Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Leveraging natural history collections to understand the impacts of global change

Occupancy–detection models with museum specimen data: Promise and pitfalls

Vaughn Shirey[1] 🆔 | **Rassim Khelifa[2]** 🆔 | **Leithen K. M'Gonigle[3]** 🆔 | **Laura Melissa Guzman[3,4]** 🆔

[1]Department of Biology, Georgetown University, Washington, DC, USA

[2]Department of Zoology, University of British Columbia, Vancouver, BC, Canada

[3]Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada

[4]Marine and Environmental Biology Section at the Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

Correspondence
Vaughn Shirey
Email: vms55@georgetown.edu

Abstract

1. Historical museum records provide potentially useful data for identifying drivers of change in species occupancy. However, because museum records are typically obtained via many collection methods, methodological developments are needed to enable robust inferences. Occupancy–detection models, a relatively new and powerful suite of statistical methods, are a potentially promising avenue because they can account for changes in collection effort through space and time.

2. We use simulated datasets to identify how and when patterns in data and/or modelling decisions can bias inference. We focus primarily on the consequences of contrasting methodological approaches for dealing with species' ranges and inferring species' non-detections in both space and time.

3. We find that not all datasets are suitable for occupancy–detection analysis but, under the right conditions (namely, datasets that are broken into more time periods for occupancy inference and that contain a high fraction of community-wide collections, or collection events that focus on communities of organisms), models can accurately estimate trends. Finally, we present a case study on eastern North American odonates where we calculate long-term trends of occupancy using our most robust workflow.

4. These results indicate that occupancy–detection models are a suitable framework for some research cases and expand the suite of available tools for macro-ecological analysis available to researchers, especially where structured datasets are unavailable.

KEYWORDS
global change, hierarchical model, museum specimens, occupancy-detection models

1 | INTRODUCTION

Global change processes are contributing to the rapid restructuring of biodiversity across the planet (Araújo & Rahbek, 2006; Bellard et al., 2012). At the same time, data regarding species occurrence are becoming more widely available through the digitization of museum collections (Hedrick et al., 2020) and citizen science platforms (Dickinson et al., 2012). At the time of writing, two large databases of species occurrences, the Global Biodiversity Information Facility (GBIF) and Integrated Digitized Biocollections (iDigBio), contain over

---

2 billion records of species occurrences across the planet (both museum specimens and human observations, although we focus on the former here). These records may be rich in detailed information including date of collection, location, taxonomic determinations and collector names. The records are often our only available information about the historical presence of species. However, because these museum data are aggregated records from many different collectors and time periods, their collections usually stem from a variety of sampling designs which, for the most part, are not usually known. Consequently, they can contain numerous spatiotemporal and taxonomic biases (Isaac et al., 2014). For example, sampling of opportunistic (henceforth, 'unstructured') data is often uneven in space and time and may be spatially concentrated around regions of high human population densities (Daru et al., 2018; Mair & Ruete, 2016; Shirey et al., 2021; Tiago et al., 2017). Additionally, charismatic taxa are often oversampled with respect to the diversity of their respective clades, leading to occurrence shortfalls, or fewer than expected occurrences given the diversity of a particular clade, such as in hyperdiverse groups like arthropods (Callaghan et al., 2021; Troudet et al., 2017). In some groups, such as North American butterflies, undersampling is prolific in regions that are forecasted to experience the most dramatic changes in climate (Shirey et al., 2021). Given these biases, improper treatment of unstructured data can lead to misleading inferences (Guzman et al., 2021; Larsen & Shirey, 2021). Thus, it is imperative that we develop statistical frameworks for analysing unstructured data to afford researchers the ability to test hypotheses related to how global change is impacting species over long time-periods and large spatial extents. Here, we examine how attributes of unstructured datasets including site visitation history and collector sampling behaviour may impact model accuracy through simulation.

Species observation is rarely perfect so, even when a given species is present at a site, an observer may fail to detect it during any given survey (e.g. animals with cryptic phenotype or behaviour may be difficult to see). Occupancy–detection models are a powerful statistical framework for disentangling these imperfect processes of observation from actual species' occurrence or 'occupancy' (MacKenzie et al., 2003) and these models have been widely used for species' distribution modelling in ecology (Kéry & Schaub, 2011). Occupancy models have been extended to model entire communities ('multi-species occupancy models') and to model one or more species across multiple seasons ('dynamic occupancy models') (Kéry & Schaub, 2011). These models have improved our ability to explore spatial and temporal patterns of biodiversity and aided in identification of drivers of global change (Kéry & Royle, 2020).

Occupancy models are potentially well-suited to handle the inherent biases in species observations present in unstructured data (Erickson & Smith, 2021). A major challenge when doing so is accounting for the fact that records of when and where historical collectors sampled are rarely available. Collections that did not yield specimens, or did but the specimens were never archived in a museum, are typically unreported because no physical specimens exist in the museum collection. This shortcoming is further exacerbated by variable (and also unknown) individual collector

behaviours. For example, some collectors focus on particular taxonomic groups, while others collect entire communities (de Siracusa et al., 2020; Di Cecco et al., 2021). Species' traits may also influence the probability of detection or 'allure' and, thus, collection outcomes (Johnston et al., 2014). In all, physical specimens and their metadata represent a fragmented history of potential collection events that researchers must carefully piece together to reconstruct historical patterns.

Researchers have developed some best practices for overcoming some of these challenges for presence-only data. For instance, it might be appropriate to infer non-detections of one (or more) species at a given site based on known detections of other species at that same site (Kamp et al., 2016; Kéry, 2010; van Strien et al., 2013). Constraining detection/non-detection records only to those sites that are plausibly within the range of a species' geographical distribution has also been shown to improve model performance (Guzman et al., 2021). Combined, these steps not only improve estimates of occupancy and detection, but can also reduce the computational size of analyses. However, we lack clear guidelines that highlight best practices for avoiding the potential biases introduced by variable observer behaviour, temporal trends in visitation, and species distributions and traits.

Here, we use a simulation framework to explore the performance of occupancy–detection models when applied to data that contain many of the patterns likely present in natural history collection data. Specifically, we explore how changes in species' occupancy, species' detection and collector visit frequency impact model performance for datasets containing presence-only records across a community of species, each with its own geographical range. We consider the consequences of (a) incorporating estimated species' ranges and (b) inferring non-detections of some species from recorded detections of others on estimated trends in occupancy and detection, for different collector behaviours. We synthesize our findings by providing guidelines for steps that could be taken to reduce bias when applying similar occupancy–detection models to empirical datasets. Finally, we apply our scenarios to a real-world dataset containing detection records for eastern North American odonates (dragonflies and damselflies).

2 | MATERIALS AND METHODS

2.1 | Data simulation

We simulate a community of N species, each inhabiting some or all of J potential sites over K time periods, which we call 'occupancy intervals (OI)'. In application, occupancy intervals could correspond to single years or could span multiple years. We suppose that each occupancy interval is further broken into I visit intervals (VI), during which collectors may visit sites to sample some or all of the species present (commonly used terms in this manuscript are presented in the glossary of terms, Box 1).

Because species often differ in their spatial extent, in some of our scenarios, we simulate a unique 'geographical' range for each species. Specifically, we assume that species i potentially occupies

BOX 1 Glossary of terms commonly used throughout this manuscript

Glossary of terms

Occupancy interval: Period of time in which occupancy is estimated. Also referred to as closure period.

Visit interval: Units of time in which occupancy intervals are further divided and in which potential site visits might occur.

Visit: A species' observation at a site×time interval combination takes place when a collector searching for that species *visits* that sites in that time interval (visit interval). A visit can be inferred to have occurred if at least one species record from a site×time interval combination exists.

Site: We partition space into a spatial grid of *sites*, in which visits can then be inferred to have occurred.

Species' range: For a given species, the collection of sites where that species could plausibly be found.

Community sampling: A sampling event where a collector samples the entire community of species. Under this type of sampling, one can infer non-detections of species not found from detections of species that are found.

