Title
Inferring species distributions from semi-structured biodiversity observations

Author
Goldstein, Benjamin R

Publication Date
2023

Inferring species distributions from semi-structured biodiversity observations

by

Benjamin R Goldstein

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy, and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Perry de Valpine, Chair
Professor Steve Beissinger
Associate Professor Carl Boettiger

Spring 2023

Inferring species distributions from semi-structured biodiversity observations

Abstract

Inferring species distributions from semi-structured biodiversity observations

by

Benjamin R Goldstein

Doctor of Philosophy in Environmental Science, Policy, and Management

University of California, Berkeley

Professor Perry de Valpine, Chair

Estimating the spatiotemporal distributions of species and understanding how variation in those distributions is explained by the environment are central goals in ecology. Observations of animals generated by participatory science (or “citizen science”) are an increasingly important resource for ecologists interested in estimating species distributions because they are high-volume and high-resolution. However, statistical inference with these data is more challenging than inference with data collected under standardized sampling, because participatory science observations contain substantial unmeasured variation in sampling effort and observer behavior. Ecologists need tools and methodological guidance that support the estimation of computationally efficient, flexible statistical models useful for robust inference with participatory science data. In this dissertation, I advance the field of species distribution modeling with participatory science data via contributions across three chapters. First, I present a new software tool, nimbleEcology, that supports the efficient and flexible estimation of hierarchical ecological models, alongside a brief review of the use of such models in ecology and three worked examples of model estimation. Second, I undertake a comparison of two modeling approaches useful for estimating relative abundance from participatory science data, making practical recommendations for model selection. Finally, I apply these methodological developments to data obtained from an important participatory science dataset, eBird, to investigate how common birds respond to drought in California’s Central Valley ecoregion. This project demonstrates the application of modeling principles to an important ecological case study and produces new evidence to characterize critical dimensions of birds’ drought responses.

This dissertation is dedicated to my mom and dad.

Contents

Acknowledgments

It is impossible to imagine having written this dissertation without the support of each person listed below (and many more than I can name here). Graduate research is an interdependent and all-consuming process, depending on collaboration and also on a network of personal support, mentorship, stress release, and love.

Thanks first and foremost to my Ph.D. advisor, Perry de Valpine, who has shown incredible thoughtfulness and care in five years of mentoring and research. Without Perry, none of this dissertation would have been possible, and I am deeply grateful to have had the opportunity to learn from him.

Many thanks to committee members Carl Boettiger, who was endlessly supportive dealing with ecological research, computational challenges, and academic bureaucracy; and Steve Beissinger, whose mentorship at critical moments during my Ph.D. helped me define my academic identity. Thanks to other faculty and research mentors who supported me with their collaboration and teaching during my doctoral dissertation, including Damian Elias, Tim Bowles, Albert Ruhi, Will Fithian, Danny Karp, and Ashley Larsen. Thanks to collaborator Brett Furnas for critical mentorship and for providing me with opportunities to learn about professional wildlife studies.

Thanks to my undergraduate mentors Kiho Kim, Richard Sha, Arthur Shapiro, and Andrea Tschemplik, who always believed in me. I am always aiming to make you all proud and make the world a better place.

Thanks to the ESPM staff, including Bianca Victorica, my graduate advisor for many years, for making it possible to navigate the campus' bureaucratic maze.

Thanks to the Boettiger lab—Millie Chapman, Kari Norman, Marcus Lapeyrolerie, Felipe Montealegre, Abby Keller, and Hopper—for quiet company and for helping deal with existential angst on Friday afternoons.

Thanks to the ESPM 2018 cohort for your kindness and patience, which made me feel comfortable and happy in California. Throughout my time at Berkeley, my favorite thing about the graduate program was that it afforded me the opportunity to chat with and learn from brilliant and inspiring folks across a dizzying range of disciplines. The 2018 cohort embodies that special quality of our department.

Thanks to the wildlife research group for helping me figure out my interests and begin to cultivate a scientific identity.

Thanks especially to research collaborators Sara Stoudt, whose mentorship and brilliant research ideas were highlights of the last few years; Kendall Calhoun, who taught me a lot of what I know about disturbance ecology; Phoebe Parker-Shames, who has been an indispensable mentor and role model; and Lucas Seninge, who kept me company during lunches and who helped design several of the figures in this dissertation.

