Received: 18 August 2023 | Accepted: 26 April 2024

DOI: 10.1111/2041-210X.14356

Methods in Ecology and Evolution

RESEARCH ARTICLE

Spatio-temporal data integration for species distribution modelling in R-INLA

Fiona M. Seaton | Susan G. Jarvis | Peter A. Henrys

UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, UK

Correspondence
Fiona M. Seaton
Email: fseaton@ceh.ac.uk

Funding information
Department for Environment, Food and Rural Affairs, UK Government, Grant/Award Number: 32156; Natural Environment Research Council, Grant/Award Number: NE/R016429/1

Handling Editor: Luis Cayuela

Abstract

1. Species distribution modelling is a highly used tool for understanding and predicting biodiversity change, and recent work has emphasised the importance of understanding how species distributions change over both time and space. Spatio-temporal models require large amounts of data spread over time and space, and as such are clear candidates to benefit from model-based integration of different data sources. However, spatio-temporal models are highly computationally intensive and integrating different data sources can make this approach even more unfeasible to ecologists.
2. Here we demonstrate how the R-INLA methodology can be used for model-based data integration for spatio-temporally explicit modelling of species distribution change. We demonstrate that this method can be applied to both point and areal data with two contrasting case studies, one using the SPDE approach for modelling spatio-temporal change in the Gatekeeper butterfly (*Pyronia tithonus*) across Great Britain and the second using a spatio-temporal areal model to describe change in caddisfly (Trichoptera) populations across the River Thames catchment.
3. We show that in the caddisfly case study integrating together different data sources led to greater understanding of the change in abundance across the River Thames both seasonally and over 5 years of data. However, in the butterfly case study moving to a spatio-temporal context exacerbated differences between the data sources and resulted in no greater ecological insight into change in the Gatekeeper population.
4. Our work provides a computationally feasible framework for spatio-temporally explicit integration of data within SDMs and demonstrates both the potential benefits and the challenges in applying this methodology to real ecological data.

KEYWORDS
citizen science, data integration, integrated distribution models, integrated nested laplace approximation, spatio-temporal models, species distribution models, stochastic partial differential equation

1 | INTRODUCTION

Species distribution modelling (SDM) is a popular tool in ecological research because it provides an empirical relationship between environmental factors and the spatial preferences of a particular species (Elith & Leathwick, 2009). It therefore provides a basis for understanding the most influential factors governing species' distributions, quantifying the impacts of changes in environmental factors and for consideration of effective conservation and management strategies (Franklin, 2010). Most SDMs are purely spatial in nature, but there is increasing interest in spatio-temporal modelling of species distributions, where the spatial distribution itself is explicitly a function of time (e.g. Fidino et al., 2022; Johnston et al., 2023; Ward et al., 2015). Spatio-temporal models help to understand change in a spatially explicit manner and therefore, the areas that are changing most, those that are more at risk, and those where certain factors are having disproportionate effect (Ward et al., 2015). However, a major barrier to the uptake of spatio-temporal SDMs is the availability of data with sufficient spatial and temporal coverage. Most SDMs are based on single sources of empirical data, and this can be limiting, with estimated relationships within the model and specific parameters only as good as the data underpinning them (Fourcade et al., 2018; Guillera-Arroita et al., 2015).

In recent years, there has been a proliferation of ecological data from a variety of sources, with increasing uptake and delivery of citizen science schemes as well as new technologies (Johnston et al., 2023). Concomitantly, there has been increasing interest in how to integrate these disparate data sources to increase our insight into species distributions (Isaac et al., 2020; Miller et al., 2019). While traditional, purely spatial, species distribution models stand to benefit significantly from data integration, spatio-temporal models have an even stronger requirement for data (Bakka et al., 2018). This is because the same data requirements exist as for the purely spatial case, but that the gradient coverage is also required over time as well, hence implying that good spatial coverage is needed for each individual time point. Data sources used for the modelling of species distributions vary in their observation intensity and accuracy across space and time, particularly so within unstructured, opportunistically collected data but also within many structured, professionally collected data sources (Binley & Bennett, 2023; Johnston et al., 2023). The ability to integrate different data sources should, therefore, offer significant potential for the modelling of species distributions across time and space. However, careful consideration throughout this process needs to be applied to thinking about the differential coverage and biases across the data sources, as in many cases, integration alone cannot be enough to integrate out bias in the data collection (Simmonds et al., 2020). This is particularly important when considering spatio-temporal models, as data sources can show biases across time that vary by space, which will impact not just the confidence of the model but also the predictions and inferences gained from the model.

There is a growing need for methods that facilitate the integration of data from diverse sources within a spatio-temporal framework. While recent work by Fidino et al. (2022) has demonstrated the benefits of integrating data for spatio-temporally explicit modelling of species distributions to understand human–wildlife conflicts, the MCMC-based method they employed is computationally intensive. When dealing with opportunistic species data sets, where sample sizes can easily be in the tens of thousands, this computational intensity can pose a significant constraint on analysis. There is a clear demand for more practical and efficient approaches to spatio-temporal data integration. In this study, we propose a method for integrated modelling of spatio-temporal change using the R-INLA methodology, which uses Integrated Nested Laplace Approximations for Bayesian inference, allowing for efficient estimation of highly complex spatio-temporal models (Bakka et al., 2018; Rue et al., 2009). This is the first time this methodology has been used in the application of spatio-temporal integrated modelling. R-INLA has been shown to be highly accurate for these types of spatio-temporal models while also being computationally far faster than MCMC-based methods (Bakka et al., 2018; Rue et al., 2017). The methods presented here allow for the modelling of spatio-temporal change including interactions between space and time such that different spatial regions can change in different ways (Blangiardo et al., 2013; Cameletti et al., 2013). We apply this integrated spatio-temporal model to two case studies, the first examining the spread of the Gatekeeper butterfly (*Pyronia tithonus*) across Great Britain according to two separate recording schemes and the second evaluating change in caddisfly (Trichoptera) populations across the River Thames comparing governmental monitoring to governmental monitoring plus citizen science schemes. The first case study lends itself to a pointwise spatio-temporal modelling approach and also provides an opportunity to consider how to integrate two spatially biased but information-rich data sources, whereas the second case study lends itself to an areal spatio-temporal modelling approach and offers an opportunity to consider how to integrate data sources where one is far more patchy spatially than the other.