Targeted sampling: A sampling event where a collector primarily seeks a single species. Under this type of sampling, records of recorded species do not provide information on the absence of non-recorded species. Targeted and community sampling events are two ends of a continuous spectrum. We do not consider intermediate combinations of these strategies here.

sites $R[i] = \{s_{i,1}, s_{i,2}, \ldots, s_{i,r[i]}\}$, where $r[i] \in \{1, 2, \ldots, J\}$ denotes the total number of sites within the range of species i and we refer to $R[i]$ as the 'range' of species i. To simulate each species range, we generated polygons by drawing a convex hull around a series of random vertices on a defined grid a[i], where:

$$a[i] \sim 1 - (0.1 \times \text{Poisson}(1)). \tag{1}$$

For site j within the range of species i, we simulate occupancy in an occupancy interval k by drawing from a Bernoulli distribution with success probability $\psi[i, j, k]$, where

$$\psi[i, j, k] = \text{expit}(\psi_0 + \psi_{\text{sp}}[i] + \psi_{\text{OI}}[i] \times k). \tag{2}$$

Here ψ_0 denotes the baseline occupancy (on the linear scale), $\psi_{\text{sp}}[i]$ is a random species-specific intercept and $\psi_{\text{OI}}[i]$ is a random species-specific effect of occupancy interval (a positive value of $\psi_{\text{OI}}[i]$ would indicate that species i is increasing in occupancy through time). We assume that $\psi_{\text{sp}}[i]$ and $\psi_{\text{OI}}[i]$ are normally distributed, such that:

$$\psi_{\text{sp}}[i] : \mathcal{N}(0, \sigma_{\psi,\text{sp}})$$
$$\psi_{\text{OI}}[i] : \mathcal{N}(\mu_{\psi,\text{OI}}, \sigma_{\psi,\text{OI}}). \tag{3}$$

Here, $\sigma_{\psi,\text{sp}}$ denotes the interspecific variability in occupancy, $\mu_{\psi,\text{OI}}$ denotes the mean effect of occupancy interval on species occupancy and $\sigma_{\psi,\text{OI}}$ denotes the variability in species' temporal occupancy trends.

Because sampling effort can change through time, we assume that the probability that a site is visited in a visit interval can change through time (i.e. across occupancy intervals). Furthermore, because some collectors may record all species that they detect, while others may record only a single species, all site visits are not equivalent. If a collector records all species (we call this 'community sampling'), then a visit to a site in a given visit interval is a relevant visit *for all species*. However, if a collector only records a single species, ignoring others (we call this 'targeted sampling'), then a visit to a site in a given visit interval is only a meaningful visit for the recorded species. In our data simulation, we suppose that, in occupancy interval k, site j receives a 'community sampling' visit in each visit interval with probability

$$v[j, k] = \rho_{\text{com}} \cdot \text{expit}[\eta_0 + \eta_{\text{OI}} \times k] \tag{4}$$

and when such a visit occurs, each species is recorded as either detected or not detected. Here, η_0 denotes the baseline probability of a community sampling visit (on the linear scale) and η_{OI} is the effect of occupancy interval on visitation probability. This latter term allows the probability of site visitation to change systematically through time (i.e. across occupancy intervals). ρ_{com} lets us tune the fraction of community sampling visits relative to targeted sampling events; if $\rho_{\text{com}} = 1$, all visits are community sampling events, whereas, if $\rho_{\text{com}} = 0$, all sampling events are targeted sampling events. By parameterizing ρ_{com} as a function of model parameters, one could also allow collector behaviour to change across space or through time, however, that is beyond our scope here.

Only visit intervals (at each site) that do not receive a community sampling visit can be subjected to targeted sampling visits (a targeted sampling visit will not add any additional information if a community sampling visit also occurs in the same visit interval, so can be ignored). Targeted sampling visits are, by their nature, species specific. The probability that, in occupancy interval k, site j receives a targeted sampling visit for species i is given by

$$v[i, j, k] = (1 - \rho_{\text{com}}) \cdot \text{expit}[v_0[i] + v_{\text{OI}}[i] \times k] \tag{5}$$

where $v_0[i]$ denotes the baseline probability of a targeted sampling visit searching for species i (on the linear scale) and $v_{\text{OI}}[i]$ denotes the effect of occupancy interval on the probability of targeted sampling visits searching for species i. We assume that both $v_{\text{sp}}[i]$ and $v_{\text{OI}}[i]$ are normally distributed, such that

$$v_0[i] : \mathcal{N}(\mu_{v,0}, \sigma_{v,\text{sp}})$$
$$v_{\text{OI}}[i] : \mathcal{N}(\mu_{v,\text{OI}}, \sigma_{v,\text{OI}}), \tag{6}$$

where $\mu_{v,0}$ denotes the mean probability, across species, of a targeted sampling visit (in the first occupancy interval), $\mu_{v,\text{OI}}$ denotes the mean

effect, also across species, of occupancy interval on the probability of targeted sampling visits, and $\sigma_{v,\text{sp}}$ and $\sigma_{v,\text{OI}}$ denote the interspecific variation in visitation probability and variation in temporal change in visitation, respectively. In our simulations, we set $\eta_0 = \mu_{v,0}$ and $\eta_{\text{OI}} = \mu_{v,\text{OI}}$ so that both the probability of community and targeted sampling events increase or decrease simultaneously, but this does not need to be the case. For example, the probability of community sampling events could increase through time while the probability of targeted sampling events might decrease. Exploring such a scenario would likely be interesting, but is beyond the scope of our paper here. Because η_{OI} is equal to $\mu_{v,\text{OI}}$, we indicate values used for both parameters by only providing the latter (therefore η_{OI}) will simply be labelled as $\mu_{v,\text{OI}}$.

Finally, for each species, we simulate detection during the visit intervals where visits occurred. For site $j \in R[i]$, we draw species' i detections across the visits that occurred and could have detected that species (e.g. community sampling visits or targeting sampling visits aimed at species i) in occupancy interval k from a Bernoulli distribution with probability $p[i,j,k]$ where

$$p[i,j,k] = \text{expit}\left[p_0 + p_{\text{sp}}[i] + p_{\text{site}}[j,k] + p_{\text{OI}} \times k\right], \qquad (7)$$

where p_0 is the baseline detection probability (on the linear scale), $p_{\text{sp}}[i]$ is a random species-specific intercept, $p_{\text{site}}[j,k]$ is a random site-specific intercept that varies by occupancy interval and p_{OI} is an overall effect of occupancy interval. $p_{\text{site}}[j,k]$ allows spatiotemporal variability in detection probability (changing across sites and occupancy intervals), and helps account for the variation that is inherent in sample effort across space and time in unstructured historical datasets. p_{OI} allows detection to change systematically through time, as has likely occurred in many groups as sampling techniques have improved, for example. We assume that both p_{sp} and p_{site} are normally distributed, such that:

$$\begin{aligned} p_{\text{sp}}[i] &: \mathcal{N}(0, \sigma_{p,\text{sp}}) \\ p_{\text{site}}[j,k] &: \mathcal{N}(0, \sigma_{p,\text{site}}), \end{aligned} \qquad (8)$$

where $\sigma_{p,\text{sp}}$ denotes the interspecific variation in detection and $\sigma_{p,\text{site}}$ the spatiotemporal (site and occupancy interval) variation in detection.

We simulate our data under a scenario comprising 10 occupancy intervals and 3 visit intervals and then we further consider (a) five different fractions of community sampling events (ρ_{com}), (b) three different trends in visitation probability ($\mu_{v,\text{OI}}$; increasing, decreasing and stable) and (c) identical versus variable species ranges for a total of 270 simulation scenarios (specific parameter values are shown in Table S1 and Figure 1). We conducted all data simulations in R (code is available at https://github.com/lmguzman/occ_historical).