A huge thank you to the tens of thousands of dedicated eBird observers across California whose checklists were used in this study.

Thanks to Yvonne Socolar, Kenzo Esquivel, and Mickey Boakye for keeping my spirits high and for hours of conversation, commiseration, and music.

Madison Brown, Maddy Bossi, Emily Drummond, Danny Kirsch, and Myriam Lapierre will forever have my undying gratitude and love for making a community with me. Thanks for thinking this is cool.

A special thanks to Alex Kushner, one of the most caring and thoughtful people I've ever known. I'm so grateful for all the time we spent together during the long COVID years and for our continuing friendship.

Thanks to Aaron Josephs, Cameron Roman, and Stephen Masson. Your camaraderie, reliability, creativity, and Star Wars opinions were a ray of light every week.

Thanks to my mom and dad and to my sisters Maya and Rebecca for setting this all up from day one. Thanks to my grandparents, whose love and support have always been a blessing.

Finally, to Maddie Wood for getting me through it, I can never say thank you enough.

This dissertation was supported in part by an NSF Graduate Research Fellowship and by an NSF Traineeship.

Chapter 1

Introduction

Estimating how wildlife species are distributed in space and time and explaining those distributions with environmental features are central goals of ecology [57, 42]. To achieve these goals, ecologists use quantitative models to analyze observational data, linking spatiotemporally tagged detections or counts of species with covariate data. As sampling and computational technology improve, the profile of the typical observational dataset is changing, complicating the task of species distribution modeling [12]. Rather than intensive in-person trapping or observational surveys, ecologists often use observations made by camera traps [119], passive acoustic monitors [48], and citizen science programs ("citizen science data" or "participatory science data") [65, 127]. These new approaches to field sampling generate data at higher volume and lower cost compared to classical observational or trapping surveys.

Participatory science data in particular have seen rapid growth in both data volume and scientific interest in recent years. Citizen science data are extremely high-volume and generate observations of species with unprecedented spatiotemporal resolution and coverage. However, citizen science data contain extra variation due to differences in observer skill and preference, non-standardized opportunistic sampling methodology, and false positive and false negative detection observations [69]. Participatory science data most useful for scientific analysis are "semi-structured," meaning that the data generating process does not follow a fixed sampling protocol but that some information on the sampling process is collected. For analyzing these data, ecologists need robust but highly flexible models that can accommodate variation in semi-structured data generation.

Recent advances in hierarchical modeling, and in software for model estimation, have set the stage for robust statistical inference of species distributions based on participatory science data. Broadly, ecologists have adopted hierarchical models as their main tool for accounting for multiple sources of heterogeneity in data while inferring properties of natural systems. One core statistical challenge—that of accounting for varying imperfect detection in the observation process—is of interest in many species distribution modeling contexts, not just in dealing with participatory science. There exist well-studied modeling frameworks to address this issue [114, 87]. Ecologists can accommodate the extra variation that arises from opportunistic sampling common to participatory science data with careful data filtering and appropriate statistical models [70]. Statisticians have also demonstrated the importance of accounting for observer-to-observer variation, which can be accomplished in several ways. Accounting for various sources of nonindependence and overdispersion in participatory science data is necessary to avoid the misattribution of variation when modeling these data [70]. Analyzing participatory science data often requires a bespoke approach where modeling priorities, assumptions, and interpretations are evaluated on a case-by-case basis, so maximally flexible modeling software that supports the estimation of arbitrary models is crucial for the uptake of citizen science data in ecology.

The primary goal of this dissertation is to advance the field of species distribution modeling with semi-structured data. I aim to address the need in ecology for clearer modeling guidance and more effective software for modeling species distributions with semi-structured data. This set of contributions will help ecologists do species distribution modeling with participatory science data.

