

Probabilistic species distributions from nine large-scale gridded atlases over five decades

Florencia Grattarola^{1*}, Gabriel Ortega-Solís¹, Carmen Diana Soria¹, Melanie Tietje¹, Friederike Johanna Rosa Wölke^{1,2,3}, Richard Fox^{4,5}, Lisbeth Hordley⁴, Sergi Herrando^{6,7,8}, Petr Keil¹

¹ Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Czech Republic

² Department of Ecology, Environment and Plant Sciences, Stockholm University, Svante Arrhenius väg 20 A 106 91 Stockholm, Sweden

³ Bolin Centre for Climate Research, Stockholm University, Svante Arrhenius väg 8, Stockholm, Sweden

⁴ Butterfly Conservation, Manor Yard, East Lulworth, Wareham, Dorset, BH20 5QP, UK.

⁵ Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, Cornwall, TR10 9FE, UK.

⁶ European Bird Census Council (EBCC), Prague, Czech Republic

⁷ CREAF, Cerdanyola del Vallès, Barcelona, Spain

⁸ ICO, Natural Sciences Museum of Barcelona, Barcelona, Spain

*Corresponding author: flograttarola@gmail.com

Abstract

Information on the long-term dynamics of species distributions is essential for assessing global biodiversity change and its causes, and for informed conservation decisions. A valuable source of such historical information are atlases. They provide spatial standardisation, extensive geographic coverage, temporal replication, documented sampling protocols, and span multiple species. Here we present model-based occupancy probabilities for 1,055 bird and 58 butterfly species, derived from nine gridded atlas datasets spanning several editions over the past 50 years. These are: Japan, New Zealand, New York state (US), the states of Alberta, Quebec, and The Maritimes (Canada), Czech Republic, United Kingdom, and Europe. Despite standardisation and coordination during data collection, the raw atlas data still contain some variation in sampling effort. We accounted for this using a two-step process: (1) we estimated a proxy for sampling effort in each grid cell within each atlas period using the Frescalo Algorithm, and (2) we ran a Bayesian spatial occupancy model for each species and atlas period, using the estimated sampling effort as the single predictor of detection in the observation process of the model. As the main product, we provide occupancy probabilities for species across grid cells and atlas periods. We also report uncertainty around each probability estimate and flag ~21% (rare) species per atlas for which the model did not fit well. The probability distributions are fit for purpose and served in a user-friendly database with standardised geographic projection, metadata, and taxonomy.

Background & Summary

The distribution of species across time and space is one of the core pieces of information in biodiversity science¹. Access to these data in an open digital form is more important than ever. Thus, it has been identified as one of the 2030 targets of the Kunming- Montreal Global Biodiversity Framework²; Target 21: to improve accessibility to relevant biodiversity data, information and knowledge. A particular type of these data are atlases³⁻⁵, which record species presences across contiguous grids, often temporally replicated over time in what are called atlas editions^{6,7}. Atlases are typically compiled at national or state levels and are coordinated by national government agencies, research institutions or not-for-profit organisations. To create them, the survey area is divided into grid cells (e.g., 10 x 10 km) and (mostly volunteer) surveyors visit these cells and record species following a set of recording guidelines. For example, participants might be asked to “complete two seasonal visits on three randomly selected squares and collect occurrence data in 10-minute intervals for a total of one hour”⁸. In other cases, recording effort is left to the discretion of individual surveyors, but standardisation is maintained through the types of data collected and guidance encouraging complete recording of all species present. Atlas projects typically include rigorous data verification steps to ensure the accuracy of the generated data. The outcome is a list of expert-validated species presences per grid cell, collected over a period of time (e.g., four years) and, in many cases, repeated periodically using the same basic protocols.

Atlases are, however, not exempt from variation in recording effort^{9,10}. For example, a species can be missed in some grid cells because of relatively low recording effort, variation in recording strategies, or in recorders’ expertise. In some cases, there is also a lack of metadata on survey methods or the number of visits, making it difficult to account for the variation in effort in analyses. As a result, both spatial patterns^{11,12} and temporal trends¹³ derived from the raw atlas data can be inaccurate. To address this issue, several strategies have been proposed, such as thinning the data¹⁴, or including sampling effort as a covariate in species distribution models (SDM)^{15,16}. The latter has been implemented in a category of SDMs called *hierarchical occupancy models*^{17,18}, which model the imperfect observation (detection) process jointly with the true distribution of the species. These models then predict the *probability of species occurrence or occupancy* (presence).

The estimated probabilistic distribution for each species in each region and time period can then be used in both basic and applied biodiversity research. For one, they represent an Essential Biodiversity Variable (EBV)^{1,19}, i.e. one of the core types of ecological information which directly underpins a range of biodiversity indicators. The probability distributions can be used, for example, to monitor temporal trends in species occupancy^{20,21}, or to test ecological hypotheses about the drivers of these changes²². They can likewise serve as empirical benchmarks for mechanistic models of biodiversity²³, or as the basis for informed large-scale biodiversity assessments²⁴. Finally, the probabilities can also be combined with functional traits or phylogenetic information to study spatial patterns and temporal trends relevant to ecosystem functioning and to test evolutionary hypotheses^{25,26}.

Here we present model-based occupancy probabilities for 1,055 bird and 58 butterfly species derived from 9 gridded atlas datasets – from Japan, New Zealand, New York state (US), the states of Alberta, Quebec, and The Maritimes (Canada), Czech Republic, United

Kingdom, and Europe, that span several editions over the past 50 years (Table 1, Figure 1). We derived these probabilities using two recent major advances, specifically the realisation that sampling effort can be approximated by local frequency of common (benchmark) species²⁷, and the emergence of fast and user-friendly implementation of the Bayesian occupancy models²⁸, which can combine the sampling effort with species detections to predict the probabilistic maps. The probabilities are presented in a common standardised format with resolved taxonomy, accompanied by full grid polygon files. This substantially reduces the cost of data management and cleaning, particularly when combining maps from multiple atlases.