2 | MATERIALS AND METHODS

2.1 | Modelling approach

There are multiple ways of integrating multiple data sources in a single model, and here we focus on shared modelling where joint likelihoods are used to model the different data sources (Pacifici et al., 2017). Previous studies have examined how methods such as point process models can be used to integrate different data sources together in a spatially explicit manner to infer species distributions (e.g. Fletcher et al., 2019). The concept behind model-based data integration is that different data sources observe some latent state, such as the true species distribution, using a variety of different observation models. Therefore, by jointly modelling the data sets together with their observation models, the latent species distribution can be inferred. Here we model the latent species distribution as a distribution over space and time, which requires explicit consideration of the dependence structure of observations over these dimensions. Doing so enables any spatial and/or temporal effects unaccounted for by model covariates, to be represented within the model. From a modelling perspective, this is achieved by including spatio-temporal random effects within the model as follows. Suppose we observe a species count Y at location s at time t, and this is assumed to follow a negative binomial distribution (overdispersion parameter n) with mean given by the product of the expected count ($E_{s,t}$) and the relative suitability for the species ($\mu_{s,t}$):

$$Y(s,t) \sim \mathrm{NB}\left(E_{s,t} \cdot \mu_{s,t}, n\right) \tag{1}$$

Then we can model the (log) relative suitability as a function of K spatially and temporally indexed covariates ($z_{s,t}$) and a spatio-temporal random effect $\omega_{s,t}$:

$$\log\left(\mu_{s,t}\right) = \alpha_0 + \sum_{k=1}^{K} \alpha_k z_{k,s,t} + \omega_{s,t} \tag{2}$$

In practice, the spatio-temporal random effect can be a sum of spatial and temporal structured and unstructured effects plus an interaction. The random effect structure can allow for consistent change over time across the whole spatial field or allow different parts of the spatial field to change in different directions over time—that is, a spatio-temporal interaction. Both the spatial and the temporal effects, as well as their interaction, can be specified in different ways, such as the difference between specifying change over time as a random walk or as an autoregressive process. Therefore, the term spatio-temporal models represent a diverse set of different model configurations that incorporate differing assumptions and constraints on the model fit. We detail the model configurations we have chosen to use within this manuscript below. However, our approach to using INLA for integrating data in spatio-temporal models can be generalised to other model configurations supported by INLA.

When integrating data from different sources, the same principles extend from the purely spatial case. Following the principles set out by Simmonds et al. (2020), the concept of shared covariate effects and shared random fields in the linear predictor extend to the spatio-temporal case, so for two data sources, P and Q, we model as a function of its own intercept (α_0), shared covariate effects (α_K) and a shared spatio-temporal effect ($\omega_{s,t}$), as in Equations (3) and (4). Data set-specific spatial (ϕ_s), or spatio-temporal ($\phi_{s,t}$), effects can also be added where appropriate to account for spatial or spatio-temporal biases within the data sources, as in Equation (4). Data set-specific intercepts and variance parameters, where appropriate, are included in the models to account for differences in observation processes between the data sets, as is recommended in the purely spatial case.

$$\log\left(\mu_{s,t,P}\right) = \alpha_{P,0} + \sum_{k=1}^{K} \alpha_k z_{k,s,t} + \omega_{s,t} \tag{3}$$

$$\log\left(\mu_{s,t,Q}\right) = \alpha_{Q,0} + \sum_{k=1}^{K} \alpha_k z_{k,s,t} + \omega_{s,t} + \phi_{s,t} \tag{4}$$

The spatio-temporal random effect(s) can take multiple forms depending on the structure of the data. Spatial structures commonly used within the SDM literature are approaches that use Gaussian random fields to model species occurrence over a defined region based on estimating the spatial covariance between points across the field. Alternatively, data that are provided on an areal basis can be modelled using (intrinsic) conditional auto-regressive models, where the model includes an effect of the values of neighbouring regions. The INLA methodology can be used to fit models, using both of these methods, is often used in integrated species distribution models and is less computationally intensive than MCMC-based methods (Blangiardo et al., 2013; Isaac et al., 2020).

2.1.1 | Point data

Here we model the spatial dependency between point data for any given timepoint ($\omega_{s,*}$ in (2)) through the stochastic partial differential equations (SPDE) approach for approximating a Gaussian random field (Lindgren et al., 2011). This approximates the Matérn covariance function across a defined region in a more computationally efficient way compared to the estimation of the entire covariance function so that computational time does not scale as a power law of the number of locations. The Matérn covariance function defines how much spatial relationship there is between points, given by a function of σ (the marginal standard deviation), ρ (the range parameter) and ν (a smoothness parameter):

$$c_\nu(r;\sigma,\rho) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu}\frac{r}{\rho}\right)^{\nu} K_\nu\left(\sqrt{8\nu}\frac{r}{\rho}\right) \tag{5}$$

where K_ν is the modified Bessel function of the second kind, order ν. Within INLA (and elsewhere), the simplification for specific values of ν is often used, as when $\nu = p + 1/2, p \in \mathbb{N}^+$, this covariance function can be expressed as a product of an exponential and a polynomial of order p. Therefore, ν is usually fixed to some positive half integer and not estimated as part of the model fitting process.

Fuglstad et al. (2019) reparameterised this function to enable the introduction of a penalised complexity (PC) prior, as introduced by Simpson et al. (2017). These PC priors involve setting a base model that has the highest probability weight upon it and enforcing a constraint such that the probability declines the further away from the base model you get. Within the SPDE approach, the base model is represented by a model with infinite range and zero covariance—that is, no spatial pattern. Within a two-dimensional space, this penalised complexity prior corresponds to

$$\pi(\sigma,\rho) = \lambda_\rho \lambda_\sigma \rho^{-2} \exp\left(-\lambda_\rho \rho^{-1} - \lambda_\sigma \sigma\right), \quad \sigma > 0, \rho > 0 \tag{6}$$

where $\mathrm{P}\left(\rho < \rho_0\right) = \alpha_1$ and $\mathrm{P}\left(\sigma > \sigma_0\right) = \alpha_2$ are achieved by

$$\lambda_\rho = -\log\left(\alpha_1\right)\rho_0 \quad \text{and} \quad \lambda_\sigma = -\frac{\log\left(\alpha_2\right)}{\sigma_0} \tag{7}$$

The reason for this PC parameterisation is that the user must specify both a limit and the total probability that is given to be above (for σ) or below (for ρ) this limit.

For any given point in space (denoted by $*$), the change by year is modelled as an order one autoregressive process as in Equation (8), with a PC prior upon the autoregressive parameters with the base model being no change over time (Sørbye & Rue, 2017). As the stochastic element of this model (ϵ) is different at every point in space and time, this therefore does not enforce that the spatial field must change in the same way across the whole field and instead allows for the direction and magnitude of the change over time to vary across the spatial field and between time points.

$$\omega_{*,t=1} \sim \mathcal{N}\left(0, \left(\tau_y\left(1-r^2\right)\right)^{-1}\right) \tag{8}$$

$$\omega_{*,t=i} = r\omega_{*,t=i-1} + \epsilon_{*,i}, \quad \epsilon \sim \mathcal{N}\left(0, \tau_y^{-1}\right) \quad i = 2, \dots, n$$

where the lag one correlation $|r| < 1$. Within INLA, this is parameterised with the hyperparameters θ_1, θ_2, where θ_1 is the log of the marginal precision (Equation 9), and θ_2 is the logit of the lag one correlation (Equation 10).

$$\theta_1 = \log\left(\tau_y\left(1-r^2\right)\right) \tag{9}$$

$$\theta_2 = \log\left(\frac{1+r}{1-r}\right) \tag{10}$$

Within the data integration, there is assumed to be a shared spatio-temporal field between the two data sets, with each data set having its own intercept, survey effort parameters and family-specific scale parameters (i.e. θ within the negative binomial family).