Throughout the dissertation, I consider the eBird dataset as a particularly interesting case study for species distribution modeling with semi-structured data. Hosted by the Cornell Lab of Ornithology, eBird is an online platform that collects observations of birds made by the public [124, 125]. eBird observations are opportunistic, meaning that observers collect data whenever and wherever they choose, but they are reported with a number of useful metadata, including the time and location of each survey and several indices of sampling effort. Additionally, nearly 90% of eBird sampling events are "complete checklists," during which the observer reports every species they identify. This means that any unreported species on a checklist is implicitly associated with a zero count, which is critical for many species distribution modeling approaches. The eBird dataset is also enormous, comprising more than 1 billion observations of birds [5]. eBird data are often considered an exemplar of participatory science data, and scientific interest in eBird data for use in distribution modeling and monitoring of birds has grown in parallel with the size of the dataset (Figure 1.1). Due to the outstanding scientific need for robust quantitative tools for dealing with these data, analysis of the eBird dataset is a core theme of this dissertation.

In Chapter 2, I present nimbleEcology, an R package meant to support the implementation of hierarchical ecological models in the NIMBLE hierarchical modeling software system (R package `nimble`, [133]). nimbleEcology provides six major ecological model constructions—hidden Markov and dynamic hidden Markov models, occupancy and dynamic occupancy models, N-mixture models, and Cormack-Jolly-Seber mark-recapture models—as marginalized probability distributions. The nimbleEcology package benefits ecologists estimating models by simplifying model specification, making possible maximum likelihood estimation of a wider set of ecological models, and in some cases improving model estimation efficiency. This chapter contains an introduction to hierarchical ecological models, an explication of the concept of marginalization, a detailed report on nimbleEcology's functions and use, and three comprehensive worked examples using nimbleEcology to simulate and analyze ecological data.

In Chapter 3, I focus on the task of relative abundance estimation from participatory science data. Considering eBird data as an interesting case study, I investigate the relative

