# Modelling Invasive Insect Spread with Citizen Science Data: A Dynamic Occupancy-Detection Model

*Yufan Zheng*[1,2,*], *Elle Saber*[2,3], *Bernd Gruber*[1], *Helen Nahrung*[4], *Angus Carnegie*[5], *Richard P. Duncan*[1,2]

[1]Centre for Conservation Ecology and Genomics, Institute for Applied Ecology, University of Canberra, Canberra ACT, Australia

[2]Australian Research Council Training Centre in Plant Biosecurity, Canberra ACT, Australia

[3]Biology Teaching & Learning Centre, Australian National University, Canberra ACT, Australia

[4]Forest Research Institute, University of the Sunshine Coast, Sunshine Coast QLD, Australia

[5]Forest Science, Fisheries & Forestry, New South Wales Department of Primary Industries and Regional Development, Sydney NSW, Australia

*Corresponding author: Yufan Zheng (yufan.zheng@canberra.edu.au)

## Acknowledgements

The PhD research on the Data Driven Decision Model for Forest Biosecurity is supported by Forest & Wood Products Australia (FWPA). This project is also supported by the Australian Government through the Australian Research Council Training Centre in Plant Biosecurity (IC230100027) and the Australian Government Research Training Program Scholarship (https://doi.org/10.82133/C42F-K220). Insect images in Figure 1 from iNaturalist are used under the following licenses: CC BY-NC 4.0 (Harsi S. Parker, Jessee J. Smith, dbeadle, ildikorab, Sergio Ibarra, Riki Miller), CC BY 4.0 (Michael Knapp), and CC0 1.0 (Katja Schulz).

## Author contributions

Under the guidance of R. P. D., Y. Z. conceived the ideas for the algorithm, collected the data, and wrote the source code for modelling; Y. Z. evaluated the model, analyzed and interpreted results advised by R. P. D., E. S., and B. G.; Y. Z. wrote the first draft of the manuscript and all co-authors reviewed and edited the manuscript and gave final approval for publication. R. P. D. contributed to project supervision and funding acquisition.

## Data availability statement

The processed data and source code are available at xxx.

## Conflict of interest statement

The authors do not have any conflict of interests to report.

# Modelling Invasive Insect Spread with Citizen Science Data: A Dynamic Occupancy-Detection Model


## Abstract

1. Human-mediated biological invasions are accelerating worldwide, but detecting early incursions and tracking their spread remains difficult. Opportunistic citizen science platforms, such as iNaturalist and eBird, generate large volumes of presence-only records, offering new opportunities to detect and monitor invasive species.
2. However, these data come with uneven sampling effort, strong spatial and taxon biases, and lack explicit non-detections, complicating efforts to infer species absences. To address these challenges, we develop a dynamic occupancy-detection model (DODM) that estimates colonisation, local extinction, and occupancy dynamics for invasive species using presence-only data.
3. The model infers detection probability as a function of background sampling effort from related taxa, enabling separation of non-detection from true absence. This detection process, combined with a temporal correlation structure and spatial dependency, is used to infer spatiotemporal changes in species occupancy using a hierarchical Bayesian model that captures uncertainty in how invasions unfold over space and time.
4. Using iNaturalist observations, we apply this model to infer detection probability, occupancy trajectories, dynamic colonisation and extinction probabilities for eight recently introduced insect species in eleven locations worldwide. Our results show that a well-structured DODM can unlock dynamic information from presence-only data, offering a flexible and generalizable framework for monitoring and risk mapping biological invasions.



## Keywords

Presence-only records; Detection probability; Sampling effort; Spatiotemporal dynamics.


## 1. Introduction

Human activities have substantially reshaped the geographic ranges of many species over decades to centuries (Boivin et al., 2016; Faurby & Svenning, 2015). Invasive species, introduced through human-mediated pathways such as trade and travel, impose high economic, environmental and social costs by impacting primary production, disrupting natural ecosystems, and spreading diseases (Crowl et al., 2008; Hulme, 2009). Invasive species often spread rapidly yet unevenly across regions, making their spatiotemporal dynamics difficult to monitor, predict, and manage (Baker & Bode, 2016; Buchadas et al., 2017). Effective monitoring is therefore critical for informing timely response and management. (Reaser et al., 2020). Traditional monitoring relies on planned field surveys and structured visits, which yield high-quality observations but are costly and logistically demanding at the spatial and temporal scales required (Holden *et al.*, 2016; Jardine and Sanchirico, 2018). In contrast, citizen science platforms (e.g., iNaturalist and eBird) have grown dramatically and now generate large volumes of species occurrence records across broad regions (Kosmala et al., 2016). These data are increasingly used to complement traditional surveillance and can expand the spatial and

temporal coverage available for invasive-species monitoring (Hulbert et al., 2023; Pocock et al., 2024; Roe et al., 2024). However, because they rely on opportunistic observations, citizen science data exhibit uneven sampling effort and strong spatial, temporal, and taxonomic biases, and they lack explicit non-detections (Johnston et al., 2018; Mengersen et al., 2017). These limitations complicate inference about species occupancy and distribution. While a sighting provides evidence of presence, the absence of a record is ambiguous because it may reflect either genuine absence or a failure to detect or report the species under uneven sampling effort. As a result, recent research has focused on extracting reliable distributional information from opportunistic observations while accounting for imperfect detection (Altwegg & Nichols, 2019; Guillera-Arroita, 2017; Probert et al., 2022; van Woensel et al., 2025).

Several approaches have been developed to infer species distributions and occupancy from opportunistic citizen science data, with robust inference depending on how sampling bias and imperfect detection arising from non-random sampling, observer behaviour, and uneven effort are handled (Isaac et al., 2014; Isaac & Pocock, 2015; Powney & Isaac, 2015). Recent work has focused on quantifying how geographic and environmental sampling bias, variable detectability, and uneven effort distort occupancy and range estimates, including optimized background sampling, spatial thinning or weighting, explicit effort proxies, and bias-correction methods (Brizuela-Torres et al., 2024; Coomber et al., 2021; Erickson & Smith, 2021; Fajgenblat et al., 2025a; Oliveira et al., 2024). However, these approaches address different components of the observation process. Methods that correct sampling bias can reduce distortions in estimated distributions but do not necessarily distinguish changes in occupancy from changes in observation effort or detectability. Process-explicit approaches have therefore advocated for dynamic phenomena such as biological invasions and range shifts, but their application can be constrained by demanding data requirements, particularly where repeated surveys or reliable measures of sampling effort are unavailable (Briscoe et al., 2019). This problem becomes more acute when reconstructing historical changes in occupancy because observation intensity and reporting effort often vary substantially through time, while records are typically sparser in earlier periods. Consequently, an apparent range expansion may reflect colonisation, increased observation, or sampling effort.

Dynamic occupancy models (DOMs) provide a framework for addressing this problem by separating the underlying ecological state from the observation process, allowing estimation of colonisation and local extinction while explicitly modelling imperfect detection. They are particularly applicable to modelling invasion dynamics because they can incorporate the spatial and temporal dependencies that arise as invasions unfold across space and time (Bled *et al.*, 2013). DOMs that model colonisation through local dispersal processes can further capture spatially structured species spread (Bled *et al.*, 2011; Broms *et al.*, 2016; Rushing *et al.*, 2019). Some occupancy models have also been developed for large spatial extents and non-standard survey designs, including settings with alternative forms of replication or single-visit data (Bled *et al.*, 2013; Sadoti *et al.*, 2013; Peach *et al.*, 2017). When applied to opportunistic citizen science data, DOMs can recover long-term trends that align with standardized monitoring while accounting for biases in detection, observation, and reporting (van Strien *et al.*, 2013a; Molinari-Jobin *et al.*, 2018). More recently, integrated frameworks combined presence-only

(PO) and presence-absence (PA) data to estimate temporal change while incorporating sampling effort and spatial structure (Grattarola *et al.*, 2023). Evidence from community- and continental-scale syntheses further highlights the feasibility and value of inferring dynamic processes even when data are biased (Grattarola, *et al.*, 2024; Fajgenblat *et al.*, 2025). Nevertheless, applying these approaches to reconstruct invasion dynamics from PO citizen science archives remains challenging because PO records provide no explicit non-detections and observation effort can vary across space and time.

Addressing this challenge requires a framework that can distinguish changes in occupancy from changes in observation while simultaneously representing the spatial and temporal dependence characteristic of invasive spread. To this end, we propose a Dynamic Occupancy-Detection Model (DODM) designed specifically for PO data. Building on advances in existing DOMs (Bled et al., 2013; Broms, Hooten, Johnson, et al., 2016; Outhwaite et al., 2018), our model explicitly represents the observation process to account for imperfect detection and uneven reporting effort, links occupancy states through time to reconstruct temporal dynamics, and incorporates neighbourhood-based spatial dependence to represent local spread. We first describe the modelling framework and outline how it addresses the challenges of reconstructing invasion dynamics from PO data. We then demonstrate its application by reconstructing the spread of invasive insect species across multiple locations worldwide using citizen science observations.

## 2. Methods

### 2.1. A framework for modelling spatiotemporal changes in species occupancy

Our objective is to reconstruct spatiotemporal changes in occupancy for a target species using PO records containing geographic coordinates and observation dates. We assign each observation to a grid cell and then estimate the probability that each cell was occupied in each period, thereby reconstructing how occupancy changes across space and time. Our approach integrates several elements that have proven effective in previous dynamic occupancy models (Bled et al., 2013; Broms, Hooten, Johnson, et al., 2016; Outhwaite et al., 2018). Specifically, in periods where no observations occur in a grid cell – an ambiguous situation in which the species may be truly absent or present but unrecorded – we estimate colonisation and occupancy probability using all available information: spatial dependence on neighbouring cells; temporal dependence on the cell's previous occupancy states; and a detection function that gives the probability of recording the species conditional on presence, modelled using background sampling effort. Spatial and temporal dependencies borrow strength from adjacent cells across space and time, improving estimates in regions with sparse observations. Together, these components allow us to infer likely occupancy from PO data, along with colonisation rates and extinction probabilities, even in the absence of explicit non-detections.

To reconstruct spatiotemporal occupancy patterns from PO data, we used two key inputs: (i) PO records of the target species, grouped by grid cell and time period, to infer occupancy dynamics; and (ii) the total number of PO records of phylogenetically related taxa (e.g., species in the class) in each grid cell and time period, used as a proxy for sampling effort directed towards the target species. Because PO data do not directly record sampling effort or non-detection, we use records of other insect taxa from the same citizen science platform as a proxy for spatiotemporal variation in observer activity. This follows the logic of target-group background approaches in species distribution models, in which records from taxa sharing similar observation processes are used to characterise spatial sampling bias (Barber et al., 2022; Phillips et al., 2009). We assume that visits generating many observations of related taxa reflect high sampling effort for the target species, whereas locations or periods without such records indicate limited observer activity. Restricting the sampling effort proxy to related taxa, rather than all observations on the citizen science platform, provides a measure more closely aligned with the search behaviour and habitats relevant for detecting the target species. We then incorporate the total number of PO records of related taxa into a detection function when fitting the dynamic occupancy model, allowing us to distinguish low occupancy from low detection arising from uneven citizen science sampling effort.

We first partition the landscape into grid cells, indexed by $i = 1,2,3 \ldots N$, and divide time into periods (e.g., months and years), indexed by $t = 1,2,3 \ldots T$. For each grid cell $i$, we use the PO data to determine whether the target species was recorded during each period $t$. This results in a set of binary observations $y_{it} \in \{0,1\}$, where $y_{it} = 1$ if the species was recorded in grid cell $i$ in period $t$, and $y_{it} = 0$ otherwise. An observation of $y_{it} = 1$ confirms the species was present in the grid cell $i$ during period $t$. However, an observation of $y_{it} = 0$ is ambiguous, as

it can arise in two ways: (i) the species was truly absent from the grid cell during that period, or (ii) the species was present but went unrecorded.

To account for both possibilities, we model $y_{it}$ as a Bernoulli distributed random variable with probability $z_{it}p_{it}$, where $z_{it} \in \{0,1\}$ is a latent binary variable indicating latent true presence (1 if present, 0 if not), and $p_{it}$ is the detection probability: the probability the species was recorded given presence:

$$y_{it} \sim \text{Bernoulli}(z_{it}p_{it}). \quad (1)$$

We modelled the probability of detecting the target species, conditional on its presence, as a function of sampling effort in grid cell $i$ during period $t$. To quantify sampling effort, we used the insects recorded in each grid cell and period, using the same broad taxonomic group as an effort proxy for all target species. This assumes that a greater number of insect records corresponds to greater observer activity and, consequently, a higher probability of recording the target species, provided it is present. We refer to the relationship between detection probability $p_{ij}$ and sampling effort $x_{ij}$ as the detection function. We modelled this function as a sigmoidal response to the logarithm of the total number of insect records in grid cell $i$ during period $t$:

$$p_{it} = \frac{\beta_1}{1 + e^{\beta_2 - \beta_3 \log(x_{it}+1)}}, \quad (2)$$

Here, $\beta_1$ represents the upper asymptote of the curve, constrained such that $0 < \beta_1 < 1$. Parameters $\beta_2$ and $\beta_3$ control the shape and steepness of the curve. This detection function accommodates scenarios where detection is unlikely at low sampling effort and where detection remains imperfect even at high sampling effort ($\beta_1 < 1$), as can be the case for cryptic or hard-to-observe species.

We modelled the latent presence of a target species in grid cell $i$ during period $t$ as a Bernoulli random variable:

$$z_{it} \sim \text{Bernoulli}(\psi_{it}), \quad (3)$$

where $\psi_{it}$ denotes the occupancy probability (i.e., the probability that the species was truly present in that cell and period). When a cell is known or assumed to be occupied ($z_{it} = 1$), $\psi_{it}$ should be close to 1, but not exactly 1, as occupancy is treated probabilistically in this framework.

We then modelled occupancy probability on the logit scale using two state components:

$$\text{logit}(\psi_{it}) = \alpha_t + S_i, \quad (4)$$

where $\alpha_t$ captures temporal variation in period $t$, and $S_i$ captures spatial variation among grid cells. We further decomposed this state process into two stages: an initialisation stage and subsequent colonisation-extinction dynamics.

