Received: 24 March 2022 | Accepted: 7 October 2022

DOI: 10.1111/2041-210X.14030

Methods in Ecology and Evolution

PERSPECTIVE

A practical guide to understanding and validating complex models using data simulations

Graziella V. DiRenzo[1] | Ephraim Hanks[2] | David A. W. Miller[3]

[1]U. S. Geological Survey, Massachusetts Cooperative Fish and Wildlife Research Unit, University of Massachusetts, Amherst, Massachusetts, USA

[2]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, USA

[3]Department of Ecosystem Science and Management, Pennsylvania State University, University Park, Pennsylvania, USA

Correspondence
Graziella V. DiRenzo
Email: gdirenzo@umass.edu

Funding information
U.S. Government

Handling Editor: Holger Schielzeth

Abstract

1. Biologists routinely fit novel and complex statistical models to push the limits of our understanding. Examples include, but are not limited to, flexible Bayesian approaches (e.g. BUGS, stan), frequentist and likelihood-based approaches (e.g. packages LME4) and machine learning methods.
2. These software and programs afford the user greater control and flexibility in tailoring complex hierarchical models. However, this level of control and flexibility places a higher degree of responsibility on the user to evaluate the robustness of their statistical inference. To determine how often biologists are running model diagnostics on hierarchical models, we reviewed 50 recently published papers in 2021 in the journal *Nature Ecology & Evolution*, and we found that the majority of published papers did *not* report any validation of their hierarchical models, making it difficult for the reader to assess the robustness of their inference. This lack of reporting likely stems from a lack of standardized guidance for best practices and standard methods.
3. Here, we provide a guide to understanding and validating complex models using data simulations. To determine how often biologists use data simulation techniques, we also reviewed 50 recently published papers in 2021 in the journal *Methods Ecology & Evolution*. We found that 78% of the papers that proposed a new estimation technique, package or model used simulations or generated data in some capacity (18 of 23 papers); but very few of those papers (5 of 23 papers) included either a demonstration that the code could recover realistic estimates for a dataset with known parameters or a demonstration of the statistical properties of the approach. To distil the variety of simulations techniques and their uses, we provide a taxonomy of simulation studies based on the intended inference. We also encourage authors to include a basic validation study whenever novel statistical models are used, which in general, is easy to implement.
4. Simulating data helps a researcher gain a deeper understanding of the models and their assumptions and establish the reliability of their estimation approaches. Wider adoption of data simulations by biologists can improve statistical inference, reliability and open science practices.

KEYWORDS
Bayesian, frequentist, goodness-of-fit, hierarchical models, occupancy model, power analysis, statistical properties, study design

1 | INTRODUCTION

Ecologists and evolutionary biologists increasingly use complex hierarchical models to answer novel questions of theoretical and practical importance (e.g. Conn et al., 2018; Hooten & Hobbs, 2015; Kéry & Royle, 2016, 2021; Kéry & Schaub, 2012). For example, 52% of 50 recently published papers in 2021 in the journal *Nature Ecology and Evolution* use a hierarchical model to analyse their data (Table S1). Examples of hierarchical models include generalized linear mixed models (GLMMs, Bolker et al., 2009), latent state models with some observation process (e.g. occupancy models MacKenzie et al., 2002; Tyre et al., 2003) and mixture models (Kéry & Schaub, 2012). Several factors have contributed to the increased application of hierarchical models in ecology and evolutionary biology (e.g. Bolker et al., 2009; Kéry & Royle, 2016, 2021; Kéry & Schaub, 2012). First, with greater access to community science data, open-source datasets, genomic data and long-term ecological research (e.g. Dryad®, GenBank®, TreeBASE®), biologists can ask bigger and more complicated questions, which typically lead to the use of more complicated modelling methods. Second, more biologists are learning to use flexible programming languages that facilitate writing tailored complex hierarchical models (e.g. BUGS Sturtz et al., 2005, JAGS Plummer, 2003, stan Carpenter et al., 2017, package LME4 Bates et al., 2015, machine learning methods Joseph, 2020). Lastly, policy and conservation decision-makers are increasingly relying on the insights from complex datasets to guide their actions (Runting et al., 2020).

However, when practitioners fit custom-built hierarchical models, their methods are often largely untested (e.g. Conn et al., 2018; Hooten & Hobbs, 2015). As the complexity of hierarchical models increases, it becomes increasingly difficult to intuitively understand the assumptions, uncertainty and potential biases of the specified model. For example, a recent paper published in *Science* used a hierarchical statistical model to examine the effects of climate change on bumble bee occupancy (Soroye et al., 2020), and a follow-up study using data simulations showed that the hierarchical model used was not robust to violations of model assumptions (Guzman et al., 2021). Although such cases may seem rare, it is likely more common than appreciated. To determine how often biologists are validating the results of the hierarchical models they use to analyse data, we reviewed 50 recently published papers in 2021 in the journal *Nature Ecology & Evolution*, and we found that the majority of published papers that used hierarchical models did *not* report any validation of the models (5 of the 26 papers checked the diagnostics and fit of hierarchical models; 19%; Table S1). Similarly, in the journal *Ecology*, only 25% of articles routinely report model diagnostics (Conn et al., 2018). Even more rarely do biologists report an evaluation of the soundness of their code or the reliability of their novel statistical models (i.e. are the statistical models unbiased? how robust are the models to violations in assumptions? Brown et al., 2018; Link et al., 2018). One reason for this lack of quantitative rigour is the absence of standard guidelines that would make it easy for biologists to evaluate their statistical models (Barraquand et al., 2014; Conn et al., 2018).

The goal of this paper is to lay out a framework for validation when complex hierarchical models are used. By validation, we mean, 'are the estimates we generate from a statistical model providing sound inferences (i.e. can we generalize the results)?' Thus, validation includes everything from whether code is correct, to whether parameters are identifiable and estimates unbiased, to whether our model can be robustly applied when assumptions are violated or new data are collected. In the real world, however, we rarely know true values of ecological parameters of interest (e.g. Kéry & Schaub, 2012); thus, in most cases, our ability to test and validate statistical methods relies on simulating datasets where truth is set and known by the user (e.g. Kéry & Royle, 2016; Kéry & Schaub, 2012). Simulated data provide an opportunity to compare different properties of our statistical estimators to the true parameter values used to generate them and to evaluate model behaviour or performance. To determine how often biologists simulate data, we reviewed 50 recently published papers in 2021 in the journal *Methods Ecology & Evolution* (Table S2), and we found that 78% of the papers that proposed a new estimation technique, package or model used simulations or generated data in some capacity (18 of 23 papers). However, even in this journal the approaches used by authors varied greatly. For example, only five of the 23 papers included a basic demonstration that code can recover realistic estimates for a dataset with known parameters. Similarly, only nine of the 23 papers included simulations that demonstrate the statistical properties of an approach (i.e. quantifying accuracy, precision, bias and coverage of the estimator). As demonstrated by our review of papers published in *Nature Ecology & Evolution*, validation is even less common in journals not focused on methods development, despite most applications of complex hierarchical models using novel methods.

While simulation studies are a natural tool for understanding and validating the statistical properties of a method, model or analysis, there is no clear standard for when ecologists can use simulation studies, and which simulation studies are useful in different scenarios (e.g. Olivetti et al., 2021; Rossman et al., 2016; Smith et al., 2021; Tingley et al., 2020). Therefore, in this paper, we provide a guide to simulation studies for biologists. Specifically, we present a taxonomy of simulation study types based on the intended inference, with two broad divisions: (1) *study-specific simulations* (i.e. studies focused on a particular ecological system, such as an analysis of an ecological dataset aimed at answering a scientific question relevant to that ecological system) and (2) *general property simulations* (i.e. studies focused on methods and guidelines for adoption in future studies). We provide general guidelines on what questions each simulation study can help answer, and we encourage authors to at a minimum include

an easily implementable basic validation simulation whenever novel statistical models are used. In an effort to facilitate the implementation of these methods, we provide a running example throughout the text with fully reproducible code in R (R Core Team, 2021) and Nimble (de Valpine et al., 2017, 2022a, 2022b). This running example takes advantage of a common hierarchical model in ecology—the occupancy model (MacKenzie et al., 2002; Tyre et al., 2003)—in which the ecological process is decomposed from the sampling process. We suggest that new statistical models be accompanied by data simulations to avoid erroneous conclusions and to avoid the use of biased models in policy and decision-making, just as it has become standard practice that field studies are accompanied by laboratory experiments to validate conclusions (Kéry & Royle, 2016).