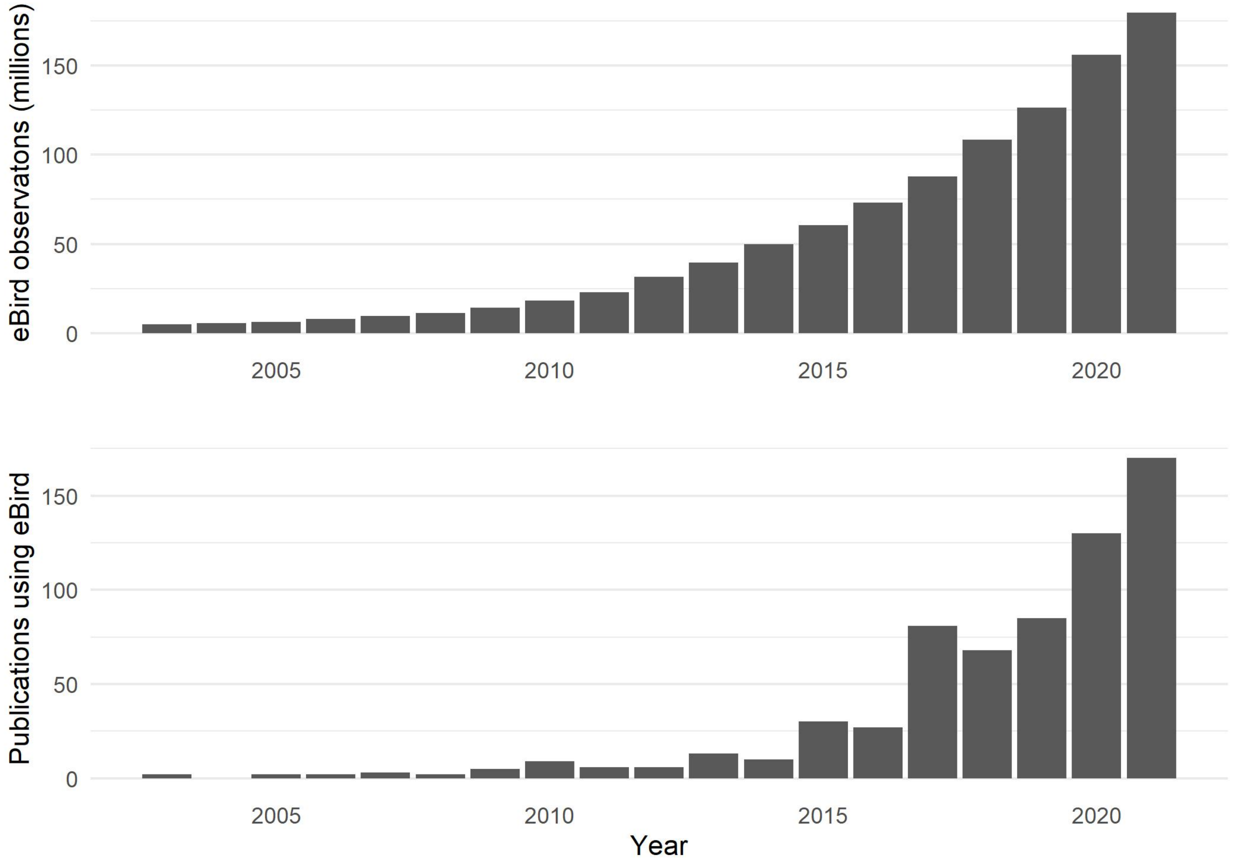

Figure 1.1: Top panel: number of species observations submitted to eBird each year [5]. Bottom panel: number of peer-reviewed publications using eBird data each year [105].

utility of two major hierarchical model families—generalized linear mixed models (GLMMs) [12] and N-mixture models [114]—for relative abundance estimation. Both N-mixture models and GLMMs have the potential to estimate patterns in relative abundance, but each has advantages and drawbacks. I give a conceptual comparison of these two models, unifying them under a single perspective to better highlight their differences. Then, I compare their relative performance across regional single-species subsets of the eBird dataset. I show that N-mixture models tended to outperform GLMMs for relative abundance estimation, but that the pattern in comparative performance is variable enough to recommend that ecologists weigh both modeling approaches in many applications.

In Chapter 4, I build on methodological findings from previous chapters to analyze eBird data and answer an outstanding ecological question. I ask whether eBird counts were systematically different between periods of severe drought and periods of non-drought, and whether species exhibit habitat-specific drought responses, taking advantage of eBird's unmatched volume and resolution to distinguish between drought responses across habitat types and environmental mechanisms. To answer these questions, I develop a novel multi-model framework that uses Bayesian counterfactual methodology to predict birds' drought responses. I show that changes in habitat associations, more than changes in overall abundance, defined birds' drought responses, with species shifting from natural to human-modified habitats during drought. I also show that temperature and precipitation variables best explained birds' drought responses. This chapter demonstrates the utility of eBird for answering important ecological questions and produces new evidence to characterize critical dimensions of birds' drought responses.

I conclude the dissertation with a brief reflection on outstanding challenges in species distribution modeling with participatory science data.

Chapter 2

Marginalizing hierarchical ecological models in NIMBLE

2.1 Introduction

Many basic goals in ecological research, such as estimating population dynamics, modeling distributions of wildlife populations in space, and characterizing the behavior of individual organisms, share an underlying goal: to understand patterns in natural systems and processes. To address each of these fundamental goals, ecologists have developed field methods to collect data on a dimension of the ecological process of interest. For example, population dynamics may be partially revealed using mark-recapture data of individual animals over time; distributions may be inferred from presence-absence data generated by surveys of a target organism across a landscape; and patterns in behavior can be inferred from location data as an organism moves about its environment [84, 42, 94]. In each case, ecologists aim to attribute variation in their data to a natural process of interest.

In practice, ecological data on natural systems contain variation arising from processes other than the processes of interest. One reason is that ecological systems are by nature difficult to observe. Organisms camouflage, avoid humans, and move around [76, 94]. Many ecosystems are difficult to access, so sampling is often uneven. Ecological field surveys require substantial person-hours or expensive equipment, so data sample sizes are often bounded by cost, and data are often hierarchically grouped rather than fully independent [13]. Processes of interest are often impossible to observe directly even under ideal circumstances; for example, complete censuses of populations are rarely feasible, so the true total abundance of many populations is not directly observable. For these reasons, ecological data nearly always contain multiple sources of variation.