For initialisation (period $t = 1$), we set $\alpha_1$ as a global intercept representing the baseline probability of occupancy across all grid cells. To allow for spatial variation in occupancy during this first period, we modelled the spatial effect $S_i$ using a conditionally autoregressive (CAR) prior:

$$S_i \sim \text{Normal}\left(\frac{1}{\sum_k w_{ik}} \sum_k w_{ik} S_k, \frac{\sigma_S^2}{\sum_k w_{ik}}\right), \quad (5)$$

reflecting the assumption that occupancy probabilities are spatially correlated. Here, $S_i$ denotes the spatial effect for grid cell $i$, modelled as conditionally dependent on the spatial effects of its neighbouring cells $k$. The weights $w_{ik}$ are binary indicators taking the value 1 when cell $k$ is a neighbour of cell $i$, and 0 otherwise. We defined neighbours as the eight cells immediately surrounding a target cell based on queen contiguity. Under this CAR prior, the expected value of $S_i$ is the mean of the spatial effects of its neighbouring cells. The variance term $\sigma_S^2$ controls how much $S_i$ may deviate from this local mean, scaled by the inverse of the number of neighbours. This structure captures spatial autocorrelation in occupancy, consistent with the ecological expectation that initial species distributions tend to be spatially aggregated rather than randomly dispersed. It also allows the model to borrow strength across neighbouring cells, smoothing over data gaps and improving baseline estimates of occupancy in areas with sparse observations.

Having initialized occupancy in period $t = 1$, we modelled subsequent dynamics (periods $t > 1$) using colonisation and extinction processes. Following Eq. 3, the occupancy probability was defined as:

$$\psi_{it} = z_{i,t-1}(1 - \varepsilon_{it}) + \left(1 - z_{i,t-1}\right)\gamma_{it}, \quad (6)$$

where $\varepsilon_{it}$ is the extinction probability and $\gamma_{it}$ is the colonisation probability.

Colonisation and extinction probabilities were modelled on the logit scale, with temporal trends represented by first-order random walk (RW) effects and neighbourhood dependence in colonisation captured by the proportion of occupied neighbouring cells in the previous period (Broms, Hooten, Johnson, et al., 2016; Outhwaite et al., 2018). Specifically, we expressed extinction and colonisation probabilities ($\varepsilon_{it}$ and $\gamma_{it}$) as:

$$\text{logit}(\varepsilon_{it}) = \sigma_\varepsilon \eta_{\varepsilon t} + \mu_\varepsilon, \quad (7)$$

$$\text{logit}(\gamma_{it}) = \sigma_\gamma \eta_{\gamma t} + \beta_n \theta_{i,t-1} + \mu_\gamma. \quad (8)$$

The parameters $\mu_\varepsilon$ and $\mu_\gamma$ are intercepts that represent baseline extinction and colonisation probabilities. In particular, $\mu_\gamma$ captures the background colonisation rate across all cells. The terms $\eta_{\varepsilon t}$ and $\eta_{\gamma t}$ are temporal effects that allow extinction and colonisation probabilities to vary among periods to capture period-to-period changes in spread driven by unmeasured time-varying processes (e.g., environmental variation, management, or propagule pressure). $\sigma_\varepsilon$ and $\sigma_\gamma$ scale these temporal effects, determining the magnitude and direction of temporal variation, respectively. Colonisation is further influenced by a neighbourhood occupancy effect through

$\beta_n\theta_{i,t-1}$, where $\theta_{i,t-1}$ denotes the occupancy status of neighbouring cells around cell $i$ in the previous period $t-1$. The coefficient $\beta_n$ quantifies how strongly neighbour occupancy influences colonisation. Positive values of $\beta_n$ indicate that empty cells are more likely to be colonised when more neighbouring cells were occupied in the preceding period, reflecting local dispersal processes. Together, these parameterisations allow extinction and colonisation rates to vary through time while incorporating both neighbourhood occupancy effects and a background rate of colonisation, thereby capturing key features of invasion dynamics.

To allow colonisation and extinction probabilities to vary smoothly over time, we modelled their temporal effects using first-order RW processes, following the approach in Outhwaite *et al.*, 2018. RW processes smooth temporal variation by borrowing strength across adjacent periods, reducing noise and improving the precision of estimated trends. The temporal effects of colonisation and extinction were defined as:

$$\eta_{\varepsilon,2} \sim \text{Normal}(0,1), \qquad \eta_{\gamma,2} \sim \text{Normal}(0,1)\ (t=2). \tag{9}$$

$$\eta_{\varepsilon,t} \sim \text{Normal}\left(\eta_{\varepsilon,t-1}, 1\right), \qquad \eta_{\gamma,t} \sim \text{Normal}\left(\eta_{\gamma,t-1}, 1\right)\ (t \geq 3). \tag{10}$$

To capture local spatial dependence in colonisation (i.e., neighbour-occupancy effects), we followed Broms *et al.* (2016) by modelling colonisation probability as an increasing function of the occupancy status of neighbouring sites. This is useful for modelling invasive spread because many invasive populations expand primarily via local dispersal, such that colonisation is more likely when nearby sites are already occupied (Solà et al., 2024; Takahashi & Park, 2020). In our model, we defined the neighbour-occupancy effect $\theta_{i,j-1}$ for cell $i$ in the previous period $t-1$ as:

$$\theta_{i,t-1} = \begin{cases} \dfrac{\sum_j w_{ij} z_{j,t-1}}{d_i}, & d_i > 0, \\ 0, & d_i = 0. \end{cases} \tag{11}$$

Where $\theta_{i,t-1}$ represents the proportion of neighbouring cells that were occupied. The weights $w_{ij} \in \{0,1\}$ take the value 1 if cells $i$ and $j$ are neighbours and zero otherwise, with neighbourhoods based on queen-contiguity. Thus, $d_i = \sum_j w_{ij}$ is the number of neighbouring cells of cell $i$, irrespective of their occupancy status. In our model, colonisation probability can increase as $\theta_{i,t-1}$ becomes larger, reflecting the expectation that empty cells surrounded by more occupied neighbours are more likely to be colonised.

We fitted two dynamic occupancy-detection models (DODMs) to assess the influence of the neighbourhood occupancy effect. The first model, DODM1, used the colonisation process specified in Eq. 8, including the neighbourhood occupancy term $\beta_n\theta_{i,t-1}$, the temporal change in colonisation $\sigma_\gamma\eta_{\gamma t}$ and the background colonisation term $\mu_\gamma$. The second model, DODM2 omitted the neighbourhood occupancy term $\beta_n\theta_{i,t-1}$, meaning colonisation was not explicitly linked to neighbourhood occupancy.

These two models were otherwise identical, retaining the same parameterization for colonisation and extinction processes, and adopting the hierarchical modelling structure of the DOMs. By comparing the fit of these two models to the data, we could assess whether including the neighbour occupancy term improved model fit and provided evidence for local spatial dependency in spread.

## 2.2. Study area and species

To illustrate how our proposed DODM can be applied to opportunistic citizen science data, we developed a series of case studies focused on invasive insects. Although, like other DOMs, our DODM is broadly applicable across taxa and ecosystems (Guillera-Arroita, 2017; Outhwaite et al., 2018), invasive insects offer a particularly informative test case (de Groot et al., 2022). This is because invasive insects pose substantial biosecurity risks, often undergo rapid range expansion following introduction (Crowl et al., 2008; Jardine & Sanchirico, 2018), and are frequently first detected in and around urban areas where introduction pathways and human activity are concentrated (Branco et al., 2019; Gaertner et al., 2017). Urban environments also exhibit sharp spatial variation in climate, land use, and habitat structure over short distances (Gaertner et al., 2017), making them well suited for examining colonisation, extinction, and neighbourhood occupancy effects. For these reasons, we assembled a dataset of insect incursions covering multiple continents using iNaturalist records (iNaturalist community, 2025) through the rinat R package.

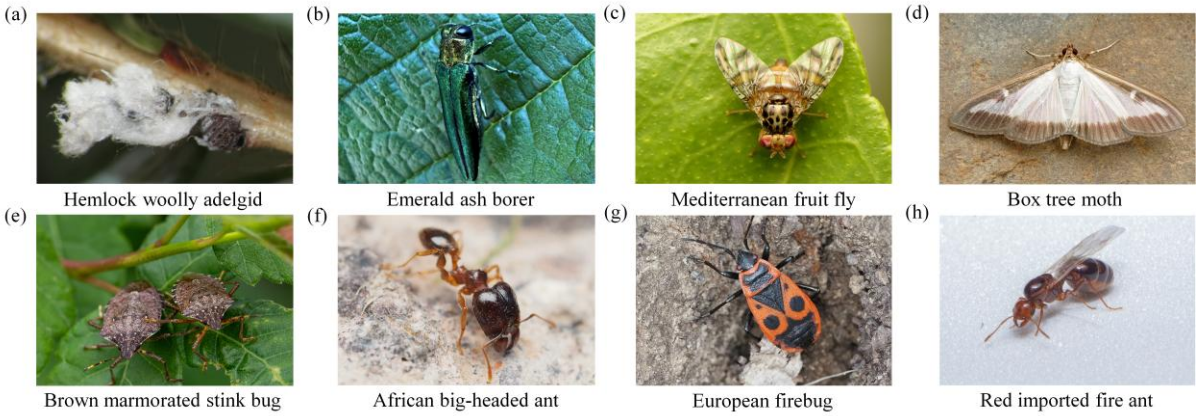



**Figure 1. Species selected for this study and their common names (from iNaturalist).** (a) Hemlock woolly adelgid (Harsi S, 2023); (b) Emerald ash borer (Smith, 2026); (c) Mediterranean fruit fly (Schulz, 2018); (d) Box tree moth (dbeadle, 2018); (e) Brown marmorated stink bug (ildikorab, 2024); (f) African big-headed ant (Sergio, 2025); (g) European firebug (Knapp, 2025); (h) Red imported fire ant (Miller, 2023). Image credits are listed in the Acknowledgments.

Our dataset comprised sixteen species-location combinations spanning eleven cities and eight invasive insect species across three continents (North America, Europe, and Australia), using records collected between July 2018 and July 2025. Species-location combinations were

selected by first identifying established invasive insect populations from the literature and official reports, and subsequently screening iNaturalist for candidate locations with that met prespecified minimum record thresholds over the study period. All observations were sourced from iNaturalist, and we chose this time window to capture a period of consistently high iNaturalist activity across the urban areas, avoiding earlier years with sparse reporting. We prioritized invasive taxa and, where possible, included species found in two continents to capture variation in climatic and geographic contexts. This design allowed us to assess whether species exhibited consistent occupancy dynamics across locations. The selected locations encompassed a wide range of urban forms and climates, ensuring that inferences about colonisation, extinction, and neighbourhood effects were not restricted to a single environmental context. Table 1 summarizes information for the eight target species (Figure 1). We provide the invasion history of each target species in the selected study locations, including the first recorded or reported year, likely introduction pathway, and supporting references in Appendix A.1.

Within each location, we examined the spatial distribution of records for each target species and identified the area of incursion as of July 2025. We placed a rectangular boundary around the area of incursion, including a buffer, to define the study extent. Each species-location pair met two criteria: (i) ≥ 100 presence records for the target species during the study period, ensuring sufficient data for modelling, (ii) ≥ 100,000 total insect records within the same period, providing a robust proxy for sampling effort to estimate detection functions.

For comparability across locations, we used a consistent sampling framework. After identifying the rectangular study area based on the area of incursion, we overlaid a 10 × 10 km grid over the study area, which defined the grid cells used for modelling species occupancy. We then partitioned the study period (July 2018 to July 2025) into fourteen six-month intervals. For each cell and period, we used iNaturalist data to (i) determine whether the target species was recorded, and (ii) calculate the total number of insect records as a proxy for sampling effort. We used all insect records rather than order- or family-level records because they provide a broader and less sparse measure of observer activity across space and time. In the Melbourne (*P. apterus*) case study, all insect records closely tracked the temporal pattern of family-level Pyrrhocoridae records, supporting their use as a background measure of sampling effort (Appendix A.2).

**Table 1. Study areas and species observation summary.** For each of the eight target invasive insect species, we identified two locations that had been recently colonised and had sufficient iNaturalist records for the period 2018-2025. The year indicates the first reported occurrence of the species in each location, based on published literature or official reports, and is not restricted to iNaturalist observations. We extracted iNaturalist PO records for each species within a rectangular study area defined by the longitude and latitude ranges, which encompassed the extent of spread of each species in each location. The number of target species records refers to the total number of presence records for the target species within the study area. The total number of insect records refers to the total number of presence records for all

insect taxa within the same study area, which defined the background sampling effort. Each study area was partitioned into 10 km × 10 km grid cells: the total number of grid cells was those retained after masking to include only terrestrial areas.

| Target species | Location (Year of first reported occurrence) | Number of target species records | Total number of insect records | Longitude range | Latitude range | Total number of grid cells |
|---|---|---|---|---|---|---|
| Hemlock woolly adelgid (*Adelges tsugae*) | Nova Scotia (2017) | 494 | 44,725 | [-66.2, -64.2] | [43.7, 45] | 182 |
| | Boston (≈1960s) | 788 | 573,902 | [-71.9, -70.5] | [41.8, 43] | 153 |
| Emerald ash borer (*Agrilus planipennis*) | Toronto (2007) | 1,027 | 1,270,572 | [-81.5, -77] | [42, 44.5] | 617 |
| | Dallas (2022) | 237 | 915,659 | [-98, -96.2] | [31, 33.6] | 548 |
| Mediterranean fruit fly (*Ceratitis capitata*) | Los Angeles (1975) | 725 | 775,133 | [-119, -116.5] | [33, 34.5] | 278 |
| | Lisboa (≈1898) | 105 | 133,031 | [-9.5, -8.7] | [38, 39.5] | 103 |
| Box tree moth (*Cydalima perspectalis*) | Toronto (2018) | 1,358 | 1270,572 | [-81.5, -77] | [42, 44.5] | 617 |
| | Paris (2008) | 215 | 147,684 | [1.4, 3.2] | [48, 49.5] | 238 |
| Brown marmorated stink bug (*Halyomorpha halys*) | Los Angeles (2006) | 1,005 | 775,133 | [-119, -116.5] | [33, 34.5] | 278 |
| | Milan (2013) | 831 | 175,244 | [8.5, 9.8] | [45, 46.3] | 136 |
| African big-headed ant (*Pheidole megacephala*) | Brisbane (≈1912) | 315 | 590,600 | [152.2, 153.6] | [-28.2, -26.3] | 256 |
| | Miami (1933) | 235 | 212,700 | [-81.2, -79.9] | [25.3, 27.2] | 249 |
| European firebug (*Pyrrhocoris apterus*) | Melbourne (2018) | 325 | 595,137 | [143, 146] | [-39, -36.5] | 635 |
| | Toronto (2017) | 1,599 | 1,270,572 | [-81.5, -77] | [42, 44.5] | 617 |
| Red imported fire ant (*Solenopsis invicta*) | Brisbane (2001) | 368 | 590,600 | [152.2, 153.6] | [-28.2, -26.3] | 256 |
| | Los Angeles (1998) | 199 | 775,133 | [-119, -116.5] | [33, 34.5] | 278 |

## 2.3. Model fitting and evaluation

We fitted DODM1 and DODM2 to each species-location dataset. Comparing the fit of these two models allowed us to assess whether incorporating neighbourhood occupancy improved model performance and provided evidence for local spatial dependence in spread.