2.1.2 | Areal data

We model areal data using the Besag York Mollié (BYM) model, which is a combination of the Besag model and an independent random noise (IID) model (Besag et al., 1991). The IID model simply defines $\omega_{s,*}$ as a vector of independent and Gaussian-distributed observations with precision τ_1.

Within the Besag model, each $\omega_{i,*}$ is defined as dependent on the sum of the weighted values of its neighbours:

$$\omega_{i,*} \mid \omega_{j,*}, i \neq j, \tau_2 \quad \sim \quad \mathcal{N}\left(\frac{1}{n_i}\sum_{i \sim j}\omega_{j,*}, \frac{1}{n_i\tau_2}\right) \tag{11}$$

where n_i is the number of neighbours of node i, $i \sim j$ indicates that the two nodes i and j are neighbours.

The BYM model combines this model together such that

$$\omega_{s,*} = \begin{pmatrix} \text{besag} + \text{iid} \\ \text{besag} \end{pmatrix} \tag{12}$$

Within this manuscript, we use the parameterisation of the BYM model explored by Riebler et al. (2016) within the PC framework, where the base model has no spatial pattern or differences between areas. Therefore, the parameters that we can define priors over this BYM model for (referred to by INLA as hyperparameters) are the marginal precision (τ_b) and the proportion of the variation explained by the spatial component of the model (ϕ). Priors are defined on these on the log and logit scales, respectively. The default PC priorto the precision is a type-2 Gumbel distribution:

$$\pi\left(\tau_b\right) = \frac{\lambda}{2}\tau_b^{-3/2}\exp\left(-\lambda\tau_b^{-1/2}\right), \quad \tau_b > 0, \lambda > 0 \tag{13}$$

where $\lambda = -\ln(\alpha)/U$.

The default prior within INLA is $\alpha = 0.01$ and $U = 1$, which corresponds to a standard deviation of around 0.3 (Simpson et al., 2017). The PC prior to ϕ is dependent on the graph within the BYM model and is derived for each model with a computational cost that scales to the cube of the number of graph components. The parameters that can be specified for this prior are u and α, where the probability that $\phi < u$ is equal to α. The defaults within INLA (at the time of writing) are $u = 0.5$ and $\alpha = 0.5$ which give equal weight to the spatial component explaining less than or more than half of the marginal variation. We model change over time as an interaction with the spatial model, with the change over time modelled as an autoregressive model of order one and assuming no effect of the past state of neighbouring regions upon the current state of the area.

In addition to the spatio-temporal model, we can also add other model elements, such as a seasonal term and/or some covariates (e.g. survey effort, environmental conditions). The seasonal model with periodicity m for some random vector $\left(x_1, \dots, x_n\right), n > m$ can be obtained by assuming the sums $x_i + x_{i+1} + \dots x_{i+m-1}$ are independent Gaussian with precision τ_s. The density for $\mathbf{x}$ is derived from the $n - m + 1$ increments as in Equation (14).

$$\pi\left(\mathbf{x}\mid\tau_s\right) = \tau_s^{\frac{n-m+1}{2}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right\} \tag{14}$$

where $\mathbf{Q} = \tau_s\mathbf{R}$ and $\mathbf{R}$ is the structure matrix representing the neighbourhood structure of the seasonal model (e.g. that March follows February, which in turn follows January), and the default prior on τ_s is a PC prior on precision as in the BYM model.

Within the data integration, there is assumed to be shared spatio-temporal and seasonal models between the two data sets, with each data set having its own intercept, survey effort covariates if appropriate, and family-specific scale parameters (i.e. θ within the negative binomial family). One data set can also be assumed to have a separate spatial model (that does not update with year) to represent the differing spatial bias within the data set (i.e. there is also a $\phi_{s,t}$ term within the model).

2.2 | Case studies

2.2.1 | Gatekeeper butterfly

To demonstrate the method upon point data, we modelled the spread of the Gatekeeper butterfly across Great Britain from 2005 to 2014, using the UK Butterfly Monitoring Scheme (UKBMS) and non-avian data collected by the British Trust for Ornithology plus partner organisations (BTO) as part of their bird monitoring. The UKBMS data consisted of abundance at transects that are regularly surveyed by volunteers for all butterflies, while the BTO data consisted of observations of butterflies submitted by citizen scientists as part of their Garden BirdWatch and BirdTrack schemes that are reported as presence-only records of butterflies. The UKBMS gatekeeper abundance data and the BTO presence data consisted of ~6000 measurements each once data processing as detailed below was carried out, with both data sources having more observations to the south of GB and showing some evidence of increasing numbers of observations over time (Figures S1 and S2).

We aggregated the BTO data to the 10 km grid square level, with the total number of gatekeeper observations within that square within the year being modelled as a function of the total list length (i.e. unique number of butterfly species) observed in that square within that year. Aggregating to the 10 km grid square level allowed us to identify and remove the regions with a very low effort during the peak flight period of the Gatekeeper butterfly, with square/year combinations that had 10 or fewer observations across all butterfly species within July and August being removed from the analysis. The 10 km scale also resulted in a similar data set size for the BTO and the UKBMS measurements, meaning there did not need to be any consideration of weighting of likelihoods to account for the different data set sizes. This resulted in a count response with a range from 0 to 19 and a median of 2, which then allowed the modelling of both the BTO and UKBMS data as count variables.

UKBMS data were represented as the generalised abundance index, which gives an annual abundance value that accounts for whether the timings of site visits across the year corresponded with the seasonal pattern of butterfly emergence (Botham, Brereton, Harrower, et al., 2020; Dennis et al., 2016). Sites that had no Gatekeeper butterflies observed at any point were not reported with a Gatekeeper abundance value, so we had to estimate whether these sites had enough visits within the Gatekeeper flight period to allow us to assume zero abundance. We did this by assuming that if the site had enough data to estimate abundance for all recorded species, it would have also had enough data to estimate abundance for Gatekeeper. Therefore, if the site was not recorded as missing data for any other butterfly, then it was added to the data set as a zero. In this process, we ignored missing data from nine common butterflies with differing flight periods to the Gatekeeper, as they would contain no information on a number of visits within the Gatekeeper flight period. The nine common butterflies with differing flight period to the Gatekeeper were the Orange-tip (*Anthocharis cardamines*), Peacock (*Aglais io*), Green-veined White (*Pieris napi*), Speckled Wood (*Pararge aegeria*), Brimstone (*Gonepteryx rhamni*), Small White (*Pieris rapae*), Large White (*Pieris brassicae*), Small Tortoiseshell (*Agalis urticae*) and the Common Blue (*Polyommatus icarus*). This abundance index is reported as a count variable, with a range within our data of between 0 and ~3000 and a median of 53. The differing spatial scales and sampling methodologies of the UKBMS and aggregated BTO data were accounted for in the model by incorporating different intercepts and overdispersion parameters for the two data sets.