2.2 | Occupancy–detection models

We built an occupancy model that reflects the structure of the simulated datasets (i.e. modelling occupancy and detection probability). We do not incorporate data about site visitation history or community sampling information into our analyses, as this information is typically
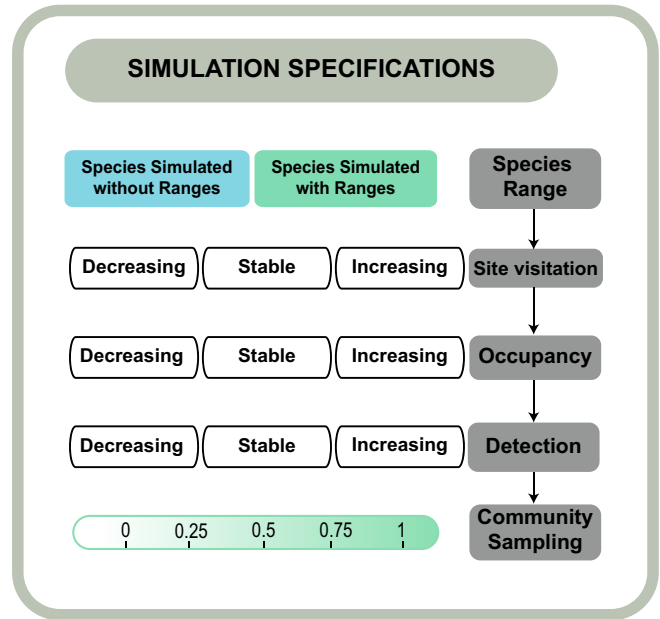

FIGURE 1 Schematic depicting data simulation decisions. Visit, occupancy and detection probabilities all have three levels: Decreasing, stable and increasing with time (−0.1, 0, 0.1). The probability that visits are community sampling events is: 0, 0.25, 0.50, 0.75 and 1. All five parameters were varied sequentially, which resulted in 270 multi-species detection–non-detection combinations ($2 \times 3 \times 3 \times 3 \times 5$)

difficult or impossible to obtain for historical datasets. How we infer site visitation history, which is part of our data filtering (rather than statistical model), is described below. We provide a detailed description of our occupancy models in Supporting Information.

Using our simulated datasets, we contrast the performance of six different data processing workflows one might follow when applying an occupancy model to a historical dataset. We provide an overview of these below and they are also shown graphically in Figure 2.

- $WF_{\text{all,all}}$: we model all sites for all species across all visit intervals, even when some of those visit intervals do not contain any known species detections.

- $WF_{\text{all,detected}}$: we model all sites for all species, but each site is only modelled over the visit intervals where at least one species was detected. Here, we infer non-detections of non-observed species when at least one other species was detected at a site in a given visit. This is a fair assumption when most collectors sample the entire community (ρ_{com} near 1), but not when they mostly target individual species (ρ_{com} near 0).

- $WF_{\text{all,visits}}$: we model all sites for all species, but each site only over the visit intervals where we know visits occurred (i.e. we model the true site visitation history which we are only able to do because we are working with simulated data). This case provides a benchmark for a 'best information' workflow.

- $WF_{\text{range,all}}$: we model each species over only the sites that fall within its inferred range and across all visit intervals (i.e. an analogue to $WF_{\text{all,all}}$ above, but where species' ranges are estimated and then incorporated).
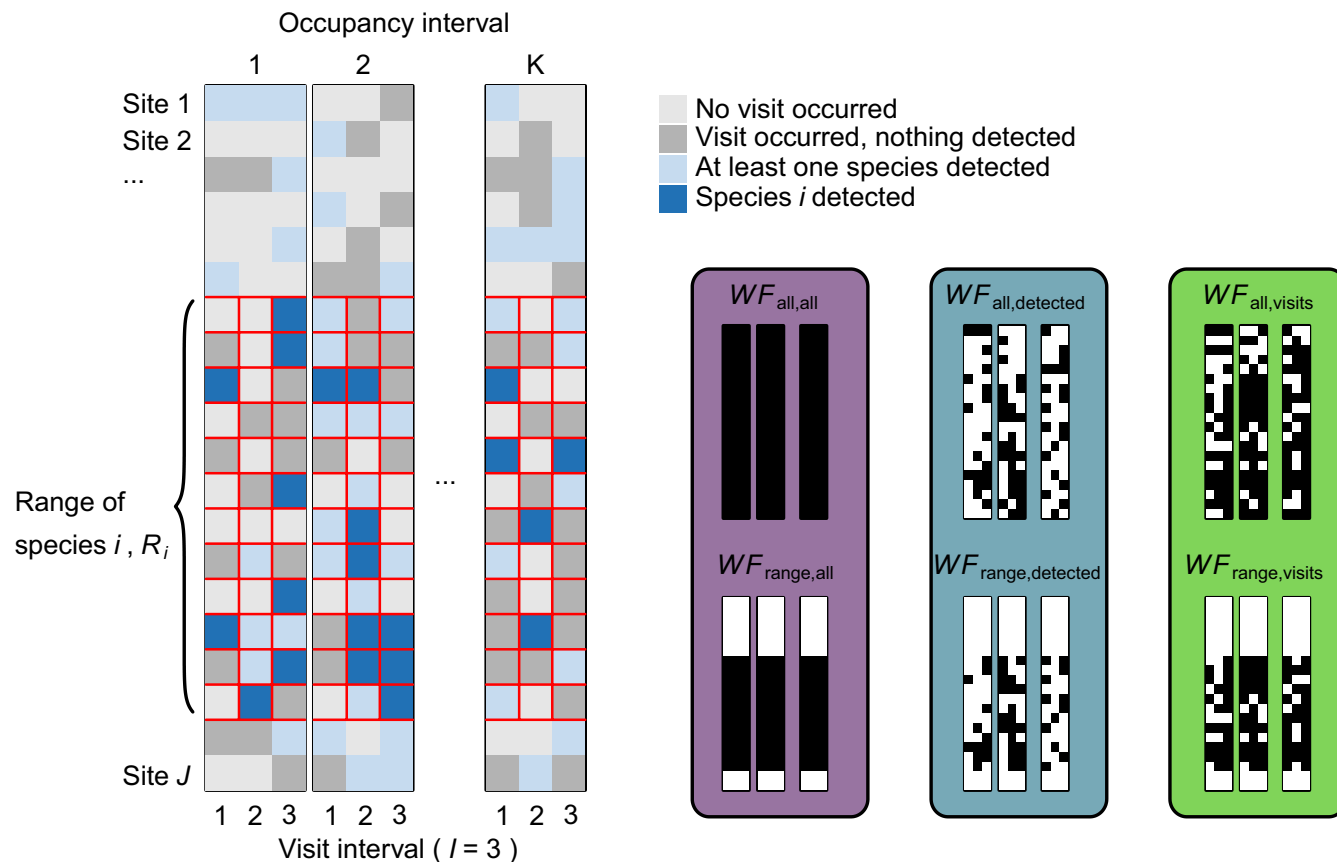
FIGURE 2 Schematic depicting data processing workflows. Workflows differ in which sites and time intervals they model. Small black and white panels show, for species *i*, which site × visit interval combinations from the larger panels are modelled in each workflow, with black denoting a modelled site × visit interval combination and white a non-modelled combination. For example, $WF_{range,all}$ models all visit intervals for every site within the range of species *i*, but none of the sites outside of it. Colours in large boxes behind small panels correspond to colours in subsequent figures (purple indicates workflows that model all sites regardless of visit; blue indicates workflows that model all sites with positive detections; green indicates workflows that model the true visit history). The values of both *K* and *I* are choices one must make when collapsing data prior to analysis. In our analyses, we use $K = 2, 5, 10$ and $I = 3$

• $WF_{range,detected}$: we model each species over only the sites that fall within its inferred range and, for each of those sites, only over visit intervals where at least one species was detected (this is an analogue to $WF_{all,detected}$ above, but where species' ranges are estimated and then incorporated).

• $WF_{range,visits}$: we model each species over only the sites that fall within its inferred range, but each site only over the visit intervals where we know visits occurred (this is an analogue to $WF_{all,visits}$ above, but where species' ranges are estimated and then incorporated).

In addition to these data processing workflows, we evaluated the effect of binning our data into different numbers of occupancy intervals. We simulated all datasets assuming 10 occupancy intervals and 3 visit intervals (so that the total amount of 'raw data' is always the same) and then, to investigate the impact of occupancy interval number, we re-binned the data to construct datasets containing $K = 2$, $K = 5$ and $K = 10$ occupancy intervals (always using $I = 3$ visit intervals). Note that visit intervals are 'longer' when we use fewer occupancy intervals for the same total amount of data (e.g. for a 12-year empirical dataset, we could bin data into

6 occupancy intervals, each comprising 2 one-year visit intervals or, alternatively, into 3 occupancy intervals, each comprising 2 two-year visit intervals). We compared different data processing workflows by assessing accuracy of model-estimated parameter values against their true values using root mean squared error (RMSE) across 10 replicates.