We used Markov chain Monte Carlo methods for model inference implemented in NIMBLE (version 1.3.0) (de Valpine et al., 2017) run through R (version 4.5.2). Parameter estimates were obtained by cyclically sampling from each parameter's full conditional distribution. For each species and each model, we ran three Markov chain Monte Carlo chains of 80,000 iterations, discarding the first 20,000 as burn-in and retaining every third sample. We assessed parameter convergence using the Gelman-Rubin statistic (R-hat, $\hat{R}$), which compares within-chain and between-chain variance (Gelman & Rubin, 1992). We considered convergence

adequate for Rhat <1.1. Being a Bayesian model, we assigned prior distributions for all unknown parameters, for which we specified relatively uninformative priors:

$$\beta_1 \sim \text{Uniform}(0,1), \beta_2 \sim \text{Normal}(0,3), \beta_3 \sim \text{Normal}(0,3)$$

$$\sigma_\varepsilon \sim \text{Uniform}(0,1), \mu_\varepsilon \sim \text{Normal}(0,2)$$

$$\sigma_\gamma \sim \text{Uniform}(0,1), \mu_\gamma \sim \text{Normal}(0,2)$$

$$\beta_n \sim \text{Normal}(0,2)$$

We evaluated model performance for each species to identify the best fitting model. To assess how well each model inferred the underlying occupancy process, we used two metrics: the log score (LS) and success-rate at coverage level 5% (SR@5). The LS, a logarithmic scoring rule widely used in occupancy model comparisons (Broms, Hooten, & Fitzpatrick, 2016; Broms, Hooten, Johnson, et al., 2016), quantifies how well posterior predictive probabilities align with observed presence records. It is computed as the negative log posterior predictive probability, such that smaller LS values indicate that the model assigns higher predictive probability to observed presences. For PO data, the model estimates the latent occupancy state $z_{it}$ and detection probability $p_{it}$ for cell $i$ at time $t$. The LS for PO data is then calculated as:

$$\hat{y}_{it} = \Pr(y_{it} = 1 \mid \text{data}) = \mathbb{E}[z_{it}p_{it} \mid \text{data}] \approx \frac{1}{S}\sum_{S=1}^{S} z_{it}^{(s)} p_{it}^{(s)}, \quad (15)$$

$$\text{LS} = -\frac{1}{N_+}\sum_{(i,t)\in\mathcal{P}} log\, \hat{y}_{it}. \quad (16)$$

For PO data, only positive records are available. Accordingly, $\mathcal{P}$ denotes the set of true presences (positive records), $N_+ = |\mathcal{P}|$ is the total number of observed presences, and $S$ is the number of Markov chain Monte Carlo samples. The quantity $\hat{y}_{it}$ is the posterior predictive probability of observing a presence, which is the expected value of $z_{it}p_{it}$, integrating occupancy and detection.

To evaluate model fit in terms of how well it captured spatial patterns, we used SR@5, a summary statistic derived from the success-rate curve at a specified coverage level. We defined the high detection area as the top 5% of grid cells ranked by the posterior predictive probability of observing a presence ($\hat{y}_{it}$). SR@5 measures the proportion of observed presences that fall within this top 5% of cells. In plain terms, SR@5 tells us how well the model concentrates known occurrences into the small set of locations it thinks are most likely to contain the species.

Given the set of observed presences $\mathcal{P}$ and the total number of presences $N_+ = |\mathcal{P}|$, we rank all grid cells in descending order of their posterior predictive probability $\hat{y}_{it}$. SR@k is then defined as the proportion of observed presences that fall within the top k% highest-ranked cells:

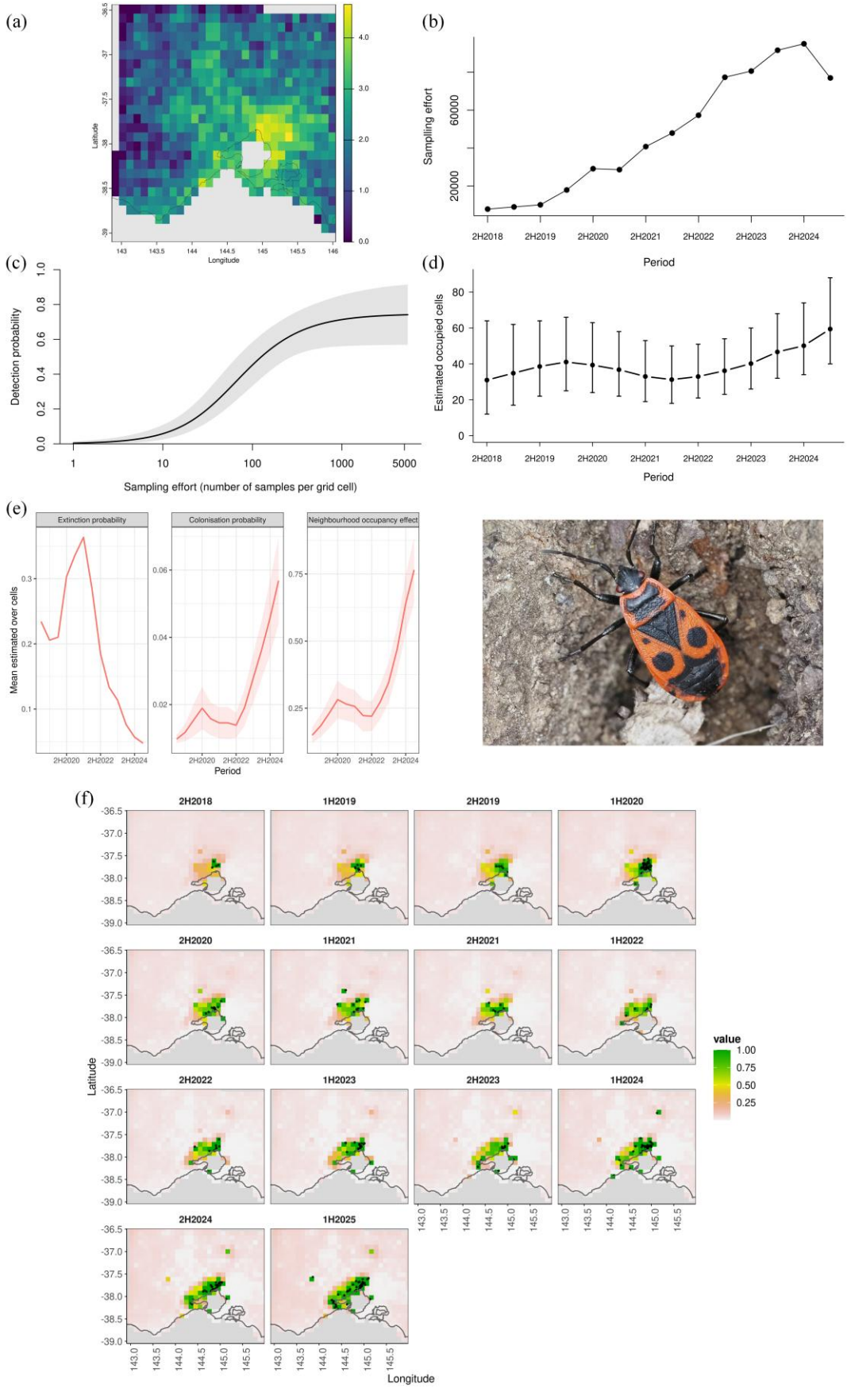
$$\text{SR@k} = \frac{\#\{i \in \mathcal{P} \mid \text{rank}(\hat{y}_{it}) \le k\%\}}{N_+}, \quad (17)$$

Here, $\text{rank}(\hat{y}_{it})$ is the rank position of the cell $i$ when all cells are sorted in descending order of $\hat{y}_{it}$ and $\#\{.\}$ denotes the number of elements satisfying the condition.

Larger SR@5 values indicate stronger alignment between the model's estimated presence surface and the spatial distribution of observed presences. That is, the model assigns higher probabilities to cells where presences occur. For example, SR@5%=0.8 means that 80% of observed presences fall within the top 5% of cells with the highest predicted probability of observing a presence. Metrics of this form have been widely used in species distribution models and DOMs to assess whether models successfully concentrate high-probability predictions in the right places (Fleishman et al., 2003; Southwell et al., 2024; Tozer et al., 2020). We also assessed goodness-of-fit of DODM using posterior predictive checks, as described in Appendix A.3.

# 3. Results

## 3.1. Case study



436 **Figure 3. A case study of European firebug incursion in Melbourne, Australia.** (a) Total
437 sampling effort (total number of insect records log10 transformed) in each 10 x 10 km cell
438 centred around Melbourne recorded in iNaturalist from July 2018 to July 2025; (b) Variation
439 in sampling effort over time during the same period in 6 monthly intervals (1H: January-June;
440 2H: July-December); (c) Estimated detection function for European firebugs. The x-axis is the
441 sampling effort per grid cell (log-transformed total number of insect records) in each 6-month
442 period; (d) Change in the total number of occupied cells estimated from the model fit. Points
443 are the mean of the posterior distribution and bars show the 95% credible intervals (CRIs); (e)
444 Time-varying changes of model parameters (extinction probability $\varepsilon_{it}$, colonisation probability
445 $\gamma_{it}$, and neighbour occupancy effect $\beta_n\theta_{i,t-1}$) in the best fitting model. Lines are cell-averaged
446 posterior means and ribbons are 95% CRIs; (f) Heat map of estimated occupancy probability
447 during each period. Black dots show the locations of European firebug records in iNaturalist
448 for each period.

449

450 To illustrate the model outcomes, we present one species-location case study in detail, with
451 details of additional case studies provided in the Supplementary Material. We then compare
452 key model outputs across species and locations. We chose the European firebug incursion in
453 Melbourne, Australia, as our detailed case study (Mata et al., 2022). The European firebug is
454 an invasive insect, native to Europe and western Asia (Käfer *et al.*, 2023). It is a Palearctic
455 seed-feeding true bug that commonly forms conspicuous aggregations around host plants,
456 which are widespread in urban environments and considered a nuisance pest (Mata et al., 2022;
457 Rojas & Jackson, 2018). Outside its native range, it has established in North America and was
458 recently detected in Melbourne, Australia, where it subsequently spread quickly (Mata et al.,
459 2022). Longer-range spread is likely helped by accidental human transport, such as moving
460 with ornamental plants, soil, or leaf litter (Gippet et al., 2019; Mata et al., 2022).

461

462 The European firebug was first recorded in Melbourne's western suburbs as early as April 2018
463 by the Australian Government's Border Surveillance Group, whereas the earliest iNaturalist
464 record was made on December 2018 in Sunshine, City of Brimbank (Mata et al., 2022). Over
465 the study period from July 2018 to July 2025, there were a total of 325 records within the study
466 area centred on the first iNaturalist observation. During the study period, a total of 595,137
467 other insect observations were recorded in iNaturalist within the study area. Sampling effort
468 was highly heterogeneous across both space and time. Spatially, effort (measured as log10
469 transformed insect records per grid cell) was concentrated in the urban core, with much lower
470 coverage toward the location fringe and in non-residential areas (Figure 3a). Because record
471 generation reflects where people live and choose to search, the raw spatial pattern of European
472 firebug records is likely to be influenced by variation in sampling effort as well as occupancy.
473 The detection function of the DODM accounts for this variation using all insect records within
474 each grid cell and period as a proxy for realised observer activity. Temporally, sampling effort
475 was low and almost flat from 2018 to 2019, increased sharply from 2020, peaked in the second
476 half of 2024, and declined thereafter (Figure 3b).

477

Of the two models, DODM1 provided the best fit to the European firebug data (Table 2), indicating that spread across neighbouring cells was occurring. This interpretation is reinforced by the strongly positive neighbourhood spread parameter in DODM1, $\beta_n = 7.92$ (95% CRIs: 6.04-9.79). The SR@5 value was 1 (Table 2), meaning the model perfectly concentrated its highest predicted probabilities in the locations where firebugs were observed.

Posterior predictive checks (PCCs) further supported the goodness-of-fit of DODM1, providing an assessment of whether the fitted model reproduced the observed patterns of spread. The conditional PCCs for the detection function produced a Freeman–Tukey Bayesian posterior predictive value (BPV) of 0.54, indicating no strong evidence of systematic lack of fit. Marginal PCCs showed close agreement between observed and replicated spread summaries, including mean first detection period (BPV=0.424) and turnover between consecutive periods (BPV=0.345). These results indicate that DODM1 adequately captured both detection patterns and the temporal dynamics of spread. Full details of these PCCs are provided in Appendix A.3.

The estimated detection function (Figure 3c) increased with sampling effort (log-transformed) but reached an asymptote at around 800 records per grid cell per period, implying little to no further gain in detection probability beyond this level of effort. Notably, the function plateaued at a detection probability of ~0.7, implying that even under high sampling effort, European firebugs would be detected only about 70% of the time when they occupied a grid cell.

Inferred occupancy shows a gradual expansion of the European firebug in Melbourne from July 2018 to July 2025. The estimated number of occupied cells remains essentially constant between 2H2018 (the second half of 2018) and 1H2019 (the first half of 2019), then increases from 2H2019 onward, reaching its highest level in 1H2025, where it occupied 58 (95% CRIs: 40-88) grid cells representing 9.1% of the total study area (Figure 3d).