The resulting dataset is fit-for-purpose and ready for immediate use²⁹: it is open access (CC BY), available for direct download in machine-readable formats, temporally replicated, and spatially comprehensive, with explicit spatial resolution, standardised taxonomy, and detailed metadata. Together, these properties make the dataset a uniquely interoperable resource for large-scale biodiversity research.

Methods

Atlas data

We obtained raw data from eight breeding bird atlases located in temperate regions (Japan, New Zealand, New York, Alberta, Quebec, The Maritimes, Czech Republic, and Europe), and one butterfly atlas (United Kingdom) (Table 1). We considered species detections as presences and assumed non-detections as absences. Some of the (original) datasets are openly available; others should be requested to the data owners. We provide this information below and in Table 1.

Data characteristics

All atlas data consist of surveys conducted across standard grid cells that cover most of the country or state. The surveys were carried out by volunteers and professionals, following a specific protocol for each atlas (see below). Each sampling period within an atlas spans multiple years with repeated visits to most cells (Table 1). All bird species' original names were harmonised according to HBW/BirdLife Taxonomic Checklist 9 (HBW & BirdLife International, 2024), while for butterfly species we retained the taxonomy used in the atlas, which follows Agassiz et al.³⁰. It should be noted, however, that taxonomy is far from constant over time and species splits can make comparisons between atlas periods extremely difficult. Although this type of taxonomic consideration should be carefully addressed when using repeated atlas work, it usually affects only a small minority of species. We recommend checking the species accounts in the original atlas works for further details.

Even though gridded atlases represent high-quality biodiversity data, there is inherent variation in field methods and the level of survey standardisation achievable, depending on the taxa studied and the geographic location of the grid cells. Here we present the main particularities of each dataset.

- Japan: It combines fieldwork records and questionnaire survey results. The fieldwork was done by 1 or 2 people walking along a pre-established survey route of ~3 km and performing 2 x 30-minute point counts. This information was complemented by questionnaires sent to researchers and by records gathered from journals and online platforms. Most of the country cells were sampled.
- New Zealand: Volunteers visited more than 96% of the grid cells in both atlas editions. People were asked to indicate whether their survey was partial or complete in terms of habitat and species coverage when submitting their recording sheets to the atlas coordinators. The return of complete sheets varied across the country.
- New York: Volunteers selected a priority block to survey and were tasked with visiting each habitat within it. Paid workers were sent to remote areas to ensure adequate survey coverage. The goal was to record at least 76 species in every sampled block with breeding evidence reported for half of them.
- Alberta, Quebec, and The Maritimes: The data were collected across a 10x10 km grid in each study area. Volunteers were asked to spend ~20 hours total surveying a given cell. However, the ~20 hours of sampling served more as guidance than as a hard target, and variable survey durations are reported as sampling effort measures for each atlas.
- Czech Republic: Volunteers were required to cover all habitats within their assigned grid cell and record evidence of breeding for each observed species. The latest atlas also includes opportunistic records with identifiable breeding evidence from online portals and standardised 1-hour walks (timed-species-lists) in randomly chosen subsquares. The atlas covers most of the grid-cells in the country, except for a few remote locations along the country's borders.
- United Kingdom (butterflies): Volunteers are encouraged to survey butterflies in all 10x10km grid cells over a 5-year period, as well as reporting opportunistic sightings. Almost all observations are reported at 1x1km grid resolution or finer. These data are combined with records from standardised butterfly monitoring (e.g. timed counts, fixed-route transect counts) into the Butterflies for the New Millennium recording scheme. The 5-year surveys run consecutively, and all records are verified by volunteer taxon experts. Only resident and regular migrant butterfly species are considered, while (re)introduced species and scarce migrant, vagrant and adventive species are excluded.
- Europe (EBBA): The data are compiled from country-level atlases, monitoring schemes, casual observations, and targeted surveys performed by the EBBA team. The data were standardised and assigned to a matching ~50x50 km grid cell. National coordinators of data-providing countries were asked to evaluate the breeding categories for each species, adding an extra layer of data quality assessment.

After data collection, atlases have a vetting committee that evaluates the reliability of the records and seeks additional expert advice to ensure the final dataset is trustworthy. These semi-structured sampling procedures across a grid, accompanied by basic survey protocol guidance and post hoc evaluation, make gridded atlases suitable for pooling and comparison despite potential differences among them.

Data filtering and cleaning

In the context of the planned follow-up analyses (not reported in this paper), we made the following decisions. The first atlas period for the Czech Republic (1973–1977) was not considered because the grid is not compatible with subsequent periods. Similarly, Japan's most recent atlas period (2016–2021) was omitted given that the volume of voluntary reports was substantially higher and targeted different species relative to earlier survey events (which is not possible to correct using Frescalo). For all atlases, we retained only grid cells surveyed across all periods. This is because occupancy estimation relies on survey effort across periods, and a cell with zero effort in any one period would introduce an artefact into the occupancy calculation. For cells partially truncated by coastlines or administrative boundaries, we excluded those in which more than three-quarters of the cell area lay outside of the atlas survey area.