2.3 | Case study: Eastern north American odonates

2.3.1 | Study organism

Odonates are globally widespread aquatic insects that live in a wide range of freshwater systems. They are charismatic, easy to detect and identify, and popular among nature enthusiasts. Both larvae and adults are generalist predators and interact with a large number of aquatic and terrestrial taxa. In North America, there are 462 species (136 Zygoptera and 326 Anisoptera), with 336 species on the east coast (Paulson, 2011). Odonates are sensitive to environmental change, particularly climate warming, and evidence suggests that

they respond by shifting their phenology and geographical distribution (Hassall et al., 2007; Hickling et al., 2005), making them interesting sentinels of climate change.

2.3.2 | Odonate dataset

We selected 27 states in Eastern North America from GBIF (https://doi.org/10.15468/dl.cabqrc). The data comprise 47,378 records of 368 species observed during 1970–2019 and show a clear increase in the number of records through time, with 70.3% of the records occurring in the last two decades (2000–2019). We partitioned records into 'sites' by overlaying a grid of 100×100 km cells and filtered to include only species with more than 100 occurrences (195 species) across 265 sites and five occupancy intervals (decades), each further

partitioned into ten 1-year visit intervals. We also considered a scenario comprising 5-year occupancy intervals (for a total of $K = 10$) which we present in Figure S5. We decided to use a single year as a visit interval (for both 5-year and 10-year occupancy intervals).

2.3.3 | Analysis of the odonate dataset

We provide guidelines for researchers when analysing historical museum records with occupancy models (Figure 3) and we demonstrate application of these steps using the above-described odonate dataset. Specifically, we use a multispecies occupancy model, as described in Supporting Information. For each species, we construct the geographical range by drawing a convex polygon around all occurrences records. Because this method might overestimate
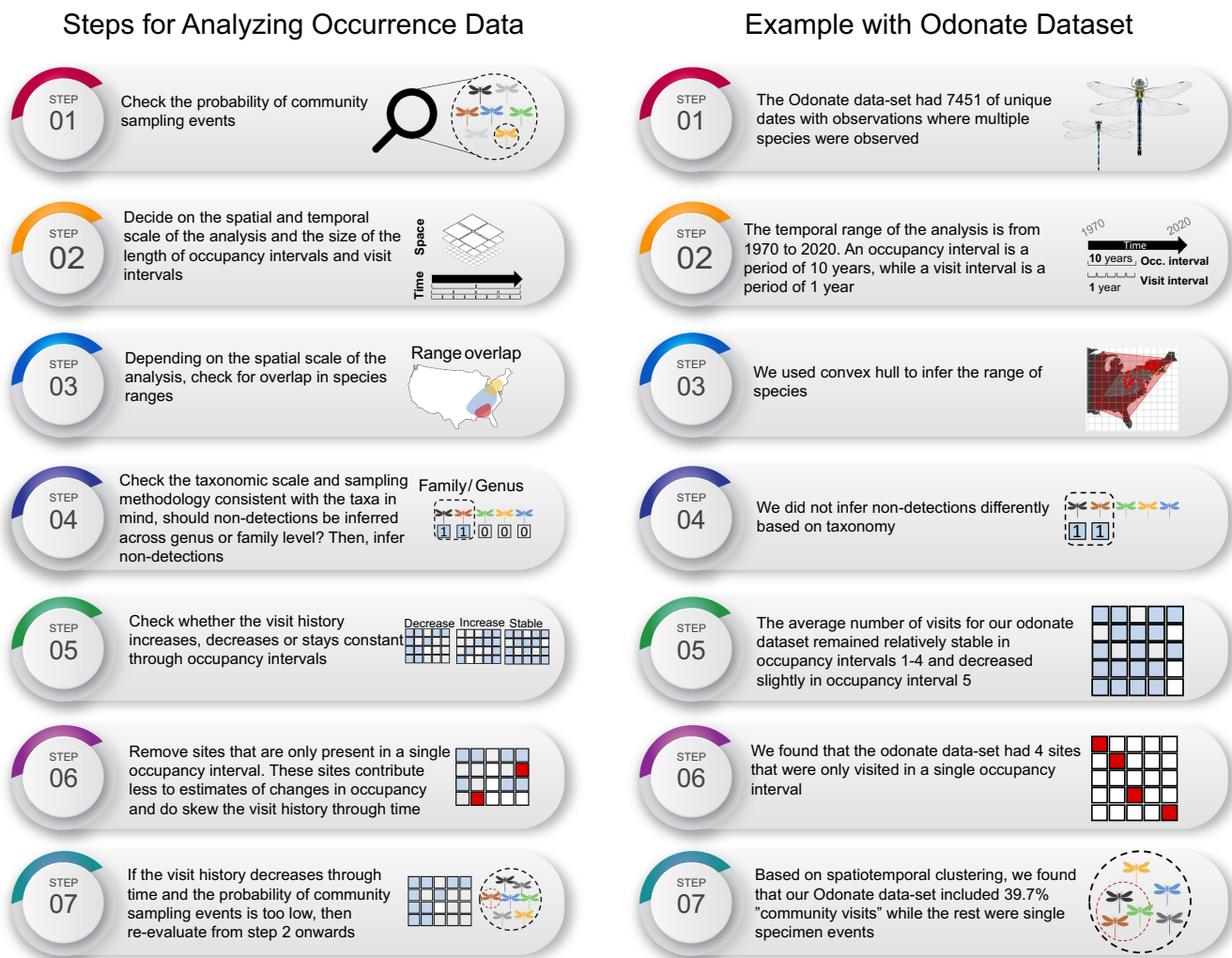

FIGURE 3 Steps that researchers can take to analyse historical record data using occupancy models, illustrated using a dataset of odonate occurrences. (1) eEstimate the probability of community sampling events. We used a spatiotemporal clustering procedure for our odonate data. (2) Decide on the spatiotemporal scope of the analysis. We recommend choosing 5 or more occupancy intervals. (3) Decide whether restricting inference to species ranges is important for your model. (4) Decide if non-detections should be inferred at a particular taxonomic level. (5) Check if the site-visitation history is increasing, decreasing or staying the same. (6) Remove sites that may only be present in a single occupancy interval. (7) Re-evaluate from Step 2 onwards if estimated community visitation probability and visit histories do not meet recommended thresholds

or underestimate ranges for some species, we suggest that when applying our methods, users should also rely on other available information such as expert range maps, if available. We process the data under the $WF_{range,detected}$ (best case scenario on simulated data, excluding $WF_{range,visits}$ which is not possible for empirical data), $WF_{all,detected}$ (ignores underlying species ranges) and $WF_{range,all}$ (models all visits, regardless of whether or not visits likely occurred). We also assessed how many of the records in our dataset likely originated via community sampling vs. targeted sampling visits. To do this, we grouped records collected on the same day and within 1 km of one another. We assumed that a community sampling event occurred if more than one species was present in each grouping. While this is likely a rough approximation of the actual fraction of community sampling events, it lets us quickly approximate this quantity without having to use more time-consuming methods like a detailed analysis of collector identities, etc.

We performed all analyses with R4.1.2 (R Core Team, 2021). We used the nimble package to compute our models (de Valpine et al., 2017). All of the code utilized in our study is available at https://github.com/lmguzman/occ_historical. A static repository for the code and data can be found via DataDryad by the following DOI: https://doi.org/10.5061/dryad.s1rn8pk9q (Shirey et al., 2022a); code also accessible via Zenodo at https://doi.org/10.5281/zenodo. 6439325 (Shirey et al., 2022b). The odonate dataset was downloaded from GBIF and can be found by the following DOI: https://doi.org/10.15468/dl.cabqrc (GBIF, 2021).