The time-varying colonisation and extinction probabilities from DODM1 (Figure 3e) help explain these dynamics. Extinction probability is relatively high during the early and middle of the study period before declining substantially. This pattern is consistent with colonisation of sites by small transient populations prone to stochastic extinction during the early stages of invasion. As a core population establishes and grows, populations arising from subsequent neighbourhood spread are more likely to persist due to spillover and rescue effects from neighbouring grid cells. In line with this, both colonisation probability and the neighbour occupancy effect increase markedly from the middle of the study period onward. Consequently, local extinction becomes less frequent while colonisation, particularly colonisation from occupied neighbouring cells, becomes increasingly prominent, driving overall occupancy upward. Consistent with this interpretation, the posterior occupancy maps (Figure 3f) show initial high-probability hotspots in the city centre that progressively expand into adjacent cells and eventually form a south-east oriented high-occupancy belt, suggesting that European

519 firebugs spread primarily through short-distance expansion rather than long-distance jumps or
520 new introductions.

521

522 **3.2. Comparative analyses of model performance across species and locations**

523 **Table 2. Performance comparison of candidate models.** Better fitting models have lower
524 values of log score and higher values of SR@5. For each species-location combination and for
525 each measure of model fit, the better performing model is highlighted in bold.

| | | Log score (↓) | | SR@5 (↑) | |
|---|---|---|---|---|---|
| Species | Location | DODM1 | DODM2 | DODM1 | DODM2 |
| Hemlock woolly adelgid (*Adelges tsugae*) | Boston | **0.82** | 0.88 | **0.32** | 0.32 |
| | Nova Scotia | **0.53** | 0.61 | **0.55** | 0.55 |
| Emerald ash borer (*Agrilus planipennis*) | Dallas | **1.43** | 1.45 | **0.91** | 0.91 |
| | Toronto | **1.20** | 1.32 | **0.64** | 0.60 |
| Mediterranean fruit fly (*Ceratitis capitata*) | Lisboa | **1.20** | 1.53 | **0.76** | 0.55 |
| | Los Angeles | **0.48** | 0.69 | **0.53** | 0.40 |
| Box tree moth (*Cydalima perspectalis*) | Paris | 1.56 | **1.53** | 0.62 | **0.64** |
| | Toronto | **0.69** | 0.76 | **0.84** | 0.84 |
| Brown marmorated stink bug (*Halyomorpha halys*) | Los Angeles | **0.48** | 0.71 | **0.46** | 0.37 |
| | Milan | 0.91 | **0.79** | 0.22 | **0.23** |
| African big-headed ant (*Pheidole megacephala*) | Brisbane | **0.83** | 0.91 | **0.93** | 0.92 |
| | Miami | **1.31** | 1.34 | **0.55** | 0.51 |
| European firebug (*Pyrrhocoris apterus*) | Melbourne | **0.75** | 1.10 | **1.00** | 0.99 |
| | Toronto | **0.31** | 0.44 | **0.98** | 0.98 |
| Red imported fire ant (*Solenopsis invicta*) | Brisbane | **0.58** | 0.73 | **0.96** | 0.94 |
| | Los Angeles | **1.24** | 1.49 | **0.77** | 0.64 |

526

527 For each species-location combination, we fitted two models and identified the best fitting
528 model based on the two evaluation criteria (Table 2). Across the 16 species-location
529 combinations, DODM1 provided the best fit (lower LS and higher SR@5) in 14 cases (87.5%),
530 while DODM2 provided a better fit in one case (Brown marmorated stink bug in Milan). For
531 the Box tree moth in Paris, the two models had similar performance (identical SR@5 and only
532 small differences in LS). Overall, these results reveal that explicitly modelling neighbourhood
533 occupancy generally improves model fit, increasing the probability assigned to observed
534 presences (lower LS) and generating more concentrated spatial rankings (higher SR@5).
535 However, the magnitude of this improvement varied across species and locations. The one case
536 where DODM1 did not outperform DODM2 (Brown marmorated stink bug in Milan) suggests
537 that, in this case, dispersal was not strongly driven by neighbourhood occupancy.

538

539 The models generally performed well in the spatial ranking of the probability of observing
540 presences. For most species, the best fit model achieved SR@5 values exceeding 0.5, meaning
541 that more than 50% of observed presence cells fell within the top 5% of estimated high
542 detection probability cells. Nevertheless, the low performance in some cases (SR@5 < 0.5)

indicates that the high detection area is more spatially diffuse, with presences scattered across many cells rather than concentrated in a few hotspots (e.g., Brown marmorated stink bug).

### 3.3. Estimated detection function across species

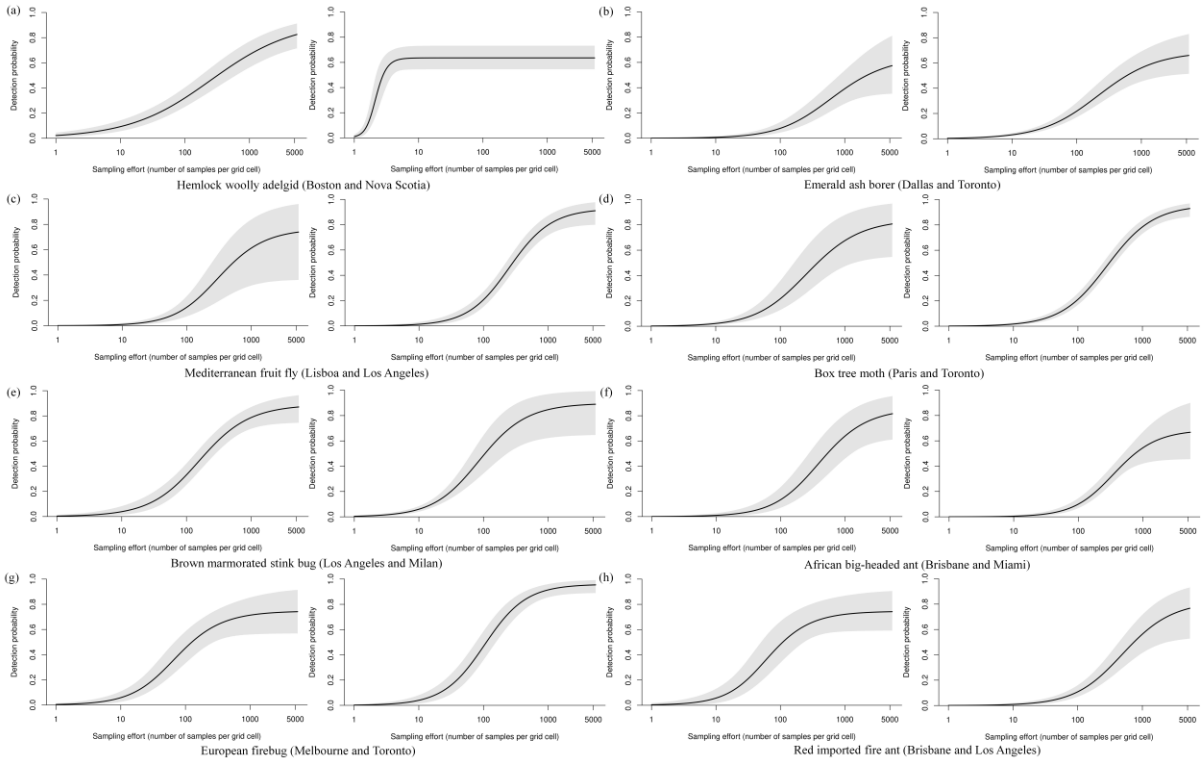



**Figure 2. Estimated detection functions for each species-location combination under the best model.** Lines represent the posterior mean detection probability with 95% CRIs.

Detection probability increased with sampling effort for all species, but both the horizontal position of the curves (the effort required to reach a given detection probability) and their slopes (effort sensitivity) varied among locations (shown in Figure 2). For the same species, detection probability curves were often displaced along the effort axis. For example, the Emerald ash borer in Toronto becomes moderately detectable (detection probability>0.4) at effort between 100 and 1000 records, whereas the Dallas curve increases more slowly and only approaches similar probabilities at much greater effort (>1000 records). Hemlock woolly adelgid shows a similar pattern, with Nova Scotia reaching high detection probabilities at lower effort than Boston. In Brisbane, the Red imported fire ant, during an active eradication program, attains moderate detection (detection probability between 0.4 and 0.6) at intermediate effort (100-1000 records), while Los Angeles requires substantially higher effort to achieve comparable probabilities. These differences indicate that even for the same target species, local citizen science efforts can vary widely in the sampling effort required to achieve a given detection probability.

Within the same location, detection functions also differ among species. In Toronto, the European firebug becomes moderately detectable at relatively low effort (<100 records), whereas the Box tree moth requires higher effort before detection probability rises. In Los Angeles, the detection probabilities of Mediterranean fruit fly, Brown marmorated stink bugs, and Red imported fire ant differ at a sampling effort of 100 records. These patterns are consistent with the idea that species-specific traits, such as visibility, host aggregation, and ease of identification, may affect detectability under otherwise similar urban and observer conditions, making some species easier to detect than others.

### 3.4. Comparison of key parameters across species-location combinations

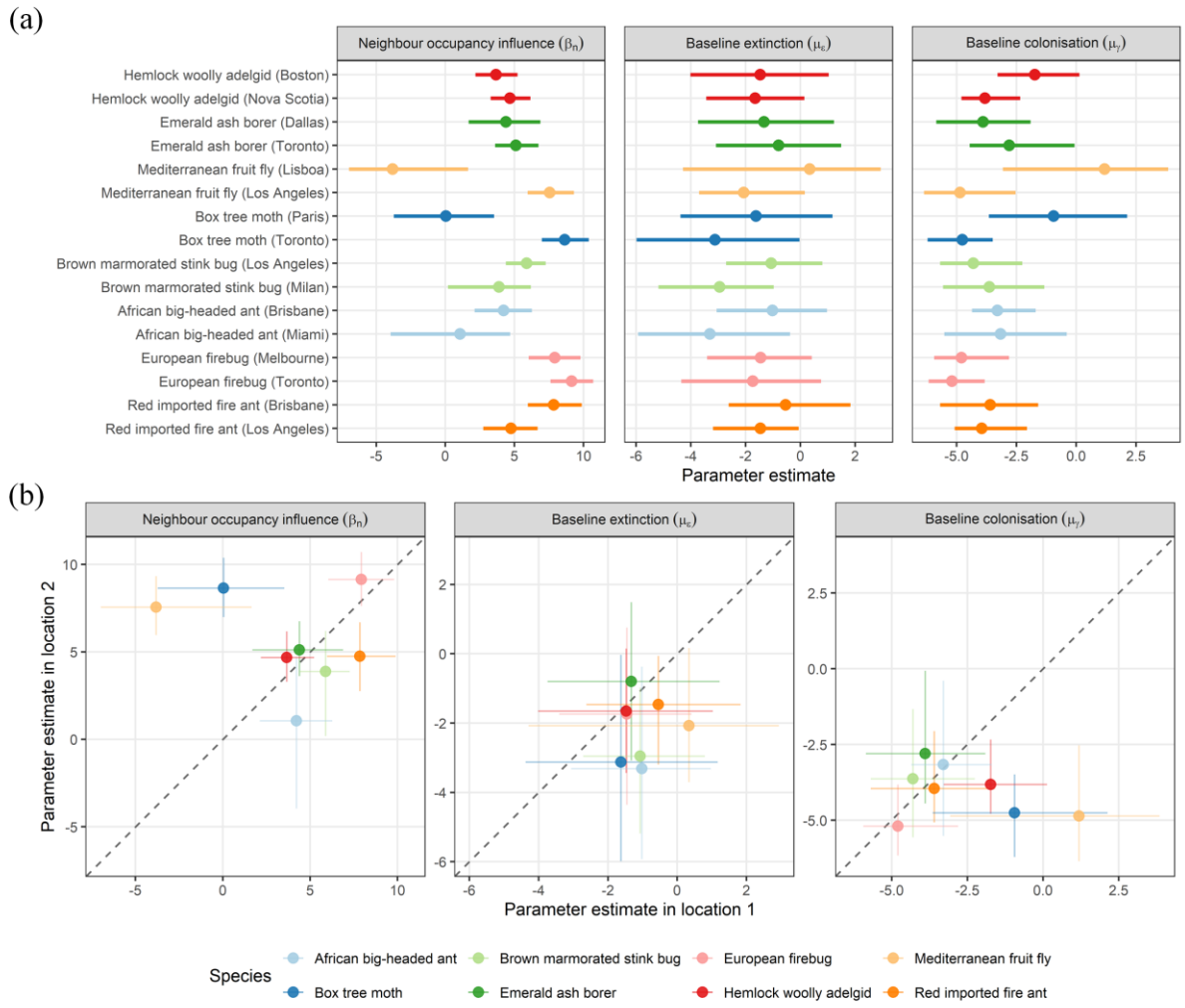



**Figure 4 Posterior parameter estimates for DODM1 across species-location combinations.** (a) Forest plot showing individual parameter estimates for each species-location combination. Points indicate medians, and horizontal lines indicate 95% CRIs. (b) Paired scatter plots comparing parameter estimates between two locations for the same species. The dashed line represents the x=y identity line.

583 Posterior estimates from the DODM1 show that the neighbour occupancy parameter $\beta_n$ is
584 mostly positive across all species-location combinations, and for most species, the 95% CRIs
585 lie well above zero (Figure 4a). This result reinforces local spatial dependency in spread
586 dynamics: colonisation probability is typically higher when neighbouring grid cells are
587 occupied. Neighbour-occupancy effects are particularly strong for Hemlock woolly adelgid,
588 Emerald ash borer, European firebug, and Red imported fire ant, all of which have higher $\beta_n$
589 estimates. In contrast, the Mediterranean fruit fly in Lisboa has a weaker neighbour-occupancy
590 effect with CRIs extending towards zero. Overall, $\beta_n$ estimates were similar between paired
591 locations for 5 of the 8 species (Figure 4b, 95% CRIs that overlap with the 1:1 identity line
592 indicates the estimated parameters for the same species are similar across different geographic
593 locations). This provides some evidence that patterns of neighbourhood spread may be
594 somewhat predictable and likely a consequence of species-specific dispersal processes.

