

Sampling effort required for fitting spatially explicit models of species distribution dynamics

Oriol Solà^{a,*}, Marc Kéry^c, Núria Aquilué^d, Lluís Brotons^{b,d,e}

^a Universitat Autònoma de Barcelona (UAB) Departament de Biologia Animal, de Biologia Vegetal i d'Ecologia, Spain

^b Centre for Ecological Research and Forestry Applications (CREAF), Autonomous University of Barcelona (UAB), Cerdanyola del Vallès, Spain

^c Swiss Ornithological Institute, Sempach, Switzerland

^d Forest Science and Technology Centre of Catalonia (CTFC), Solsona, Spain

^e CSIC, Cerdanyola del Vallès, Spain

ARTICLE INFO

Keywords:

Colonization
Detection
Distribution change
Dynamic occupancy models
Extinction
Model bias
Neighborhood connectivity
Sampling coverage

ABSTRACT

Understanding and accurately predicting species distribution dynamics is essential for effective biodiversity conservation and management. Spatial dynamic occupancy models (SpDynOcc models) provide a valuable framework for analyzing temporal changes in species occurrence but require substantial data, making it critical to understand their data needs for reliable estimates. In this study, we use a simulation approach to investigate the role of survey effort, both in terms of study duration and spatial coverage, in obtaining accurate predictions from a generic *SpDynOcc* model. We also test the efficacy of two alternative sampling designs compared to random sampling. We varied multiple factors influencing species occurrence and detection (initial occupancy, occupancy dynamics, and probability of detection) to study the way in which they affect the data requirements for accurate parameter estimation. Models performed best with longer study durations, higher spatial coverage, and higher effective probability of detection (i.e., over all survey occasions). Nevertheless, the specific minimum sampling coverage needed notably varied based on initial occupancy and on occupancy dynamics scenarios. Preferential habitat sampling performed particularly well for low initial occupancy and high-decrease scenarios. These results indicate that tailored survey strategies are essential and must be informed by the specific ecological context. Our findings provide guidance on the survey designs needed to obtain accurate SpDynOcc model predictions, aiding researchers in the effective application of these models for studying species spatial occupancy dynamics.

1. Introduction

Understanding and predicting species distribution and range dynamics is crucial for ecological and conservation efforts (Krebs, 2001). From assessing habitat suitability to monitoring population dynamics, accurate predictions are required for informed decision-making. However, obtaining adequate inferences about spatiotemporal changes in species distributions is challenging, as these estimates are almost always only derived from a partial survey of the populations of interest. In addition, measurement errors such as false negatives and false positives need to be accounted for (Yoccoz et al., 2001; Chapter 1 in Kéry and Royle, 2016). Statistical models play a crucial role in confronting these challenges, as they help identify patterns and generalize the factors that influence population dynamics, while correcting for measurement errors

as needed. Incorporating ecological theory into statistical models can strengthen these predictions, enhancing their relevance and applicability for conservation management (Austin, 2002).

Among the large array of models available for estimating species distributions from survey data, dynamic occupancy models (MacKenzie et al., 2003; Royle and Kéry, 2007), hereafter *DynOcc* models, stand out for their simple, yet mechanistic description of changing species occurrence over time (Briscoe et al., 2021). Unlike static distribution methods, *DynOcc* models offer a closer depiction of the underlying processes driving range dynamics by explicitly modeling species colonization and local extinction events over time. In this, they resemble very much the classical extinction/colonization model of metapopulation ecology (Hanski 1998). Moreover, when multiple detection/non-detection records are available for at least some sites

* Corresponding author.

E-mail address: o.sola@creaf.uab.cat (O. Solà).

<https://doi.org/10.1016/j.ecolmodel.2024.110998>

Received 18 September 2024; Received in revised form 12 December 2024; Accepted 12 December 2024

Available online 19 December 2024

0304-3800/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

within survey seasons, and the occurrence state at a site is assumed to remain constant throughout the season, DynOcc models can account for false-negative errors (i.e., instances where a species is present but goes undetected). This will lead to estimates of occupancy, colonization, and local extinction that are more robust to the imperfect detection bias that is so ubiquitous in animal and plant data from the field (Lahoz-Monfort et al., 2014).

So far, a majority of studies working with *DynOcc* models have treated sites as independent entities (Chapter 4 in Kéry and Royle, 2021). However, for many species, colonization and extinction events are influenced by the occupancy status of neighboring patches through dispersal dynamics (García-Valdés et al., 2015; Hanski, 1998). On the one hand, for a colonization event to happen, the presence of the species in the neighbors around the focal site is often a prerequisite. On the other hand, extinction risk diminishes for sites with more occupied neighbors due to the 'rescue effect', i.e., higher chances of immediate recolonization after extinction (Brown and Kodric-Brown, 1977). A spatially explicit variant of the *DynOcc* model, the spatial autologistic dynamic occupancy model, hereafter *SpDynOcc*, extends the traditional *DynOcc* framework by explicitly considering the spatial dependencies among sites, thus enhancing ecological realism. This extension, pioneered by Royle and Dorazio (2008) and used with several variations by subsequent studies (Bled et al., 2011; Broms et al., 2016; Eaton et al., 2014; Yackulic et al., 2012), defines a neighborhood for each study site and incorporates a latent autocovariate to the colonization and extinction probability regression equations. This autocovariate, which captures the occupancy status of neighboring sites, influences a target cell likelihood of colonization or persistence, aligning with the concept of biogeographic patterns arising from dispersal constraints and local extinction dynamics (Bell and Schlaepfer, 2016). The logic of this model is that the more neighboring sites are occupied around a target cell, the more potential immigrants are available to either colonize the cell when it is empty or rescue it from extinction when it is occupied but losing its population. This model is a close relative of classical metapopulation models, but for a discrete and defined neighborhood (Risk et al. 2011).

One main challenge of the *SpDynOcc* model is that it relies on the inference of occupancy status of all cells, or patches, in a study landscape to correctly estimate the autocovariate effects on both colonization and extinction processes (Moilanen, 2000; Risk et al., 2011). However, most sampling schemes do not exhaustively cover the totality of the study landscape and thus, occurrence data will only be available from a subset of all cells. The model can overcome this limitation by using the so-called unconditional probability of occupancy, which is the likelihood that a patch is occupied regardless of whether it has been surveyed, when a neighboring patch is not surveyed. Thus, autologistic modeling can be conducted using all potential neighbor patches, irrespective of their survey status, as long as neighboring patches are identifiable and patch-specific covariates are measurable. However, when occurrence data is scarce, resulting estimates are susceptible to bias, which may jeopardize the inferences about true spatiotemporal patterns in the species distribution. This can lead to erroneous predictions, which either underestimate or overestimate the future occupancy status of the study patches (McKann et al., 2013), highlighting the need for model simplification and appropriate ecological justification in complex models (Bell and Schlaepfer, 2016).