2 | USES OF DATA SIMULATION STUDIES IN ECOLOGY AND EVOLUTIONARY BIOLOGY

Simulation studies are valuable for a wide range of analyses conducted using a wide range of statistical frameworks. Frequentist, Bayesian and optimization-based machine learning approaches all lend themselves equally well to simulation studies (Muff et al., 2020; Weber et al., 2021). Similarly, simulation studies are valuable when inference is conducted analytically (i.e. when using an analytic maximum likelihood estimator and analytic asymptotic confidence intervals) or numerically (i.e. when using numerical optimization Kendall et al., 1997; Kendall & Nichols, 1995; Mackenzie & Royle, 2005).

It is important to clearly identify the goal of any simulation study and to identify the statistics of interest that will help address this goal (see Table 1 for a list of common goals of simulations studies; Kéry & Royle, 2016). The basic steps of conducting any simulation study are to (1) simulate one to many unique datasets using a data generating model (referred to as *M*), (2) estimate the desired parameters (or other statistics) using the statistical model of interest (referred to as *A*) and (3) summarize the performance of those estimates using Monte Carlo methods (see Box 1 for a more detailed algorithm). Monte Carlo approaches are those which rely on random variables simulated from a distribution, instead of the theoretical properties of the distribution itself (see Rizzo, 2019 for an introduction). Two very common uses of Monte Carlo methods are Markov Chain Monte Carlo (MCMC) methods used primarily in Bayesian statistical analysis (Gelman et al., 1995) to draw samples from a posterior distribution, and simulation studies, which we focus on here. A simulation study is simply the process of drawing samples from a distribution of a desired statistic, and using those samples to understand the statistical properties, like bias and variance, of a statistic (a function of the random samples). The questions which can be addressed by a given simulation study depend heavily on (1) the way random samples are drawn and (2) the statistics, or quantities, of interest. In this paper, we outline a classification of simulation studies, and provide an illustration of many common types of simulation studies.

The first classification of simulation studies, which we refer to as *study-specific simulations*, are methods appropriate when the goal is to validate the analysis of a dataset already in hand and the interpretation of the ecological results of the analysis in the context of the

TABLE 1 Summary table of the taxonomy of simulation studies used to understand and validate statistical models. Under each category, there are three types of simulation studies. For each type of simulation study, we summarize the types of questions that each aid in answering alongside the goals of the simulation

Category	Type of simulation study	When to use this simulation study?	
		Questions to answer	Goal
Study-specific simulations	Basic validation simulation	Are the parameters identifiable? Is this model computationally feasible?	Explore and verify identifiability of model parameters, given available data
	Determining statistical properties	What are the basic statistical properties of my model under a standard set of conditions? How does the approach perform with respect to parameter accuracy, bias, precision and coverage? What are the computational requirements/time of running the model?	Understand statistical properties of an algorithm output
	Assessing goodness-of-fit	Does my model sufficiently and accurately explain my data?	Understand if the estimator can sufficiently reproduce the observed data
General property simulations	Simulation-based study design	How many samples will be needed to generate quality estimates? What is the optimal allocation of samples? Where and when do I collect samples?	Understand sampling design requirements for robust inference
	Assessing statistical robustness	What are the properties of the estimator across different parameter spaces? Can the approach be applied to different conditions?	Understand model performance under a wide array of parameter conditions
	Comparing the efficacy of different approaches	What happens when data violate model assumptions? How do different approaches perform under non-optimal conditions?	Understand model performance when assumptions are violated

BOX 1 General framework for data simulations

The general procedure for a simulation study can be defined by three steps (Figure B1.1).

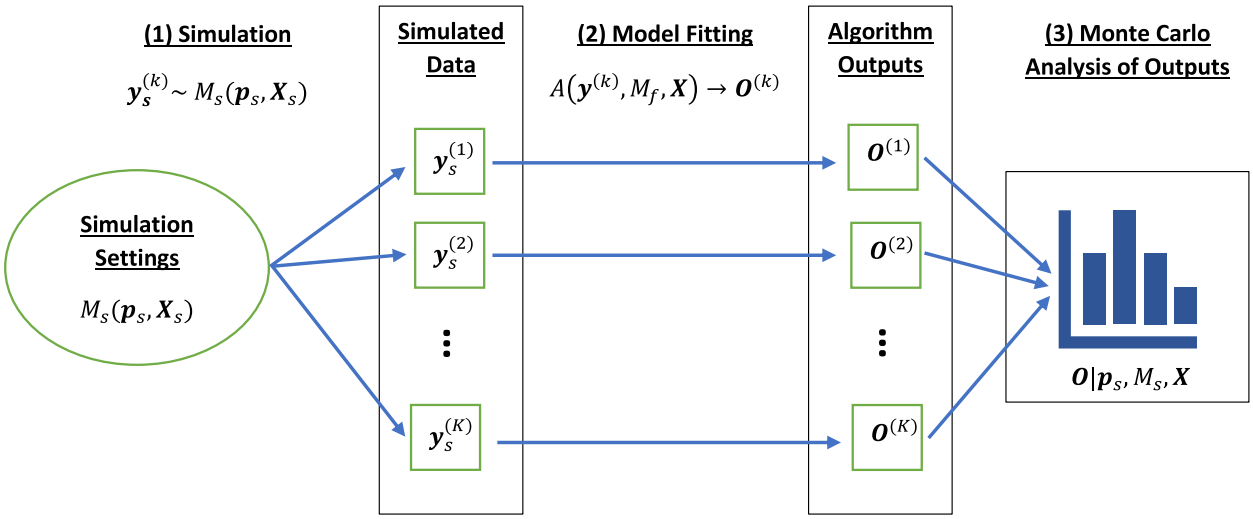

FIGURE B1.1 Graphical illustration of the steps of a simulation study. These steps are described in detail in this box. [Correction added on 14 December 2022, after first online publication: Figure B1.1 has been revised].

(1) Simulation. To simulate data, the user specifies a probabilistic model M_s for simulation, parameters p_s of that model and independent variables X_s. The ability to vary the model, parameters and independent variables determine the types of inferences that can be made from a simulation study. Thus, it is common for simulation studies to consider a range of simulation settings. We denote the kth simulated dataset as $y_s^{(k)}$, with the 's' subscript denoting that this is a 'simulated' dataset, and the '(k)' superscript denoting the replicate number, with 'k' ranging from 1 to K, the total number of simulations.

(2) Model Fitting. The goal of most simulation studies is to understand the distributional properties of a 'statistic', which is often an estimate of a parameter or a summary of a set of data. We assume an algorithm, A, is applied to a simulated dataset y_s and returns a set of statistics or outputs O_s, and the user has a lot of flexibility in defining the algorithm, A. Note that the algorithm is not always the statistical model, M_f, and the statistical model often matches the data generating model. The statistics or algorithm outputs of interest may be an estimate $\hat{\beta}$ of a parameter, the width or coverage of a credible interval or a p value or test statistic associated with a parameter.

(3) Monte Carlo Analysis. In most cases, many datasets will be simulated and analysed, and we summarize these results. To accomplish this, after simulating data $\left\{y_s^{(1)}, y_s^{(2)}, \ldots, y_s^{(K)}\right\}$ and calculating the statistics of interest $\left\{O^{(1)}, O^{(2)}, \ldots O^{(K)}\right\}$, the distributional properties of $O \mid M_s, p_s, X_s$ are explored by use of the samples $\left\{O^{(1)}, O^{(2)}, \ldots O^{(K)}\right\}$ from this distribution. Exploring or estimating properties of a distribution using random samples from that distribution is referred to as a 'Monte Carlo' analysis. For example, the mean of the statistics $E(O \mid M_s, p_s, X_s)$ could be approximated using the sample mean, $E(O \mid M_s, p_s, X_s) \approx 1/K \sum O^{(k)}$.

General considerations for simulation studies

Monte Carlo approaches are approximate approaches, and their accuracy depends on the number of simulated datasets generated (i.e. accuracy increases as the sample size increases). Often the goal of a simulation study is to estimate a property of a distribution—for example, 'what is the expected distribution of parameter estimates given some true value of the parameter?'. After conducting a simulation study, along with the estimate of interest, it is possible to compute a standard error of that estimate using the Monte Carlo samples. Following Koehler et al. (2009), the Monte Carlo standard error of $\hat{O}$ is:

$$MCSE\left(\hat{O}\right) = \sqrt{Var\left(\widehat{O^{(k)}}\right)},$$

where the variance is calculated over the K outputs from the simulation. This provides a straightforward way to assess whether or not more simulations are needed. In general, let S be the statistic of interest (i.e. the power of a test, or the upper bound of a 95% CI of a

parameter) which will be approximated using simulation *s*. An approximate 95% confidence interval of the statistic *S* is $\hat{S} \pm 1.96 * MCSE\left(\hat{S}\right)$. As the number *K* of the simulation study replicates increases, the $MCSE\left(\hat{S}\right)$ will converge to the standard error of the estimate $\hat{S}$. A good visual check of the effect of replication size *K* of the simulation study is a plot of $MCSE\left(\hat{S}\right)$ for increasing values of *K*. If the resulting plot shows convergence of the MCSE, then it is clear that the size of the simulation study is high enough that the error in estimation is due mostly to standard uncertainty that is associated with any estimate based on a finite dataset, and not strongly being driven by not having enough replicates in the simulation study to effectively quantify uncertainty.

system where the data were collected (McClintock, 2021; Palencia et al., 2021; Santos-Fernandez & Mengersen, 2021). Simulation studies of this type can range from a basic validation of the code and model to extensive explorations of the model properties as it relates to analysing the specific dataset.