To analyze data containing multiple sources of variation, ecologists often use hierarchical statistical models. In a hierarchical model, variation in data arises from multiple probability distributions that are conditionally linked [79, 42, 30]. Many hierarchical models in ecology include discrete latent states, such as the true abundance of a species at a site or the unobserved behavioral state of an animal. Hierarchical models in ecology are often designed to correspond to a data generating process, where latent states represent "true" discrete system states arising from an ecological process of interest, and data are linked to latent states via a probability distribution representing an imperfect observation process. This thinking is helpful for understanding hierarchical ecological models, but models can be useful even if their assumptions do not perfectly match the data-generating process.

Hierarchical models are useful for accommodating multiple sources of variation in data but are challenging to estimate. Most ecologists using hierarchical models depend on dedicated software. There are two main types of software that ecologists use to estimate their models: (1) software written to estimate a predefined set of models, and (2) software that gives the user tools to define and estimate arbitrary models.

The first type of statistical software for hierarchical model estimation provides the ecologist with a predefined set of models that can be fit to input data. Programs of this type include the standalone PRESENCE and MARK software, as well as R packages such as unmarked, momentuHMM, spOccupancy, and ubms [64, 82, 46, 93, 40, 77]. While a comprehensive overview of these programs is beyond the scope of this report, they together cover a wide range of model constructions, including many variants of the occupancy model, hidden Markov models, mark-recapture models, and N-mixture abundance models. Software programs that estimate predefined models are hugely popular in ecology. Because they cover a finite set of model constructions, implementations in these software tend to be computationally tuned for efficiency. They are also easy to use, replicate, and teach to new users because they require minimal understanding of model structure, code, and estimation methods. The main drawback of the predefined model paradigm in statistical software is that the set of models supported by a given software package is fundamentally limited. While most ecologists will never want to develop novel hierarchical models, a statistical software program using predefined model-fitting methods may prohibit even small deviations from the constructions provided, such as an alternative link function or the inclusion of an additive random effect.

The second type of statistical software encompasses a suite of software tools useful for estimating nearly arbitrary hierarchical models. These tools share a common workflow: first, the user uses a special coding language to specify a model; then, the user is able to apply one or more model estimation algorithms, such as Markov chain Monte Carlo (MCMC), to the specified model to obtain parameter estimates. Popular software of this type include the BUGS-family software (named for "Bayesian inference Using Gibbs Sampling," the first software project in this family), including WinBUGS, JAGS, and NIMBLE, which share a design philosophy and model specification language but differ greatly in functionality [133, 101]. Another general-model statistical software program is Stan [126]. This type of software provides the user with a substantial amount of flexibility and control over model specification. In the case of NIMBLE, the user has even greater control and can make choices about details of the model estimation algorithm. However, because predefined models are unavailable, these tools are more difficult to use and require a much greater level of expertise and knowledge about the workings of a statistical model. Additionally, because they accommodate models of arbitrary complexity, these software tools tend to support estimation methods that are more general but less computationally efficient.

Most statistical software in ecology uses one of two techniques to estimate hierarchical models: maximum likelihood estimation (MLE) or Bayesian Markov chain Monte Carlo (MCMC) estimation [102, 13, 79]. MLE refers to a class of algorithms for finding the point in parameter space that maximizes the likelihood of the parameters given the observed data and the model. MLE tends to be relatively efficient for simple models, and maximum likelihood estimates of model parameters are easy to handle using basic frequentist statistics. However, MLE requires calculating or approximating the likelihood, which is not always possible and can be hard to generalize. For this reason, most statistical software that supports estimation of predefined models uses MLE (with notable exceptions like ubms), while most flexible software programs do not. MCMC estimation refers to a set of algorithms for drawing samples approximating the joint posterior probability distribution of the parameters [16]. MCMC methods tend to be slower than MLE but accommodate a much greater range of models without special implementation [141].

In this chapter, I present nimbleEcology, an R package designed to make it easier to define and estimate common ecological models in the flexible statistical software NIMBLE. The nimbleEcology package extends NIMBLE's model specification language by providing a suite of custom distributions representing six common hierarchical ecological models. These distributions can be used in the NIMBLE environment to easily define models incorporating hierarchical structure. nimbleEcology's distributions are marginalized, meaning they eliminate the need to explicitly define or sample a latent state (see Section 2 for a discussion of marginalization). This package is useful for ecologists for several reasons. First, nimbleEcology custom distributions are reliable and error-tested, which makes the model developer's job simpler, bridging the gap between the predefined and highly customized modeling paradigms. Second, marginalized distributions sometimes, though not always, lead to more efficient model estimation in MCMC [103]. Third, by implementing marginalized distributions that eliminate the need for discrete latent states, these custom distributions enable maximum likelihood estimation for NIMBLE models.