595

596 Baseline extinction and colonisation ($\mu_\varepsilon$ and $\mu_\gamma$), reflecting the initial invasion stages, varied
597 among species and locations (Figure 4a). The baseline extinction $\mu_\varepsilon$ ranged from
598 approximately -4 to +2, with most cases showing values less than 0 or close to 0. For example,
599 the $\mu_\varepsilon$ values for the Brown marmorated stink bug in Milan and Los Angeles were negative. In
600 contrast, Emerald ash borer in Toronto, and Mediterranean fruit fly in Lisboa have $\mu_\varepsilon$ closer to
601 zero or slightly positive, indicating higher baseline extinction probabilities. This result is
602 consistent with the relatively low extinction probability and more persistent neighbour
603 occupation in the initial stages of species invasion. Baseline colonisation $\mu_\gamma$ is negative for
604 most species-location combinations (between -5 and 0), indicating that colonisation of
605 previously unoccupied cells is generally infrequent in the absence of occupied neighbours, and
606 is more likely when occupied neighbours are present. Only a few cases, notably Mediterranean
607 fruit fly in Lisboa and Box tree moth in Paris, show $\mu_\gamma$ estimates near or above zero, suggesting
608 a higher background colonisation probability for these species. The comparison across location
609 pairs (Figure 4b) shows that the location-specific parameter estimates tended to fall within
610 broadly similar ranges across urban environments, while still showing species- and location-
611 specific variation.

612

## 4. Discussion

Dynamic occupancy models and species distribution models are widely used to study species distribution and occupancy and to estimate future changes. However, with the growth of citizen science data, methods that can directly accommodate presence-only data remain limited. We developed a dynamic occupancy-detection model that uses background records of the same taxonomic class species as a proxy for sampling effort in the detection function, and that incorporates a neighbour-occupancy effect in the colonisation process. Applying this model to sixteen species-location invasion case studies, we find that explicitly modelling neighbour-occupancy generally provides a closer match to the observed occupancy dynamics and improves inference of species temporal trends.

The presence of neighbour-occupancy effects reinforces the observation that many invasive insects expand primarily through local dispersal, with colonisation probabilities increasing in adjacent cells of an already occupied area (Bled et al., 2011; Broms, Hooten, Johnson, et al., 2016). However, cases where the neighbourhood-occupancy effect provided limited improvement (e.g., Emerald ash borer) suggest that local occupancy processes do not drive all invasive insect spread. The contrasting neighbour effects and baseline colonisation estimates for Mediterranean fruit fly and Box tree moth between locations further suggest that the relative importance of local and non-local colonisation may vary geographically. For taxa dominated by skip-diffusion or anthropogenic dispersal, first-order neighbourhood terms may fail to capture spread (Briscoe et al., 2019; Hulme, 2009). Future models could incorporate explicit long-distance terms, such as distance-weighted kernels or transport network covariates, to improve reconstruction for non-locally spreading species (Paini & Yemshanov, 2012; Suarez et al., 2001).

The primary challenge of PO data is the ambiguity of "no record", which may signify either genuine absence or non-detection (Altwegg & Nichols, 2019; Guillera-Arroita, 2017). We address this by adopting a target-group background approach, modelling detection probability as a function of sampling effort (Barber et al., 2022; Phillips et al., 2009). Our model ensures that "no record" provides stronger evidence against occupancy when estimated sampling effort is high, while maintaining higher uncertainty in low-effort regions to prevent citizen science activity fluctuations from being misinterpreted as species expansion or decline (Altwegg & Nichols, 2019; Isaac et al., 2014). Furthermore, our estimated detection function follows a saturated growth curve: initial increases in effort significantly boost detectability, whereas additional sampling in already well-surveyed areas yields diminishing marginal returns (Callaghan et al., 2022). Explicitly modelling these detection biases is essential to distinguish observational processes from true ecological dynamics (Dorazio, 2014; Erickson & Smith, 2021; Mengersen et al., 2017).

Detection probabilities in citizen science are inherently heterogeneous across species and locations. In addition to variation in observer activity, this heterogeneity may reflect species-specific differences in conspicuousness, identifiability, and seasonal availability, because taxa with distinctive morphological features or longer observable periods may be more likely to be

detected and reported by citizen scientists (Caley et al., 2020; Kral-O'Brien et al., 2020; van Tongeren et al., 2023). Even within the same location, the detection-effort relationship varied significantly, likely depending on location accessibility, observer behaviour and knowledge, and species-specific traits (Isaac et al., 2014; Isaac & Pocock, 2015; Johnston et al., 2018; Kosmala et al., 2016). This spatial and taxonomic heterogeneity requires the use of explicit effort models, such as accessibility-based predictors or target-group records, to accurately represent varying detection capabilities (Barber et al., 2022; Fernández & Nakamura, 2015; Phillips et al., 2009). Consequently, detection parameters inferred from one location cannot be directly transferred to others in most cases. Without accounting for these effort-dependent dynamics, PO models risk overemphasizing high-effort regions while underestimating low-effort ones, leading to biased occupancy maps and trend inferences (Altwegg & Nichols, 2019; Coomber et al., 2021; van Strien et al., 2013). A focus for future research is to investigate how these detection-effort relationships vary across diverse taxa and geographic contexts.

The choice of background group used to represent sampling effort is an important consideration when applying DODMs to PO records, because this covariate enters the detection component and determines how strongly no records are interpreted as non-detections rather than absences (Altwegg & Nichols, 2019; Guillera-Arroita, 2017). Background records should therefore approximate the observation process affecting the target species, following target-group background approaches that use taxa expected to exhibit sampling biases similar to those of the focal species (Barber et al., 2022; Fithian et al., 2015; Mateo et al., 2010; Phillips et al., 2009). However, choosing the right background record requires a balance. A narrowly defined background (e.g., similar species or same genus) may better match taxon-specific observer behaviour, but sparse records can reduce the stability of effort and model estimates. Conversely, a broad background provides greater coverage but may include records generated by different observer communities, habitats, or search behaviour, weakening its value as a bias-correction covariate (Barber et al., 2022; Fithian et al., 2015). In Melbourne (*P. apterus*) case study, all insect records were strongly correlated with family-level Pyrrhocoridae records over time, suggesting that the broader insect background captured a similar temporal pattern of sampling effort while being less sparse (Appendix A.2). Although our analysis of the Melbourne (*P. apterus*) case study suggests that all insect records provided a suitable effort proxy in this instance, future work could further evaluate the influence of alternative background definitions, such as order, family, or genus records, influence estimated detection functions and occupancy dynamics in DODMs.

Early detection and rapid response are critical for mitigating the ecological and economic impacts of invasive species (Crowl et al., 2008; Jardine & Sanchirico, 2018; Reaser et al., 2020). While structured surveys remain the gold standard, their high costs at broad spatiotemporal scales necessitate complementary monitoring strategies (Holden et al., 2016). Our findings demonstrate that PO data, when integrated with explicit observational models, may bolster early detection and rapid response systems by reconstructing dynamic occupancy surfaces, quantifying uncertainty, and identifying invasion frontiers and hotspots. This is particularly important in urban environments, which serve as primary hubs for non-native insect introduction due to high introduction pressure and links to transport networks (Branco et al.,

2019; Gaertner et al., 2017). Consequently, a strategic priority for citizen science should be the prioritization of sampling in high-value, under-sampled locations. By identifying areas with high presence probability but insufficient effort, our model provides a quantitative framework to guide citizen science activities toward the most informative locations. Optimizing such engagement will further require social science insights to develop effective participant interaction strategies (Gibson et al., 2024; Hecker & Taddicken, 2022).

To further enhance the precision and reliability of occupancy dynamic reconstructions, future research should transition toward integrating multi-source metadata into detection probability estimation. Incorporating observer activity metrics, checklist data, and species morphological traits may enhance robustness to interspecific behavioural heterogeneity and variation in detectability (Callaghan et al., 2022; Isaac & Pocock, 2015; Oliveira et al., 2024; Phillips et al., 2009). Furthermore, achieving scalable inference is essential for fully exploiting large-scale citizen science data. While Markov chain Monte Carlo provides coherent uncertainty propagation, its computational cost scales poorly with increased spatial resolution and study size. Future work should explore more scalable alternatives (e.g., approximate Bayesian inference) to maintain efficiency without sacrificing uncertainty quantification (Fajgenblat et al., 2025b; Fletcher Jr. et al., 2019). Finally, while focused on invasive species, our proposed model is broadly applicable to occupancy dynamic inference for native and threatened taxa, providing a robust tool for monitoring biodiversity as opportunistic citizen science data coverage expands globally.

## References

Altwegg, R., & Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, *10*(1), 8–21. https://doi.org/10.1111/2041-210X.13090

Baker, C. M., & Bode, M. (2016). Placing invasive species management in a spatiotemporal context. *Ecological Applications*, *26*(3), 712–725. https://doi.org/10.1890/15-0095

Barber, R. A., Ball, S. G., Morris, R. K. A., & Gilbert, F. (2022). Target-group backgrounds prove effective at correcting sampling bias in Maxent models. *Diversity and Distributions*, *28*(1), 128–141. https://doi.org/10.1111/ddi.13442

Bled, F., Nichols, J. D., & Altwegg, R. (2013). Dynamic occupancy models for analyzing species' range dynamics across large geographic scales. *Ecology and Evolution*, *3*(15), 4896–4909. https://doi.org/10.1002/ece3.858

Bled, F., Royle, J. A., & Cam, E. (2011). Hierarchical modeling of an invasive spread: The Eurasian Collared-Dove Streptopelia decaocto in the United States. *Ecological Applications*, *21*(1), 290–302. https://doi.org/10.1890/09-1877.1

Boivin, N. L., Zeder, M. A., Fuller, D. Q., Crowther, A., Larson, G., Erlandson, J. M., Denham, T., & Petraglia, M. D. (2016). Ecological consequences of human niche construction: Examining long-term anthropogenic shaping of global species distributions. *Proceedings of the National Academy of Sciences*, *113*(23), 6388–6396. https://doi.org/10.1073/pnas.1525200113

Branco, M., Nunes, P., Roques, A., Fernandes, M. R., Orazio, C., & Jactel, H. (2019). Urban trees facilitate the establishment of non-native forest insects. *NeoBiota*, *52*, 25–46. https://doi.org/10.3897/neobiota.52.36358

Briscoe, N. J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J. J., Camac, J. S., Giljohann, K. M., Holden, M. H., Hradsky, B. A., Kearney, M. R., McMahon, S. M., Phillips, B. L., Regan, T. J., Rhodes, J. R., Vesk, P. A., Wintle, B. A., Yen, J. D. L., & Guillera-Arroita, G. (2019). Forecasting species range dynamics with process-explicit models: Matching methods to applications. *Ecology Letters*, *22*(11), 1940–1956. https://doi.org/10.1111/ele.13348

Brizuela-Torres, D., Elith, J., Guillera-Arroita, G., & Briscoe, N. J. (2024). Dealing with sampling bias and inferring absence data to improve distribution models of a widely distributed vulnerable marsupial. *Austral Ecology*, *49*(1), e13474. https://doi.org/10.1111/aec.13474

Broms, K. M., Hooten, M. B., & Fitzpatrick, R. M. (2016). Model selection and assessment for multi-species occupancy models. *Ecology*, *97*(7), 1759–1770. https://doi.org/10.1890/15-1471.1

Broms, K. M., Hooten, M. B., Johnson, D. S., Altwegg, R., & Conquest, L. L. (2016). Dynamic occupancy models for explicit colonization processes. *Ecology*, *97*(1), 194–204. https://doi.org/10.1890/15-0416.1

Buchadas, A., Vaz, A. S., Honrado, J. P., Alagador, D., Bastos, R., Cabral, J. A., Santos, M., & Vicente, J. R. (2017). Dynamic models in research and management of biological invasions.

*Journal of Environmental Management*, *196*, 594–606.
https://doi.org/10.1016/j.jenvman.2017.03.060

Caley, P., Welvaert, M., & Barry, S. C. (2020). Crowd surveillance: Estimating citizen science reporting probabilities for insects of biosecurity concern. *Journal of Pest Science*, *93*(1), 543–550. https://doi.org/10.1007/s10340-019-01115-7

Callaghan, C. T., Bowler, D. E., Blowes, S. A., Chase, J. M., Lyons, M. B., & Pereira, H. M. (2022). Quantifying effort needed to estimate species diversity from citizen science data. *Ecosphere*, *13*(4), e3966. https://doi.org/10.1002/ecs2.3966

Coomber, F. G., Smith, B. R., August, T. A., Harrower, C. A., Powney, G. D., & Mathews, F. (2021). Using biological records to infer long-term occupancy trends of mammals in the UK. *Biological Conservation*, *264*, 109362. https://doi.org/10.1016/j.biocon.2021.109362

Crowl, T. A., Crist, T. O., Parmenter, R. R., Belovsky, G., & Lugo, A. E. (2008). The spread of invasive species and infectious disease as drivers of ecosystem change. *Frontiers in Ecology and the Environment*, *6*(5), 238–246. https://doi.org/10.1890/070151

dbeadle. (2018, August 28). *iNaturalist observation*. iNaturalist. iNaturalist. https://www.inaturalist.org/observations/15978957

de Groot, M., Ogris, N., van der Meij, M., & Pocock, M. J. O. (2022). Where to search: The use of opportunistic data for the detection of an invasive forest pest. *Biological Invasions*, *24*(11), 3523–3537. https://doi.org/10.1007/s10530-022-02857-9

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413. https://doi.org/10.1080/10618600.2016.1172487

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484. https://doi.org/10.1111/geb.12216

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*(9), 1341–1352. https://doi.org/10.1111/ecog.05679

Fajgenblat, M., Wijns, R., De Knijf, G., Stoks, R., Lemmens, P., Herremans, M., Vanormelingen, P., Neyens, T., & De Meester, L. (2025a). Leveraging Massive Opportunistically Collected Datasets to Study Species Communities in Space and Time. *Ecology Letters*, *28*(3), e70094. https://doi.org/10.1111/ele.70094

Fajgenblat, M., Wijns, R., De Knijf, G., Stoks, R., Lemmens, P., Herremans, M., Vanormelingen, P., Neyens, T., & De Meester, L. (2025b). Leveraging Massive Opportunistically Collected Datasets to Study Species Communities in Space and Time. *Ecology Letters*, *28*(3), e70094. https://doi.org/10.1111/ele.70094

799 Faurby, S., & Svenning, J.-C. (2015). Historic and prehistoric human-driven extinctions have
800 reshaped global mammal diversity patterns. *Diversity and Distributions*, *21*(10), 1155–1166.
801 https://doi.org/10.1111/ddi.12369

802 Fernández, D., & Nakamura, M. (2015). Estimation of spatial sampling effort based on
803 presence-only data and accessibility. *Ecological Modelling*, *299*, 147–155.
804 https://doi.org/10.1016/j.ecolmodel.2014.12.017