The second classification of simulation studies, which we refer to as *general property simulations*, are methods used when determining the efficacy of applying a novel analytical approach for the design and analysis of future studies (Bellier et al., 2016; Rossman et al., 2016; Tingley et al., 2020; Zipkin et al., 2017). Simulations in this category can validate model performance across a broad parameter space, guide data collection and study design, determine how robust an approach will be to assumption violations and provide guidance regarding the relative performance of multiple analytical approaches.

In the following sections, we describe three types of *study-specific simulations* and three types of *general property simulations*. For each type of simulation, we include a worked example using detection/non-detection data for the spatial distribution of Cape Weavers in South Africa (Clark & Altwegg, 2019b; see Box 2 for a description of the dataset). We provide code to reproduce all model fitting and simulation studies discussed in the text, which can be found in Supplement S1 (DiRenzo, Hanks, et al., 2022). In choosing an example dataset we are left with the challenge of choosing an example that easily illustrates concepts, while also meeting our definition of a complex hierarchical model. Our example dataset consists in its simplest form as an exercise in building a regression model to predict the expected probability of observing a specific type of bird at a specific location given specific survey conditions. It also includes two types of structure typical of many hierarchical models. First, there is non-independence in the data that must be controlled for using a random-effects structure. Second, the true state of the system is not observable (i.e. we can only observe whether the bird is detected and not whether it is actually present at a site) and thus it includes a latent variable that is linked to data through an observation model. It mirrors the structure of many other models used by ecological and evolutionary researchers such as those used for GLMMs (Harrison et al., 2018), phylogenetic analyses (Revell & Harmon, 2022), for hierarchical data collection (Miller & Grant, 2015) or for predicting system dynamics (Buderman et al., 2020).

Lastly, we note that in this paper we only focus on simulation studies as a tool for understanding and validating statistical models and methods. Simulations are also a critical component of the standard parameter estimation toolkit (e.g. bootstrap approaches to hypothesis tests, simulation-based inference such as particle filter) (Lahiri, 2005; Loh & Stein, 2004) and a critical tool for prediction and forecasting (Bergmeir et al., 2018; Pagel & Schurr, 2012), neither of which are addressed here.

3 | STUDY-SPECIFIC SIMULATIONS

Study-specific simulations studies are appropriate when the focus of the scientific study is the analysis of a single dataset with the goal to understand the system being studied. Below we review study-specific simulations that accomplish three goals: (1) basic validation simulation, (2) determining statistical properties and (3) assessing goodness-of-fit.

3.1 | Basic validation simulation

3.1.1 | Objective

The goal of a basic validation simulation is to determine whether the model, fitting algorithm and code can generate realistic parameter estimates for an observed dataset (Table 1). This serves as a bare minimum check of code and model validity (Kéry & Schaub, 2012). It also provides a description for the data generating model *M* (Box 1) that can be used to evaluate the model assumptions and be used as a template for more extensive validation methods, such as evaluating bias and interval coverage statistical properties. A basic validation simulation can help confirm parameter identifiability given the available data, illuminate weaknesses in the model and fitting algorithms for a given dataset, identify when major issues (e.g. coding errors or model identifiability) have occurred, and provide a minimum threshold of evidence that these issues are not likely to exist in a particular study. In addition, inclusion of a basic validation simulation in a published paper facilitates a more open, transparent and reproducibility approach for the implementation of novel analytical methods. Therefore, a basic validation can be an important contribution when a novel model is fit or when new code is developed for a model. A basic validation can be comprised of relatively minor computing, as it is comprised of only fitting the statistical model twice (once for the analysis of the observed data and once for a single simulated dataset).

BOX 2 Spatial occupancy modelling of Cape Weaver in South Africa

As part of the second Southern African Bird Atlas project, community scientist birders were asked to spend at least 2 h on a checklist, recording all species they observed and the order in which they were observed. Here, we consider only the data related to the Cape Weaver *Ploceus capensis* collected by this community science project, as made available by Clark and Altwegg (2019b) in Clark and Altwegg (2019a). A total of 9356 recorded detection/non-detection observations are available for this species (Figure B2.1a,b). Clark and Altwegg (2019b) use two principal components (PC) to summarize multiple spatial covariates, with PC1 interpretable as a temperature-related factor (Figure B2.1c) and PC2 interpretable as a measure of climate intensity (Figure B2.1d). As a measure of observation accuracy of each individual birder, Clark and Altwegg (2019b) used the total number of species observed by the birder as a covariate on detection probability.

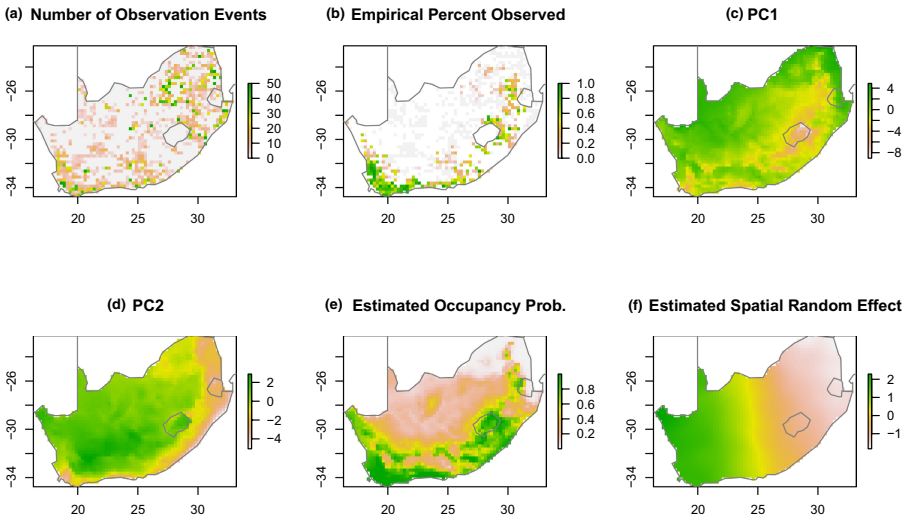

FIGURE B2.1 Summary of the Cape Weaver dataset from South Africa: (a) Depicts the number of observation events per location. (b) Shows the empirical percent observed. (c) Shows the spatial mapping of principal component PC1, and (d) shows the spatial mapping of principal component PC2. (e) Shows the estimated occupancy probability of the Cape Weaver across South Africa using our spatial occupancy model (Equations (1)–(5)), and (f) shows the estimated spatial random effect across South Africa. [Correction added on 14 December 2022, after first online publication: Figure B2.1 has been revised].

In the following, we give a brief ecological description of the model and its structure. The goal of the model is to estimate the probability the Cape Weaver occurs at a given location across the study area. The model we describe will include two hierarchical components, each of which add greater structure and complexity as compared to a standard logistic regression analysis. First, the probability of observing a Cape Weaver during a survey is a function of not only whether the species is present, but also whether it is detected given it does occur at the location. To account for this, a nested model is used to estimate the probability of observing the species as a function of both whether it is present and whether it is detected given it is present. Second, we want to account for spatial dependency among observations and this is done by including a spatial random effect. However, as noted in the text, accounting for spatial dependency in an unbiased manner is a non-trivial problem.

