stein M. (2017). Should annual maximum temperatures follow a generalized extreme value distribution? *Biometrika*, *104*(1), 1–16. https://doi.org/10.1093/biomet/asw070

Tikhonov G., Opedal Ø. H., Abrego N., Lehikoinen A., de Jonge M. M., Oksanen J., & Ovaskainen O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, *11*(3), 442–447. https://doi.org/10.1111/mee3.v11.3

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. https://doi.org/10.1111/jpe.2013.50.issue-6

Zammit-Mangion A., Sainsbury-Dale M., & Huser R. (2025). Neural methods for amortized inference. *Annual Review of Statistics and Its Application*, *12*, 311–335. https://doi.org/10.1146/annurev-statistics-112723-034123

https://doi.org/10.1093/jrsssa/qnaf058
Advance access publication 14 May 2025