The goal of this chapter is to present the main features of the nimbleEcology package and demonstrate its use. The chapter is organized as follows. First, I present the motivating concept of "marginalization" in hierarchical models and discuss its use in nimbleEcology's implementation. I briefly discuss NIMBLE, highlighting how NIMBLE's extensibility creates a niche for helper software that bridges the gap between arbitrary and pre-written models. I then discuss the structure and use of the nimbleEcology package and give a brief overview of each of the six ecological model families represented in nimbleEcology. Finally, I present three worked examples using nimbleEcology distributions in NIMBLE models to demonstrate (1) maximum likelihood estimation with NIMBLE, (2) data simulation and MCMC estimation, and (3) evaluating MCMC estimation efficiency.

2.2 Marginalized probability distributions in NIMBLE

In this section, we present a brief conceptual overview of marginalization, discuss the potential benefits and drawbacks of marginalization in Bayesian model estimation, and describe how nimbleEcology takes advantage of NIMBLE infrastructure to implement marginalized distributions.

What is marginalization?

Consider a hierarchical ecological model with discrete latent states. The model can be written generically as

$$X \sim P_x(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$$
$$Z \sim P_z(\mathbf{z}|\boldsymbol{\theta})$$
$$(\Theta \sim P_\theta(\boldsymbol{\theta}))$$

The probability density of a set of observations, $\mathbf{x}$, is a function $P_x(\cdot|\mathbf{z}, \boldsymbol{\theta})$ of a vector of parameters, $\boldsymbol{\theta}$, and a set of latent states, $\mathbf{z}$. The probability that the discrete latent states $\mathbf{z}$ take a particular value is itself a function of that set of parameters, $P_z(\cdot|\boldsymbol{\theta})$. In a Bayesian framework, the vector $\boldsymbol{\theta}$ also has a prior distribution $P_\theta(\cdot)$.

To marginalize the model, the likelihood of the parameters given a set of observations depending on shared latent states needs to be analytically calculated or approximated, summing or integrating over the latent state space [141, 130]. The likelihood is given by summing over the product of the data and latent state probability densities for each possible value of the set of latent states, as

$$\mathcal{L}(\theta|\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} \Big\{ P_x(\mathbf{x}|\mathbf{z}, \theta) P_z(\mathbf{z}|\theta) \Big\}$$

where $\mathcal{Z}$ is the set of discrete values that z can have. This gives the likelihood of the parameters solely determined by the data $\mathbf{x}$. In MCMC estimation, marginalization is an alternative to sampling the posterior distributions of the latent states.

We can illustrate marginalization more concretely using the single-species site occupancy model (hereafter “occupancy model”). In the occupancy model, a series of visits to each of several sites produces “detection histories,” vectors of detection and nondetection data indicating whether a species of interest was observed on each visit. The site is assumed to have a true occupancy status (either occupied or not) represented by a latent state in the hierarchical formulation. Then, a detection of the species indicates that the site was occupied, while a nondetection arises due to the site being unoccupied or as a result of imperfect detection. The model for observations made at one site is given by

$$z|\psi \sim \text{Bernoulli}(\psi)$$

$$x_j|z,p \sim \text{Bernoulli}(zp) \text{ for } j = 1...J$$

where $x_j = 1$ if the species was observed on the jth replicate survey and $x_j = 0$ if it was not; ψ is the probability that the site is occupied; p is the probability that the species is detected given that it is occupied; and z is a latent state describing whether or not the species is in fact present.

For estimating the model with MCMC, one might typically define nodes representing the latent occupancy state z to be sampled along with other parameters. This model could be specified in NIMBLE code as

```
nimbleCode({
    z ~ dbern(psi)
    for (j in 1:J) {
        x[j] ~ dbern(z * p)
    }

    # Priors
    psi ~ dunif(0, 1)
    p ~ dunif(0, 1)
})
```

where z is a latent state whose value (1 or 0) would be sampled.