Both the BTO and the UKBMS data sets show a clear bias in the location of measurements, in particular with more measurements towards the south of Great Britain (Figures S1 and S2). This spatial pattern of citizen science engagement is likely related to a variety of factors including, but not limited to, population density, level of education, income, accessibility and the density of scenic or high biodiversity areas. If we wish to robustly estimate the changes in the range of the Gatekeeper butterfly over time, we would need to account for all of the factors that influence both observation intensity and Gatekeeper butterfly abundance within the model. That work is outside the scope of this case study example, so instead, we use a commonly used approach of including survey effort parameters within the model and demonstrate the potential pitfalls of this approach. For the BTO data, we include list length (i.e. the total number of species recorded within the 10 km grid square that year) as a predictor of abundance, while within the UKBMS data site, abundances were modelled as a function of transect length.

The mesh within the SPDE model was set up to cover the entirety of Great Britain with a maximum edge length 50 km and to have a surrounding buffer area with maximum edge length 300 km (Figure S3). This mesh parameterisation was chosen as it balanced creating a mesh of regular density across GB with computational feasibility. The spatial field is updated every year with an autoregressive process of order one, no finer temporal resolution was considered due to the UKBMS abundance indices being given as annual values. PC priors for the spatial covariance were specified such that only 1% of probability mass for the scale parameter was below 100 km, and 1% probability mass that the standard deviation was greater than 5. The ν parameter was fixed to the INLA default of 1. For the autoregressive process priors, 80% of the probability was set for the correlation between 0.8 and 1.0, and only 1% probability that the precision was greater than 0.25.

2.2.2 | Caddisfly

We modelled caddisfly abundance across the River Thames catchment from 2015 to 2019 in a spatio-temporal areal model, combining together structured surveys of the freshwater macroinvertebrate communities run by governmental agencies with a citizen science scheme. Riverfly (including caddisflies, stoneflies and mayflies) abundance is an important indicator of river quality, and, as such, is monitored by the Environment Agency (EA) in the UK. There is also an active citizen science community that regularly monitors sites for riverfly abundance under the Anglers' Riverfly Monitoring Initiative (ARMI) run by the Riverfly Partnership. Both schemes use standardised kick sampling to collect riverflies and other taxa from the water. To account for differences in taxonomic resolution between schemes, all caddisfly records were combined to calculate an overall caddisfly count per sample.

Representing the structure of a river is a challenging prospect within analyses of river quality, as movement can be constrained to the river and may show different patterns for moving upstream versus downstream. Our study organism, the caddisfly, we assume finds it as easy to move upstream as it does downstream for the ease of analysis. Here we aggregated stretches of river into a smaller number of connected components and supplied the edgelist of connections between this subset of river components into the BYM model. For ease of computation, we limited this to 137 areas across the entire Thames catchment, each of which can contain multiple sites. These components were created by running a fast greedy modularity optimisation algorithm upon the whole Thames river network using the igraph package (Clauset et al., 2004; Csardi & Nepusz, 2006). A neighbourhood matrix could then be constructed based on these components and which other components they were connected to (Figure S11). We found that the EA data sampled from all of the components, whereas the ARMI data were limited to a small number of components but involved far more visits on average per site—with ~7 visits a year on average in the ARMI data compared to ~2 in the EA data. Both surveys sampled relatively evenly throughout the year (with a slight dip in the summer months in the EA survey), but there was, on average, a larger gap between revisits of EA sites compared to ARMI sites. In addition to the spatio-temporal BYM model, which updates every year, we included a seasonal component to the model, which depended upon the month of the year, as we expected higher counts of caddisflies in the spring and summer, and we had the precise dates of survey for both data sources. Within the integrated model, the spatio-temporal and seasonal components were assumed to be shared between the EA and ARMI data. The intercepts and overdispersion parameters for the negative binomial distribution for the two sources of data were modelled separately.

Unlike in the butterfly case study above, in this case study, we have a data set that shows minimal spatio-temporal bias (i.e. the EA data). If we assume that the factors driving sampling intensity are unrelated to the factors driving caddisfly abundance, we can therefore model the EA data without accounting for any confounding factors. The ARMI data do show some bias across the Thames catchment, with certain areas showing far more sites than others (Figure S12). However, we do not need the ARMI data to contribute to the understanding of the spatial process as we have a reliable source within the EA data, and therefore, we can include a separate ARMI-specific spatially varying component that will prevent the biased ARMI data from overly influencing the overall spatial model estimate. We assume that the spatial bias is constant over time and therefore do not add an ARMI-specific temporal effect. The additional spatial field is modelled as an IID model instead of a spatially explicit model because few contiguous river sections are included in the ARMI data. Survey-specific intercepts and variance parameters are included to account for any systematic differences in survey methodologies (e.g. on average surveying at different times of day).

For the three precision parameters, we used the PC prior with $\alpha = 0.01$ and u set to 0.75 for the seasonal component, 0.5 for the marginal variation of the BYM model and 0.25 for the temporal component. This represents a decreasing level of variation expected to be explained by the three components. The mixing parameter (ϕ) within the BYM model was set to $u, \alpha = 0.5$, that is, giving equal probability to ϕ being higher or lower than 0.5. The lag one correlation parameter within the AR1 model (ρ) was given the PC prior with base model assuming no change, and with $u, \alpha = 0.9$, that is, giving 90% of the probability that the temporal autocorrelation is >0.9 as we assume minimal change per year across the time modelled at the broad spatial scales of interest. The PC prior to the precision of the ARMI-specific IID model had $u = 0.5$.

3 | RESULTS

3.1 | SPDE—Gatekeeper butterfly integration

The BTO-specific, UKBMS-specific and integrated joint models all showed similar broad-scale spatial patterns of predicted Gatekeeper abundance, with potentially some evidence of sharper transitions from area to area within the UKBMS model, which is based on more clustered data (Figure 1, Figures S3–S5). As the overall scales of the two data sets are different, due to UKBMS being an abundance index, while BTO being a count of presence-only observations, the greater range within the UKBMS predictions was expected, with UKBMS having both a wider range of mean predictions across space and time and also a higher interquartile range. The interquartile range is calculated based on the fitted spatial field for the given time point and presented as a metric of model certainty due to its robustness to the distribution of the variable measured. Despite the data sets reflecting slightly different aspects of the butterfly population, the joint model did prove able to fit a spatio-temporal field that was roughly intermediate between the two data sources and with a smaller interquartile range than the other two data sets. The interquartile range and standard deviation of the spatio-temporal field were higher within the UKBMS model than the BTO and joint model, with limited variation across the years (Figure 1, Figure S6). The two data set-specific models show differing spatial covariance parameters, with the UKBMS model having a range of 239 km (±34), and the BTO model having a much higher range of 460 km (±120). The joint model shows a range closer to the UKBMS model range of 281 km (±35). The marginal standard deviation for the Matérn covariance function was larger for UKBMS, at 6.025 (±0.868) compared to the BTO-only model having 2.923 (±0.706) and the joint model having 3.342 (±0.368). The level of temporal autocorrelation was consistently high in all models, with the UKBMS model being 0.999 (±0.007), the BTO-only model having 0.999 (±0.004) and the joint model having 0.998 (±0.003). This corresponds to very limited change per year overall across the broad spatial field used within the model, as the total spatial field incorporates large areas of no change (e.g. the buffer zone over the waters around Great Britain) and even those areas within England, which show some change mostly show less than one unit of change (log-scale) across the 10-year period (Figure 2).