In the present study, we investigate the minimum survey coverage characteristics, in terms of both spatial and temporal extent, needed to accurately infer the effects of neighbors' occupancy status on site-specific colonization and persistence probabilities, and thus to accurately model species distribution change with a *SpDynOcc* model. Additionally, we explore the efficacy of two alternative sampling designs in improving the precision of parameter estimates compared to a basic design using randomly selected sampling locations. We hypothesize that the minimum sampling coverage required to fit a spatial dynamic occupancy model and achieve accurate estimates will vary depending on several species characteristics, including (1) detection

probability, (2) the initial range of the species, (3) colonization and persistence patterns, and (4) survey design. Despite these variations, we anticipate a general inverse relationship between study duration and spatial survey coverage, where longer study durations allow for reduced spatial coverage to maintain prediction accuracy, and vice versa.

While dynamic occupancy models are widely applied, this study provides the first simulation-based assessment of spatial dynamic occupancy models (*SpDynOcc*) to quantify the role of sampling constraints in obtaining reliable predictions of species distribution dynamics. By establishing a simulation framework, we provide foundational guidance for designing effective survey strategies for *SpDynOcc* models, which can be readily adapted to diverse ecological contexts. This framework also lays a basis for future research on *SpDynOcc* model applications with additional ecological variables and complexities.

2. Methods

2.1. Model specification

We used the *SpDynOcc* model as described in Kéry and Royle, 2021 (chapter 9) to simulate multivariate occupancy time-series data. This model builds on the basic formulation of the *DynOcc* model (MacKenzie et al., 2003), where the occurrence status at the first time period ($z_{i,1}$) of a site i is determined by the initial probability of occurrence ($\psi_{i,1}$), and the occurrence status at subsequent time periods ($z_{i,t+1}$) is defined by the presence/absence state at the previous time step and the probabilities of colonization (γ) or persistence (ϕ), or alternatively extinction (ϵ), which is $1 - \phi$. The *SpDynOcc* model is a spatially explicit extension of the *DynOcc* model, incorporating the proportion of occupied neighbors of a site i at time t ($x_{i,t}$ in Eq. 3–5) to influence colonization or persistence rates during the interval from t to $t + 1$. In our simulation study, sites are arrayed in a grid, and we define a cell's neighborhood ($\mathcal{N}_i$) as all its adjacent cells. Therefore, each site has eight neighbors ($n_i = 8$), except for those on the edge of the domain, which have fewer neighbors. Below, we specify the mathematical expressions of the model.

$$z_{i,1} \sim \text{Bernoulli}(\psi_{i,1}) \quad (1)$$

$$z_{i,t+1} \sim \text{Bernoulli}(z_{i,t} * \phi_{i,t} + (1 - z_{i,t}) * \gamma_{i,t}) \quad (2)$$

Here, $\psi_{i,1}$ is the probability of occurrence at time 1, and $z_{i,t}$ is the binary (realized) presence/absence of cell i at time t . The probability of colonization (γ) and persistence (ϕ) can be defined as linear functions of covariate as in any generalized linear model (McCullagh and Nelder, 1989). One of these covariates (and in our study, the only covariate) will be the autocovariate in the *SpDynOcc* model, described above. Thus, we will have:

$$\text{logit}(\phi_{i,t}) = \beta_0 + \beta_1 x_{i,t} \quad (3)$$

$$\text{logit}(\gamma_{i,t}) = \alpha_0 + \alpha_1 x_{i,t} \quad (4)$$

$$x_{i,t} = \frac{1}{n_i} \left(\sum_{j \in \mathcal{N}_i} z_{j,t} \right) \quad (5)$$

Here, α and β are regression coefficients that are estimated.

2.2. Data simulation

We simulated multivariate occupancy time-series data for 15 consecutive years in a 50×50 grid (i.e., with 2500 cells or sites) using the *SpDynOcc* model for 12 combinations of 3 scenarios of the initial number of occupied sites ("ini.occ" scenarios) and 4 scenarios of colonization or extinction ("colext" scenarios); see below.

We used a Gaussian Random Field (GRF) as a covariate to determine the initial probability of occupancy (Eq. (6)), so the initial species

distribution of presence/absence was made more realistic by including some spatial autocorrelation structure. Thus, for each of the 3 “ini.occ” scenarios, we selected a set of initial probability of occurrence ($\psi_{i,1}$ parameters) so the mean simulated initial percentage of occupied sites was 1 % (low), 10 % (moderate), and 90 % (high) for each scenario, respectively (appendix S1).

$$\text{logit}(\psi_{i,1}) = \delta_0 + \delta_1 \text{GRF}_i \quad (6)$$

For each “ini.occ” scenario we chose 4 different sets of values of the colonization (γ) and persistence (ϕ) model parameters, to yield 4 different “colex” scenarios: high occupancy increase, moderate occupancy increase, high occupancy decrease, and moderate occupancy decrease (see Appendix S1 for the detailed parameters and simulated occupancy trajectories of each “colex” scenario). For the first two “colex” scenarios, the probability of persistence was set to 1 (that is, there was no extinction) and colonization probability was made dependent on the proportion of occupied neighbors, whereas for the latter two, colonization rate was set to 0 (i.e., there was no colonization) and the probability of persistence was made dependent on the proportion of occupied neighbors. Scenarios of high occupancy increase had higher probabilities of colonization compared to scenarios of moderate occupancy increase, whereas scenarios of high occupancy decrease had lower probabilities of persistence compared to scenarios of moderate occupancy decrease.

We ran 25 replicates of each combination of “ini.occu” and “colex” scenarios, resulting in a total of 300 occupancy datasets. For each occupancy dataset we simulated detection/non detection data under the usual binomial observation model (Eq. (7)) for exactly two visits within every year for 3 different scenarios of effective probability of detection: 0.2, 0.5, and 0.8. We denote the effective probability of detection as p^* according to Eq. (8).

$$y_{i,t,j} \sim \text{Bernoulli}(z_{i,t}^* p) \quad (7)$$

$$p^* = 1 - (1 - p)^2 \quad (8)$$

For each detection/non-detection dataset we fitted models for the combination of 3 scenarios of study duration (i.e., temporal survey effort) and 4 or 5 scenarios of spatial coverage (i.e., spatial survey effort). Scenarios of study duration were defined as the first 3, 6, or 12 consecutive years of data. Scenarios of spatial coverage, were defined as 5, 10, 20, 40, and 80 % of randomly selected cells in the landscape. We did not consider the 12 years scenario combined with the 80 % spatial coverage scenario in order to reduce extra computing time, and since, based on the other scenarios, it became clear that for this combination very good results were guaranteed. In total we fitted 3 (“ini.occu” scenarios) * 4 (“colex” scenarios) * 25 (replicates) * 3 (detection probability scenarios) * 4 (spatial survey scenarios) * 3 (temporal survey scenarios) = 10,800 models + 1800 (spatial survey of 80 % and temporal survey of 3y and 6y) = 12,600 models.