3 | RESULTS

3.1 | Simulation study

The choice of data processing workflow directly affects inferences made by occupancy models. Not surprisingly, only including site×visit interval combinations where visits actually occurred ($WF_{all,visits}$) yields better estimates than approaches that infer site visitation, and this is invariant to the fraction of visits that represent community sampling events (Figure 4). Performance of workflows that infer species' non-detections by either modelling all visit intervals ($WF_{all,all}$) or only visit intervals where at least one species was detected ($WF_{all,detected}$) depends on the fraction of visits that are community sampling events. Specifically, when all site visits correspond to community sampling events ($\rho_{com} = 1$), inferring species non-detections based on recorded detections of other species is sensible and, consequently, $WF_{all,detected}$ performs as well as $WF_{all,visits}$ (Figure 4). In contrast, when all site visits are targeted sampling events, ($\rho_{com} = 0$), detections of one species provide little information about non-detections of other species and, consequently, both $WF_{all,detected}$ and $WF_{all,all}$ yield biased estimates of occupancy change through time (purple and blue curves are much higher than green curve for small values of ρ_{com} in Figure 4).

Model error increases as the minimum fraction of visits that correspond to community sampling events decreases and, furthermore, the extent of this error depends on (a) the number of occupancy

FIGURE 4 Root mean squared error (RMSEs) across replicate simulated datasets for model estimated temporal trends in occupancy, $\mu_{\psi,OI}$ (left) and detection, p_{OI} (right). When we only model time intervals where visits actually occurred (green, $WF_{all,visits}$), estimates of change in occupancy ($\mu_{\psi,OI}$) and detection (p_{OI}) are generally more accurate (lower root mean square error RMSE) compared to workflows where we either model only visits with at least one species detected (blue, $WF_{all,detected}$) or where we model all time intervals (purple, $WF_{all,all}$). This difference is largest when the fraction of community sampling events is low (small ρ_{com}). Across the board, models that comprise a greater number of occupancy intervals generally yield more precise estimates (compare rows and the different range of values for the y-axis). $\mu_{v,OI} = 0$, species were simulated to occur in all sites, all other model parameters can be found in Table S1
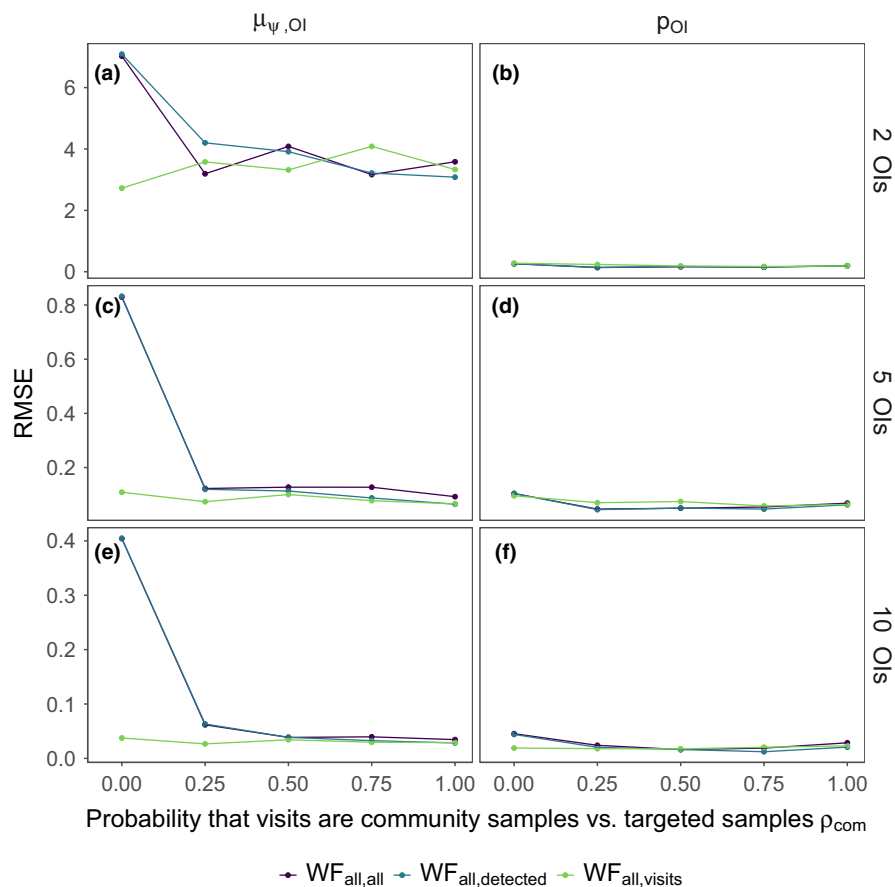
intervals, (b) whether species' ranges are incorporated into the analysis, (c) the extent to which site visitation probability changes through time, and (d) the extent to which occupancy and/or detection probability changes through time. We will discuss these processes next.

In general, estimated temporal changes in occupancy and/or detection probability through time are more accurate when data are binned as finely as possible (i.e. into a greater number of occupancy intervals; larger K) and this conclusion is true across data processing workflows (compare rows in Figure 4). Binning data into 10 occupancy intervals led to better estimates of changes in occupancy and detection through time than binning into five or two occupancy intervals (Figure 4, note the different scales on the y-axes). Estimates for changes in detection through time are, however, less sensitive to both the fraction of visits that comprise community sampling events and to the workflow (Figure 4).

We simulated datasets either by assuming that every species could potentially occupy every site or by assuming species were constrained to different (possibly overlapping) geographical ranges. We subsequently differentiate our data processing workflows into those that ignore potential differences in species ranges ($WF_{all,all}$, $WF_{all,detected}$ and $WF_{all,visits}$, which we collectively label as $WF_{all,X}$) and those where each species is modelled over only the sites in its range ($WF_{range,all}$, $WF_{range,detected}$ and $WF_{range,visits}$, which we collectively label as $WF_{range,X}$). When all species could plausibly occupy all sites, $WF_{all,X}$ and $WF_{range,X}$ become equivalent. Workflows that correctly match

the workflow with the underlying species' ranges (either $WF_{all,X}$ when all the species were simulated to plausibly occupy all sites or $WF_{range,X}$ when species ranges were simulated) generally had lower error rates in changes of occupancy through time than the case where the data processing workflow mismatched the underlying species ranges (using $WF_{all,X}$ when species ranges were simulated) (Figure 5a,e vs. c).

When the probability of site visitation decreased through time, we found that workflows that attempted to infer and model species' detections based on recorded presences of other species (e.g. $WF_{range,detected}$) needed a higher probability of community sampling events to perform well (Figure 6a,b), compared to cases where the probability of site visitation increase or stayed constant through time (Figure 6c–f). Generally, when a lower fraction of the visits correspond to community sampling events. For instance, models yield accurate estimates of occupancy and detection when only 25% of site visits are community sampling events, as long as the site visitation probability increased through time, the number of occupancy intervals was high ($K = 10$), the data processing workflow took into account the underlying species ranges and species' non-detections were only inferred when other species were detected (i.e. $WF_{range,detected}$).