To apply marginalization to this model, consider that for observation vector $\mathbf{x}$, with parameters $\theta = \{\psi, p\}$ and latent state z, the model likelihood for one site is given by

$$\mathcal{L}(\psi, p|\mathbf{x}) = \sum_{z=\mathcal{Z}} \Big\{ P_x(\mathbf{x}|z,p) \times P_z(z|\psi) \Big\}$$

Since z only takes values 0 and 1, we can analytically derive the likelihood of the occupancy model by summing across the possible values of z, giving

$$L(\psi, p|x) = (1-\psi) \prod_j \Big\{ \mathbf{1}_0(x_j) \Big\} + \psi \prod_j \Big\{ p^{x_j}(1-p)^{1-x_j} \Big\}$$

where $1_0(x) = 1$ if $x = 0$ and $1_0(x) = 0$ otherwise. In plain language, the site is either occupied and the data came from observations made at an occupied site (meaning any nondetection arises from a failure to observe the species despite it being present), or the data came from observations made at an unoccupied site (meaning all nondetections are guaranteed and detections are impossible). By summing these two conditions, the latent state z has been marginalized over and is no longer represented. This marginalized construction can be written out explicitly in NIMBLE model code or more simply using a NIMBLE custom distribution (see “NIMBLE custom functions”).

Benefits and drawbacks of marginalization in model estimation

The most decisive benefit of a marginalized implementation of a hierarchical ecological model is the possibility for maximum likelihood estimation. Finding the maximum likelihood requires that the likelihood is calculated or approximated numerically, marginalized across all latent states. The maximum likelihood estimate of a hierarchical model does not pertain to specific values of latent states (which are not model parameters), but rather corresponds to the most likely values of parameters considering the full distribution of possible latent state values. Indeed, all domain-specific MLE software, such as unmarked, PRESENCE, and momentuHMM, depend on marginalized implementations of model likelihoods [46, 64, 93].

For ecologists estimating Bayesian models, marginalization is not strictly necessary for model estimation, but it can in some situations improve estimation efficiency. Bayesian models tend to be relatively slow to estimate by MCMC [141], and improvements to mixing efficiency are important for the feasibility of Bayesian model estimation. In MCMC estimation, the general strategy is to iteratively draw samples that approximate the joint posterior distribution of the parameters and latent states given the data. Estimation efficiency is determined by two factors: the computational time per sample and the quality of mixing. The former is defined as the amount of real time needed to draw a number of joint posterior samples. The second factor, quality of mixing, describes the degree of independence between sequential samples of the posterior. A good metric for determining quality of mixing is the effective sample size (ESS), an estimate of the number of independent draws that would contain the same amount of information as the (non-independent) MCMC samples [103]. Better mixing produces more independent samples, so the effective sample size of a given number of real samples increases as mixing quality improves. In a study of MCMC efficiency, Ponisio et al. compared marginalized implementations to latent state formulations of a number of realistic ecological models [103]. The authors found that marginalization had a strong effect on sampling efficiency, but that the direction of its effect depended both on model type and on the degree of hierarchy in the model.

A final benefit of marginalization is that marginalized BUGS code is simpler to write. Predefined, marginalized probability distributions are easier to use than custom models with explicitly coded latent states and leave less room for user error. Therefore the model specification and debugging process becomes more efficient when using marginalized distributions such as those provided by nimbleEcology.

Marginalizing over latent states in a Bayesian model has an important drawback: by marginalizing latent states out of the model, ecologists lose the ability to monitor the posterior distributions of the latent states. While latent states are not, strictly speaking, the subject of model estimation, ecologists are sometimes interested in examining properties of the posterior distributions of latent states or derived products of those latent states. For example, in the hierarchical occupancy model, one might be interested not only in the average occupancy probability ψ, but in the fraction of sites that are truly occupied as predicted by the model, so the posterior distribution of $\sum_i z_i$ might be of interest. It is possible to recover posterior distributions of latent states when using marginalization [141], but this functionality is not yet supported by nimbleEcology without manual implementation.