To fit data, we considered a spatial occupancy model for this dataset, with binary observations y_{si} being the ith observation at spatial location s. As noted, above, the probability $y_{si} = 1$ depends both on whether the Cape Weaver is present and whether it is observed. Therefore, we model the probability of detecting a bird during a survey as:

$$y_{si} \sim \text{Bern}(z_s * p_{si}), \tag{1}$$

where $z_s = 1$ if Cape Weaver occupies the sth spatial location and $z_s = 0$ if not (i.e. z_s is the latent true occupancy), and p_{si} is the probability of detection for the ith observation at location s. We model detection probability using a probit regression model and as a function of observer experience, with

$$\Phi^{-1}(p_{si}) = \alpha_0 + \alpha_1 w_{o(si)}, \quad (2)$$

where $w_{o(si)}$ is the total number of species observed in the second South African Bird Atlas project by the observer o who made the ith observation at spatial location s (see Clark & Altwegg, 2019b for additional explanation). The second component of the probability of observing a bird is whether it is actually present at the location, which also happens to be the variable we are most interested in estimating. This probability is latent, and can be modelled using a probit regression model where:

$$z_s \sim \text{Bern}(u_s), \quad (3)$$

$$\Phi^{-1}(u_s) = \beta_1 + \beta_2 x_{1s} + \beta_3 x_{2s} + \eta_s, \quad (4)$$

where x_{1s} and x_{2s} are the first two principal components described above (Figure B2.1c,d). Spatial autocorrelation in occupancy is modelled by a spatial random effect η_s, which we model using a basis function approach (Cressie et al., 2022) with basis vectors constructed from the first M eigenvectors of the inverse of an Intrinsic Conditional Auto-Regressive (ICAR) precision matrix. This differs slightly from Clark and Altwegg (2019b), who used similar eigenvectors, but first removed some correlation between the spatial random effect and the fixed effects.

$$\eta_s = \sum_{m=1}^{M} \gamma_m \mathbf{v}_m, \gamma_m \sim N(0, \sigma_m^2), \quad (5)$$

where $\mathbf{v}_m$ is the mth eigenvector and σ_m^2 is the corresponding mth eigenvalue. All regression parameters are assigned diffuse (variance = 100) zero-mean Gaussian priors.

We specified diffuse Gaussian priors on all regression parameters, and diffuse half-normal priors on all variance parameters, and fit this BHM using MCMC. All computing was done using the NIMBLE R-package (de Valpine et al., 2017, 2022a, 2022b). We ran the MCMC sampler for 20,000 iterations and removed the first 10% of the chain as burn-in. The posterior mean occupancy probabilities are shown in Figure B2.1e, and the estimated spatial random effect is shown in Figure B2.1f.

3.1.2 | Simulation settings

A basic validation simulation consists of first fitting the statistical model, A, used for the scientific analysis using the observed data (y^* and X^*) to generate parameter estimates ($\hat{p}^*$; Kéry & Schaub, 2012). Then, the parameter estimates are used to simulate a single new dataset from the simulation distribution, such as $y_s \sim M(\hat{p}^*, X^*)$. This simulated dataset is the same size as the observed data and uses the same covariates, spatial locations and settings (X^*) as the observed data.

3.1.3 | Model fitting

The same statistical model used for the original fit, A, is then fit to this simulated dataset (Kéry & Schaub, 2012). The parameter estimates, confidence or credible intervals, and other model diagnostics are checked to make sure that the results are reasonable given the parameter values. For example, do most parameter estimate include the true value in the 95% CI? or, are credible interval widths narrow enough to suggest that there is sufficient power to estimate a parameter? etc.

3.1.4 | Example

Here, we perform a basic validation simulation of a spatial occupancy model that is fit to the Cape Weaver dataset (Box 2). First, we fit the spatial occupancy model described in Box 2 (Equations (1)–(5)) to the observations of Cape Weavers, and our model parameter estimates (specifically the posterior means) are shown graphically in Figure 1a as vertical dashed grey lines. For brevity and highlighting interesting results, we only display parameters β_2 and β_3 in Figure 1. These two parameters capture the relationship between local climate and the presence of the Cape Weaver. These parameters are both interesting as they capture important ecological relationships and because they have a high degree of spatial structure, which as we explain below is relevant to our ability to estimate parameters. Full results are found in Appendix Figure A1.

After fitting the statistical model to the observation data, we simulated data based on the parameter estimates from the field data to perform a basic validation. Note that simulating data presented a specific challenge in our case, which was to capture the spatial random process estimated in our model. As our spatial occupancy model includes a latent spatial random effect, our simulation was done hierarchically by:

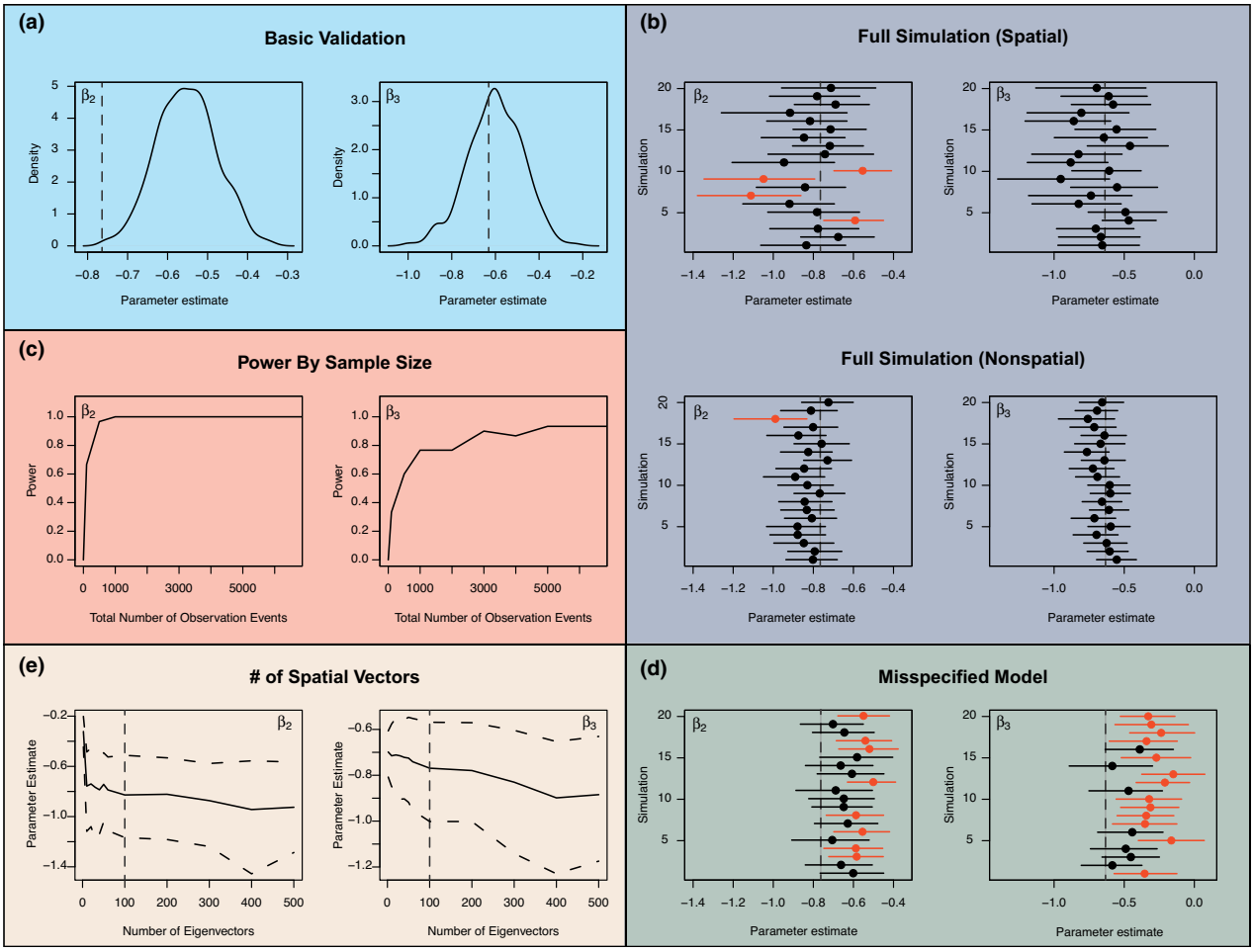

FIGURE 1 Simulation study results. Panel (a) shows the density plots of the posterior distributions of model estimated parameters from a simulated dataset demonstrating a basic validation simulation. The vertical dashed grey line represents the model parameter estimates when fitting the Cape Weaver dataset using a spatial occupancy model. Results suggest that the spatial occupancy model may have difficulty recovering parameter β_2. Panel (b) shows the 95% CIs of the posterior distributions of β_2 and β_3 for the first 20 of the simulated datasets under the spatial and nonspatial models. The point ranges highlighted in red show simulation runs that do not overlap with the true parameter value, which is represented by the horizontal vertical line. Point ranges in black represent simulation runs that do overlap with the true parameter value. Results suggest that β_2 suffers from identifiability due to spatial confounding. Panel (c) shows a power analysis for recovering estimates of β_2 and β_3. The results show that β_3 requires more independent samples than β_2 for consistent estimation. Panel (d) shows the results of a simulation study with model misspecification. Point ranges in red and black represent those that did not and that did overlap with the true parameter value, respectively. We simulated data with observer heterogeneity in the detection process and analysed the data using a model that assumes homogeneity in detection. Results show that ignoring heterogenous detection probabilities can lead to bias in the estimates of β_2 and β_3, as shown by the lack of overlap between the 95% CIs of the posterior distributions of β_2 and β_3 with the true parameter value (grey dashed vertical line). Panel (e) is a comparison of models with different numbers of spatial basis functions. We find that inference on β_2 and β_3 is relatively stable when more than 100 basis functions are used. [Correction added on 14 December 2022, after first online publication: Figure 1 has been revised].