Because real-world changes in quantities such as site visitation rate, occupancy, detection and collector behaviour are potentially changing through time, we ran a full suite of occupancy models that included 375 workflow-simulation combinations with 10 replicates
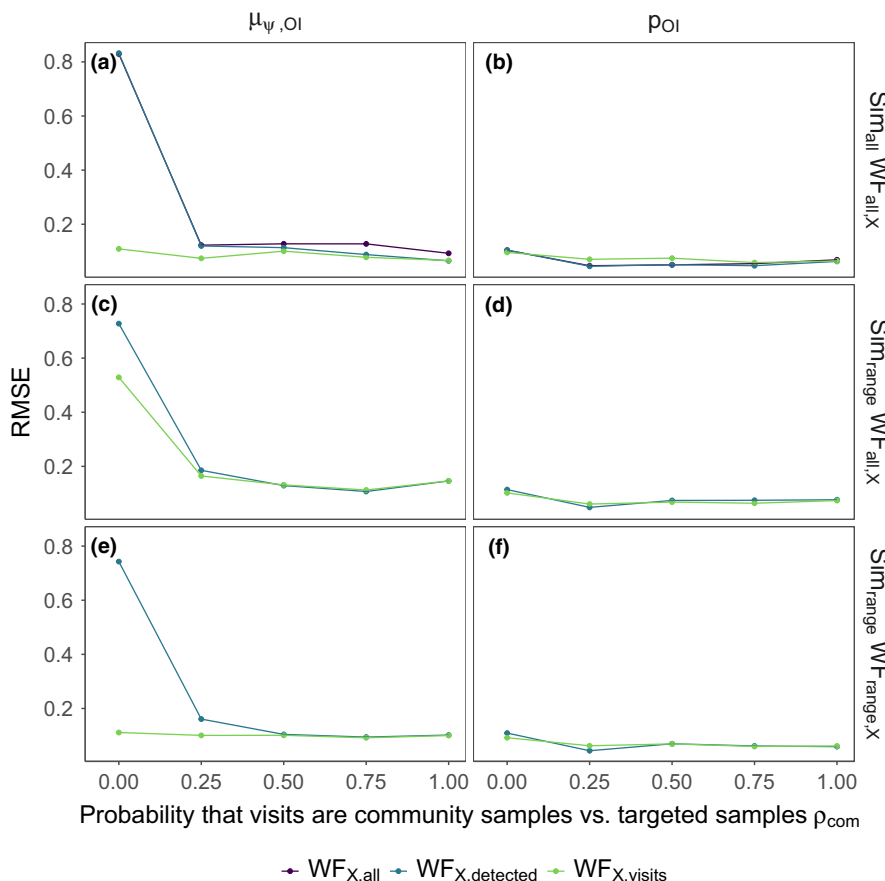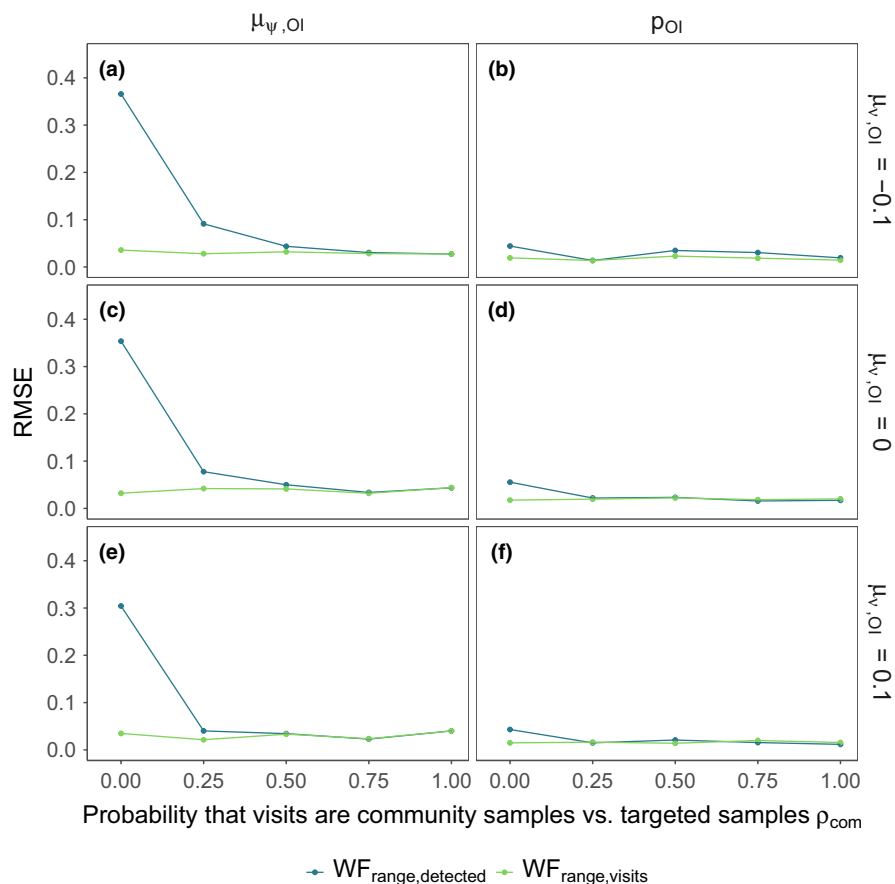

FIGURE 5 Root mean squared error (RMSEs) across replicate simulated datasets for model estimated temporal trends in occupancy, $\mu_{\psi,OI}$ (left) and detection, p_{OI} (right). Coloured lines denote workflows (in the legend, X denotes either 'range' or 'all'). The different rows show the effect of ignoring species ranges depending on whether ranges were simulated or not. In panels (a) and (b) (Sim$_{all}$), all species could plausibly occupy all sites (each species' range comprised all the sites), whereas in panels (c)–(f) (Sim$_{range}$) species were simulated to have ranges. $WF_{all,X}$ is the data processing workflow where all sites are included in the occupancy model (c and d, i.e. the ranges modelled do not match the ranges simulated) and $WF_{range,X}$ is the data processing workflow where the only sites that are included in the model are the ones that fall within a species range (e and f, i.e. only sites that fall within a species range are modelled so the modelling matches the ranges simulates). $\mu_{v,OI} = 0$, number of modelled occupancy intervals is $K = 5$

per combination for a total of 3,750 simulation runs. Root-mean squared errors (RMSEs) from this full set of workflow-simulation configurations revealed that overestimation of occupancy declines is likely to occur when the fraction of visits that are community focused is low (0–0.25) and this is especially true when detection and visit probability decrease through time (Figures S1–S3).

3.2 | Case study: Eastern north American Odonates

To examine occupancy trends for eastern North American odonates from the 1970s through 2010s, we ran occupancy models on data prepared using each of the processing workflows $WF_{range,detected}$, $WF_{all,detected}$ and $WF_{range,all}$. We found that the proportion of site×occupancy intervals×visit interval combinations that were likely community samples in this data was ~60.3%. To increase this measure, we applied stricter filtering criteria to the original dataset. Specifically, we restricted our records to only those from site×occupancy interval×visit interval combinations where at least 50% of sampling events were inferred to be community sampling events (i.e. multiple species detected on the same day within 1 km×1 km from each other). This led to an estimated global proportion of community samples equal to 89.4%. Even though we removed sites that did not have enough community sampling events, we still used the same range regardless of the sites included based on the proportion of

community events. We then re-ran an occupancy model using our best performing workflow ($WF_{range,detected}$) on this restricted dataset.

In general, we found that mean occupancy increased over the past four decades. However, species trends were highly variable, with many increasing while others decreased. Using a range-restricted workflow, our odonate results became less extreme in terms of overall, species-specific occupancy change from the 1970s to 2010s (Figure 7). For example, the $WF_{all,detected}$ workflow yields 72 species that are either increasing or decreasing by at least 25%, whereas $WF_{range,detected}$ reduces this figure to 39 species (5 species for $WF_{range,detected}$ when we filtered the data to have a higher proportion of community sampling events; Figure 7d).