2.3. Alternative sampling designs

To evaluate the impact of different sampling strategies on the precision of parameter estimates, we examined two alternative sampling designs and compared them to a basic design with randomly selected sampling locations.

2.3.1. Habitat-preferential sampling

In the first alternative design, we preferentially sampled sites with high habitat suitability. Data were collected at locations where the initial probability of occurrence was high (i.e., GRF covariate > 0). This approach assumes prior knowledge of the species’ habitat. The rationale for this design is that by sampling high-quality habitat sites, we increase the likelihood of detections, which may lead to a greater number of observed extinction and colonization events.

2.3.2. Sequential sampling design

In the second design, we employed a sequential sampling approach, using two sets of randomly selected sampling sites that alternated each year. This approach aims to maximize the spatial extent of sampled cells across the landscape by covering different sites each year, allowing for broader spatial coverage over time.

These alternative sampling designs were evaluated with 10 % spatial coverage, a 6-year study duration, and an effective detection probability of 0.8.

2.4. Model fitting

The 12,600 models were fitted using the Bayesian general-purpose modelling software JAGS (Plummer, 2003) via the R package `jagsUI` (Kellner, 2024), using vague priors for all the model parameters (see the Appendix S2 for a description of the model, including the priors used, in the BUGS language). We initially chose minimal Markov chain Monte Carlo (MCMC) settings for all models, ran 2000 iterations, with a burn-in of 1000 iterations and a thinning rate of 2. This process took 24 days on an Ubuntu machine with Ampere® Altra® CPU, which uses Arm64 architecture with 80 cores and 128 GB RAM. We checked model convergence and for those models that did not converge we ran 3000 additional iterations using the `update()` function of the `jagsUI` package. For models that did still not converge after the combined 2000+3000 iterations, if they obtained accurate predictions (as defined in the “Evaluation of predictive model performance” section), we assumed that models would eventually converge yielding similar predictive performance, whereas for models with inaccurate predictions that had not converged, we ran 5000 more iterations (for a total of 10,000 iterations), and then either we retained their results (if they were accurate and/or had converged) or we discarded the results from these data sets. We report all models that did not converge in Appendix S3.

2.5. Evaluation of predictive model performance

Model evaluation focused on assessing the model’s ability to capture and predict range-wide temporal trends in occupancy. To evaluate the predictive performance of the model, we compared the predicted occurrence data to the true occurrence data. Predictions were generated by simulating the binary (realized) presence/absence state of all cells in the landscape over a 15-year period. For each fitted model, we ran 1000 simulations using the posterior mean estimates for the parameters of initial occupancy, colonization, and persistence functions.

To assess the bias of the model predictions, we calculated the total number of occupied sites at time $t + 1$ minus the total number of occupied sites at time t , and we averaged this difference over the 15-year period. We refer to this metric as “mean occurrence change”. We chose this metric because it directly measures the year-to-year change in occupancy, which is a key indicator of how well the model captures dynamic processes such as colonization and extinction. Traditional metrics like root mean square error (RMSE) might not adequately reflect the direction of occupancy change (i.e., increase or decrease). Specifically, RMSE could classify models as having accurate predictions even if they predicted the wrong sign of occupancy change (predicting an increase when there is a decrease, and vice versa). By focusing on “mean occurrence change,” we ensure that the model’s predictions align with the actual trend in occupancy dynamics, thus providing a more intuitive and relevant measure of predictive performance for ecological applications. However, we note that this metric is only appropriate for cases where changes are unidirectional and non-periodic, as is the case in our simulations.

A model was considered to produce accurate predictions if the difference in ‘mean occurrence change’ between its prediction (the average across all 1000 simulations) and the true data was smaller than the largest difference observed among the “mean occurrence change” values of the true data replicates within the specific scenarios of initial

occupancy and type of occupancy dynamics. While this classification is somewhat arbitrary, it provides a clear metric to classify model prediction performance in a binary manner. By setting this threshold, we ensured reliable predictions while allowing for some bias.

Finally, we modeled the probability of the models correctly predicting occupancy dynamics across the different scenarios using a logistic mixed effects model implemented with the glmer function in the lme4 package (Bates et al., 2015). We included spatial coverage, study

duration, and effective probability of detection as fixed effects, while the interaction of initial occupancy and type of occupancy scenarios was incorporated as a random intercept. This approach allowed us to generalize the required survey effort for different effective probabilities of detection to obtain accurate predictions, while accounting for the variability within occupancy scenarios.

To address potential edge effects, we also evaluated model predictions using a larger grid (10,000 cells) for a subset of the scenarios