1. Simulating the parameters in the spatial random effect $\gamma_m \sim N(0, \sigma_m^2)$ and then creating the simulated spatial random effect with $\eta_s = \sum_{m=1}^{M} \gamma_m \mathbf{v}_m$.
2. Using this spatial random effect and the existing covariates, we simulated first true spatial occupancy (z) and then observations of detection/non-detection. These simulations used Equations (1)–(5) in Box 2 with all α and β parameters being set to their posterior means from our statistical model fit using a spatial occupancy model to the Cape Weaver dataset.

These specifications create a simulated dataset with the exact same scientific settings as our observed data. We then completed the loop by fitting our spatial occupancy model to this new simulated data, using the same MCMC algorithm used to fit the model to the observed detection/non-detection data.

Density plots of the posterior distribution of model parameters, given the simulated data, are shown in Figure 1a, with full results in Appendix Figure A1a–f. The posterior distributions for most model parameter estimates overlapped the value used to simulate

the data. This is what we hope to see with a basic validation, that our model fit to simulated data generates results that are consistent with the parameters used to simulate that data. However, the posterior distribution for β_2 does not overlap the value used to simulate the data, suggesting that the spatial occupancy model may have difficulty in estimating this parameter. This basic validation has no replication, so it is not immediately clear if the results we see are indicative of something systematically wrong with our analysis (i.e. non-identifiability of parameters, an error in our code), or just an extreme resulting from normal sampling variation. Comparing the true occupancy probabilities used to simulate datasets with the corresponding values estimated by the statistical model indicates that we are able to estimate the parameters used to generate the simulated data reasonably well (Figure A1f). Given these results, our basic validation study indicates that we may have some difficulty with the identifiability of β_2, given the available data. Note our use of the term 'indicates' in describing these results. Confirming the fit and identifiability of the statistical model for the simulated datasets entails simulating more than a single dataset, as shown in the next section. By completing the basic validation, we have greatly increased the likelihood of identifying major coding errors, issues of identifiability, model misspecification and potential power issues. The next sections outline how that inference can be strengthened using a more comprehensive approach to estimate the distribution of outcomes expected for an estimation approach and to identify whether real-world data used to fit the model are consistent with the assumed distributions underlying our approach.

3.2 | Determining statistical properties

3.2.1 | Objective

A more robust exploration would help determine the full statistical properties of an estimation approach, such as when there are concerns of parameter identifiability, the behaviour of model parameters (e.g. estimating bias, accuracy and precision) or when there are questions about whether interval coverage is well calibrated (Table 1). To do this, we would use a full simulation study. A 'full simulation study' means that 100s to 1000s of simulations are performed, and Monte Carlo methods can be used to understand estimator properties of the simulated datasets, which ought to resemble the observed dataset in hand (e.g. DiRenzo, Miller, et al., 2022; Rossman et al., 2016; Tingley et al., 2020).

When evaluating statistical properties, there are several metrics of potential interest: accuracy, precision, bias and coverage. Each metric has lots of ways of being calculated. Here, we present definitions and a couple of ways to calculate each metric. Accuracy answers the question 'how close are model estimates to true values?' and can be quantified multiple ways. For example, accuracy can be calculated as the mean error by taking the absolute difference between the model estimate and truth. Alternatively, accuracy can be calculated using mean squared error (MSE) methods, giving greater weight to big differences when assessing performance. Precision answers the question 'how large is the 95% credible interval?'. Again, multiple measures of precision exist, including calculating CI width (1) by subtracting the lower 95% CI estimate from the upper 95% CI estimate or (2) by estimating standard error of an estimate. Bias answers the question 'what are patterns of parameter over- versus under- estimation?' For simulation methods, bias can be estimated by subtracting the average model estimate across many simulated datasets from the true parameter value. Lastly, coverage answers the question 'how often does the true parameter value fall within the range of the 95% CI?', and it can be obtained by calculating the proportion of simulations where the true value fell within the 95% CI of the model estimate.

3.2.2 | Simulation settings

The process for simulating data to determine statistical properties is identical to the process of simulating data for a basic validation above, except 100s to 1000s of unique simulated datasets are generated using the estimated parameter values ($\hat{p}^*$) rather than the single dataset (e.g. DiRenzo, Miller, et al., 2022; Rossman et al., 2016; Tingley et al., 2020).

3.2.3 | Model fitting

Again, the statistical model is fit to each of the simulated datasets (e.g. DiRenzo, Miller, et al., 2022; Rossman et al., 2016; Tingley et al., 2020). Once all datasets are fit, Monte Carlo methods are used to examine the frequentist properties of the 100s to 1000s of simulated datasets.

3.2.4 | Example

Continuing the spatial occupancy model example from above, we next conducted a full simulation study to explore the statistical properties of model parameters. We focus again on our ability to estimate the relationship between each of our covariates and the probability, a Cape Weaver occurs at a location. As was suggested by Clark and Altwegg (2019b), and also explored by Hanks et al. (2015), Hodges and Reich (2010), and Paciorek (2010), and others, parameter identifiability in spatial regression models when the predictor variables are spatially structured (or spatially autocorrelated) can be challenging and our basic validation suggested that identifiability may be an issue in our case (Figure 1a). Thus, we simulated 100 datasets from the spatial occupancy model using the posterior means obtained for parameters when the model was fit to the observed dataset as the 'true' value. We also simulated 100 datasets from the spatial occupancy model with the spatial component set to zero. Given our suspicion that spatial structure would be an issue, this second set of simulations gave us a

reference to test whether this was the case and to determine if the biases we observed are in fact due to spatial non-independence. Next, we fit the spatial occupancy model to each of these 200 simulated datasets. Figure 1b shows the 95% CIs of the posterior distributions of β_2 and β_3 for the first 20 of the simulated datasets under the spatial model and the corresponding 95% CIs for β_2 and β_3 estimated from the simulated datasets under the nonspatial model. Our expectation if our estimates were unbiased and we correctly were estimating precision was that on average 19 out of 20 times the true value would occur in the 95% CI.

When data are simulated without spatial autocorrelation, we see that the 95% CIs are well-calibrated, with the posteriors for α_2, β_2 and β_3 all overlapping the true parameter used for simulation a large proportion of the time (93%, 96% and 98%, respectively; Figure A1j–l). However, when the data are simulated with spatial autocorrelation, we see that the credible intervals for β_2 and β_3 often do not overlap the true parameter (Figure A1g–i). This simulation study illuminates how spatial confounding, as described by Hodges and Reich (2010), Hanks et al. (2015), Silk et al. (2020) and others, can result in biased parameter estimates, especially when covariates are spatially smooth, as are the temperature and climate covariates associated with β_2 and β_3. The reason for this confounding boils down to the correlation between the covariate and the spatial autocorrelation (Hanks et al., 2015). The detection covariate associated with α_2 is much less spatially smooth, as multiple different individuals (each with different levels of the detection covariate w_o) often make observations at locations close in space. We see from the simulation results that spatial confounding is much less pronounced when covariates have less spatial structure, like the detection covariate in this example.

3.3 | Goodness-of-fit assessments

3.3.1 | Objectives

Simulation studies also have an important role in goodness-of-fit assessment. Goodness-of-fit assessments are used to determine if the statistical model applied in the analysis can generate the observed data. In the case of goodness-to-fit assessments, the simulated data are compared to the observed data to determine whether the data fit model assumptions. Examples of commonly used goodness-of-fit assessments include Bayesian posterior predictive checks (Kéry & Schaub, 2012) and those used in the DHARMA package (Hartig, 2020) in R for generalized mixed effects models. Note that the focus on fit using simulations has shifted from whether the estimates are reasonable given the true values of the parameters (step 3 in Figure B1.1) to whether the simulated data are a reasonable approximation of our true data.

Assessing the goodness-of-fit allows biologists to answer the following questions:

- Can my model replicate or reproduce the patterns in my observed data?
- Does my model do an adequate job of representing my observed data?

Note that lack of fit does not always mean an estimator will perform poorly (or vice versa) in part because goodness-of-fit assessments are highly dependent on sample size to identify lack of fit. If lack of fit is identified, simulation studies can be used to determine how robust the estimator is to violation of assumptions (see 'Assessing statistical robustness' section below).