805 Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution
806 models: Pooling survey and collection data for multiple species. *Methods in Ecology and*
807 *Evolution*, *6*(4), 424–438. https://doi.org/10.1111/2041-210X.12242

808 Fleishman, E., Nally, R. M., & Fay, J. P. (2003). Validation Tests of Predictive Models of
809 Butterfly Occurrence Based on Environmental Variables. *Conservation Biology*, *17*(3), 806–
810 817. https://doi.org/10.1046/j.1523-1739.2003.02113.x

811 Fletcher Jr., R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio,
812 R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*,
813 *100*(6), e02710. https://doi.org/10.1002/ecy.2710

814 Gaertner, M., Wilson, J. R. U., Cadotte, M. W., MacIvor, J. S., Zenni, R. D., & Richardson,
815 D. M. (2017). Non-native species in urban environments: Patterns, processes, impacts and
816 challenges. *Biological Invasions*, *19*(12), 3461–3469. https://doi.org/10.1007/s10530-017-
817 1598-7

818 Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple
819 Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

820 Gibson, D. J., Benda, C. D., Felsl, I., & Kiefer, G. (2024). Challenges in the establishment of
821 a rare plant species monitoring program using community science volunteers. *Ecological*
822 *Solutions and Evidence*, *5*(4), e12400. https://doi.org/10.1002/2688-8319.12400

823 Gippet, J. M., Liebhold, A. M., Fenn-Moltu, G., & Bertelsmeier, C. (2019). Human-mediated
824 dispersal in insects. *Current Opinion in Insect Science, Global Change Biology • Molecular*
825 *Physiology*, *35*, 96–102. https://doi.org/10.1016/j.cois.2019.07.005

826 Grattarola, F., Bowler, D. E., & Keil, P. (2023). Integrating presence-only and presence–
827 absence data to model changes in species geographic ranges: An example in the Neotropics.
828 *Journal of Biogeography*, *50*(9), 1561–1575. https://doi.org/10.1111/jbi.14622

829 Grattarola, F., Tschernosterová, K., & Keil, P. (2024). A continental-wide decline of
830 occupancy and diversity in five Neotropical carnivores. *Global Ecology and Conservation*,
831 *55*, e03226. https://doi.org/10.1016/j.gecco.2024.e03226

832 Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and
833 communities under imperfect detection: Advances, challenges and opportunities. *Ecography*,
834 *40*(2), 281–295. https://doi.org/10.1111/ecog.02445

835 Harsi S, P. (2023, January 6). *iNaturalist observation*.
836 https://www.inaturalist.org/observations/184782673

Hecker, S., & Taddicken, M. (2022). Deconstructing citizen science: A framework on communication and interaction using the concept of roles. *Journal of Science Communication*, *21*(1), A07. https://doi.org/10.22323/2.21010207

Holden, M. H., Nyrop, J. P., & Ellner, S. P. (2016). The economic benefit of time-varying surveillance effort for invasive species management. *Journal of Applied Ecology*, *53*(3), 712–721. https://doi.org/10.1111/1365-2664.12617

Hulbert, J. M., Hallett, R. A., Roy, H. E., & Cleary, M. (2023). Citizen science can enhance strategies to detect and manage invasive forest pests and pathogens. *Frontiers in Ecology and Evolution*, *11*. https://doi.org/10.3389/fevo.2023.1113978

Hulme, P. E. (2009). Trade, transport and trouble: Managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, *46*(1), 10–18. https://doi.org/10.1111/j.1365-2664.2008.01600.x

ildikorab. (2024, May 28). *iNaturalist observation*. iNaturalist. iNaturalist. https://www.inaturalist.org/observations/218970041

iNaturalist community. (2025). *Observations of Adelges tsugae, Agrilus planipennis, Ceratitis capitata, Cydalima perspectalis, Halyomorpha halys, Pheidole megacephala, Pyrrhocoris apterus, and Solenopsis invicta from selected locations in Australia, Canada, France, Italy, Portugal, and the United States, observed between July 2018 and July 2025* [Dataset]. iNaturalist. https://www.inaturalist.org

Isaac, N. J. B., & Pocock, M. J. O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, *115*(3), 522–531. https://doi.org/10.1111/bij.12532

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*(10), 1052–1060. https://doi.org/10.1111/2041-210X.12254

Jardine, S. L., & Sanchirico, J. N. (2018). Estimating the cost of invasive species control. *Journal of Environmental Economics and Management*, *87*, 242–257. https://doi.org/10.1016/j.jeem.2017.07.004

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*(1), 88–97. https://doi.org/10.1111/2041-210X.12838

Käfer, H., Kovac, H., & Stabentheiner, A. (2023). Habitat Temperatures of the Red Firebug, Pyrrhocoris apterus: The Value of Small-Scale Climate Data Measurement. *Insects*, *14*(11), 843. https://doi.org/10.3390/insects14110843

Knapp, M. (2025, February 28). *iNaturalist observation*. iNaturalist. iNaturalist. https://www.inaturalist.org/observations/263424292

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/10.1002/fee.1436

876 Kral-O'Brien, K. C., Karasch, B. M., Hovick, T. J., Moranz, R. A., & Harmon, J. P. (2020).
877 Morphological traits determine detectability bias in North American grassland butterflies.
878 *Ecosphere*, *11*(12), e03304. https://doi.org/10.1002/ecs2.3304

879 Mata, L., Vogel, B., Palma, E., Malipatil, M., & Krüger, S. (2022). The Arrival and Spread of
880 the European Firebug (Pyrrhocoris apterus) in Australia as Documented by Citizen Scientists.
881 *Urban Naturalist Notes*, *N3*, 1–7.

882 Mateo, R. G., Croat, T. B., Felicísimo, Á. M., & Muñoz, J. (2010). Profile or group
883 discriminative techniques? Generating reliable species distribution models using pseudo-
884 absences and target-group absences from natural history collections. *Diversity and*
885 *Distributions*, *16*(1), 84–94. https://doi.org/10.1111/j.1472-4642.2009.00617.x

886 Mengersen, K., Peterson, E. E., Clifford, S., Ye, N., Kim, J., Bednarz, T., Brown, R., James,
887 A., Vercelloni, J., Pearse, A. R., Davis, J., & Hunter, V. (2017). Modelling imperfect presence
888 data obtained by citizen science. *Environmetrics*, *28*(5), e2446.
889 https://doi.org/10.1002/env.2446

890 Miller, R. (2023, June 8). *iNaturalist observation*. iNaturalist. iNaturalist.
891 https://www.inaturalist.org/observations/166169201

892 Molinari-Jobin, A., Kéry, M., Marboutin, E., Marucco, F., Zimmermann, F., Molinari, P.,
893 Frick, H., Fuxjäger, C., Wölfl, S., Bled, F., Breitenmoser-Würsten, C., Kos, I., Wölfl, M.,
894 Černe, R., Müller, O., & Breitenmoser, U. (2018). Mapping range dynamics from
895 opportunistic data: Spatiotemporal modelling of the lynx distribution in the Alps over 21
896 years. *Animal Conservation*, *21*(2), 168–180. https://doi.org/10.1111/acv.12369

897 Oliveira, U., Soares-Filho, B., & Nunes, F. (2024). Controlling the effects of sampling bias in
898 biodiversity models. *Journal of Biogeography*, *51*(9), 1755–1766.
899 https://doi.org/10.1111/jbi.14851

900 Outhwaite, C. L., Chandler, R. E., Powney, G. D., Collen, B., Gregory, R. D., & Isaac, N. J.
901 B. (2018). Prior specification in Bayesian occupancy modelling improves analysis of species
902 occurrence data. *Ecological Indicators*, *93*, 333–343.
903 https://doi.org/10.1016/j.ecolind.2018.05.010

904 Paini, D. R., & Yemshanov, D. (2012). Modelling the Arrival of Invasive Organisms via the
905 International Marine Shipping Network: A Khapra Beetle Study. *PLOS ONE*, *7*(9), e44589.
906 https://doi.org/10.1371/journal.pone.0044589

907 Peach, M. A., Cohen, J. B., & Frair, J. L. (2017). Single-visit dynamic occupancy models: An
908 approach to account for imperfect detection with Atlas data. *Journal of Applied Ecology*,
909 *54*(6), 2033–2042. https://doi.org/10.1111/1365-2664.12925

910 Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S.
911 (2009). Sample selection bias and presence-only distribution models: Implications for
912 background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.
913 https://doi.org/10.1890/07-2153.1

914 Pocock, M. J. O., Adriaens, T., Bertolino, S., Eschen, R., Essl, F., Hulme, P. E., Jeschke, J.
915 M., Roy, H. E., Teixeira, H., & de Groot, M. (2024). Citizen science is a vital partnership for

invasive alien species management and research. *iScience*, *27*(1), 108623. https://doi.org/10.1016/j.isci.2023.108623

Powney, G. D., & Isaac, N. J. B. (2015). Beyond maps: A review of the applications of biological records. *Biological Journal of the Linnean Society*, *115*(3), 532–542. https://doi.org/10.1111/bij.12517

Probert, A. F., Wegmann, D., Volery, L., Adriaens, T., Bakiu, R., Bertolino, S., Essl, F., Gervasini, E., Groom, Q., Latombe, G., Marisavljevic, D., Mumford, J., Pergl, J., Preda, C., Roy, H. E., Scalera, R., Teixeira, H., Tricarico, E., Vanderhoeven, S., & Bacher, S. (2022). Identifying, reducing, and communicating uncertainty in community science: A focus on alien species. *Biological Invasions*, *24*(11), 3395–3421. https://doi.org/10.1007/s10530-022-02858-8

Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-Rodríguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: A conceptual framework and federal capacities assessment. *Biological Invasions*, *22*(1), 1–19. https://doi.org/10.1007/s10530-019-02156-w

Roe, A. D., Greenwood, L. F., & Coyle, D. R. (2024). Catching invasives with curiosity: The importance of passive biosecurity surveillance systems for invasive forest pest detection. *Environmental Entomology*, *53*(6), 881–893. https://doi.org/10.1093/ee/nvae082

Rojas, P. J. O., & Jackson, M. D. (2018). *Pyrrhocoris apterus* L. (Hemiptera: Pyrrhocoridae), a newly introduced family, genus, and species to Ontario and Canada. *The Journal of the Entomological Society of Ontario*, *149*, 27–32. https://doi.org/10.21083/jeso.v149i0.4406

Rushing, C. S., Royle, J. A., Ziolkowski, D. J., & Pardieck, K. L. (2019). Modeling spatially and temporally complex range dynamics when detection is imperfect. *Scientific Reports*, *9*(1), 12805. https://doi.org/10.1038/s41598-019-48851-5

Sadoti, G., Zuckerberg, B., Jarzyna, M. A., & Porter, W. F. (2013). Applying occupancy estimation and modelling to the analysis of atlas data. *Diversity and Distributions*, *19*(7), 804–814. https://doi.org/10.1111/ddi.12041

Schulz, K. (2018, January). *iNaturalist observation*. iNaturalist. iNaturalist. https://www.inaturalist.org/observations/9423340

Sergio, I. (2025, September 9). *iNaturalist observation*. iNaturalist. iNaturalist. https://www.inaturalist.org/observations/312661119

Smith, J. J. (2026, March 29). *iNaturalist observation*. iNaturalist. iNaturalist. https://www.inaturalist.org/observations/346017406

Solà, O., Aquilué, N., Fraixedas, S., & Brotons, L. (2024). *Evaluating the influence of neighborhood connectivity and habitat effects in dynamic occupancy species distribution models*. https://doi.org/10.1111/ecog.06985

Southwell, D. M., Smart, A., Merson, S. D., Selwood, K. E., & Macgregor, N. A. (2024). Using power analysis and spatial prioritization to evaluate the design of a forest bird monitoring programme. *Oryx*, *58*(4), 522–531. https://doi.org/10.1017/S0030605323001382

955 Suarez, A. V., Holway, D. A., & Case, T. J. (2001). Patterns of spread in biological invasions
956 dominated by long-distance jump dispersal: Insights from Argentine ants. *Proceedings of the*
957 *National Academy of Sciences*, *98*(3), 1095–1100. https://doi.org/10.1073/pnas.98.3.1095

958 Takahashi, D., & Park, Y.-S. (2020). Spatial heterogeneities of human-mediated dispersal
959 vectors accelerate the range expansion of invaders with source–destination-mediated
960 dispersal. *Scientific Reports*, *10*(1), 21410. https://doi.org/10.1038/s41598-020-78633-3

961 Tozer, D. C., Stewart, R. L. M., Steele, O., & Gloutney, M. (2020). Species-Habitat
962 Relationships and Priority Areas for Marsh-Breeding Birds in Ontario. *The Journal of*
963 *Wildlife Management*, *84*(4), 786–801. https://doi.org/10.1002/jwmg.21840

964 van Strien, A. J., van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science
965 data of animal species produce reliable estimates of distribution trends if analysed with
966 occupancy models. *Journal of Applied Ecology*, *50*(6), 1450–1458.
967 https://doi.org/10.1111/1365-2664.12158

968 van Tongeren, E., Sistri, G., Bonifacino, M., Menchetti, M., Pasquali, L., Salvati, V., Balletto,
969 E., Bonelli, S., Cini, A., Portera, M., & Dapporto, L. (2023). Unstructured citizen science
970 reduces the perception of butterfly local extinctions: The interplay between species traits and
971 user effort. *Biodiversity and Conservation*, *32*(14), 4701–4718.
972 https://doi.org/10.1007/s10531-023-02721-9

973 van Woensel, E., van der Werf, W., Parnell, S., van Vliet, A., & Douma, B. (J. C. ). (2025).
974 Exploring the contribution of citizen science to statistically sound and risk-based surveillance
975 of insect pests in the EU. *EFSA Supporting Publications*, *22*(2), 9186E.
976 https://doi.org/10.2903/sp.efsa.2025.EN-9186