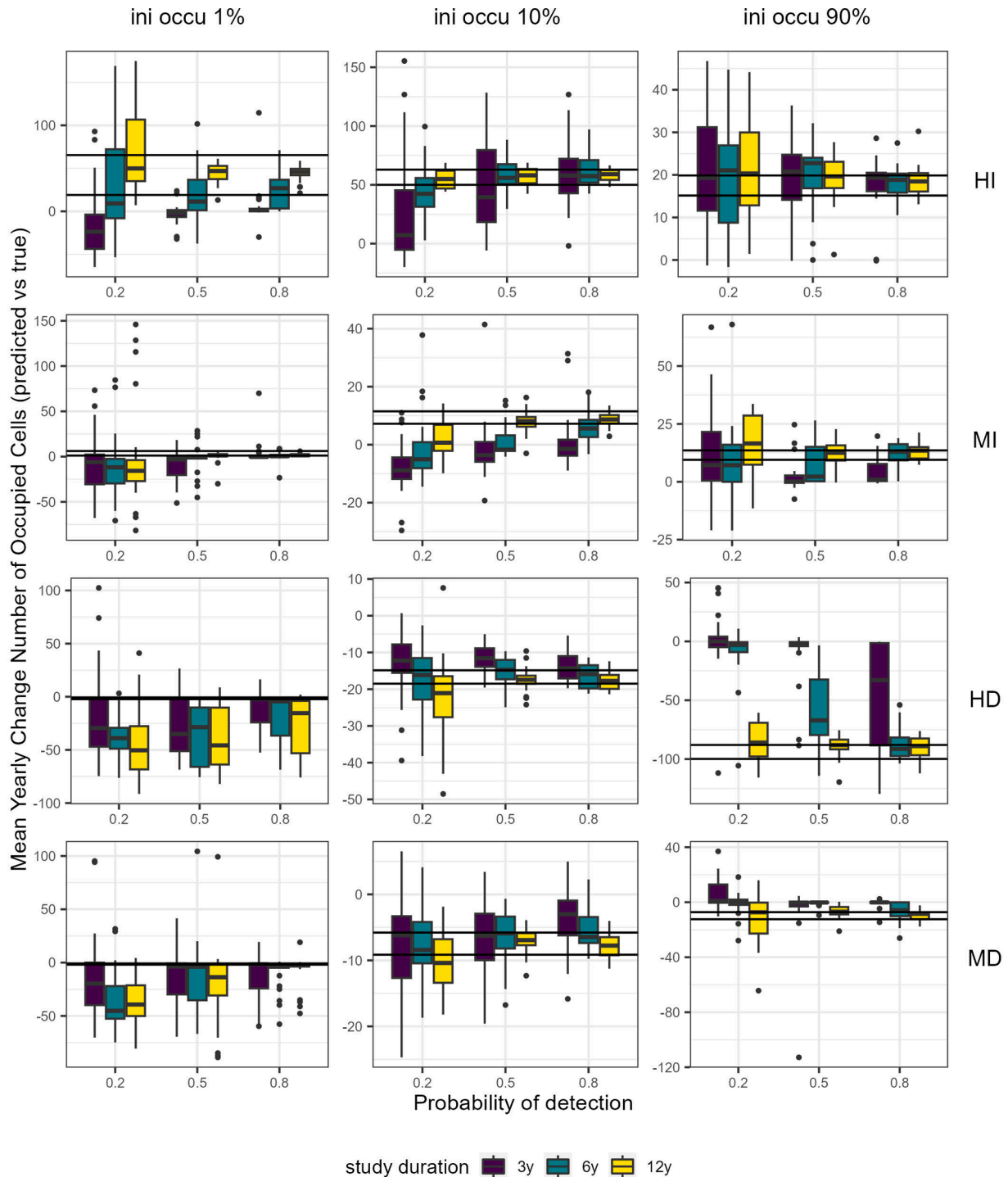


Fig. 1. Mean annual change in the total number of occupied cells at 10 % of spatial coverage. Boxplots represent the predicted values among replicates of each scenario. Horizontal black lines indicate the maximum and minimum values of the true data among the 25 replicates. Columns represent the different initial occupancy scenarios, and rows type of occupancy dynamics scenarios. Scenarios of occupancy dynamics are denoted as follows: HI - High occupancy increase, MI - Moderate occupancy increase, HD - High occupancy decrease, and MD - Moderate occupancy decrease.

analyzed. This supplementary analysis allowed us to verify that edge effects did not influence the results, as detailed in the Appendix S6.

3. Results

3.1. Prediction bias

For scenarios where models failed to obtain accurate predictions, we observed distinct bias patterns depending on the initial occupancy and type of occupancy dynamics (Fig. S4 and Fig. 1 for bias at 10 % spatial coverage). In scenarios of increasing occupancy, with initial occupancy at 1 % or 10 %, the predicted “mean occurrence change” in most replicates with poor model predictions was below the true occurrence change, indicating an underestimation of the number of colonization events (Fig. S4). Some replicates even showed negative “mean occurrence change” values, indicating a predicted decrease in occupancy instead of an increase, especially in scenarios of moderate occupancy increase (Fig. S4). For initial occupancy at 90 %, in scenarios of increasing occupancy with biased model predictions, the “mean occurrence change” was more evenly distributed around the true data among replicates, but most replicates exceeded the maximum and minimum true values, indicating both overestimation and underestimation of the number of colonization events (Fig. S4).

In scenarios of decreasing occupancy, when initial occupancy was 1 %, the predicted “mean occurrence change” in most scenarios with poor model predictions had lower values when compared to the true data, indicating an overestimation of the number of extinction events (Fig. S3). When initial occupancy was 10 %, models with biased predictions tended towards overestimation of the number of extinction events for 12-year study durations, while for 3-year study durations, there was a tendency to underestimate the number of extinction events. For 6-year study durations, no clear trend was observed. Lastly, when initial occupancy was 90 %, the “mean occurrence change” in scenarios with poor model predictions was higher than the true data, indicating an underestimation of the number of extinction events. Some replicates even showed a positive occupancy change, predicting an increase in

occupancy instead of a decrease (Fig. S4).

3.2. Model performance across scenarios

As expected, the probability of obtaining accurate model predictions increased with longer study duration, higher spatial survey coverage, and higher probability of detection (Figs. 2 and 3). When spatial coverage was limited to 5 % or 10 %, high probabilities (>70 %) of models with accurate predictions were only achieved with a study duration of 12 years and an effective probability of detection (*eff.p.det*) of 0.8 (Fig. 2). At 20 % spatial coverage, good model performance was also observed with 6-year study duration and an *eff.p.det* of 0.8, as well as with 12-year duration and an *eff.p.det* of 0.5 (Fig. 2). With 40 % spatial coverage, accurate predictions were achieved for scenarios with 12-year duration and an *eff.p.det* of 0.2, with 6-year duration and an *eff.p.det* of 0.5, as well as with 3-year duration and an *eff.p.det* of 0.8 (Fig. 2). Conversely, scenarios with 3-year duration and values of *eff.p.det* of 0.5 or 0.2, and 6-year duration with an *eff.p.det* of 0.2 consistently showed low probabilities of obtaining accurate model predictions, even with maximum spatial coverage of 80 % (Fig. 2).