3.3.2 | Simulation settings

The simulation settings for the goodness-of-fit assessment are identical to those presented above for the 'Determining statistical properties' section. That is, many datasets are simulated from the fitted model. When Bayesian approaches are used, typically simulations are conducted using values from iterations of the posterior distribution.

3.3.3 | Model fitting

For each simulated dataset, estimates of the dataset characteristics (e.g. variance, frequency of zeros, measures of normality) are calculated and the distribution of these values is compared to the observed dataset.

3.3.4 | Example

The appropriate goodness-of-fit test to use for a dataset will vary among applications. For occupancy models, a common approach to assessing goodness of fit is to use Pearson Chi-square statistics (MacKenzie & Bailey, 2004), $X^2 = \sum_{s,i}\left(z_s - p_{si}\right)^2 / p_{si}$, where (as in Box 2) z_s is the true latent occupancy of site s, and p_{si} is the probability of detection at site s by observer i. This test focuses on determining whether the distribution of times a species is detected at a site is more variable than expected, and it can help identify unexplained variation in detection that can bias results. Previous work has shown that too much heterogeneity among sites can lead to bias in estimating occupancy probabilities (e.g. Ferguson et al., 2015; McNew & Handel, 2015). While in some cases, the distribution of X^2 is known, and the calculated value of X^2 can be compared with theoretical critical values, we instead illustrate the more general situation where the distribution of the statistic of interest is unknown, and goodness-of-fit assessment is carried out using Monte Carlo methods.

We first calculated the chi-square statistic $\widehat{X^2}$ using posterior mean estimates for all parameters to obtain $\widehat{p_{si}}$. We then simulated 1000 datasets using the posterior mean estimates of all parameters. Each of these datasets were fit using our Bayesian MCMC approach, and the resulting posterior means for each simulated dataset were used to compute corresponding Chi-square statistics.

This provides 1000 samples of the chi-square statistic under the null hypothesis that our model is correct. The rank of $\widehat{X^2}$ compared to these values provides a Monte Carlo *p*-value to test the null hypothesis that our model is correct. For this dataset, the Monte Carlo *p*-value was 0.866, indicating that we do not have strong evidence to reject the null hypothesis that our model is reasonable for this data. If the *p*-value was small (e.g. <0.05), we would have evidence that our model is missing something to accurately capture the variation in the observed dataset. *p*-values higher than 0.05 (especially *p*-values much higher than 0.05) indicate that there is no strong evidence that our model is missing something important. We note that, while we are using Bayesian methods, our approach to this simulation study example and calculating *p*-values is frequentist. We are interested in the statistical properties of a particular statistic—the posterior mean—and thus we are not conducting posterior predictive inference (Gelman et al., 1995), but rather we are taking a frequentist approach, with the statistics of interest being estimated quantities of a Bayesian posterior distribution.

4 | GENERAL PROPERTY SIMULATIONS

In this section, we review general property simulations, which are used when the goal is to make general recommendations regarding the efficacy of applying a novel analytical approach for the design and analysis of future studies.

4.1 | Simulation-based study design

4.1.1 | Objectives

There are many reasons to perform a simulation-based assessment of study design. First, a biologist may want to understand how the estimator performs across a wide range of parameter values and sample sizes when selecting a new statistical approach to design and analyse data for a new study (e.g. DiRenzo, Miller, et al., 2022). This information could assist others in deciding whether to adopt a method to analyse existing data or design new studies with the approach in mind. The second goal of simulation-based study design falls under the category of a power analysis, with a goal of understanding the effect that sample size has on our power to detect non-zero parameters when they occur, on the accuracy of model parameter estimates or on the predictive performance of the model (e.g. Guillera-Arroita & Lahoz-Monfort, 2012). Biologists are routinely interested in examining the effect of sample size during the study design phase when there are concerns about being able to collect enough observations, especially as they relate to Type I and II error, which occurs regularly during the early stages of an ecological project when designing field studies. Lastly, biologists may use simulation-based study design to evaluate model performance under different study designs (e.g. Wright et al., 2022). In this case, we can explore how varying sampling designs, such as stratified vs random sampling design, affect the sample size to ensure against Type I and II error, the model estimates of parameters in terms of accuracy and bias, and ecological inference across space and time.

4.1.2 | Simulation settings

Depending on the objectives of the simulation study, the values to be varied when simulating datasets may include the model parameters (*p*), the explanatory variables (*X*) or the sample sizes of the datasets (*n*). Varying model parameters allows for evaluation of how the estimator will perform across different ecological scenarios, while varying the distribution of the explanatory variables and the sample sizes will provide inference about optimal study design. Typically, a finite set of values across the range of the parameters are chosen, and many datasets (from 100s to 1000s) are simulated at each of these values. Alternatively, 'space-filling' designs can be used to sample across many combinations of parameter values (Carnell, 2022; DiRenzo, Miller, et al., 2022).

4.1.3 | Model fitting

The simulated datasets are fit to the model that the researcher plans to use for the analysis of the observations. The chosen metric for performance (e.g. root MSE as a measure of accuracy or standard error as a measure of precision) is calculated for each simulated dataset and summarized across the different parameter values that were varied.

4.1.4 | Example

In the context of our spatial occupancy model and Cape Weaver dataset, we consider a simulation study aimed at understanding how sample size of a spatial occupancy dataset effects estimator performance, which can be used to inform future studies. Here, we were interested in determining the degree to which reducing sampling effort will affect inference from our hierarchical model. Understanding how sample size influences our ability to estimate parameters can guide future survey efforts to ensure that limited monitoring resources are properly allocated. For our simulation study, we considered simulations that included only a subset of the observations in our Cape Weaver dataset. First, we randomly subsampled *N* of the 9356 observations in the detection/non-detection dataset without replacement, with *N* varying from 100 up to 7000. We then simulated 100 independent occupancy datasets at each value of *N*. For each simulated dataset, we fit the spatial occupancy model and estimated the power for each parameter. We define power as the proportion of the posterior 95% credible intervals for each parameter that did not overlap zero and showed the same sign (i.e. positive or negative) as the true, simulated parameter value.

The results for varying dataset sample sizes are shown in Figure 1c and Figure A1m. We see that the effect of heterogenous

detection (α_2) and the temperature-related principal component (β_2) are estimated well with a relatively small sample size ($N < 1000$), while the effect of the second principal component (β_3) suggests that a relatively larger sample size is needed to consistently observe an estimate that does not include zero in the CI.

4.2 | Assessing statistical robustness

4.2.1 | Objectives

Simulation studies often focus on simulating datasets using a data generation model that matches the statistical model (e.g. if residual variation in the statistical model is assumed to be normally distributed, the data generating model for the simulations will use a normal distribution). However, it is also important to understand the effect of model misspecification, which occurs when the data generating model does not match the model used for statistical analysis (e.g. Dennis et al., 2019; Dey et al., 2022; DiRenzo, Miller, et al., 2022). Examples include cases where the dataset has extra sources of heterogeneity, there are distributional mismatches between the statistical model and the data generating model, or extra explanatory variables are not included in the statistical model. The goal of this type of simulation study is to determine how impactful such misspecifications are on our desired scientific inference (Table 1). These simulation study approaches address misspecification by considering the case where we define a particular form of misspecification (i.e. ignoring an important explanatory variable). This can provide insight into what forms of misspecification are most impactful on the specific aims of a given study.

4.2.2 | Simulation settings

Data are simulated using a data generating model that does not match the structure of the statistical model used for the scientific analysis. In most cases, these data are compared to data from a data generating model that matches the statistical model as a baseline for comparison. Similarly, it is possible to vary parameter values in the simulations (see 'Simulation-based study design' section) to determine how the study design affects the robustness of the estimator.

4.2.3 | Model fitting

The model fitting procedure is the same as the one described in the 'Simulation-based study design' section above.

4.2.4 | Example

To illustrate how a simulation study can help understand the effects of model misspecification, we conducted a simulation study to explore the effect that ignoring heterogeneity in detection probabilities has on estimating the occupancy parameters using our spatial occupancy model applied to the Cape Weaver dataset. We simulated 100 datasets from our fitted model but where we also included heterogeneity in detection probabilities for each observer. We then fit our spatial occupancy model to this simulated data that assumed homogeneous detection probabilities (i.e. all observers had equal probability of observing a species when present).

Figure 1d shows the posterior 95% credible intervals for 20 such simulated datasets. In the figure, we see that the credible intervals for most simulated datasets do not overlap the true value used for simulation, indicating that the model does a poor job of recovering true parameter values when we ignore heterogeneity in detection probability. This result highlights the importance of testing for detection heterogeneity using goodness-of-fit methods (see Section 3 'Study-specific simulation studies') and for models that address heterogeneity when it occurs.