977

978

# Appendix

## 1. History of target species invasions

**Table A.1. Summary of target species and their incursion histories.**

| Target species | Location | Year of first reported occurrence | Incursion pathway | Reference |
|---|---|---|---|---|
| Hemlock woolly adelgid (*Adelges tsugae*) | Nova Scotia | 2017 | Likely secondary spread via infested hemlock nursery stock, logs, firewood or other wood products; natural dispersal may also contribute. | (Ellison et al., 2018; Roscoe et al., 2025) |
| | Boston | ≈1960s | | |
| Emerald ash borer (*Agrilus planipennis*) | Toronto | 2007 | Human-assisted movement of infested ash material, especially firewood, logs, branches, nursery stock, wood chips and other ash wood products; original North American entry likely via solid wood packaging from Asia; Local spread by adult flight. | (Animal and Plant Health Inspection Service, 2026; Canadian Food Inspection Agency, 2007; Herms & McCullough, 2014; Texas A&M Forest Service, 2026) |
| | Dallas | 2022 | | |
| Mediterranean fruit fly (*Ceratitis capitata*) | Los Angeles | 1975 | Infested fruit pathway, especially larvae in smuggled or illegally imported fruit, fruit cargo, or fruit carried in passenger baggage. | (Carey, 1991; Malacrida et al., 1998; Szyniszewska et al., 2016) |
| | Lisboa | ≈1898 | | |
| Box tree moth (*Cydalima perspectalis*) | Toronto | 2018 | Likely ornamental *Buxus* plant trade / movement of infested box plants within Europe. | (Bras et al., 2022; Canadian Food Inspection Agency, 2019; Jean-François et al., 2009) |
| | Paris | 2008 | | |
| Brown marmorated stink bug (*Halyomorpha halys*) | Los Angeles | 2006 | Hitchhiker pathway via transported goods, especially containers, vehicles or machinery. | (Hodges, 2026; Ingels & Varela, 2014; Lupi et al., 2017; Maistrello et al., 2014; Roscoe et al., 2025) |
| | Milan | 2013 | | |
| African big-headed ant (*Pheidole megacephala*) | Brisbane | ≈1912 | Human commerce / tramp-ant pathway, likely transported in soil, potted plants, nursery material, wood products, packaging, cargo or other goods. | (D. Hoffmann, 1998; Hoffmann & Parr, 2008; Smith, 1933) |
| | Miami | 1933 | | |
| European firebug (*Pyrrhocoris apterus*) | Melbourne | 2018 | Likely human-assisted movement with imported plant material or soil; possible airport-associated introduction route. | (Mata et al., 2022; Rojas & Jackson, 2018) |
| | Toronto | 2017 | | |
| Red imported fire ant (*Solenopsis invicta*) | Brisbane | 2001 | Human-assisted movement of infested soil, nursery potting media, potted plants, sod/turf, hay, machinery or cargo; possible international entry via contaminated containers/cargo. | (Gutrich et al., 2007; Health (PLH) et al., 2023; Nattrass & Vanderwoude, 2001) |
| | Los Angeles | 1998 | | |

## 2. Background sampling efforts

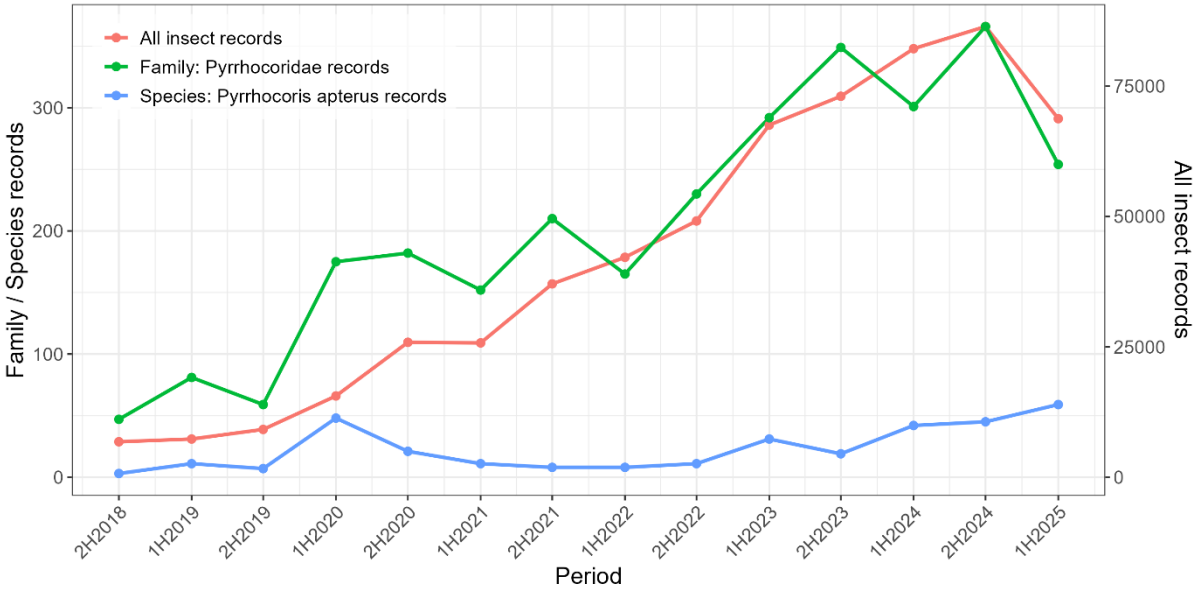



**Figure A.1. Temporal trends in iNaturalist records for species *Pyrrhocoris apterus*, its family Pyrrhocoridae, and all insect records in Melbourne from July 2018 to July 2025.** The left y-axis shows record counts for the species and family, while the right y-axis shows all insect records. The time interval is every 6 months (1H: January-June; 2H: July-December).

**Table A.2. Correlation coefficients comparing temporal trends in record numbers among all insect, family Pyrrhocoridae, and species *Pyrrhocoris apterus* records.**

| Comparison | Pearson correlation | Spearman correlation |
|---|---|---|
| All records vs. Family records | 0.945 | 0.952 |
| All records vs. Species records | 0.580 | 0.597 |
| Family vs. Species records | 0.594 | 0.681 |

According to iNaturalist data, all insect records, records of the family Pyrrhocoridae, and records of the species *Pyrrhocoris apterus* in the Melbourne area all increased between July 2018 and July 2025 (Figure A.1). Correlation analysis showed a strong positive correlation between all insect records and Pyrrhocoridae records, with a Pearson correlation coefficient (PCC) of 0.945 and a Spearman correlation coefficient of 0.952 (Table A.2). Record numbers for both groups peaked in the second half of 2024 (2H2024). By contrast, records of *P. apterus* increased more slowly and showed moderate positive correlations with both all insect records and Pyrrhocoridae records, with PCC values of 0.580 and 0.594, respectively. Although records of *P. apterus* increased briefly in the first half of 2020, sustained growth was not evident until after 2023. These results indicate that the increase in Pyrrhocoridae records was likely driven by the broader expansion of insect observations. However, the distinct temporal pattern observed for *P. apterus* may reflect species-specific population dynamics, differences in detectability, or observer preferences.

In our DODMs, we selected all insect records as the background sampling effort because they provide a more comprehensive estimate of overall observer effort in the Melbourne area, while still closely tracking the temporal pattern observed for Pyrrhocoridae records. The strong correlation between total insect records and Pyrrhocoridae records suggests that these two measures capture similar temporal variation in observer activity. Therefore, using all insect records as the background sampling effort in the detection function is unlikely to change the inferred sampling effort pattern relative to using family records alone.

## 3. Posterior predictive checks

### 3.1. Designing posterior predictive checks for goodness-of-fit

The structure of the dynamic occupancy-detection model (DODM) required posterior predictive checks (PPCs) that distinguish lack of fit in the detection process from lack of fit in the latent occupancy dynamics, thereby providing a flexible framework for assessing DODM goodness-of-fit (Gelman, Meng, and Stern, 1996). In this study, we evaluated DODM1 fitting to European firebug data from Melbourne area as a case study. DODM1 contains binary latent occupancy states, binary detection observations, and imperfect detection. Accordingly, we implemented two PCC approaches. The first approach is conditional PPCs, which condition on the posterior draws of the latent occupancy states and evaluate the detection function alone. The second approach is marginal PCCs, which regenerate occupancy histories using the fitted DODM1 and evaluated whether the fitted DODM1 could reproduce the observed temporal patterns of detection.

### 3.2. Conditional posterior predictive checks: detection function

The first set of PPCs focused on the detection function in DODM1. Posterior Markov chain Monte Carlo draws were indexed by $m = 1, \dots, M$. For each posterior draw, the latent occupancy states were held fixed at their sampled posterior values, $z_{i,t}^{(m)}$, and the detection probability $p_{i,t}^{(m)}$ was calculated from the corresponding posterior draw of the detection parameters and the observed sampling effort. Replicated detection data $y_{i,t}^{rep,(m)}$ were then simulated as

$$y_{i,t}^{rep,(m)} \sim Bernoulli\left(z_{i,t}^{(m)} p_{i,t}^{(m)}\right).$$

By conditioning on latent occupancy states $z_{i,t}^{(m)}$, these checks evaluate whether the detection function can reproduce the replicated detections $y_{i,t}^{rep,(m)}$ given the model's estimate of where the species was present, rather than confounding detection error with errors in the occupancy dynamics.

We used two conditional discrepancy measures. First, we calculated a Freeman–Tukey statistic (Freeman and Tukey, 1950; Brooks, Catchpole, and Morgan, 2000),

$$T_{FT}(y, \mu^{(m)}) = \sum_{i,t} (\sqrt{y_{i,t}} - \sqrt{\mu_{i,t}^{(m)}})^2,$$

where $\mu_{i,t}^{(m)} = z_{i,t}^{(m)} p_{i,t}^{(m)}$ is the expected observation probability. This statistic was summed across all site-period cells and was calculated for both the observations $y_{i,t}$, and the posterior replicated observations $y_{i,t}^{rep,(m)}$, generated from $Bernoulli\left(\mu_{i,t}^{(m)}\right)$.

Second, we compared the observed and replicated numbers of detected sites in each period. Period-specific Bayesian posterior predictive p-values (BPVs) were calculated as the

proportion of posterior draws in which the replicated count for that period was greater than to the observed count.

For each discrepancy measure, we calculated a BPV $p_B$ as the proportion of posterior draws for which the replicated discrepancy $S^{rep,(m)}$ was greater than to the observed discrepancy $S^{obs,(m)}$,

$$p_B = \frac{1}{M}\sum_{m=1}^{M} I(S^{rep,(m)} > S^{obs,(m)}),$$

where $I(\cdot)$ is an indicator function. Values of $p_B$ near 0.5 indicate that the model can reproduce the observed pattern for the statistic being tested, whereas values near 0 or 1 indicate a systematic mismatch. For scalar count-based summaries, the direction of the BPV indicates whether the model tends to over- or under-predict the statistic. For period-specific count checks, BPV was calculated as the proportion of posterior draws in which the replicated count was greater than to the observed count.

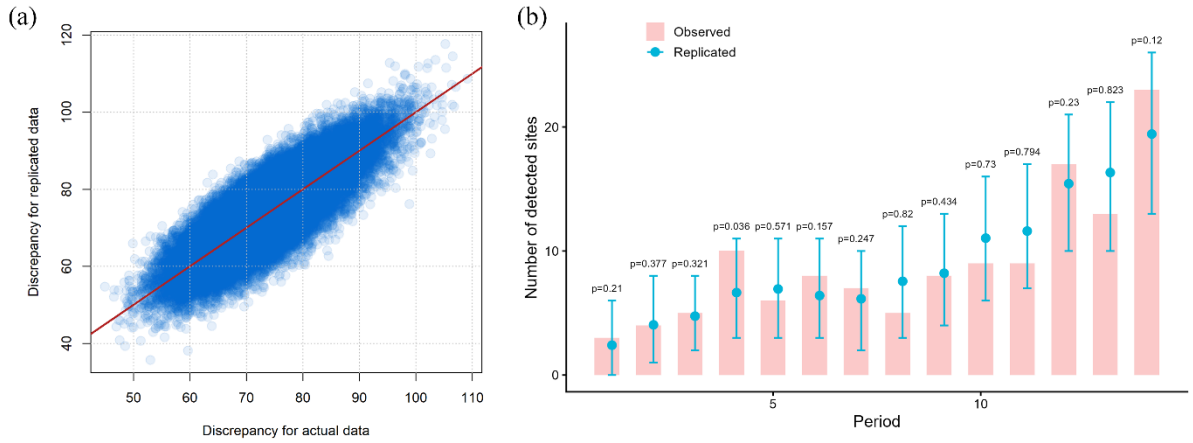



**Figure A.3.1. Conditional posterior predictive checks for the detection function in DODM1.** (a) Scatter plot of Freeman–Tukey discrepancies for observed versus replicated data; the diagonal line indicates the 1:1 reference line. (b) Observed numbers of detected sites per period are shown as bars, and posterior predictive means and 95% intervals are shown as points and error bars, respectively. Individual BPVs are displayed above each period.

Conditional PPC results indicated that the detection function of the DODM1 generally reproduced the observed detection patterns. For the Freeman–Tukey discrepancy, posterior replicated discrepancies were closely aligned with the observed discrepancies and were distributed around the reference line with a conditional BPV of 0.54, suggesting no strong systematic lack of fit in the overall detection process (Figure A.3.1a). This indicates that, conditional on the sampled latent occupancy states, the fitted detection function was able to reproduce the aggregate site-period detection structure well.

The period-specific count checks provided further support for the adequacy of the detection function, although some temporal discrepancies were evident (Figure A.3.1b). For most periods, the observed numbers of detected sites fell within the 95% posterior predictive intervals ($0.025 \leq BPV \leq 0.975$). This suggests that the model generally captured temporal variation in the number of detected sites. Overall, these checks suggest that the detection function provided an adequate representation of the observation process, with only limited period-specific lack of fit.

### 3.3. Marginal posterior predictive checks: dynamic occupancy-detection model

The second set of PCCs evaluated the dynamic occupancy process and the detection process jointly in DODM1. For each posterior draw, we generated a new latent occupancy history under the fitted DODM1. The initial occupancy state was fixed to the posterior draw for the first period, $z_{i,1}^{rep,(m)} = z_{i,1}^{(m)}$. For subsequent periods, replicated occupancy states were simulated as

$$z_{i,t}^{rep,(m)} \sim Bernoulli\left(\psi_{i,t}\right),$$

$$z_{i,t}^{rep,(m)} \sim Bernoulli\left(z_{i,t-1}^{rep,(m)}\left(1-\varepsilon_{i,t}^{(m)}\right)+\left(1-z_{i,t-1}^{rep,(m)}\right)\gamma_{i,t}^{(m)}\right),$$

where $\epsilon_{i,t}^{(m)}$ is the extinction probability between periods and $\gamma_{i,t}^{(m)}$ is the colonisation probability. When neighbourhood occupancy effect was included, the colonisation probability was updated using the simulated neighbourhood occupancy proportion at the previous time step.