However, these general trends varied depending on the scenario of initial occupancy and the type of occupancy dynamics (Figs. 2 and 3). On the one hand, for scenarios of increasing occupancy, differences in model performance between scenarios of initial occupancy were minimal, although models with an initial occupancy of 90 % performed slightly worse for *eff.p.det* 0.5 and 0.2 (Fig. 3). On the other hand, scenarios with high occupancy increase consistently required less survey coverage and lower effective probability of detection for good predictions compared to scenarios with moderate occupancy increase (Figs. 2 and 3). In contrast, for scenarios of decreasing occupancy, marked differences were observed between initial occupancy scenarios, while differences between scenarios of high and moderate occupancy decrease were less noticeable within each initial occupancy scenario (Figs. 2 and 3). Models performed much better with an initial occupancy of 10 % compared to 90 %, especially for study durations of 6 and 3 years (Fig. 3). Notably, scenarios with 1 % initial occupancy and

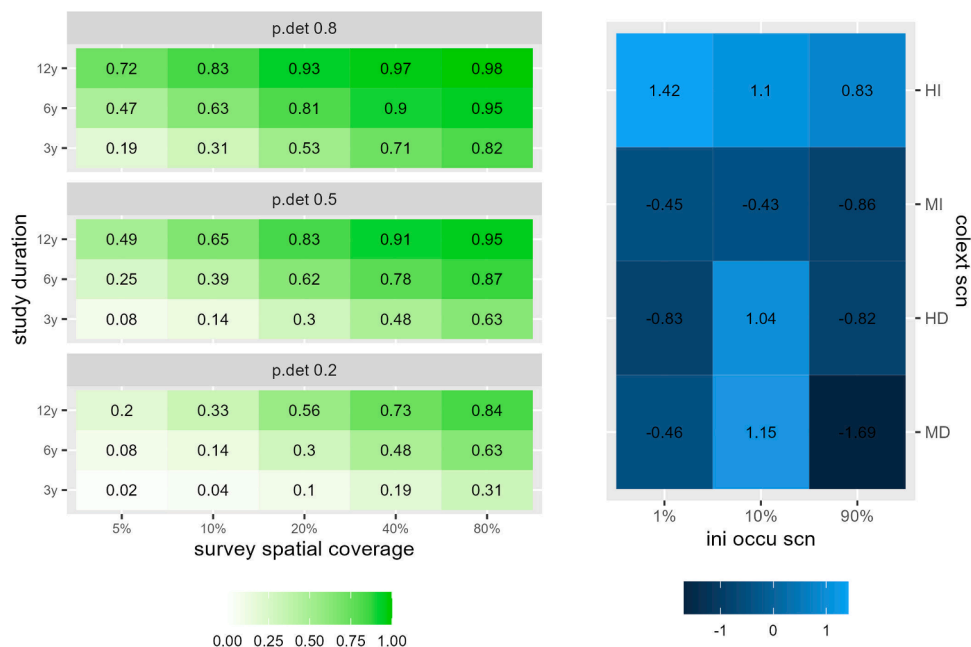


Fig. 2. Probability of obtaining accurate predictions of occupancy dynamics according to the mixed effects model of the *SpDynOcc* model fitting results. The figure on the left displays the probability for all combinations of fixed effects in the model: effective probability of detection (top, middle and bottom panel), study duration (rows), and survey spatial coverage (columns). The figure on the right shows the random intercepts for each combination of initial occupancy (columns) and type of occupancy dynamics scenario (rows) on the logit scale. Scenarios of occupancy dynamics are denoted as follows: HI - High occupancy increase, MI - Moderate occupancy increase, HD - High occupancy decrease, and MD - Moderate occupancy decrease.

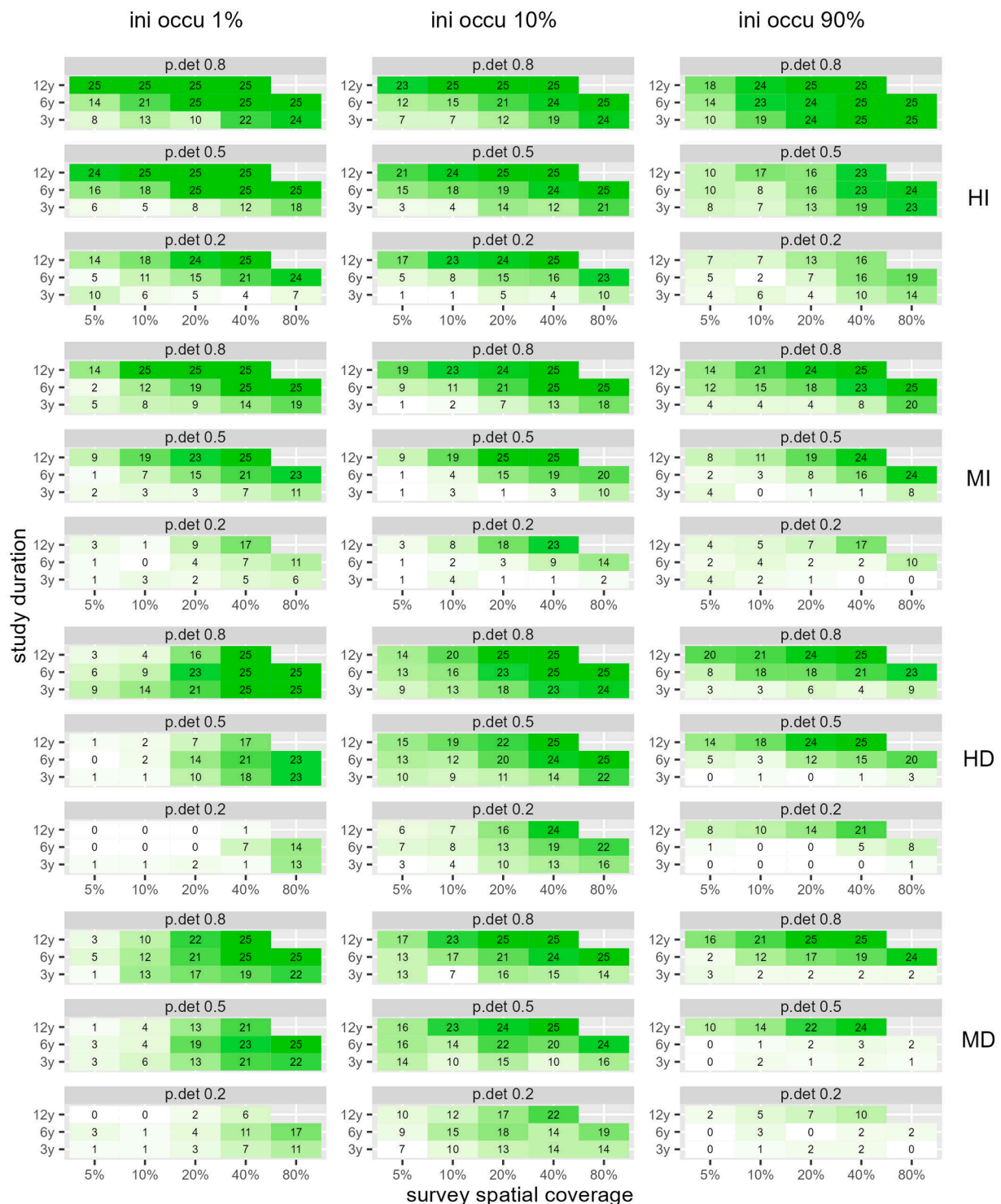


Fig. 3. Number of accurate predictions for each scenario out of the 25 replicates for each combination. Accurate predictions are classified based on the predicted mean occurrence change as described in the Methods section. Columns represent the different initial occupancy scenarios, and rows type of occupancy dynamics scenarios. Scenarios of occupancy dynamics are denoted as follows: HI - High occupancy increase, MI - Moderate occupancy increase, HD - High occupancy decrease, and MD - Moderate occupancy decrease. For each occupancy scenario, there are three panels representing scenarios with effective probabilities of detection of 0.8, 0.5, and 0.2 (top, middle, and bottom panels, respectively). Within each panel, rows represent study duration, and columns represent survey spatial coverage. Note that the 'maximum combination' of 12 years and 80 % spatial coverage was not included due to its high computational cost.

decreasing occupancy dynamics (both high and moderate decrease) were unique. For these scenarios, study duration did not significantly impact prediction accuracy, indicating that more years of surveys did not necessarily lead to better model predictions (Fig. 3).