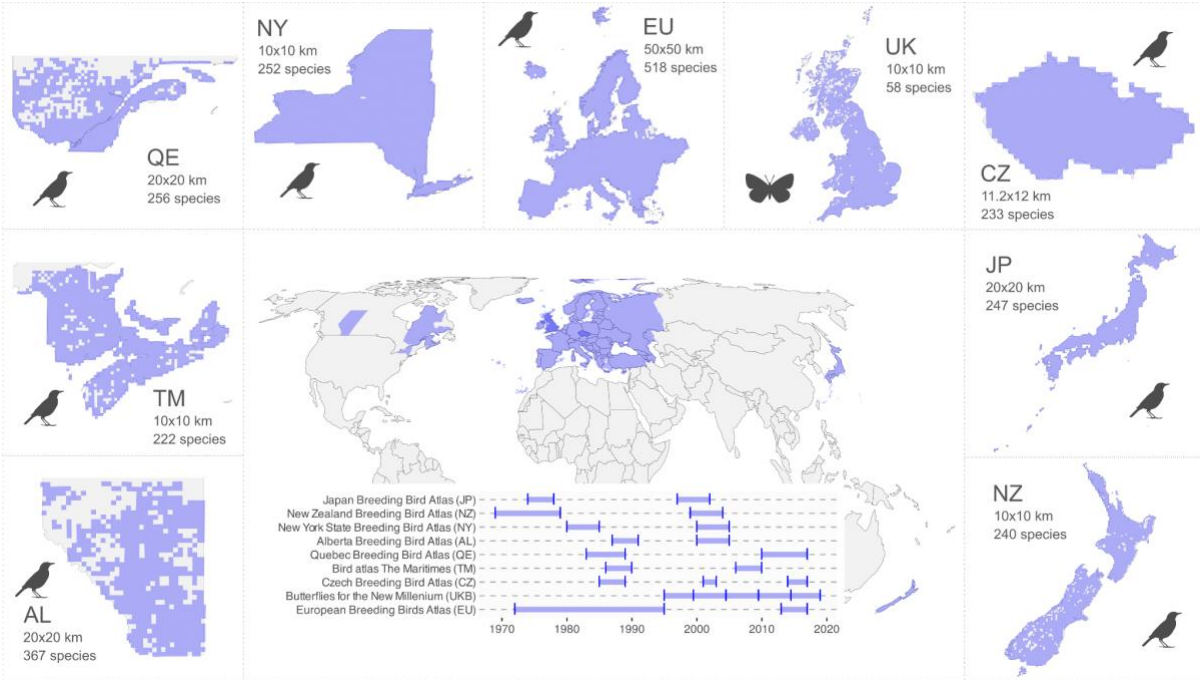


Figure 1. Temporal and spatial coverage of the atlases used for which we provide the probabilistic distributions. Each dataset spans a period between 14 and 31 years, over an extent of between 78,000 and 5,864,590 km² (Table 1). Areas in darker blue (the Czech Republic and UK) are part of two different atlases, one at the continental and one at the national level.

Table 1. Temporally replicated gridded atlases for which sampling effort and species occupancy probabilities were estimated. The Dataset ID corresponds to the number in the internal Modelling of Biodiversity (MOBI) Lab database and was used for the file names.

Dataset ID	Atlas name	Atlas code	Taxa	Number of species	Original grain (km)	Grain of analysis (km)	Number of grid cells (km ²)	Temporal periods	Temporal span (years)	Raw data accessibility	References
5	Czech Breeding Bird Atlas	CZ	Birds	233	11.2 x 12 (10' x 6')	11.2 x 12	615 (78,000)	1985-1989, 2001-2003, 2014-2017	29	Open-access (https://doi.org/10.5281/zenodo.18503904)	8,31,32
6	New York State Breeding Bird Atlas	NY	Birds	252	5 x 5	10 x 10	1,363 (126,933)	1980-1985, 2000-2005	20	Open-access (https://dec.ny.gov/nature/animals-fish-plants/birds/breeding-bird-atlas/past-data-maps)	33,34
7	Alberta Breeding Bird Atlas	AL	Birds	367	10 x 10	20 x 20	623 (245,464)	1987-1992, 2000-2005	14	Open access (https://www.natureconservation.ca/)	6,7
13	Japan Breeding Bird Atlas	JP	Birds	247	20 x 20	20 x 20	1,003 (359,020)	1974-1978, 1997-2002	24	Request to Mutsuyuki Ueta (Japan Bird Research Association)	35,36
17	New Zealand Breeding Bird Atlas	NZ	Birds	240	10 x 10	10 x 10	2,574 (243,124)	1969-1979, 1999-2004	28	Request to Susan Walker or Adrian Monks (Landcare Research)	37,38
20	Quebec Breeding Bird Atlas	QE	Birds	256	10 x 10	20 x 20	901 (338,027)	1984-1989, 2010-2014	28	Open access (https://www.natureconservation.ca/)	39
23	Bird atlas The Maritimes	TM	Birds	222	10 x 10	10 x 10	1,340 (120,314)	1986-1990, 2006-2010	20	Open access (https://www.natureconservation.ca/)	40
26	European Breeding Birds Atlas	EU	Birds	518	50 x 50	50 x 50	2,506 (5,864,590)	1972-1995, 2013-2017	31	Request at https://ebba2.info/data-request/	41
36	Butterflies for the New Millennium	UKB	Butterflies	58	1 x 1	10 x 10	2,689 (231,920)	1995-1999, 2000-2004, 2005-2009, 2010-2014, 2015-2019	20	Request to Richard Fox and Lisbeth Hordley (Butterfly Conservation)	42,43

208

209 *From raw data to probabilities*

210 We converted the raw atlas data (species detections) into occupancy probabilities using a
 211 two-step process. First, we estimated sampling effort for each grid cell within each atlas
 212 period, and then we ran a Bayesian spatial occupancy model for each species and atlas
 213 period, using the estimated sampling effort as the only predictor of detection in the
 214 observation process of the model. As the main output, we provide the estimated occupancy
 215 probability for each species across grid cells for each atlas period, along with the uncertainty
 216 around each estimate.

217

218 **Sampling effort calculation**

219 Even though some of the atlases included information on the number of visits or hours spent
 220 recording in the grid cell, these values do not necessarily reflect sampling effort, as some
 221 recorders may be able to record more species than others within the same time span. Thus,
 222 to accurately estimate sampling effort across our datasets, we ran the Frescalo algorithm ²⁷
 223 for each atlas and atlas period using the R package *sparta* ⁴⁴. This method uses the
 224 frequency of a set of "benchmark" species (i.e., species that are widespread and assumed to
 225 be consistently recorded), and compares their presence in a focal grid cell to their frequency
 226 in neighbouring cells, weighted by habitat similarity and physical distance. See Goury et al. ⁴⁵
 227 for a practical guide on the use of Frescalo.