4 | DISCUSSION

We assessed the performance of different data-processing workflows when applying multi-species occupancy models to simulated datasets to identify a set of best practices guidelines one might follow when analysing historic museum records or unstructured data. Our work here confirms the importance of restricting species-specific analyses across large geographical scales to those species' ranges. Furthermore, inferring non-detections of species from other species' detections emerged as a better approach than simply modelling all possible visit intervals at all sites. Our analysis also determines the extent to which patterns in data sampling, such as changing collector behaviour, can undermine
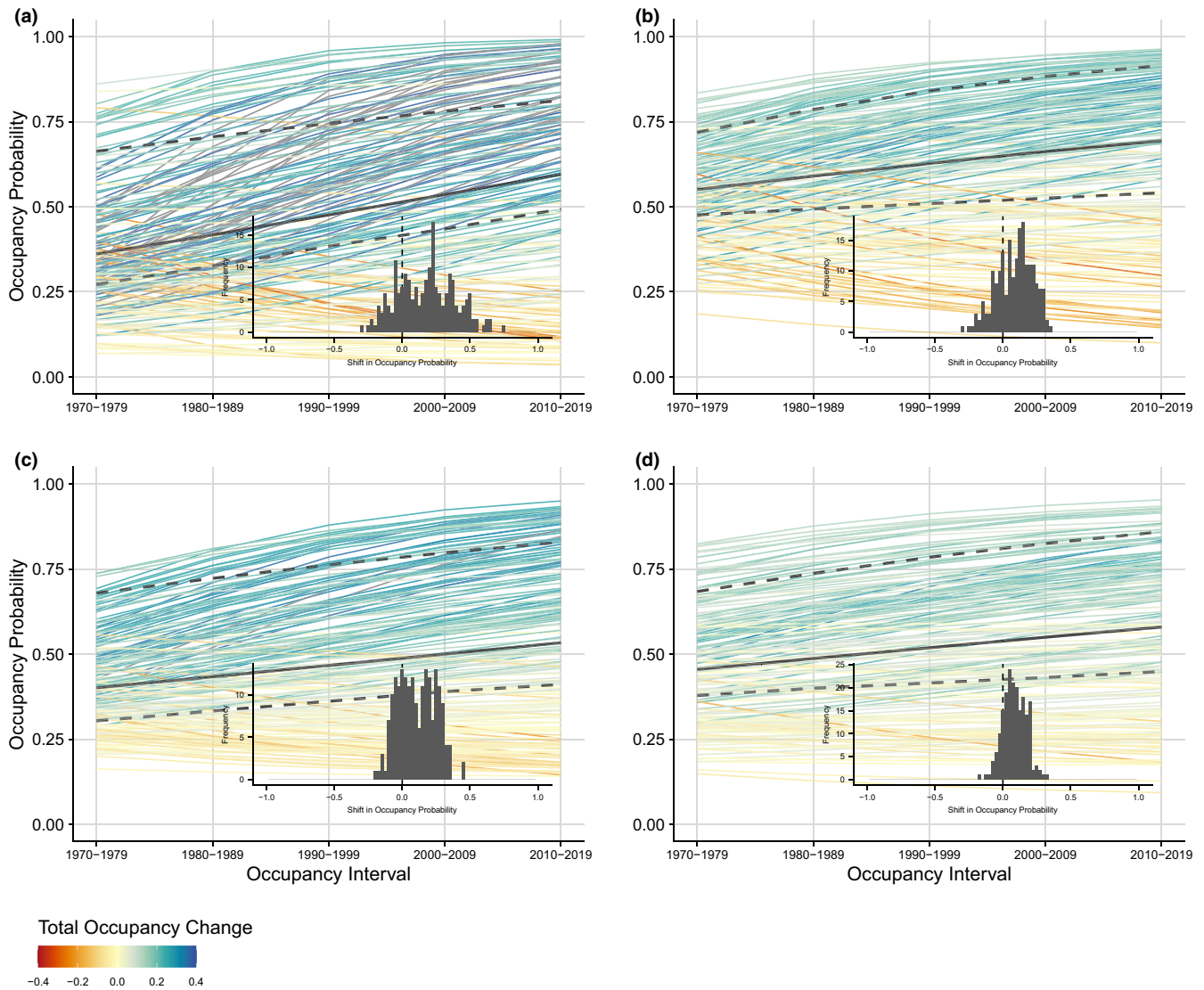
FIGURE 7 Species-specific (coloured lines) and community (black line) trends for eastern North American odonates under three data processing workflows: (a) $WF_{all,detected}$, (b) $WF_{range,all}$, (c–d) $WF_{range,detected}$. In (d), we further restrict analyses to only site × occupancy interval × time interval combinations where the proportion of clustered collection events containing two or more species exceeds 0.5. Lines are coloured to indicate occupancy trend through time (difference from occupancy interval 1 to occupancy interval 5). Inset histograms show these occupancy differences. Black dashed lines denote 95% Bayesian credible intervals

our ability to infer trends using occupancy models from historic records. In addition, workflows that bin occurrence data into fewer occupancy intervals are generally less accurate.

Inferring non-detections of some species based on recorded observations of other species is becoming a more frequent practice when working with presence-only occurrence data, but specific approaches differ. For example, van Strien et al. (2013), Kamp et al. (2016) and Powney et al. (2019) inferred non-detections for a particular species if a different species was observed at that same site on the same date. Outhwaite et al. (2019) used the same approach, but only inferred non-detections for species from the same taxonomic group as the observed individual(s). Analyses using these approaches have mostly taken place in Europe and with taxonomic groups where community science approaches yield structured datasets, for example those from regional monitoring schemes

(van Swaay et al., 2008). For example, Johnston et al. (2021) found that well-curated checklists combined with incomplete checklists were sufficient to enable reasonable estimates of occupancy, but not encounter rates. Here we have quantified the extent to which a dataset can comprise incomplete checklists or simply individual observations while still producing good parameter estimates. We found that, for the cases considered here, our workflows produced reliable estimates of occupancy trends through time when at least 50% of the samples derive from community sampling events.

Obtaining an approximate estimate of how frequently collectors sample the entire community, versus only collect individual species, is a necessary step in assessing whether the methods we use here are likely to be appropriate. Unfortunately, direct measurement of this quantity is often impossible for large historic datasets and, thus, we suggest that researchers approximate this metric to the best of

their ability. In our odonate analysis, we grouped records in space and time and assumed that specimens collected on the same day and nearby to one another were part of the same collection event. Based on our simulated datasets, we conclude that using two occupancy intervals is generally a bad practice as estimates of occupancy trends have high error. In workflows that span a greater number of occupancy intervals, the probability of required community visits could conceivably be as low as 25 %; however, this likely depends on other data-specific trends and parameter estimates, and we would suggest that authors conducting such analyses always use simulation to guide their analyses and extensive examination of the patterns in the raw data (i.e. potential visitation trends). Our examination of interactions between changes in occupancy and detection probability over time as well as changes in visitation probability revealed that when detection probability or when visit probability decrease over time, overestimation of occupancy declines will likely occur in datasets where the probability of community visits is less than 0.5 (Figures S8–S10).

In applying our different workflows to a large historic dataset for eastern North American odonates, we demonstrate that neglecting to incorporate estimates of species' ranges demonstrates how different workflow choices can impact downstream results with a real-world dataset. In particular, the results from this analysis show how data filtering can reduce the estimation of dramatic occupancy increases and declines (Figure 7a vs. Figure 7b–d). Overestimation of occupancy change is a known behaviour of occupancy–detection models when heterogeneous observer effort is not accounted for (Kéry & Royle, 2020). Our findings generalize this conclusion to show that similar overestimation is possible when species are modelled outside of their ranges (using the odonate example), when species are modelled in all time intervals, or when there is too few community sampling events (using simulated data).

Using the best workflow to analyse the odonate dataset, we found that the majority of species have exhibited range expansions while a few have exhibited range contractions in the eastern United States. This is mostly concordant with the assessment of the conservation status and population trends of the IUCN Red List at the global scale which lists all studied species as Least Concern with stable or increasing population trends. Odonates have a high ability to colonize new suitable habitats following the warming of cool areas (Hickling et al., 2005) or the creation of artificial sites (e.g. ponds for irrigation) (Simaika et al., 2016). The observed low-range contractions might be due to their ability to rapidly recolonize areas where they were temporarily extirpated (Shiffer & White, 2014), which is often observed at the local scale in the studied area (Shiffer et al., 2014, 2015). Studies using occupancy models to assess the temporal pattern of the geographical distribution of odonates at a large scale are rare (Rocha-Ortega et al., 2020); however, they may be possible under the occupancy–detection framework.

While we only use 'single-season' occupancy models in our analyses here, we also explored (but do not present) dynamic occupancy models which link each occupancy interval via extinction–colonization dynamics (Royle & Kéry, 2007). In general, we found

that these models performed poorly unless the probability of community visitation was high. Missing visits/occupancy intervals in the dynamic framework created multiple possibilities for parameter estimation, where a single occupancy interval with missing data could represent extinction followed by colonization, or persistence; multiple consecutive occupancy intervals with missing visits simply compounds the number of possibilities. We expect that many historical datasets likely do not have enough data to make dynamic models feasible. However, for many questions, single-season models parameterized with appropriate fixed and random temporal effects should suffice.

Processes which generate unstructured occurrence data are complex and it is likely that the fraction of community sampling events (vs. targeted sampling) varies across occupancy intervals and through time. This may be especially relevant if researchers are interested in utilizing combined museum specimen and community science records for inference. Individual collector behaviours may be important in models using these combined datasets given that collector behaviour has a potentially large effect on community science outcomes (Di Cecco et al., 2021). Incorporating these processes is an increasingly important task, as community science records far outpace the digitization and collection of specimens via other methods in the last decade (Shirey et al., 2021; Spear et al., 2017).