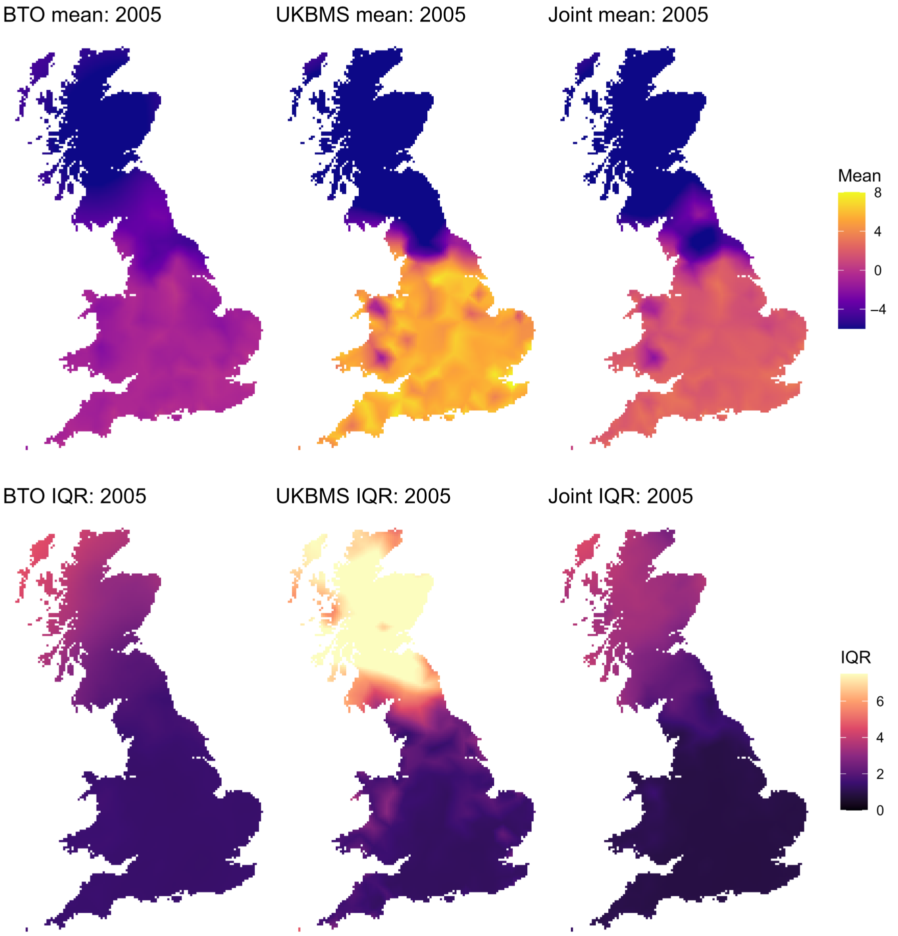

FIGURE 1 Predicted mean spatial field (top) and interquartile range (IQR, bottom) for relative Gatekeeper abundance for all models in 2005. Mean and IQR are calculated on the log scale, and with mean values lower than −6 shown as −6 and IQR values greater than 7.5 shown as 7.5.

The survey effort parameters were associated with increased Gatekeeper numbers. However, the effect of transect length upon abundance within UKBMS was weaker, and the 95% quantile crossed zero in both the UKBMS-only model (0.745 ± 0.999) and the joint model (0.83 ± 0.999). The estimate for the effect of list length upon BTO gatekeeper abundance was consistently positive (BTO-only model: 2.188 ± 0.054, joint model: 2.306 ± 0.055). The UKBMS intercept was predicted to be higher than the BTO intercept within the joint model (2.899 ± 0.643 compared to −4.03 ± 0.644). The UKBMS data had a lower 1/overdispersion than the BTO data, 0.943 ± 0.024 compared to 13.605 ± 1.889.

Despite the similarly high levels of temporal autocorrelation predicted by the model, it can be seen that the two different data sets show very different patterns of change across GB. The UKBMS model shows a large increase in Northern England, whereas BTO shows a much smaller increase across all regions other than the South West and Southern Wales, with the joint model showing an intermediate pattern (Figure 2). The pattern of change for each model varies year on year, reflecting the flexibility of the model to fit the changes present in the data (Figures S7–S9). The increase in Northern England in UKBMS is largely driven by a cluster of sites in southern Cumbria that show increased abundance over the whole 10-year period (Figure S1). It should be noted that these changes are all plotted on the log scale, and the increase in North-Western England and Southern Scotland in UKBMS is both highly uncertain and represents an ecologically unnoticeable increase in predicted abundance due to both the start and endpoints corresponding to less than 0.01 on the response scale (Figure 1). Plotting these changes on the log scale allows us to see the way in which the joint model tries to fit the two highly disparate data sources by taking an intermediate route between the two.

3.2 | Areal data—Caddisfly abundance

The spatial spread of the ARMI caddisfly data was too limited to fit a citizen science-only model (Figure S12), so here we compare only the EA-only model and the joint model. The spatial pattern of caddisfly abundance was predicted to be similar between the two models (Figure S13). The difference between the models is seen in how much change from year to year occurs, with the joint model showing higher levels of change, although the direction, strength and duration of the changes vary by river section (Figure 3). This is reflected by the lower temporal autocorrelation parameter within the joint model compared to the EA-only model, with it being estimated as 0.862 ± 0.023 within the joint model compared to 0.925 ± 0.021 in the EA-only model. The proportion of the variance explained by the spatial effect within the BYM model is low in both models, indicating little linkage between river sections at this scale, and this was somewhat lower in