3.3. Effectiveness of alternative sampling strategies on model prediction accuracy

Preferential habitat sampling yielded a slight improvement in the number of accurate model predictions compared to random sampling

(mean 16.50 vs. 15.08 for random sampling), though this advantage was not consistent across all occupancy scenarios (Table 1). This strategy performed particularly well in scenarios with an initial occupancy of 1 % (mean 17.25 vs. 13.5) and high occupancy decrease scenarios (mean 20 vs. 14.3). In contrast, sequential sampling did not enhance model predictions compared to random sampling (Table 1), and it resulted in slightly fewer accurate fits across most occupancy scenarios (mean 13.42 vs. 15.08).

4. Discussion

Spatial dynamic occupancy models are designed to understand and predict species spatial dynamics by integrating ecological realism through the consideration of dispersal processes driving range changes. These models can offer valuable insights into the spatial population dynamics of many species (Bled et al., 2013; Eaton et al., 2014). However, fitting these models to often scarce data carries the risk of obtaining inaccurate parameter estimates, leading to biased or inadequate predictions (Mckann et al., 2013), even when the model structure aligns closely with the (unknown) data-generation mechanism. In this study, we examined how varying temporal and spatial survey efforts impact the accuracy of spatial dynamic occupancy models. Our findings highlight the critical balance needed between study duration and spatial coverage to achieve reliable predictions, offering insights that are broadly applicable to any species distribution model incorporating explicit colonization and extinction functions. Our analysis revealed an inverse relationship between the minimum study duration and spatial survey coverage that was required to achieve satisfactory model predictions across various scenarios of occupancy and detection. Furthermore, our results underscore that species with low detection probabilities require higher sampling coverage for accurate model estimates, emphasizing the importance of optimizing survey techniques to improve detectability. This study aims to guide researchers in effectively utilizing spatial dynamic occupancy models for studying species occupancy dynamics, ensuring that predictions are as unbiased as possible, and thereby enabling informed decisions in biodiversity management and conservation efforts (Briscoe et al., 2019; Guisan et al., 2013).

In recent years, substantial research has been conducted to address sampling bias in species distribution models (SDMs). For instance, Rana

et al. (2024) highlighted the non-stationarity of SDMs when faced with varying data types and sampling extents, while Boyd et al. (2024) explored the trade-offs between spatial resolution and accuracy in estimating species occupancy from geographically biased samples. However, most of these studies focus on static models and do not account for imperfect detection, which limits their applicability to dynamic systems where temporal changes and species detectability play a significant role. In contrast, our work specifically targets the spatial dynamics of species occupancy over time and evaluates the minimum survey effort required for reliable parameter estimation in SpDynOcc models.

Previous findings suggested a minimum of 120 unique sampling locations for accurate estimation of colonization and extinction parameters in non-spatial dynamic occupancy models (McKann et al., 2013). In contrast, our study reveals far more nuanced insights into the data requirements for spatial dynamic occupancy models. When spatial coverage was the most limited in our study, corresponding to 125 sites, accurate predictions were only possible for a study duration of 12 years and for only certain occupancy scenarios among the 12 scenarios that we studied. For study durations of 6 years, at least 500 sites were needed to obtain accurate predictions, and for 3 years, 1000 or even more sites were required. By quantifying these relationships across diverse occupancy and detection scenarios, our results provide a framework that can inform survey designs for a wide range of ecological systems. These findings highlight the limitations that many monitoring programs may currently face in applying SpDynOcc models, and the need for more unique sampling locations to effectively capture spatial effects in model predictions. Moreover, our results emphasize that the number of sampling locations should not be fixed for any specific landscape but should be higher when the survey duration is shorter, and vice versa.

We hypothesized that the minimum sampling coverage needed to fit a spatial dynamic occupancy model and achieve accurate estimates varies as a function of several factors, including the initial range of a species, occupancy dynamics, and the effective probability of detection. Our results confirmed all of these. Notably, the effective probability of detection emerged as a pivotal factor. For species with low detection probabilities, higher sampling coverage was necessary to obtain accurate estimates. Encouragingly, species detectability can be improved by optimizing survey techniques, such as selecting the best time of day, using more experienced observers, or employing modern detection technologies like camera traps or acoustic recorders (O'Connor et al., 2017). These enhancements can lead to more reliable model performance without increasing spatial or temporal survey coverage, which are the costliest features of a sampling design. In addition, the effective probability of detection can be pushed up also by simply visiting each site more often.

Moreover, the initial number of occupied cells and the type of occupancy dynamics also influenced model performance. These factors affected both the minimum survey effort needed for accurate predictions and the bias patterns in models with inaccurate predictions. For scenarios of increasing occupancy, models performed better with the same sampling coverage in scenarios of high occupancy increase compared to scenarios of moderate occupancy increase. This suggests that for species undergoing rapid range expansion, predicted trajectories will generally be more reliable than for species with slower expansion. Our results show that for species with moderate occupancy increases, insufficient data can lead to models predicting a decrease in occupancy, and similarly, in scenarios with moderate occupancy decreases, models with insufficient data can predict an increase in occupancy. Thus, when species trends are not clear, it is crucial to ensure that we have sufficient data to trust model predictions and avoid misleading conclusions. For scenarios of decreasing occupancy with an initial occupancy of 1 %, longer study durations did not necessarily lead to better predictions. In these cases, maximizing spatial coverage and ensuring high species detectability is crucial.

Our exploration of alternative sampling strategies suggests that preferential habitat sampling offers a modest improvement over random

Table 1
Number of accurate predictions for each sampling strategy out of the 25 replicates for each scenario. Accurate predictions are classified based on the predicted mean occurrence change as described in the Methods section. Models were evaluated with 10 % spatial coverage, a 6-year study duration, and an effective detection probability of 0.8.