4.3 | Comparing the efficacy of different modelling approaches

4.3.1 | Objectives

In many cases, multiple modelling approaches will be available to estimate parameters of interest, and simulations as a tool can be used to compare the efficacy and robustness of the different modelling approaches. Here, a common goal is to compare a proposed approach with an existing approach (i.e. a more simplified version of a model) for estimating parameters or testing hypotheses (Table 1). Often, this is in conjunction with studying the effects of model misspecification (see 'Assessing statistical robustness' above), as we are considering comparing existing approaches to a novel approach that models more complexity.

4.3.2 | Simulation settings

The user can set the simulation settings to compare the efficacy of different modelling approaches to match the most important outcomes of the planned study. For example, simulation settings can follow those of the section 'Determining statistical properties' if interest is in a single set of parameters, 'Simulation-based study design' if interest is in performance across different parameters, or 'Assessing statistical robustness' if interest is in determining whether one approach is more robust to violations.

4.3.3 | Model fitting

For each dataset that is simulated, the data are analysed using multiple statistical models (e.g. DiRenzo, Miller, et al., 2022). Results across simulated datasets are summarized for each statistical model and their output is compared. The comparison among statistical

models is dependent on the practitioner and their needs. Widely used measures of model performance include: MSE (where the mean square difference between truth and a point estimate are calculated), mean square predictive error (where the mean square predictive difference between truth at an unobserved point and the statistical estimate at the unobserved point are calculated) and coverage (where a '1' is assigned if the 95% CI covers the true value and '0' if it does not).

4.3.4 | Example

In our case study of the Cape Weaver dataset, we have focused on a single specific model to account for spatial autocorrelation in the data. To this point, we modelled spatial autocorrelation in occupancy probabilities using a zero-mean Gaussian spatial random effect with an intrinsic conditional autoregressive (ICAR) precision matrix (Box 2). There are many ways to model spatial correlation and even within the method we chose, the settings can be varied. We now consider different spatial models for our dataset, to see how performance varies based on our choice of statistical model. As is common in the spatial statistical literature, we considered a basis function approximation to this spatial random effect (Cressie et al., 2022) with basis functions being the first m eigenvectors of the inverse of the ICAR precision matrix. These m basis vectors capture the most possible variation in the spatial random effect using only m vectors. For our initial data analysis, we chose $m = 100$ basis vectors (shown in the vertical line in Figure 1e). This choice was based on a comparison of the results from a varying set of basis vectors. We let m vary from 2 to 500 and fit the model.

Figure 1e shows the posterior mean and 95% credible intervals for different model parameters as a function of m. As m increases, the computational complexity of the model increases and the time to fit increases as well (see Appendix Figure A1t). These figures show that when $m < 50$, the posteriors for multiple parameters are very different than for larger values of m, but when $m > 75$, the posteriors are all reasonably similar. This simulation study shows that our approximate approach to modelling spatial autocorrelation by keeping only m eigenvectors provides a good approximation to the full model when $m > 75$.

5 | CONCLUSIONS

We delineate the uses and purposes of simulation studies to understand and validate hierarchical models. In doing so, we propose a new *status quo* for reporting of the properties of new and complex statistical models, as well as when applying them to routinely used statistical models. We encourage all studies that use a novel estimation procedure—whether it be an extension of existing methods, development of new code or employing a new procedure for fitting the model—to include at least a basic validation simulation. This step will help avoid many potential pitfalls in fitting a new model (e.g. error in coding, parameter identifiability issues). In addition, the inclusion of model code increases reproducibility and transparency in an age where open science in ecology and evolutionary biology is gaining traction (Powers & Hampton, 2019). By also including code that simulate a dataset and fits them to a statistical model, it will open doors to understanding the assumed data generation process underlying our statistical inferences. Simulations have an integral role in testing the reliability and limits of statistical inference, providing information about a statistical model's ability to recover accurate, precise, unbiased parameter estimates (Kéry & Royle, 2016, 2021; Kéry & Schaub, 2012). Simulations can also help answer many common questions asked by biologists, such as 'how many samples do I collect?', 'which model do I use to analyse my data?', 'does my model do an adequate job of representing my data?'. Simulations may therefore become an integral part of a biologist's tool kit.

AUTHOR CONTRIBUTIONS

Graziella V. DiRenzo, Ephraim Hanks and David A. W. Miller worked together to write the first draft. All co-authors edited the manuscript.

ACKNOWLEDGEMENTS

We thank L. M. Browne for constructive comments on a previous version of this manuscript. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

CONFLICT OF INTEREST

We have no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14030.

DATA AVAILABILITY STATEMENT

All data for analyses can be found in the Dryad Data Repository: https://doi.org/10.5061/dryad.jt2002k (Clark & Altwegg, 2019a). All code for analyses can be found in the Supplement S1 R code and: https://doi.org/10.5066/P99B0IJ7 (DiRenzo, Hanks, et al., 2022).

ORCID

Graziella V. DiRenzo https://orcid.org/0000-0001-5264-4762
Ephraim Hanks https://orcid.org/0000-0003-0345-7164
David A. W. Miller https://orcid.org/0000-0002-3011-3677

REFERENCES

Barraquand, F., Ezard, T. H. G., Jørgensen, P. S., Zimmerman, N., Chamberlain, S., Salguero-Gómez, R., Curran, T. J., & Poisot, T. (2014). Lack of quantitative training among early-career ecologists: A survey of the problem and potential solutions. *PeerJ*, *2014*(1), 1–14. https://doi.org/10.7717/peerj.285

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bellier, E., Kéry, M., & Schaub, M. (2016). Simulation-based assessment of dynamic N-mixture models in the presence of density dependence and environmental stochasticity. *Methods in Ecology and Evolution*, *7*(9), 1029–1040. https://doi.org/10.1111/2041-210X.12572

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, *120*, 70–83. https://doi.org/10.1016/j.csda.2017.11.003

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2563–2570. https://doi.org/10.1073/pnas.1708279115

Buderman, F. E., Devries, J. H., & Koons, D. N. (2020). Changes in climate and land use interact to create an ecological trap in a migratory species. *Journal of Animal Ecology*, *89*(8), 1961–1977.

Carnell, R. (2022). Lhs: Latin hypercube samples. R package version 1.1.5.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Clark, A. E., & Altwegg, R. (2019a). Data from: Efficient Bayesian analysis of occupancy models with logit link functions. *Dryad, Dataset*. https://doi.org/10.5061/dryad.jt2002k

Clark, A. E., & Altwegg, R. (2019b). Efficient Bayesian analysis of occupancy models with logit link functions. *Ecology and Evolution*, *9*(2), 756–768. https://doi.org/10.1002/ece3.4850

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to Bayesian model checking for ecologists. *Ecological Monographs*, *88*(4), 526–542. https://doi.org/10.1002/ecm.1314

Cressie, N., Sainsbury-Dale, M., & Zammit-Mangion, A. (2022). Basis-function models in spatial statistics. *Annual Review of Statistics and Its Application*, *9*(1), 1–28. https://doi.org/10.1146/annurev-statistics-040120-020733

de Valpine, P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, Wehrhahn Cortes C, Rodriguez A, Temple Lang D, Paganin S (2022a). *NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling*. https://doi.org/10.5281/zenodo.1211190, R package version 0.12.2, https://cran.r-project.org/package=nimble.

de Valpine, P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, Wehrhahn Cortes C, Rodriguez A, Temple Lang D, Paganin S. (2022b). *NIMBLE user manual*. https://doi.org/10.5281/zenodo.1211190, R package manual version 0.12.2, https://r-nimble.org.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*, 403–413. https://doi.org/10.1080/10618600.2016.1172487

Dennis, B., Ponciano, J. M., Taper, M. L., & Lele, S. R. (2019). Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution*, *7*, 372. https://doi.org/10.3389/fevo.2019.00372

Dey, S., Bischof, R., Dupont, P. P. A., & Milleret, C. (2022). Does the punishment fit the crime? Consequences and diagnosis of misspecified detection functions in Bayesian spatial capture–recapture modeling. *Ecology and Evolution*, *12*(2), 1–15. https://doi.org/10.1002/ece3.8600

DiRenzo, G. V., Hanks, E., & Miller, D. A. W. (2022). *A practical guide to understanding and validating complex models using data simulations. Version v1.0.0*. U.S. Geological Survey Software Release. https://doi.org/10.5066/P99B0IJ7

DiRenzo, G. V., Miller, D. A. W., & Grant, E. H. C. (2022). Ignoring species availability biases occupancy estimates in single-scale occupancy models. *Methods in Ecology and Evolution*, *13*, 1790–1804. https://doi.org/10.1111/2041-210X.13881

Ferguson, P. F. B., Conroy, M. J., & Hepinstall-Cymerman, J. (2015). Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. *Methods in Ecology and Evolution*, *6*(12), 1395–1406. https://doi.org/10.1111/2041-210X.12442

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Guillera-Arroita, G., & Lahoz-Monfort, J. J. (2012). Designing studies to detect differences in species occupancy: Power analysis under imperfect detection. *Methods in Ecology and Evolution*, *3*(5), 860–869. https://doi.org/10.1111/j.2041-210X.2012.00225.x

Guzman, L. M., Johnson, S. A., Mooers, A. O., & M'Gonigle, L. K. (2021). Using historical data to estimate bumble bee occurrence: Variable trends across species provide little support for community-level declines. *Biological Conservation*, *257*(July 2020), 109141. https://doi.org/10.1016/j.biocon.2021.109141

Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, *26*(4), 243–254. https://doi.org/10.1002/env.2331

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794. https://doi.org/10.7717/peerj.4794

Hartig, F. (2020). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R Package Version 0.3.3.0.