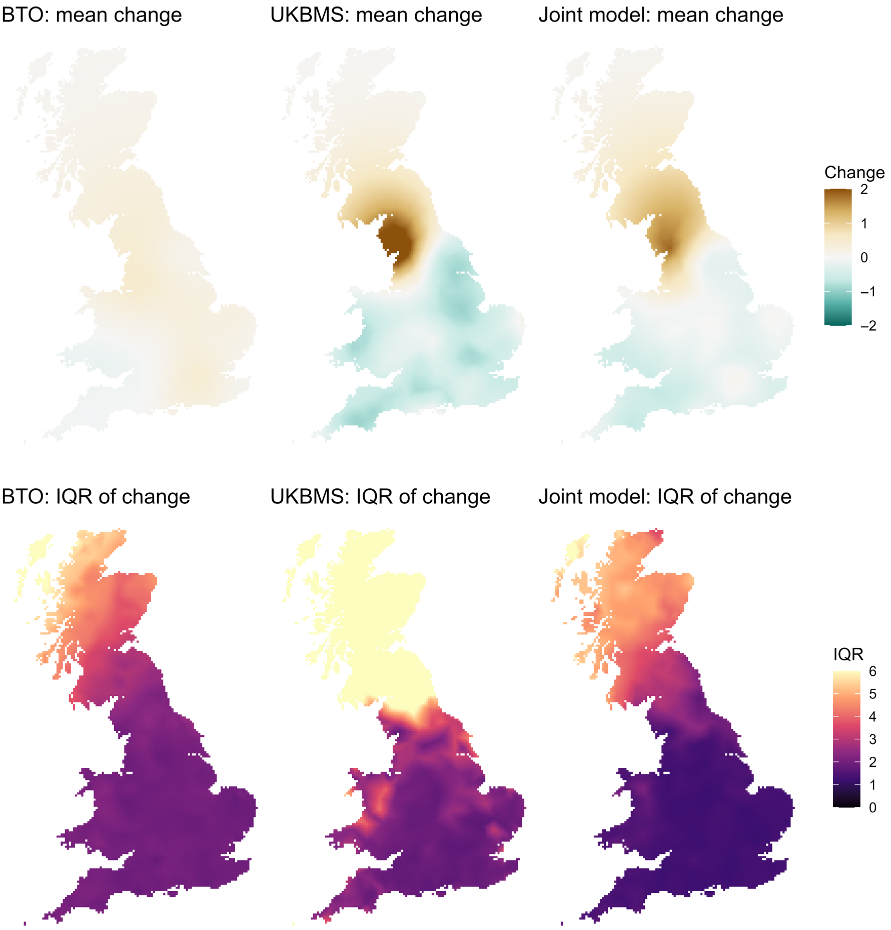

FIGURE 2 Total change from 2005 to 2014 on the log scale according to the model for BTO (left), UKBMS (centre) and the integrated model (right), with both the mean estimated change (top) and the interquartile range of the estimated change (bottom). For ease of interpretation, absolute values of the mean change greater than 2 are shown as 2, and IQR values greater than 6 are shown as 6.

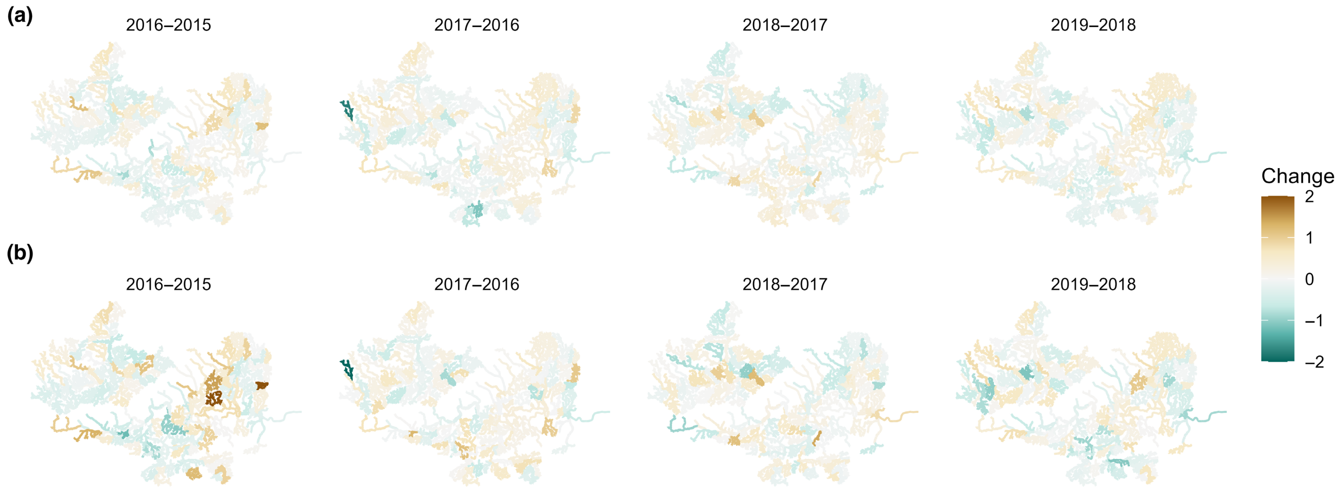

FIGURE 3 Change from year to year in the EA-only model (a) and the joint model (b). EA, Environment Agency.

the joint model (0.068 ± 0.045) compared to the EA-only model (0.147 ± 0.072). The standard deviation of the spatially structured effect was also higher within the joint model than the EA-only model, while the deviation of the area-specific effect was similar across models (Figure S14). This similarity between models is likely due to the ARMI-specific spatial field containing any differences in areas within the ARMI data compared to the EA data (Figure S15). The EA survey generally had higher caddisfly counts than the ARMI survey (EA intercept in joint model: 4.108 ± 0.108, ARMI intercept in joint model: 2.628 ± 0.166). Note that these results are given on the log scale and correspond to mean counts of ~61 and ~14, respectively. However, both surveys showed similar levels of overdispersion (1/overdispersion, EA: 0.64 ± 0.015, ARMI: 0.614 ± 0.012). The precision for the seasonal effect was lower in the joint model (3.38 ± 2.41) than in the EA-only model (10.7 ± 6.48), and examining the predicted effect of each season in the two models shows that the joint model gives a much more biologically realistic result (Figure 4).

4 | DISCUSSION

Our results show the potential benefits and difficulties when jointly modelling two (or more) data sets in a spatio-temporal context. We demonstrate that INLA provides a framework to enable spatio-temporal integration of both point and areal data, which is far faster and covers more spatio-temporal data types than the MCMC-based method for spatio-temporal data integration already present in the literature (Fidino et al., 2022). This approach should therefore be useful for applied researchers understanding change in species distributions. The caddisfly case study shows how when the citizen science data supplements an aspect of a structured monitoring survey, then the joint model can better evaluate the aspect of the data provided in more detail by the citizen science data, in that case, the change over time and seasonal trends in caddisfly abundance. This shows the potential gains to be had from integrating together multiple data sets to evaluate change in species occupancy and abundance over time, particularly as modelling change in space over time is more data intensive than modelling a spatial pattern. However, the Gatekeeper butterfly example shows that moving to a spatio-temporal context can reveal or exacerbate differences between data sources. The Gatekeeper butterfly spatial pattern was relatively consistent across the two surveys, however, the trends over time were very different, with strong evidence of range expansion within the Northwest of England only within the UKBMS data. The joint model still proved able to use information from both data sets to create a map of population change that showed an intermediate pattern between the two data sets, however, relating this to the true population change would require further investigation of how the data collection methodologies interact with the ecological dynamics of the Gatekeeper butterfly. Integration of different data sources within this spatio-temporal context requires careful consideration of the bias and representativeness of the data and the model for the question of interest.