228

229 We calculated the weights required to run Frescalo using the euclidean distances between
 230 grid cells and the mean elevation per cell (as a habitat similarity proxy), downloaded using
 231 *geodata* ⁴⁶. These weights ensure that, within the neighbourhood of selected cells, species
 232 lists from cells that are physically more distant or less similar in elevation contribute less than
 233 those from nearer or more similar cells. We used 20 as the number of neighbours to include
 234 after ranking by distance, and 10 as the number of neighbours to include after ranking by
 235 similarity. Finally, we ran the algorithm and calculated the proportion of benchmark species in
 236 each grid cell and sampling period using the function *frescS_it*, developed by Auffret & Arnell
 237 ⁴⁷. We used this proportion of benchmark species as our proxy for sampling effort.

238

239 **Occupancy models**

240 To account for the effect of varying sampling effort in the raw data, we ran a spatial
 241 occupancy model for each species recorded in each atlas-period combination using the
 242 *spPGOcc* function from *spOccupancy* ²⁸. This function fits a single-species spatial occupancy
 243 model using Nearest Neighbor Gaussian Processes (NNGP) and Pólya-Gamma data
 244 augmentation for fast and efficient computation (for more details, see Doser et al., 2022).
 245 The model separates the observation process from the true occupancy. Thus, given that we
 246 have estimated the sampling effort per grid cell from the data (through Frescalo), we can use
 247 this variable to model the observation process, while assuming the true occupancy is the
 248 result of the observation process, given the area of the grid cell.

249

250 The occupancy of a species at a grid cell in a specific time period is therefore conditioned on
 251 the presence of the species, the area of the cell, and the estimated recording effort at that
 252 location. We did not include environmental covariates in the occupancy model for three
 253 reasons. First, identifying relevant predictors would require researching the habitat

preferences of thousands of species individually. Second, using generic predictors (e.g., annual temperature, precipitation seasonality) would bias occupancy estimates toward species that respond strongly to those particular variables. Third, since we aim to test what drives occupancy change, including environmental predictors would introduce circularity into the analyses.

Model description

Let z_j be the true presence (1) or absence (0) of a species at a site j , with $j=1, \dots, J$. We assume this latent occurrence variable arises from a Bernoulli process following,

$$z_j \sim \text{Bernoulli}(\psi(s_j)), \quad (\text{eq. 1})$$

$$\text{logit}(\psi(s_j)) = x(s_j)^T \beta + \omega(s_j), \quad (\text{eq. 2})$$

where $\psi(s_j)$ is the probability of occupancy at site j with coordinates s_j , which is a function of site-specific covariates x and a vector of regression coefficients (β), T denotes transposition of column vector $x(s_j)$, and where $\omega(s_j)$ is a realisation from a zero-mean spatial Gaussian process. In our case, the only site-specific occurrence covariate is the area of the grid cell, which accounts for the fact that, although the grids are almost regular, not all grid cells are exactly the same, some are cropped or elongated.

The Gaussian process $\omega(s_j)$ introduces autocorrelation in the $\psi(s_j)$, based on the empirically well-supported assumption that most real-world species distributions are autocorrelated^{48,49}; this is also the reason why the model can separate the detection process (described below) from the occupancy $\psi(s_j)$, something that would otherwise require repeated visits. It is, however, also the reason the model has not converged for some of the extremely rare species (see the “Flagged species” section below).

The zero mean Gaussian process considered assumes that,

$$\omega(s) \sim N(0, \Sigma(s, s', \theta)) \quad (\text{eq. 3})$$

where $\Sigma(s, s', \theta)$ is a $J \times J$ covariance matrix that is a function of the distances between any pair of site coordinates s and s' , and a set of parameters (θ) that govern the spatial process according to a spatial correlation function. We used the exponential function $\theta = \{\sigma^2, \phi\}$, where σ^2 is the spatial variance parameter and ϕ is a spatial decay parameter.

Now let y_j be the observed detection (1) or nondetection (0) of a species of interest at site j . We see the detection-no detection y_j data as arising from a Bernoulli observation process:

$$y_j \sim \text{Bernoulli}(p_j z_j), \quad (\text{eq. 4})$$

$$\text{logit}(p_j) = v^T \alpha, \quad (\text{eq. 5})$$

where p_j is the probability of detecting a species at site j (given it is present at site j), which is a function of site-specific covariates v , with T again denoting transposition of v , and a vector of regression coefficients (α). In our case, the only site-specific detection covariate is the sampling effort in the grid cell.

We assign multivariate normal priors for the occurrence (β) and detection (α) regression coefficients.

Model fitting

We fitted the model using ‘exponential’ as the spatial correlation function and 5 nearest neighbours. We used 300,000 samples per chain (12,000 batches of length 25), burn-in 100,000, thinning rate 500, and 3 chains. The total posterior samples were 1,200 for each parameter and for each $\psi(s_j)$. These large numbers of iterations ensured convergence of the spatial covariance parameters. Finally, we stored the mean and the standard deviation (SD) of the occurrence probability of each species per grid cell and atlas period, among other parameters. See the *Data Records* section for a full description of the data.

The analyses were run in parallel on a server with two CPU AMD of 128 cores and 256 threads each and a memory of 550 GB. Using 150 cores, model fitting took on average 27.8 hours per atlas, depending on the number of species, grid cells and periods considered. In the case of the Czech Republic (233 species, 615 grid cells, and 3 periods), the models took around 11.4 hours, while on the other extreme, Europe (with 518 species, 2506 grid cells, and 2 periods) took 68.2 hours.