After simulating the replicated occupancy history, we generated replicated observation data as

$$y_{i,t}^{rep,(m)} \sim Bernoulli\left(z_{i,t}^{rep,(m)} p_{i,t}^{(m)}\right).$$

These marginal PPCs therefore assess whether the fitted DODM1, could reproduce the observed temporal structure of observations while accounting jointly for latent occupancy dynamics and imperfect detection.

We summarised the temporal detection histories using three statistics. First, we calculated the first detection period at each site. Second, we calculated the detection trajectory, defined as the number of detected sites in each period. Third, we calculated turnover, defined as the total number of observation state switches between consecutive periods across all sites.

For first detection timing, we constructed observed and posterior replicated histograms and compared their shapes using a paired chi-square-like discrepancy statistic. For each bin, the expected value was defined as the element-wise average of the observed and replicated counts, with a small constant added to avoid division by zero. The detection trajectory was summarised as the number of detected sites in each period, and the observed and replicated trajectory

vectors were compared using the same paired chi-square-like discrepancy statistic. We also compared scalar timing summaries, including the mean first detection period, calculated across sites where the species was detected at least once. Finally, total turnover was calculated as the total number of observation state switches between consecutive periods across all sites. BPVs were calculated as the proportion of posterior draws for which the replicated statistic was greater than to the observed statistic.

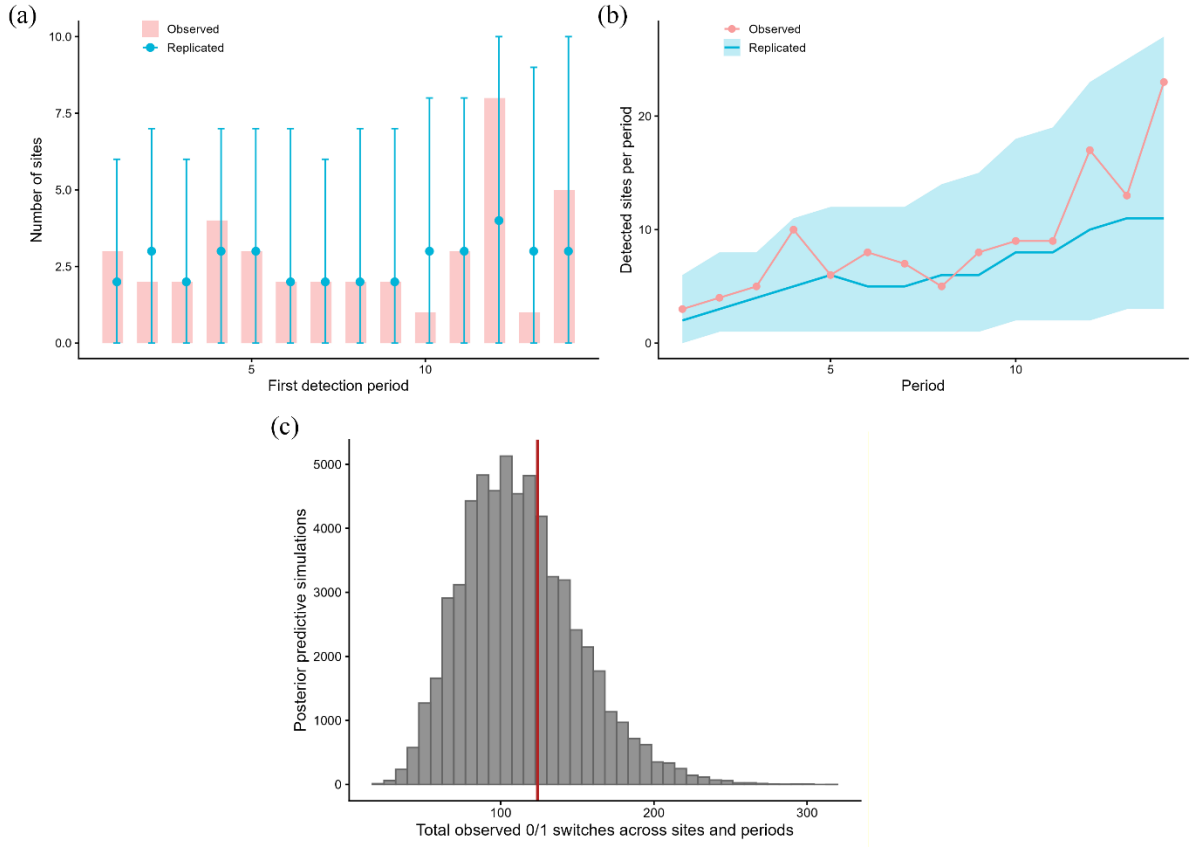



**Figure A.3.2. Marginal posterior predictive checks under DODM1.** (a) PPC for the first detection timing. Observed numbers of initial detections per period are shown as bars and are compared with posterior predictive medians and 95% posterior predictive intervals generated under DODM1. (b) PPC for the detection trajectory. Temporal trend in the total number of detected sites per period across the study duration. The observed number of detected sites per period is compared with the posterior predictive median and 95% posterior predictive interval across the study period. (c) PPC for the turnover. The histogram shows the posterior predictive distribution of the total number of observation state switches across consecutive periods and sites, with the vertical line indicating the observed value.

**Table A.3. Summary of marginal posterior predictive checks for the DODM.** Observed number are compared with the replicated posterior means, together with BPVs.

| PPC | Observed | Replicate | BPV |
|---|---|---|---|
| Mean first detection | 8.15 | 7.95 | 0.424 |
| Turnover | 124.00 | 113.04 | 0.345 |

Marginal PCCs showed that the DODM1 generally reproduced the main temporal features of the observed detection histories (Figure A.3.2 and Table A.3). The distribution of first detection periods was well captured by the model, with the observed counts falling within the posterior predictive intervals and a BPV of 0.563. The scalar summary of mean first detection period also showed good agreement between observed and replicated data, with an observed mean of 8.15, a posterior predictive mean of 7.95, and a BPV of 0.424 (Table A.3).

The model also captured the overall detection trajectory well (Figure A.3.2b). The detection trajectory was also consistent with the posterior predictive distribution. The detection trajectory remained within the 95% posterior predictive intervals with a BPV of 0.842, indicating that the model captured the overall temporal pattern of detections reasonably well. The turnover check further indicated that the observed number of observation state switches was 124, compared with a posterior predictive mean of 113.04 and a BPV of 0.345 (Figure A.3.2c and Table A.3). Overall, the marginal PCCs suggest that the DODM1 reproduced first detection timing, first detection patterns, and the level of turnover in the observed data well.

## References


Animal and Plant Health Inspection Service. (2026). *Emerald Ash Borer.* https://www.aphis.usda.gov/plant-pests-diseases/eab

Bras, A., Lombaert, E., Kenis, M., Li, H., Bernard, A., Rousselet, J., Roques, A., & Auger-Rozenberg, M.-A. (2022). The fast invasion of Europe by the box tree moth: An additional example coupling multiple introduction events, bridgehead effects and admixture events. *Biological Invasions*, *24*(12), 3865–3883. https://doi.org/10.1007/s10530-022-02887-3

Brooks, S. P., Catchpole, E. A., & Morgan, B. J. T. (2000). Bayesian Animal Survival Estimation. *Statistical Science*, *15*(4). https://doi.org/10.1214/ss/1009213003

Canadian Food Inspection Agency. (2007, December 5). Emerald Ash Borer Confirmed in Toronto [News releases]. https://www.canada.ca/en/news/archive/2007/12/emerald-ash-borer-confirmed-toronto.html

Canadian Food Inspection Agency. (2019, April 2). *Cydalima perspectalis* (Walker) – (Box tree moth) [Reference material]. http://inspection.canada.ca/en/plant-health/invasive-species/insects/box-tree-moth

Carey, J. R. (1991). Establishment of the Mediterranean Fruit Fly in California. *Science*, *253*(5026), 1369–1373. https://doi.org/10.1126/science.1896848

D. Hoffmann, B. (1998). The Big-headed *Ant Pheidole megacephala*: A new threat to monsoonal northwestern Australia. *Pacific Conservation Biology*, *4*(3), 250–255. https://doi.org/10.1071/PC980250

Ellison, A. M., Orwig, D. A., Fitzpatrick, M. C., & Preisser, E. L. (2018). The Past, Present, and Future of the Hemlock Woolly Adelgid (*Adelges tsugae*) and Its Ecological Interactions with Eastern Hemlock (*Tsuga canadensis*) Forests. *Insects*, *9*(4), 172. https://doi.org/10.3390/insects9040172

Freeman, M. F., & Tukey, J. W. (1950). Transformations Related to the Angular and the Square Root. *The Annals of Mathematical Statistics*, *21*(4), 607–611.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Gutrich, J. J., VanGelder, E., & Loope, L. (2007). Potential economic impact of introduction and spread of the red imported fire ant, *Solenopsis invicta*, in Hawaii. *Environmental Science & Policy*, *10*(7), 685–696. https://doi.org/10.1016/j.envsci.2007.03.007

Health (PLH), E. P. on P., Bragard, C., Baptista, P., Chatzivassiliou, E., Di Serio, F., Gonthier, P., Jaques Miret, J. A., Justesen, A. F., Magnusson, C. S., Milonas, P., Navas-Cortes, J. A., Parnell, S., Potting, R., Reignault, P. L., Stefani, E., Thulke, H.-H., Van der Werf, W., Vicent Civera, A., Yuen, J., … MacLeod, A. (2023). Pest categorisation of *Solenopsis invicta*. *EFSA Journal*, *21*(5), e07998. https://doi.org/10.2903/j.efsa.2023.7998

Herms, D. A., & McCullough, D. G. (2014). Emerald Ash Borer Invasion of North America: History, Biology, Ecology, Impacts, and Management. *Annual Review of Entomology*, *59*(Volume 59, 2014), 13–30. https://doi.org/10.1146/annurev-ento-011613-162051

1191 Hodges, A. C. (2026). *Brown Marmorated Stink Bug, Halyomorpha halys (Stål) (Insecta:*
1192 *Hemiptera: Pentatomidae)*. https://ask.ifas.ufl.edu/publication/IN623

1193 Hoffmann, B. D., & Parr, C. L. (2008). An invasion revisited: The African big-headed ant
1194 (*Pheidole megacephala*) in northern Australia. *Biological Invasions*, *10*(7), 1171–1181.
1195 https://doi.org/10.1007/s10530-007-9194-x

1196 Ingels, C. A., & Varela, L. G. (2014). *Brown Marmorated Stink Bug.*
1197 https://ipm.ucanr.edu/home-and-landscape/brown-marmorated-stink-bug/

1198 Jean-François, F., Jean-Jacques, F., & Christophe, B. (2009). Premiers signalements en
1199 France de la Pyrale du Buis *Diaphania perspectalis* (Walker, 1859), espèce exotique
1200 envahissante s'attaquant aux Buis (Lepidoptera, Crambidae). https://lorraine-
1201 entomologie.org/webobs/biblio/index.php?module=biblio&action=biblio&id=1482

1202 Lupi, D., Dioli, P., & Limonta, L. (2017). First evidence of *Halyomorpha halys* (Stål)
1203 (Hemiptera Heteroptera, Pentatomidae) feeding on rice (*Oryza sativa* L.). *Journal of*
1204 *Entomological and Acarological Research*, *49*(1). https://doi.org/10.4081/jear.2017.6679

1205 Maistrello, L., Costi, E., & Vaccari, G. (2014). *Primi rinvenimenti in Italia della cimice*
1206 *esotica Halyomorpha halys, una nuova minaccia per la frutticoltura.*
1207 https://iris.unimore.it/handle/11380/1010523?mode=complete

1208 Malacrida, A., Marinoni, F., Torti, C., Gomulski, L., Sebastiani, F., Bonvicini, C., Gasperi, G.,
1209 & Guglielmino, C. (1998). Genetic aspects of the worldwide colonization process of Ceratitis
1210 capitata. *Journal of Heredity*, *89*(6), 501–507. https://doi.org/10.1093/jhered/89.6.501

1211 Mata, L., Vogel, B., Palma, E., Malipatil, M., & Krüger, S. (2022). The Arrival and Spread of
1212 the European Firebug (Pyrrhocoris apterus) in Australia as Documented by Citizen Scientists.
1213 *Urban Naturalist Notes*, *N3*, 1–7.

1214 Nattrass, R., & Vanderwoude, C. (2001). A preliminary investigation of the ecologicaleffects
1215 of Red Imported Fire Ants (*Solenopsisinvicta*) in Brisbane, Australia. *Ecological*
1216 *Management & Restoration*, *2*(3), 220–242. https://doi.org/10.1046/j.1442-
1217 8903.2001.00087.x

1218 Rojas, P. J. O., & Jackson, M. D. (2018). *Pyrrhocoris apterus* L. (Hemiptera: Pyrrhocoridae),
1219 a newly introduced family, genus, and species to Ontario and Canada. *The Journal of the*
1220 *Entomological Society of Ontario*, *149*, 27–32. https://doi.org/10.21083/jeso.v149i0.4406

1221 Roscoe, L. E., Stastny, M., Boone, C. K., Ogden, J. B., Forbes, G. B. H., & Fidgen, J. G.
1222 (2025). Phenology of invasive hemlock woolly adelgid, *Adelges tsugae* Annand (Hemiptera:
1223 Adelgidae), and the role of resident natural enemies in its population regulation in Nova
1224 Scotia, Canada. *The Canadian Entomologist*, *157*, e34.
1225 https://doi.org/10.4039/tce.2025.10018

1226 Smith, M. R. (1933). Additional Species of Florida Ants, with Remarks. *The Florida*
1227 *Entomologist*, *17*(2), 21. https://doi.org/10.2307/3492695

1228 Szyniszewska, A. M., Leppla, N. C., Huang, Z., & Tatem, A. J. (2016). Analysis of Seasonal
1229 Risk for Importation of the Mediterranean Fruit Fly, *Ceratitis capitata* (Diptera: Tephritidae),

via Air Passenger Traffic Arriving in Florida and California. *Journal of Economic Entomology*, *109*(6), 2317–2328. https://doi.org/10.1093/jee/tow196

Texas A&M Forest Service. (2026). *Emerald Ash Borer.* https://tfsweb.tamu.edu/trees/invasive-species/invasive-insects/emerald-ash-borer/