Different data sources contain within them different biases and can represent different aspects of a species behaviour due to differing sampling methodologies. Previous work has shown how bias can influence the results of joint models, and approaches for consistent evaluation of bias across studies have been proposed (Boyd et al., 2022; Simmonds et al., 2020). Different data sources could, for example, sample differing regions in space and/or time, which would influence their ability to evaluate change over time. As a function of this, data sources could be sampling different environmental spaces, with implications for the inferences drawn on the niche space occupied by the species within species modelling based on environmental covariates. Particularly problematic biases could relate to clustering of observations or when the bias and the pattern of ecological interest confound each other. In some cases, biases can be accounted for within the model, particularly when we have one source of relatively unbiased data such as within the caddisfly case study. However, the approach taken there assumed that the bias in the citizen science data would be constant over time and in many cases, further information on the drivers of bias over space and time would also need to be included in the model. Careful accounting for biases based on knowledge of both the ecological population of interest and the data collection methodologies is required for effective model-based data integration for SDM (Johnston et al., 2023; Simmonds et al., 2020).

Model-based data integration can be used to combine data sources that have few records in key regions to address ecological questions of interest. This was apparent within the Gatekeeper butterfly case study, where the northern range edge, which we might expect to have changed over time, was also where there was less data available and therefore the model estimates were more uncertain. This made evaluating range expansion challenging, and bias correction alone would not be able to resolve the differences between the UKBMS and BTO results without the introduction of more data on the Gatekeeper butterfly in the data-sparse northern range edge. The differing results could be due to the different methodologies of the surveys resulting in different aspects of the butterfly population being captured, that is, the UKBMS data were a measure of abundance across a transect, whereas the BTO data were constructed through aggregating occurrence data. However,

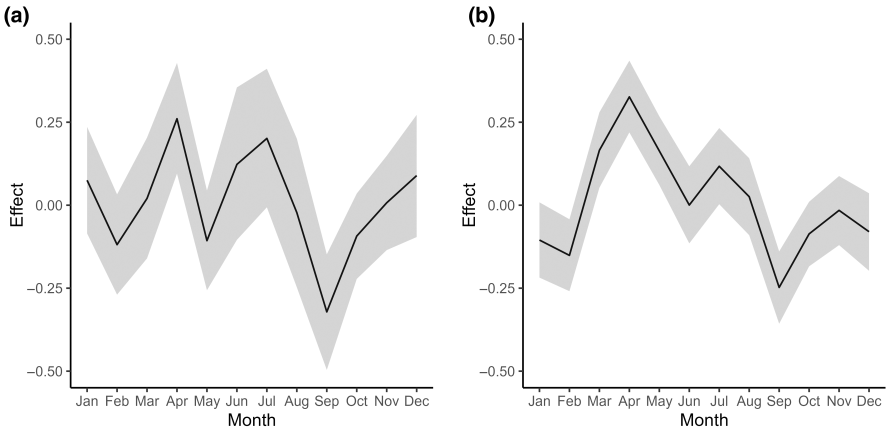

FIGURE 4 Seasonal effect upon caddisfly abundance in the EA-only model (a) and the joint model (b). EA, Environment Agency.

both transect abundance and the number of opportunistic records should be related to the true Gatekeeper abundance. Alternatively, the differences could be due to differing goals of data collectors, with opportunistic observations being taken by BTO volunteers of unusual butterflies, so capturing the change in range earlier than we might expect it to be seen in the UKBMS data (Johnston et al., 2023). The broad scale of the spatial pattern allowed within the model, chosen for computational reasons, also made it difficult to separate out the fine-scale clusters of Gatekeeper butterfly populations within the UKBMS data from more broad-scale inference about the change in range edge.

Ecological processes such as those represented by a species distribution in geographic or environmental space are scale dependent (Levin, 1992; Spake et al., 2021). In general, the scale chosen within a spatial, or spatio-temporal, model is determined based on the scale of the data available or to ensure computational feasibility. The computational cost of spatio-temporal models is much higher than spatial models, which can result in a coarsening of the scale of the model by necessity. Also, depending on the model structure and implementation, certain parts of the model could scale with a power law with the number of areas (e.g. the PC prior to the BYM model or fitting a Gaussian Process not using the SPDE approach), indicating a limit to how many areas could be incorporated into the model and it remain computationally feasible. The spatial representation of the region could also represent a simplification of the ecologically relevant distance, which occurs in the application of Euclidean space models to river data, which is why we have modelled caddisfly abundance using an areal approach rather than the SPDE approach. There are alternative approaches to modelling rivers, largely focused on modelling the distance along the river rather than Euclidean distance, but these are also currently computationally unfeasible for large numbers of sites and dependent on precise locations being used, which does not generalise well to the joint modelling process (Santos-Fernandez et al., 2022). While computational achievements are always occurring, leading to the ability to fit models to more data than ever before, these constraints on the model fitting process can lead by necessity to mismatch between the scale of the model and the scale of the data and the question that is of interest.

Within this work, we have demonstrated a methodology for model-based data integration for spatio-temporally explicit SDMs. The principles of our approach can be generalised to a variety of spatio-temporal model configurations within INLA, enabling a much wider selection of statistical analyses than just the two examples demonstrated within our two case studies. Our work has identified some potential benefits and challenges when applying this method to real ecological data and identified key considerations when applying this to other contexts. These considerations include the comparability of data set collection methodology, the comparability of the patterns shown by data sets (and drivers of any differences), and the relative sizes of the differing data sets. Some of the considerations can be addressed through building-specific model structures that account for them, such as including data set-specific spatial effects to account for spatial bias or weighting likelihoods to account for different data source sizes (Fletcher et al., 2019). However, other issues cannot be addressed purely by adjusting the model and as such need to be addressed through other methods, such as collecting data from under-sampled regions. Overall, data integration for spatio-temporal modelling of SDMs offers a promising avenue for future research if careful consideration of the biases and limitations of the data sources is given throughout the modelling process.

AUTHOR CONTRIBUTIONS

Peter A. Henrys and Fiona M. Seaton conceived the ideas and designed methodology with input from Susan G. Jarvis; Fiona M. Seaton analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This work was funded by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability and builds upon work supported by a research partnership agreement between Defra and the UK Centre for Ecology & Hydrology (UKCEH) number 32156. We would like to thank David Roy and the Biological Records Centre for their assistance in obtaining appropriate data sources for the case studies.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14356.