Modelling performance

We checked model convergence through *Rhat* statistics and Bayesian p-values. *Rhat* compares within-chain and between-chain variance across multiple Markov Chain Monte Carlo (MCMC) chains⁵⁰, while the Bayesian p-value evaluates the proportion of model-generated data (from posterior samples) where the fit statistic is larger than the one from the real data⁵¹. Specifically, we checked that the *Rhat* values were <1.1 for all estimated parameters (indicating that the chains have converged to the true posterior distribution), and that the Bayesian p-value was around 0.5, ranging from 0.1 to 0.9 (Doser et al. 2022). We also compared the raw occupancy of each species per grid cell *with* the estimated occupancy values and visually inspected the species' probabilistic maps for clear outliers or distortions.

The *Rhat* values for all estimated parameters were mostly <1.1 , except for the σ^2 and ϕ parameters (for which convergence is always challenging in spatial occupancy models). Bayesian p-values ranged between 0.485 and 0.649 (mean = 0.566, SD = 0.05) across atlases. Based on the raw occupancy vs estimated occupancy values per grid cell and visual checks of the probabilistic maps of the species, the overall convergence rate per species across all periods ranged from 55.4% (New Zealand) to 100% (UK) (Table 2).

Flagged species

We classified an occupancy value as inflated when the mean estimate (‘mean.psi’) was at least 10 times the raw occupancy (i.e., the raw number of occupied grid cells divided by the total number of grid cells). In each atlas, several species were flagged for having inflated occupancy values (marked as “poor performing” in Table 2). This issue occurred only for rare species (e.g., with raw occupancy of 1-2% of the territory) that had very sparse spatial distributions, leading to inflated estimated occupancies. We advise users to interpret estimates for these species with caution. Additionally, we advise users to always take into

account the occupancy (``mean.psi``) together with its uncertainty (``sd.psi``) and if possible, propagate the uncertainty in their analyses (e.g., Grattarola et al., ⁵²).

These inflated estimates are closely tied to the challenges of modelling rare species. Assessing convergence in spatial factor models can be a complicated task ⁵³. With so few observations, fitting any basic occupancy model is generally difficult, and thus a spatial occupancy model is even more challenging because there is limited information to estimate all parameters. The spatial model we use includes a site-level random effect that assumes occupancy is spatially correlated, meaning nearby sites tend to have similar occupancy probabilities (e.g., if a species occurs in one cell, it is more likely to occur in neighbouring cells than in distant ones). This is why a low number of records is less problematic when there is spatial autocorrelation in the detection pattern.

Table 2. Model performance. Total number of species modelled, number of species for which the spatial occupancy model performed poorly (at least in one atlas period), and percentage of species for which the occupancy model converged, for each atlas.

	CZ	NY	AL	JP	NZ	QE	TM	EU	UKB
# species modelled	233	252	367	256	240	256	222	518	58
# poor performing	54	45	97	46	107	46	54	80	0
% converged	76.8	82.1	73.6	82	55.4	82	75.7	84.6	100

Data Record

The data, including the occupancy probability for each species per grid cell within an atlas period and the uncertainty around each probability estimate, and the code used to generate them, are available at [anonymised repository: <https://anonymous.4open.science/r/gridded-atlas-probabilities-2094/>]. We provide a separate folder for each of the 9 atlases. See more details below.

Data folder structure

Each folder is named after the dataset ID in Table 1, using the format `occ_<datasetID>`. For example, “occ_5” corresponds to the Czech Republic dataset (datasetID = 5).

Data files

Inside each folder, you will find three types of files: the species distribution probabilities (``occupancy_probabilities``), the diagnostics of all species in each period, including the flagged species (``diagnostics``), and the atlas grid (``atlas_grid``). The files are named according to the ID of the dataset, using the format `<file type>_<datasetID>.<file extension>`. For example, `occupancy_probabilities_5.duckdb` contains the data on species occupancy probabilities for the Czech Republic.

1. ``occupancy_probabilities``: species grid-level data

This is a long table stored as a DuckDB database file (*.duckdb*), a self-contained, serverless database that organises data in a columnar layout optimised for efficient querying of large datasets without loading the entire table into memory⁵⁴. The file contains the occupancy probability data for a species during a given atlas period, and includes the mean occupancy probability and its standard deviation for each grid cell (*`mean.psi`*, *`sd.psi`*). Each element of the list is named using the scientific name of the species (*`sp.name`*), which matches the same value in the *`diagnostic`* object.

DuckDB database file with one row per grid cell and the following variables:

- *period*: Integer that refers to the period number (e.g., 1,2,3)
- *sp.name*: Standardised scientific name of the species
- *siteID*: Unique identifier of the grid cell (maps to *siteID* in *`atlas\_grid`*)
- *mean.psi*: Mean occurrence probability ($\psi \in [0, 1]$)
- *sd.psi*: Standard deviation of the occurrence probability
- *mean.w*: Mean spatial random effect
- *sd.w*: Standard deviation of the spatial random effect
- *mean.p*: Mean detection probability ($p \in [0, 1]$)
- *sd.p*: Standard deviation of the detection probability

2. *`diagnostics`*: **species-level model parameters and performance diagnostics**

This CSV file (*.csv*) contains species-level model parameters and estimates of model performance. Species were flagged (*`is.flagged`* = TRUE) if, in any atlas period, *`mean.psi`* was at least ten times greater than the raw occupancy (the proportion of occupied grid cells out of the total number of grid cells) in any atlas period.

CSV file with one row per species and the following variables:

- *sp.name*: Standardised scientific name of the species
- *alpha.det*: Mean detection regression coefficient (α)
- *beta.occ*: Mean occurrence regression coefficient (β)
- *bayesian.p*: Bayesian p-value from posterior predictive checks (i.e., the proportion of posterior samples in which the fit statistic of the model-generated data exceeds that of the observed data; values near 0.5 indicate a good fit)
- *rhat.beta.int*: Diagnostic statistic for the occurrence intercept
- *rhat.beta.area*: Diagnostic statistic for the occurrence coefficient for area (β)
- *rhat.alpha.int*: Diagnostic statistic for the detection intercept
- *rhat.alpha.eff*: Diagnostic statistic for the detection coefficient for effort (α)
- *rhat.sigma*: Diagnostic statistic for the spatial variance (σ^2)
- *rhat.phi*: Diagnostic statistic for the spatial range (ϕ)
- *is.flagged*: Logical (TRUE/FALSE), indicating whether the species has a poorly performing model.