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *American Statistician*, *64*(4), 325–334. https://doi.org/10.1198/tast.2010.10052

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists M. *Ecological Monographs*, *85*(1), 3–28. https://doi.org/10.1890/07-1861.1

Joseph, M. B. (2020). Neural hierarchical models of ecological populations. *Ecology Letters*, *23*(4), 734–747. https://doi.org/10.1111/ele.13462

Kendall, W., & Nichols, J. D. (1995). On the use of secondary capture–recapture samples to estimate temporary emigration and breeding proportions. *Journal of Applied Statistics*, *22*(5–6), 751–762. https://doi.org/10.1080/02664769524595

Kendall, W. L., Nichols, J. D., & Hines, J. E. (1997). Estimating temporary emigration using capture-recapture data with Pollock's robust design. *Ecology*, *78*(2), 563–578. https://doi.org/10.1890/0012-9658(1997)078[0563:ETEUCR]2.0.CO;2

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*. Academic Press.

Kéry, M., & Royle, J. A. (2021). *Applied hierarchical modeling in ecology analysis of distribution, abundance and species richness in R and BUGS: Volume 2: Dynamic and advanced models*. Academic Press.

Kéry, M., & Schaub, M. (2012). *Bayesian population analysis using WinBUGS: A hierarchical perspective*. Elsevier.

Koehler, E., Brown, E., & Haneuse, S. J. P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *American Statistician*, *63*(2), 155–162. https://doi.org/10.1198/tast.2009.0030

Lahiri, S. N. (2005). Consistency of the jackknife-after-bootstrap variance estimator for the bootstrap quantiles of a studentized

statistic. *Annals of Statistics*, *33*(5), 2475–2506. https://doi.org/10.1214/009053605000000507
Link, W. A., Schofield, M. R., Barker, R. J., & Sauer, J. R. (2018). On the robustness of N-mixture models. *Ecology*, *99*(7), 1547–1551. https://doi.org/10.1002/ecy.2362
Loh, J. M., & Stein, M. L. (2004). Bootstrapping a spatial point process. *Institute of Statistical Science, Academia Sinica*, *14*(1), 69–101.
MacKenzie, D. I., & Bailey, L. L. (2004). Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological, and Environmental Statistics*, *9*(3), 300–318. https://doi.org/10.1198/108571104X3361
MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*, 2248–2255.
Mackenzie, D.I., & Royle, J.A. (2005). Designing occupancy studies: General advice and allocating survey effort. *Journal of Applied Ecology*, *42*(6), 1105–1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x
McClintock, B. T. (2021). Worth the effort? A practical examination of random effects in hidden Markov models for animal telemetry data. *Methods in Ecology and Evolution*, *12*(8), 1475–1497. https://doi.org/10.1111/2041-210X.13619
McNew, L. B., & Handel, C. M. (2015). Evaluating species richness: Biased ecological inference results from spatial heterogeneity in detection probabilities. *Ecological Applications*, *25*(6), 1669–1680. https://doi.org/10.1890/14-1248.1
Miller, D. A., & Grant, E. H. C. (2015). Estimating occupancy dynamics for large-scale monitoring networks: Amphibian breeding occupancy across protected areas in the Northeast United States. *Ecology and Evolution*, *5*(21), 4735–4746.
Muff, S., Signer, J., & Fieberg, J. (2020). Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using Bayesian or frequentist computation. *Journal of Animal Ecology*, *89*(1), 80–92. https://doi.org/10.1111/1365-2656.13087
Olivetti, S., Gil, M. A., Sridharan, V. K., & Hein, A. M. (2021). Merging computational fluid dynamics and machine learning to reveal animal migration strategies. *Methods in Ecology and Evolution*, *12*(7), 1186–1200.
Paciorek, C. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, *25*(1), 107–125. https://doi.org/10.1214/10-STS326.The
Pagel, J., & Schurr, F. M. (2012). Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, *21*(2), 293–304. https://doi.org/10.1111/j.1466-8238.2011.00663.x
Palencia, P., Fernández-López, J., Vicente, J., & Acevedo, P. (2021). Innovations in movement and behavioural ecology from camera traps: Day range as model parameter. *Methods in Ecology and Evolution*, *12*, 1201–1212. https://doi.org/10.1111/2041-210X.13609
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vol. *124*. No. 125.10.
Powers, S. M., & Hampton, S. E. (2019). Open Science, reproducibility, and transparency in ecology. *Ecological Applications*, *29*, e01822.
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Revell, L. J., & Harmon, L. J. (2022). *Phylogenetic comparative methods in R*. Princeton University Press.
Rizzo, M. L. (2019). *Statistical computing with R*. Chapman and Hall/CRC. https://www.R-project.org/
Rossman, S., Yackulic, C. B., Saunders, S. P., Reid, J., Davis, R., & Zipkin, E. F. (2016). Dynamic N- occupancy models: Estimating demographic rates and local abundance from detection- nondetection data. *Statistical Reports*, *97*(12), 3300–3307. https://doi.org/10.1002/ECY.1598
Runting, R. K., Phinn, S., Xie, Z., Venter, O., & Watson, J. E. M. (2020). Opportunities for big data in conservation and sustainability. *Nature Communications*, *11*, 1–4.
Santos-Fernandez, E., & Mengersen, K. (2021). Understanding the reliability of citizen science observational data using item response models. *Methods in Ecology and Evolution*, *12*(8), 1533–1548. https://doi.org/10.1111/2041-210X.13623
Silk, M. J., Harrison, X. A., & Hodgson, D. J. (2020). Perils and pitfalls of mixed-effects regression models in biology. *PeerJ*, *8*, e9522. https://doi.org/10.7717/peerj.9522
Smith, S. M., Stayton, C. T., & Angielczyk, K. D. (2021). How many trees to see the forest? Assessing the effects of morphospace coverage and sample size in performance surface analysis. *Methods in Ecology and Evolution*, *12*(8), 1411–1424.
Soroye, P., Newbold, T., & Kerr, J. (2020). Climate change contributes to widespread declines among bumble bees across continents. *Science*, *367*(6478), 685–688. https://doi.org/10.1126/science.aax8591
Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*, 1–16.
Tingley, M. W., Nadeau, C. P., & Sandor, M. E. (2020). Multi-species occupancy models as robust estimators of community richness. *Methods in Ecology and Evolution*, *11*(5), 633–642. https://doi.org/10.1111/2041-210X.13378
Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, *13*(6), 1790–1801. https://doi.org/10.1890/02-5078
Weber, F., Knapp, G., Glass, Ä., Kundt, G., & Ickstadt, K. (2021). Interval estimation of the overall treatment effect in random-effects meta-analyses: Recommendations from a simulation study comparing frequentist, Bayesian, and bootstrap methods. *Research Synthesis Methods*, *12*(3), 291–315. https://doi.org/10.1002/jrsm.1471
Wright, A. D., Campbell Grant, E. H., & Zipkin, E. F. (2022). A comparison of monitoring designs to assess wildlife community parameters across spatial scales. *Ecological Applications*, *32*(6), 1–13. https://doi.org/10.1002/eap.2621
Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J., & Grant, E. H. C. (2017). Integrating count and detection–nondetection data to model population dynamics. *Ecology*, *98*(6), 1640–1650. https://doi.org/10.1002/ecy.1831

SUPPORTING INFORMATION

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

How to cite this article: DiRenzo, G. V., Hanks, E., & Miller, D. A. W. (2023). A practical guide to understanding and validating complex models using data simulations. *Methods in Ecology and Evolution*, *14*, 203–217. https://doi.org/10.1111/2041-210X.14030