DATA AVAILABILITY STATEMENT

The UKBMS data are publicly available from the NERC Environmental Information Data Centre© copyright and database right Butterfly Conservation, the Centre for Ecology & Hydrology, British Trust for Ornithology and the Joint Nature Conservation Committee (Botham, Brereton, Harrower, et al., 2020; https://doi.org/10.5285/180a1c76-bceb-4264-872b-deddfe67b3de; Botham, Brereton, Harris, et al., 2020; https://doi.org/10.5285/8a41e1c8-3018-44f1-8d0a-c1b1ad957fc9). The BTO data were provided by the British Trust for Ornithology, accessed by NBN Atlas website (CC-BY-NC); for more information, see https://registry.nbnatlas.org/public/show/dr529 (British Trust for Ornithology, 2023). The EA caddisfly data are available online from the water quality data archive © Crown Copyright 2022 (https://environment.data.gov.uk/water-quality/view/), and the ARMI caddisfly data are available online from the Riverfly Partnership (https://riverflies.org). Code for fitting these models and producing all figures are available on GitHub and archived on Zenodo https://doi.org/10.5281/zenodo.11102958 (Seaton, 2024).

ORCID

Fiona M. Seaton https://orcid.org/0000-0002-2022-7451

Susan G. Jarvis https://orcid.org/0000-0001-5382-5135

Peter A. Henrys https://orcid.org/0000-0003-4758-1482

REFERENCES

Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., & Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(6), 1–24. https://doi.org/10.1002/wics.1443

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20. https://doi.org/10.1007/BF00116466

Binley, A. D., & Bennett, J. R. (2023). The data double standard. *Methods in Ecology and Evolution*, *2023*(March), 1389–1397. https://doi.org/10.1111/2041-210X.14110

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, *4*, 33–49. https://doi.org/10.1016/j.sste.2012.12.001

Botham, M., Brereton, T., Harris, S., Harrower, C., Middlebrook, I., Randle, Z., & Roy, D. B. (2020). *United Kingdom butterfly monitoring scheme: Site location data 2019*. NERC Environmental Information Data Centre. https://doi.org/10.5285/8a41e1c8-3018-44f1-8d0a-c1b1ad957fc9

Botham, M., Brereton, T., Harrower, C., Middlebrook, I., & Roy, D. B. (2020). *United Kingdom butterfly monitoring scheme: Site indices 2019*. NERC Environmental Information Data Centre. https://doi.org/10.5285/180a1c76-bceb-4264-872b-deddfe67b3de

Boyd, R. J., Powney, G. D., Burns, F., Danet, A., Duchenne, F., Grainger, M. J., Jarvis, S. G., Martin, G., Nilsen, E. B., Porcher, E., Stewart, G. B., Wilson, O. J., & Pescott, O. L. (2022). ROBITT: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods in Ecology and Evolution*, *13*(7), 1497–1507. https://doi.org/10.1111/2041-210X.13857

British Trust for Ornithology. (2023). *Non-avian taxa (BTO+ partners). Occurrence dataset. Accessed through NBN Atlas*. National Biodiversity Network. https://doi.org/10.15468/2m9nxa

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, *97*(2), 109–131. https://doi.org/10.1007/s10182-012-0196-3

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*, 066111. https://doi.org/10.1103/PhysRevE.70.066111

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. https://igraph.org

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

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fidino, M., Lehrer, E. W., Kay, C. A. M., Yarmey, N. T., Murray, M. H., Fake, K., Adams, H. C., & Magle, S. B. (2022). Integrated species distribution models reveal spatiotemporal patterns of human–wildlife conflict. *Ecological Applications*, *32*(7), 1–12. https://doi.org/10.1002/eap.2647

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*(6), e02710. https://doi.org/10.1002/ecy.2710

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Franklin, J. (2010). Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, *16*(3), 321–330. https://doi.org/10.1111/j.1472-4642.2010.00641.x

Fuglstad, G.-A., Simpson, D., Lindgren, F., & Rue, H. (2019). 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

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., Mccarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*(3), 276–292. https://doi.org/10.1111/geb.12268

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

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/10.1111/2041-210X.13834

Levin, S. A. (1992). The problem of pattern and scale in ecology: The Robert H. MacArthur Award Lecture. *Ecology*, *73*(6), 1943–1967. https://doi.org/10.2307/1941447

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 B*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Miller, D. A. W., 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/10.1111/2041-210X.13110

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850. https://doi.org/10.1002/ecy.1710

Riebler, A., Sørbye, S. H., Simpson, D., Rue, H., Lawson, A. B., Lee, D., & MacNab, Y. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, *25*(4), 1145–1165. https://doi.org/10.1177/0962280216660421

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, *71*, 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

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*, 395–421. https://doi.org/10.1146/annurev-statistics-060116-054045

Santos-Fernandez, E., Ver Hoef, J. M., Peterson, E. E., McGree, J., Isaak, D. J., & Mengersen, K. (2022). Bayesian spatio-temporal models for stream networks. *Computational Statistics & Data Analysis*, *170*, 107446. https://doi.org/10.1016/j.csda.2022.107446

Seaton, F. (2024). *NERC-CEH/SpTempDataIntPaper: v1.0.0* (Version v1) [Computer software]. *Zenodo* https://doi.org/10.5281/zenodo.11102958

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*, 1413–1422. https://doi.org/10.1111/ecog.05146

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

Sørbye, S. H., & Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, *38*(6), 923–935. https://doi.org/10.1111/jtsa.12242

Spake, R., Mori, A. S., Beckmann, M., Martin, P. A., Christie, A. P., Duguid, M. C., & Doncaster, C. P. (2021). Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecology Letters*, *24*(2), 374–390. https://doi.org/10.1111/ele.13641

Ward, E. J., Jannot, J. E., Lee, Y. W., Ono, K., Shelton, A. O., & Thorson, J. T. (2015). Using spatiotemporal species distribution models to identify temporally evolving hotspots of species co-occurrence. *Ecological Applications*, *25*(8), 2198–2209. https://doi.org/10.1890/15-0051.1

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. UKBMS locations and Gatekeeper abundance by year.

Figure S2. BTO locations and counts of Gatekeeper observations within the 10km square by year.

Figure S3. The spatial mesh used within the SPDE model.

Figure S4. Spatio-temporal field for BTO only model.

Figure S5. Spatio-temporal field for UKBMS only model.

Figure S6. Spatio-temporal field for joint model.

Figure S7. Standard deviation of spatial field for all models in 2005 and 2010.

Figure S8. Change in spatial field from year to year in BTO only model.

Figure S9. Change in spatial field from year to year in UKBMS only model.

Figure S10. Change in spatial field from year to year in joint model.

Figure S11. Thames catchment coloured by cluster components used in modelling.

Figure S12. Total number of sites per river section per survey.

Figure S13. Spatio-temporal effect for the EA only model and the joint model.

Figure S14. Standard deviation of the spatio-temporal effect for the EA only model and the joint model.

Figure S15. ARMI specific spatial field from joint model.

How to cite this article: Seaton, F. M., Jarvis, S. G., & Henrys, P. A. (2024). Spatio-temporal data integration for species distribution modelling in R-INLA. *Methods in Ecology and Evolution*, *00*, 1–12. https://doi.org/10.1111/2041-210X.14356