3. *`atlas\_grid`*: **spatial grid for the atlas**

This is a geopackage file (*.gpkg*), a self-contained, standards-based format built on SQLite that stores vector geometries, attribute tables, and spatial reference information in a single portable file without requiring dedicated GIS software or a database server. The file stores the spatial polygons in WGS84 (World Geodetic System 1984) of the entire atlas grid. It

provides the geometric reference to which the `occupancy\_probabilities` estimates are spatially linked via siteID.

Geopackage file (POLYGON geometry) with one row per grid cell and the following variables:

- siteID: Unique identifier of the grid cell (maps to siteID in `occupancy`)
- datasetID: Identifier of the dataset
- footprintSRS: Spatial reference system of the grid cell footprint
- verbatimFootprintSRS: Original spatial reference system as provided in the source data
- area: Total area of the grid cell (km²)
- croppedArea: Area of the grid cell after cropping to the study region boundary (km²)
- areaUnit: Unit of area measurement
- croppedAreaPercent: Proportion of the grid cell area retained after cropping (%)
- geometry: Polygon geometry of the grid cell

Data Overview

To illustrate our analyses, we present the full data generation process and results using the New York state atlas as a case example. See the Code Availability section for more details.

The Frescalo run took 2 minutes. We checked whether the proportion of benchmark species calculated were comparable to the raw numbers of effort (on the original data) (Figure 2ab). We found high levels of correlation between raw and estimated sampling effort metrics ($\tau = 0.18$, $p < 0.001$; $\rho = 0.257$, $p < 0.001$; Figure 2c). These metrics are not expected to correlate perfectly, since the same number of visits across grid cells can still reflect very different levels of sampling effort²⁷. Nonetheless, the overall pattern aligns as expected: poorly sampled grid cells tend to show a lower proportion of benchmark species (Figure 2c).

The occupancy model took ~19 hours to run (252 species, over 1,363 grid cells, and atlas 2 periods). We flagged 45 species as poor performing (17.9% of the total). We performed visual checks of the probabilistic maps of the species, and present *Setophaga citrina* as an example (Figure 3).

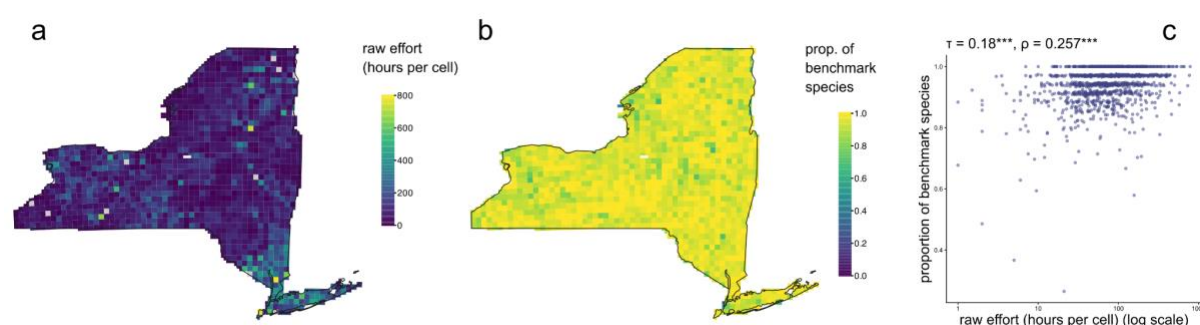


Figure 2. Association between the raw effort (number of visits as in the original data) and the proportion of local benchmark species per grid cell, using the first edition of the New York state atlas (1980-1985) as an example. (a) Distribution of the raw effort (number of hours per cell) (log scale), (b) distribution of the proportion of benchmark species (calculated using Frescalo), and (c) correlation between raw sampling effort (hours per cell) and

estimated sampling effort per grid cell (proportion of benchmark species). Kendall's τ and Spearman's ρ correlation values are shown on top of the panel. The *** symbols indicate p value < 0.001.

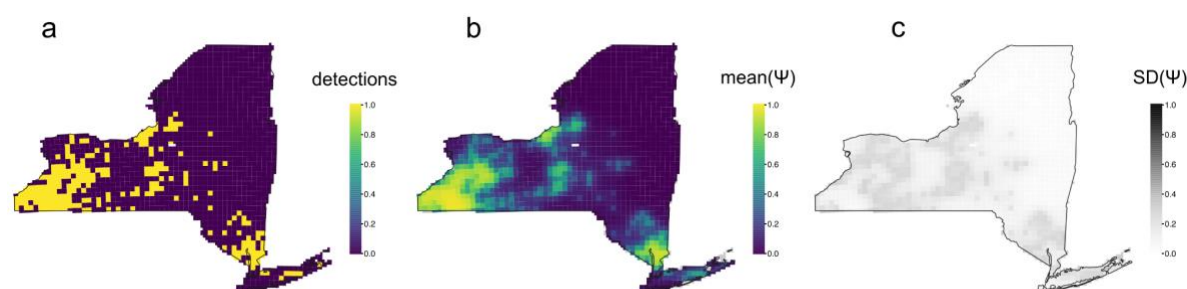


Figure 3. Example of estimated occupancy probability using *Setophaga citrina* from the first edition of the New York state (1980-1985) as an example. (a) Raw occupancy (detections) per grid cell, (b) mean occupancy probability ψ per grid cell, and (c) uncertainty (SD) of the occupancy probability per grid cell.