Initial occupancy	Occupancy dynamics	Random sampling	Preferential habitat sampling	Sequential sampling
1 %	High increase	21	24	21
10 %	High increase	15	16	18
90 %	High increase	23	21	21
1 %	Moderate increase	12	16	13
10 %	Moderate increase	11	12	10
90 %	Moderate increase	15	2	11
1 %	High decrease	9	14	7
10 %	High decrease	16	24	21
90 %	High decrease	18	22	9
1 %	Moderate decrease	12	15	12
10 %	Moderate decrease	17	13	14
90 %	Moderate decrease	12	19	4
Mean ± SD		15.08 ± 4.15	16.50 ± 6.07	13.42 ± 5.12

sampling in achieving accurate model predictions, particularly in scenarios with low initial occupancy (1 %) and high occupancy decrease. This may be the case for rare threatened species for instance and therefore the results may be of practical importance in conservation settings. This improvement may be due to the strategy's focus on higher-quality habitat patches, which likely increases the detection of occupied sites and the observation of key extinction and colonization events. In scenarios with low initial occupancy, preferentially sampling areas of suitable habitat helps ensure that even sparse populations are represented in the data, thereby enhancing model precision. This preferential habitat sampling strategy is likely to be efficient for a wide range of modeling techniques that aim to capture the distribution dynamics of rare species, further broadening its applicability beyond the SpDynOcc framework. In high-decrease scenarios, where population decline is more rapid, focusing on suitable habitat patches may also allow for more timely detection of extinction dynamics, leading to better-informed estimates. Conversely, sequential sampling did not improve prediction accuracy compared to random sampling. The switch between two random sets of sampling locations each year may have limited the ability to capture occupancy trends consistently over time, resulting in slightly lower accuracy across most occupancy scenarios. Together, these results suggest that while preferential habitat sampling can yield benefits in specific scenarios, the effectiveness of alternative designs depends heavily on occupancy patterns and population trends, underscoring the need for tailored survey strategies based on ecological context. Simulation studies to evaluate the effectiveness of different sampling strategies and identify optimal approaches could be valuable for guiding species-specific surveys (Chapter 4 in [Kéry and Royle, 2016](#); [DiRenzo et al., 2023](#)).

Our study offers a foundational assessment of data requirements in spatial dynamic occupancy models (SpDynOcc). This model choice allowed us to focus explicitly on dispersal, a key driver of species range dynamics, as a mechanistic factor in our simulations. Other studies have shown the importance of dispersal in shaping ecological patterns and highlighted the challenges it poses for parameter estimation. For instance, [Xu et al. \(2023\)](#) demonstrated how dispersal can introduce biases in estimating competitive interactions within metacommunities, complicating the interpretation of ecological processes. Similarly, [Detto et al. \(2019\)](#) revealed that dispersal-driven sampling biases could obscure the detection of negative density dependence in plant communities. These findings underscore how dispersal, while critical for understanding species dynamics, can also confound parameter estimation if not carefully accounted for in study designs. In our work, we address this challenge by explicitly incorporating dispersal processes into a dynamic occupancy modeling framework, ensuring that the effects of dispersal on colonization and extinction dynamics are explicitly modeled and quantified.

By embedding habitat as a predictor of initial occupancy, we ensured that favorable habitat indirectly influences dispersal, capturing the broader ecological realism needed in species distribution modeling. However, we acknowledge that other factors—such as species interactions and environmental gradients—are not directly included in our model. Introducing additional complexity by incorporating variables like competition, temperature, and landscape features could enhance the model's applicability to field conditions but would also necessitate further investigation into data requirements to support reliable parameter estimates. Additionally, our analysis of alternative sampling designs provides empirical insights for field researchers, offering a data-driven approach to evaluating the efficacy of different sampling schemes. This study therefore bridges a gap in the empirical testing of heuristic sampling recommendations, providing an essential framework for future research aiming to optimize sampling strategies under more complex ecological conditions. As so often for questions of sampling design, one could say "When in doubt, simulate." ([Kéry and Kellner, 2024](#)).

5. Conclusion

Effective research and management efforts focused on species occupancy dynamics must consider the limitations of sparse data in obtaining accurate and precise model estimates when using spatial dynamic occupancy models. This study provides quantitative insights into the data requirements necessary for reliable predictions, highlighting the influence of survey duration, spatial coverage, and detection probability on model accuracy. Our findings underscore that while increasing model complexity can enhance ecological realism, it also demands higher data quality and quantity. Blindly applying complex models to limited datasets risks producing biased, non-generalizable predictions—a concern that must be weighed against the intended application of the model. By evaluating the efficacy of different sampling designs, we offer researchers a data-driven framework to tailor sampling strategies to species characteristics and occupancy dynamics, thus optimizing survey design and parameter estimation. This approach illustrates the critical role of simulation studies in testing sampling designs before empirical implementation, offering a robust foundation for enhancing biodiversity management and conservation planning. We echo the value of simulation as a tool to explore and refine sample design, particularly when tackling the challenges of ecological model complexity ([DiRenzo et al., 2023](#)).

CRedit authorship contribution statement

Oriol Solà: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Marc Kéry:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Núria Aquilué:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Lluís Brotons:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Acknowledgments

This study was partially funded by MCIN/AEI through the projects GREENRISK (PID2020-119933RB-C22) and CEX-2018-000828-S “Centro de Excelencia Severo Ochoa”. O.S. was supported by a pre-doctoral grant PRE2020-092082, by the Spanish Ministry of Science, Innovation and Universities. N.A. was supported by a Juan de la Cierva fellowship of the Spanish Ministry of Science, Innovation and Universities (FCJ2020-046387-I).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.ecolmodel.2024.110998](https://doi.org/10.1016/j.ecolmodel.2024.110998).

Data availability

Script files to replicate the data simulation, model fitting, analysis, and plot generation are available from the Zenodo Digital Repository: [10.5281/zenodo.13627582](https://doi.org/10.5281/zenodo.13627582).

References

- Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Modell.* 157 (2–3), 101–118. [https://doi.org/10.1016/S0304-3800\(02\)00205-3](https://doi.org/10.1016/S0304-3800(02)00205-3).

- Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Bell, D.M., Schlaepfer, D.R., 2016. On the dangers of model complexity without ecological justification in species distribution modeling. *Ecol. Modell.* 330, 50–59. <https://doi.org/10.1016/j.ecolmodel.2016.03.012>.
- Bled, F., Nichols, J.D., Altwegg, R., 2013. Dynamic occupancy models for analyzing species' range dynamics across large geographic scales. *Ecol. Evol.* 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. *Ecol. Appl.* 21 (1), 290–302. <https://doi.org/10.1890/09-1877.1>.
- Boyd, R.J., Bowler, D.E., Isaac, N.J.B., Pascott, O.L., 2024. On the trade-off between accuracy and spatial resolution when estimating species occupancy from geographically biased samples. *Ecol. Modell.* 493, 110739. <https://doi.org/10.1016/j.ecolmodel.2024.110739>.
- 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., Veski, 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. *Ecol. Lett.* 22 (11), 1940–1956. <https://doi.org/10.1111/ele.13348>.
- Briscoe, N.J., Zurell, D., Elith, J., König, C., Fandos, G., Malchow, A., Kéry, M., Schmid, H., Guillera-Arroita, G., 2021. Can dynamic occupancy models improve predictions of species' range dynamics? A test using Swiss birds. *Glob. Chang. Biol.* 27 (18), 4269–4282. <https://doi.org/10.1111/gcb.15723>.
- 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>.
- Brown, J.H., Kodric-Brown, A., 1977. Turnover rates in insular biogeography: effect of immigration on extinction. *Ecology* 58 (2), 445–449. <https://doi.org/10.2307/1935620>.
- Detto, M., Visser, M.D., Wright, S.J., Pacala, S.W., 2019. Bias in the detection of negative density dependence in plant communities. *Ecol. Lett.* 22 (11), 1923–1939. <https://doi.org/10.1111/ele.13372>.
- DiRenzo, G.V., Hanks, E., Miller, D.A.W., 2023. A practical guide to understanding and validating complex models using data simulations. *Methods Ecol. Evol.* 14 (1), 203–217. <https://doi.org/10.1111/2041-210X.14030>.
- Eaton, M.J., Hughes, P.T., Hines, J.E., Nichols, J.D., 2014. Testing metapopulation concepts: effects of patch characteristics and neighborhood occupancy on the dynamics of an endangered lagomorph. *Oikos* 123 (6), 662–676. <https://doi.org/10.1111/oik.01008>.
- García-Valdés, R., Gotelli, N.J., Zavala, M.A., Purves, D.W., Araújo, M.B., 2015. Effects of climate, species interactions, and dispersal on decadal colonization and extinction rates of Iberian tree species. *Ecol. Modell.* 309–310, 118–127. <https://doi.org/10.1016/j.ecolmodel.2015.04.003>.
- Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. *Ecol. Lett.* 16 (12), 1424–1435. <https://doi.org/10.1111/ele.12189>.
- Hanski, I., 1998. Metapopulation dynamics. *Nature* 396 (6706). <https://doi.org/10.1038/23876>. Article 6706.
- Kellner, K. (2024). jagsUI: a wrapper around 'rjags' to streamline 'JAGS' analyses. *R Package Version 1.6.2*. <https://kenkellner.com/jagsUI/>.
- Kéry, M., Kellner, K.F., 2024. *Applied Statistical Modelling For Ecologists: A practical Guide to Bayesian and Likelihood Inference Using R, JAGS, NIMBLE, Stan and TMB*. Elsevier.
- Kéry, M., Royle, J., 2021. *Applied Hierarchical Modeling in ecology: Analysis of distribution, Abundance and Species Richness in R and BUGS: Volume 2: Dynamic and Advanced Models*. Elsevier, 1st ed.
- Kéry, M., Royle, J.A., 2016. *Applied Hierarchical Modeling in Ecology: Analysis of distribution, Abundance and Species Richness in R and BUGS: Volume 1: Prelude and Static Models*. Elsevier, 1st ed.
- Krebs, C.J., 2001. *Ecology: The experimental Analysis of Distribution and Abundance*. Benjamin Cummings, 5th ed.
- Lahoz-Monfort, J.J., Guillera-Arroita, G., Wintle, B.A., 2014. Imperfect detection impacts the performance of species distribution models. *Glob. Ecol. Biogeogr.* 23 (4), 504–515. <https://doi.org/10.1111/geb.12138>.
- MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, A.B., 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* 84 (8), 2200–2207. <https://doi.org/10.1890/02-3090>.
- McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models*. Chapman and Hall. <https://doi.org/10.1007/978-1-4899-3242-6>, 2nd Edition.
- Mckann, P.C., Gray, B.R., Thogmartin, W.E., 2013. Small sample bias in dynamic occupancy models. *J. Wildl. Manage.* 77 (1), 172–180. <https://doi.org/10.1002/jwmg.433>.
- Moilanen, A., 2000. The equilibrium assumption in estimating the parameters of metapopulation models. *J. Anim. Ecol.* 69 (1), 143–153. <https://doi.org/10.1046/j.1365-2656.2000.00381.x>.
- O'Connor, K.M., Nathan, L.R., Liberati, M.R., Tingley, M.W., Vokoun, J.C., Rittenhouse, T.A.G., 2017. Camera trap arrays improve detection probability of wildlife: investigating study design considerations using an empirical dataset. *PLoS. One* 12 (4), e0175684. <https://doi.org/10.1371/journal.pone.0175684>.
- Rana, D., Sartor, C.C., Chiaverini, L., Cushman, S.A., Kaszta, Z., Ramakrishnan, U., Macdonald, D.W., 2024. Differentially biased sampling strategies reveal the non-stationarity of species distribution models for Indian small felids. *Ecol. Modell.* 493, 110749. <https://doi.org/10.1016/j.ecolmodel.2024.110749>.
- Risk, B.B., de Valpine, P., Beissinger, S.R., 2011. A robust-design formulation of the incidence function model of metapopulation dynamics applied to two species of rails. *Ecology* 92 (2), 462–474. <https://doi.org/10.1890/09-2402.1>.
- Royle, J.A., Kéry, M., 2007. A Bayesian state-space formulation of dynamic occupancy models. *Ecology* 88 (7), 1813–1823.
- Xu, L., Clark, A.T., Rees, M., Turnbull, L.A., 2023. Estimating competition in metacommunities: accounting for biases caused by dispersal. *Methods Ecol. Evol.* 14 (1), 291–301. <https://doi.org/10.1111/2041-210X.14022>.
- Yackulic, C.B., Reid, J., Davis, R., Hines, J.E., Nichols, J.D., Forsman, E., 2012. Neighborhood and habitat effects on vital rates: expansion of the Barred Owl in the Oregon Coast Ranges. *Ecology* 93 (8), 1953–1966. <https://doi.org/10.1890/11-1709.1>.
- Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time. *Trends Ecol. Evol. (Amst.)* 16 (8), 446–453. [https://doi.org/10.1016/S0169-5347\(01\)02205-4](https://doi.org/10.1016/S0169-5347(01)02205-4).