While the current paper presents context to certain technical aspects of occupancy models that may have important consequences, there are many further complications that require additional attention. For example, restricting species' range in multi-species occupancy modelling is a good general practice when estimating occupancy and detection. However, if researchers are interested in shifts in species ranges outside of their currently considered range (e.g. range expansion), the set of modelled sites for a given species needs to be sufficiently large to allow for such range expansions. Adding a buffer around a species' inferred range is a possible solution, but such decisions should be explored via simulation in more detail.

5 | CONCLUSIONS

Our analyses of simulated data demonstrate that for some unstructured data, robust estimates of temporal trends in occupancy and/or detection can be obtained, particularly when (a) community sampling events comprise at least half of all sampling events and (b) when analyses span a greater number of occupancy intervals. We strongly encourage researchers to 'get to know' their data by assessing the ways in which it has been historically collected and curated and to perform, at a minimum, the guidelines we propose in this paper. Estimating the fraction of collection events that were community sampling events is critical for assessing how reliable occupancy–detection estimates may be. Overall, understanding when occupancy models are likely to fail and lead to incorrect inferences is a must when applying these models to historical museum records.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13896.

DATA AVAILABILITY STATEMENT

The code used to generate simulated data is available via GitHub at https://github.com/lmguzman/occ_historical. A static repository for the code and data can be found via DataDryad by the following DOI: https://doi.org/10.5061/dryad.s1rn8pk9q (Shirey et al., 2022a); code also via Zenodo at https://doi.org/10.5281/zenodo.6439325 (Shirey et al., 2022b). The odonate dataset was downloaded from GBIF and can be found by the following DOI: https://doi.org/10.15468/dl.cabqrc (GBIF, 2021).

ORCID

Vaughn Shirey https://orcid.org/0000-0002-3589-9699
Rassim Khelifa https://orcid.org/0000-0001-6632-8787
Leithen K. M'Gonigle https://orcid.org/0000-0002-6015-9748
Laura Melissa Guzman https://orcid.org/0000-0003-2198-9738

REFERENCES

Araújo, M. B., & Rahbek, C. (2006). How does climate change affect biodiversity? *Science*, 313, 1396–1397.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, 15, 365–377.

Callaghan, C. T., Poore, A. G., Hofmann, M., Roberts, C. J., & Pereira, H. M. (2021). Large-bodied birds are over-represented in unstructured citizen science data. *Scientific Reports*, 11, 1–11.

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M., & Davis, C. C. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, 217, 939–955.

de Siracusa, P. C., Gadelha, L. M., & Ziviani, A. (2020). New perspectives on analysing data from biological collections based on social network analytics. *Scientific Reports*, 10, 1–10.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26, 403–413.

Di Cecco, G., Barve, V., Belitz, M., Stucky, B., Guralnick, R., & Hurlbert, A. (2021). Observing the obserers: How participants contribute data to inaturalist and implications for biodiversity science. *Bioscience*, 71, 1179–1188.

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T., & Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, 10, 291–297.

Erickson, K. D., & Smith, A. B. (2021). Accounting for imperfect detection in data from museums and herbaria when modeling species distributions: Combining and contrasting data-level versus model-level bias correction. *Ecography*, 44, 1341–1352.

GBIF. (2021). *Gbif occurrence download*. https://doi.org/10.15468/dl.cabqrc

Guzman, L. M., Johnson, S. A., Mooers, A. O., & M'Gonigle, L. K. (2021). Using historical data to estimate bumble bee occurrence: Variable trends across species provide little support for community-level declines. *Biological Conservation*, 257, 109141.

Hassall, C., Thompson, D. J., French, G. C., & Harvey, I. F. (2007). Historical changes in the phenology of British Odonata are related to climate. *Global Change Biology*, 13, 933–941.

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A., Blackburn, D. C., Edwards, S. V., & Davis, C. C. (2020). Digitization and the future of natural history collections. *Bioscience*, 70, 243–251.

Hickling, R., Roy, D. B., Hill, J. K., & Thomas, C. D. (2005). A northward shift of range margins in British Odonata. *Global Change Biology*, 11, 502–506.

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

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

Johnston, A., Newson, S. E., Risely, K., Musgrove, A. J., Massimino, D., Baillie, S. R., & Pearce-Higgins, J. W. (2014). Species traits explain variation in detectability of UK birds. *Bird Study*, 61, 340–350.

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., & Donald, P. F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Diversity and Distributions*, 22, 1024–1035.

Kéry, M. (2010). *Introduction to WinBUGS for ecologists: A Bayesian approach to regression, ANOVA and related analyses*. Academic Press, Elsevier.

Kéry, M., & Royle, J. A. (2020). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 2: Dynamic and advanced models*. Academic Press.

Kéry, M., & Schaub, M. (2011). *Bayesian population analysis using WinBUGS: A hierarchical perspective*. Academic Press.

Larsen, E. A., & Shirey, V. (2021). Method matters: Pitfalls in analysing phenology from occurrence records. *Ecology Letters*, 24(6), 1287–1289.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local

extinction when a species is detected imperfectly. *Ecology*, *84*, 2200–2207.

Mair, L., & Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS ONE*, *11*, e0147796.

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

Paulson, D. (2011). *Dragonflies and damselflies of the east*. Princeton University Press.

Powney, G. D., Carvell, C., Edwards, M., Morris, R. K., Roy, H. E., Woodcock, B. A., & Isaac, N. J. (2019). Widespread losses of pollinating insects in britain. *Nature Communications*, *10*, 1–6.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rocha-Ortega, M., Rodríguez, P., Bried, J., Abbott, J., & Córdoba-Aguilar, A. (2020). Why do bugs perish? Range size and local vulnerability traits as surrogates of odonata extinction risk. *Proceedings of the Royal Society B*, *287*, 20192645.

Royle, J. A., & Kéry, M. (2007). A bayesian state-space formulation of dynamic occupancy models. *Ecology*, *88*, 1813–1823.

Shiffer, C. N., Leppo, B., & White, H. (2014). Odonata of Black Moshannon State Park, Centre County, Pennsylvania. *Argia*, *26*, 8–15.

Shiffer, C. N., Leppo, B., & White, H. B. (2015). Odonata of Beaver Dam, Huntingdon County, Pennsylvania: A record of faunal succession in a changing habitat. *Argia*, *27*, 12–23.

Shiffer, C. N., & White, H. B. (2014). Dragonfly and damselfly colonization and recolonization of a large, semi-permanent Pennsylvania pond. *Northeastern Naturalist*, *21*, 630–651.

Shirey, V., Belitz, M. W., Barve, V., & Guralnick, R. (2021). A complete inventory of North American butterfly occurrence data: Narrowing data gaps, but increasing bias. *Ecography*, *44*, 537–547.

Shirey, V., Khelifa, R., M'Gonigle, L., & Guzman, L. (2022a). Dryad repository containing simulated data. https://doi.org/10.5061/dryad.s1rn8pk9q.

Shirey, V., Khelifa, R., M'Gonigle, L., & Guzman, L. (2022b). Zenodo repository containing archived code. https://doi.org/10.5281/zenodo.6439325.

Simaika, J. P., Samways, M. J., & Frenzel, P. P. (2016). Artificial ponds increase local dragonfly diversity in a global biodiversity hotspot. *Biodiversity and Conservation*, *25*, 1921–1935.

Spear, D., Pauly, G., & Kaiser, K. (2017). Citizen science as a tool for augmenting museum collection data from urban areas. *Frontiers in Ecology and Evolution*, *5*, 86.

Tiago, P., Ceia-Hasse, A., Marques, T. A., Capinha, C., & Pereira, H. M. (2017). Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Scientific Reports*, *7*, 1–9.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*, 9132.

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

van Swaay, C., Nowicki, P., Settele, J., & van Strien, A. (2008). Butterfly monitoring in europe: Methods, applications and perspectives. *Biodiversity and Conservation*, *17*, 3455–3469.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Shirey, V., Khelifa, R., M'Gonigle, L. K., & Guzman, L. M. (2022). Occupancy–detection models with museum specimen data: Promise and pitfalls. *Methods in Ecology and Evolution*, *00*, 1–13. https://doi.org/10.1111/2041-210X.13896