Usage Notes

The data are available under the CC-BY license (<https://creativecommons.org/licenses/by/4.0>); reusers of the data must acknowledge the authors but are allowed to distribute, remix, adapt, and build upon the created material. For inquiries about the raw data and their accessibility, and for any further collaborations, we encourage reusers to contact each atlas' point of contact (Table 1).

To help users access and process the data, we provide an html vignette that demonstrates how to: filter atlases, atlas periods, and species (including flagged ones), and obtain their distribution probabilities to create species maps. We recommend using R packages 'sf' ⁵⁵, 'duckdb' ⁵⁴, and 'tidyverse' ⁵⁶ to manipulate the data. Importantly, the data have been made fit-for-purpose (with standardised species name and geometries), and thus do not need to be cleaned any further.

Data Availability

The database can be accessed and downloaded at <https://anonymous.4open.science/r/gridded-atlas-probabilities-2094/>, under a CC-BY license (<https://creativecommons.org/licenses/by/4.0>).

Code Availability

The code to produce the occupancy data is available at <https://anonymous.4open.science/r/gridded-atlas-probabilities-2094/>, under a GNU General Public License license (<https://www.gnu.org/licenses/gpl-3.0.html>).

The code for each function is annotated, stating the purpose, inputs, outputs, and package dependencies needed. To demonstrate how to run the Frescalo algorithm and the spatial

occupancy model, we provide a vignette with a case example using data from the state of New York (see *Data Overview*). All atlases were processed following the same method.

References

1. Jetz, W. *et al.* Essential biodiversity variables for mapping and monitoring species populations. *Nat. Ecol. Evol.* **3**, 539–551 (2019).
2. CBD Secretariat. *Kunming-Montreal Global Biodiversity Framework*. <https://www.cbd.int/conferences/2021-2022/cop-15/documents> (2022).
3. Keller, V. *et al.* *European Breeding Bird Atlas 2: Distribution, Abundance and Change*. (European Bird Census Council & Lynx Edicions, Barcelona, 2020).
4. Pototsky, P. C. & Cresswell, W. A new global review of bird atlases and their contribution to knowledge. *Bird Study* **70**, 84–98 (2023).
5. Schäfers, G. *et al.* *Atlas Der Säugetiere Hamburgs: Artenbestand, Verbreitung, Rote Liste, Gefährdung Und Schutz*. (Amt für Naturschutz, Grünplanung und Energie, Abteilung Naturschutz, Hamburg, 2016).
6. Nature Alberta. Alberta breeding bird atlas (1987–1992). (2018).
7. Nature Alberta. Alberta breeding bird atlas (2000–2005). (2018).
8. Šťastný, K., Bejček, V., Mikuláš, I. & Telenský, T. *Atlas hnízdního rozšíření ptáků v České republice 2014-2017*. (Aventinum, Praha, 2021).
9. Hugo, S. & Altwegg, R. The second Southern African Bird Atlas Project: Causes and consequences of geographical sampling bias. *Ecol. Evol.* **7**, 6839–6849 (2017).
10. Robertson, M. P., Cumming, G. S. & Erasmus, B. F. N. Getting the most out of atlas data. *Divers. Distrib.* **16**, 363–375 (2010).
11. Boakes, E. H. *et al.* Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLOS Biol.* **8**, e1000385 (2010).
12. Hughes, A. C. *et al.* Sampling biases shape our view of the natural world. *Ecography* **n/a**, (2021).

13. Rapacciuolo, G., Young, A. & Johnson, R. Deriving indicators of biodiversity change from unstructured community-contributed data. *Oikos* **130**, 1225–1239 (2021).
14. Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38**, 541–545 (2015).
15. Chauvier, Y. *et al.* Novel methods to correct for observer and sampling bias in presence-only species distribution models. *Glob. Ecol. Biogeogr.* **30**, 2312–2325 (2021).
16. Merow, C., Allen, J. M., Aiello-Lammens, M. & Silander, J. A. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. *Glob. Ecol. Biogeogr.* **25**, 1022–1036 (2016).
17. Kéry, M. & Royle, J. A. *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 1: Prelude and Static Models.* (Elsevier, 2015).
18. *Occupancy Estimation and Modeling (Second Edition).* (Academic Press, Boston, 2018). doi:10.1016/B978-0-12-407197-1.00023-5.
19. Pereira, H. M. *et al.* Essential biodiversity variables. *Science* **339**, 277–278 (2013).
20. Herrando, S. *et al.* Providing regular and frequent maps of losses and gains of farmland birds based on European monitoring data. *Conserv. Biol.* **n/a**, e70268 (2026).
21. Klinkovská, K. *et al.* Dynamics of the Czech flora over the last 60 years: Winners, losers and causes of changes. *Biol. Conserv.* **292**, 110502 (2024).
22. Montràs-Janer, T. *et al.* Anthropogenic climate and land-use change drive short- and long-term biodiversity shifts across taxa. *Nat. Ecol. Evol.* **8**, 739–751 (2024).
23. Kalyuzhny, M., Flather, C. H., Shnerb, N. M. & Kadmon, R. A framework for quantifying deviations from dynamic equilibrium theory. *Glob. Ecol. Biogeogr.* **31**, 183–195 (2022).
24. IPBES. *Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.* 56 (2019) doi:10.5281/zenodo.3553579.

25. Faurby, S., Matthews, T. J., Triantis, K. A. & Sayol, F. Quantifying the unrecorded loss of avian phylogenetic diversity. *Ecography* **2026**, e08267 (2026).
26. Matthews, T. J. *et al.* The global loss of avian functional and phylogenetic diversity from anthropogenic extinctions. *Science* **386**, 55–60 (2024).
27. Hill, M. O. Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods Ecol. Evol.* **3**, 195–205 (2012).
28. Doser, J. W., Finley, A. O., Kéry, M. & Zipkin, E. F. spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods Ecol. Evol.* **13**, 1670–1678 (2022).
29. UNEP-WCMC. *Global Biodiversity Data Fitness Assessment: Criteria for Decision-Grade Data*. 6 <https://unbiodiversitylab.org/wp-content/uploads/2021/01/Global-Biodiversity-Data-Fitness-Assessment.pdf> (2024).
30. Third update to a checklist of the Lepidoptera of the British Isles, 2013. *Entomol. Rec. J. Var.* **132**, (2020).
31. Šťastný, K., Bejček, V. & Hudec, K. *Atlas Hn'izdn'ího Rozš'ířen'í Ptáků v České Republice 1985–1989*. (Nakladatelstv'í a vydavatelstv'í H&H, Jinočany, Czech Republic, 1997).
32. Šťastný, K., Bejček, V. & Hudec, K. *Atlas Hn'izdn'ího Rozš'ířen'í Ptáků v České Republice: 2001–2003*. (Aventinum, Praha, Czech Republic, 2006).
33. New York State Department of Environmental Conservation. New York state breeding bird atlas: First atlas (1980–1985) data. (2007).
34. New York State Department of Environmental Conservation. New York state breeding bird atlas: Second atlas (2000–2005) data. (2007).
35. Biodiversity Center of Japan. *The National Survey on the Natural Environment: Report of the Distributional Survey of Japanese Animals (Birds)*. (Biodiversity Center of Japan, Fujiyoshida, Japan, 2004).
36. Environment Agency of Japan. *Breeding Distribution of Japanese Birds*. (Environment Agency of Japan, Tokyo, Japan, 1981).

37. Bull, P. C., Gaze, P. D. & Robertson, C. J. R. *The Atlas of Bird Distribution in New Zealand*. (The Ornithological Society of New Zealand, Wellington, New Zealand, 1985).
38. Robertson, C. J. R., Hyvonen, P., Fraser, M. J. & Pickard, C. R. *Atlas of Bird Distribution in New Zealand 1999–2004*. (The Ornithological Society of New Zealand, Wellington, New Zealand, 2007).
39. Québec Breeding Bird Atlas. Québec breeding bird atlas. (2024).
40. Birds Canada, Environment Canada - Canadian Wildlife Service, New Brunswick Department of Natural Resources, Nova Scotia Department of Natural Resources, & Prince Edward Island Department of Agriculture and Forestry. Maritimes breeding bird atlas database. (2012).
41. Keller, V. *et al.* *European Breeding Bird Atlas 2: Distribution, Abundance and Change*. (European Bird Census Council & Lynx Edicions, Barcelona, Spain, 2020).
42. Asher, J. *et al.* *The Millennium Atlas of Butterflies in Britain and Ireland*. (Oxford University Press, Oxford, UK, 2001).
43. Fox, R. *Atlas of UK Butterflies 2015–2019*. (Butterfly Conservation, United Kingdom, 2023).
44. Isaac, N. *et al.* sparta: Trend Analysis for Unstructured Data. (2024).
45. Goury, R. *et al.* A practical guide to species trend detection with unstructured data using local frequency scaling (Frescalo). *Ecography* **n/a**, e08270 (2026).
46. Hijmans, R. J., Barbosa, M., Ghosh, A. & Mandel, A. geodata: Download Geographic Data. (2024).
47. Auffret, A. G. & Arnell, M. AGAuffret/FrescaloFun. (2026).
48. Dormann, C. F. *et al.* Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609–628 (2007).
49. Soria, C. D. *et al.* Spatial Autocorrelation of Species Diversity and Distributions in Time and Across Spatial Scales. *Glob. Ecol. Biogeogr.* **35**, e70221 (2026).
50. Gelman, A. & Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Stat. Sci.* **7**, 457–472 (1992).

51. Gelman, A. Comment: Fuzzy and Bayesian p-Values and u-Values. *Stat. Sci.* **20**, (2005).
52. Grattarola, F., Tschernosterová, K. & Keil, P. A continental-wide decline of occupancy and diversity in five Neotropical carnivores. *Glob. Ecol. Conserv.* **55**, e03226 (2024).
53. Doser, J. W. Convergence diagnostics and other considerations when fitting spatial occupancy models. <https://doserlab.com/files/spoccupancy-web/articles/modelconsiderations#convergence-assessment-of-spatial-factor-models-and-other-spatial-occupancy-models> (2024).
54. Mühleisen, H. & Raasveldt, M. duckdb: DBI Package for the DuckDB Database Management System. (2026).
55. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *R J.* **10**, 439–446 (2018).
56. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

Acknowledgements

Thanks to Jeff Doser for important comments related to spOccupancy and spatial occupancy models. Bird atlas data are mostly collected by skilled volunteers coordinated by ornithological organisations, which are responsible for the quality of the observations, data curation, analyses, and publication. We wish to acknowledge all these fieldworkers and the organisations that made these atlas possible, i.e. Czech Society for Ornithology, New York State Breeding Bird Atlas, Nature Alberta, Environment Agency of Japan and Biodiversity Center of Japan, Ornithological Society of New Zealand, Québec Breeding Bird Atlas, Maritimes Breeding Bird Atlas, the European Bird Census Council, and Butterfly Conservation and Natural England, as well as all their partners and supporters. FG, GOS, MT, CDS, FJRW, and PK were funded by the European Union (ERC, BEAST, 101044740). LH was funded by the Heather Corrie Fund by Butterfly Conservation.

Author Contributions

Conceptualisation and Investigation: FG and PK. Resources and Data curation: FG, GOS, MT, CDS, FJRW, RF, LH, SH, and PK. Formal analysis, Methodology, Visualisation, and Writing – original draft: FG. Funding acquisition: PK. Writing – review & editing: all authors.

643 **Funding**

644 This work was funded by the European Union (ERC, BEAST, 101044740). Views and
 645 opinions expressed are however those of the author(s) only and do not necessarily reflect
 646 those of the European Union or the European Research Council Executive Agency. Neither
 647 the European Union nor the granting authority can be held responsible for them.

648

649 **Competing Interests**

650 The authors declare no competing interests.

651