NIMBLE custom functions

nimbleEcology provides custom distributions specifically for use with NIMBLE models. Fundamentally, nimbleEcology's distributions are probability mass functions for marginalized hierarchical models. They can be used in the same way as basic probability density functions like `dnorm` to declare that data follow one of a number of common hierarchical model constructions.

NIMBLE, which stands for "Numerical Inference for statistical Models using Bayesian and Likelihood Estimation", is a system for writing hierarchical statistical models and algorithms in the R interface [133]. NIMBLE belongs to the family of software tools built to specify and estimate arbitrary hierarchical models. NIMBLE uses a declarative BUGS-like language along with a compiler to specify and build graphical models, and provides a variety of additional tools for model estimation and evaluation, including but not limited to a set of functions for an MCMC model estimation workflow.

NIMBLE is extensible, meaning that users can write customized arbitrary functions, such as probability distributions, to be used in the BUGS model specification language. Custom NIMBLE functions are defined using the function `nimbleFunction` in the NIMBLE package. There are several ways to set up a NIMBLE custom function, but for the purposes of a custom probability distribution, functions are defined using the following syntax. `nimbleFunction` takes the argument `run`, which is a function following R syntax (see the NIMBLE User Manual for an overview of supported functionality in compilable nimbleFunctions). The following is an example of a custom density distribution for the Bernoulli distribution:

```
dbernCustom <- nimbleFunction(
  run = function(x = double(0),
                 p = double(0),
                 log = integer(0, default = 0)) {
    returnType(double(0))

    prob <- x * p + (1-x) * (1-p)

    if (log) return(log(prob))
    return(prob)
  }
)
```

An R function is provided to the `run` argument of `nimbleFunction`. The arguments to that R function have default values that define the typing of each parameter. NIMBLE uses static typing, which means that the types and dimensions of all arguments must be specified in advance in this way. The first argument to a density function is always `x`, which takes the datum or data for which the probability will be returned; in this case, x can take values of 1 or 0. Then, an arbitrary number of named parameters can be provided. Finally, an integer argument `log`, which takes values of 0 or 1 representing "false" or "true", indicates whether the likelihood (`log=0`, the default) or log-likelihood (`log=1`) should be returned. Note the reference to `returnType()`, which tells the compiler the type that will be returned by the function.

Alongside each density function (prefixed with "`d`"), we can define a random generator function (prefixed with "`r`"). This is required for MCMC estimation in certain cases. While the density function takes observed values and parameters and returns a likelihood or log-likelihood of those parameters, the random generation function takes parameter values as arguments and probabilistically returns a random observation or observation history. The following defines an "`r`" Bernoulli function to accompany `dbernCustom`:

```
rbernCustom <- nimbleFunction(
  run = function(n = double(0),
                 p = double(0)) {
    returnType(double(0))
    if (runif(1, 0, 1) < p) {
      return(1)
    } else {
      return(0)
    }
})
```

The first argument `n` is required by NIMBLE but usually ignored, after which all parameters are specified to match the density function. Arguments `x` and `log` are omitted. Once both distributions are defined, they can be registered for use with NIMBLE models using the function `registerDistributions`. Once registered, NIMBLE will recognize these functions automatically when called in model code. NIMBLE will also automatically attempt to register any unregistered distributions when the model is built. Custom distributions can also be called from the R environment, where they will be executed in uncompiled form, or they can be compiled directly.

By implementing the marginalized distribution in a NIMBLE custom distribution, it can be used in NIMBLE code equivalently to any other distribution like `dnorm`. Consider the marginalized occupancy model given above. We can define a NIMBLE custom distribution, `dOcc`, representing this distribution. In this case, we could write the following:

```
dOcc <- nimbleFunction(
  run = function(x = double(1),
                 probOcc = double(0),
                 probDetect = double(0),
                 len = integer(0, default = 0),
                 log = logical(0, default = 0)) {
    returnType(double(0))
    logProb_x_given_occupied <-
         sum(dbinom(x, prob = probDetect,
                    size = 1, log = TRUE))
    prob_x_given_unoccupied <- sum(x) == 0
    prob_x <- exp(logProb_x_given_occupied) * probOcc +
                prob_x_given_unoccupied * (1 - probOcc)
    if (log) return(log(prob_x))
    return(prob_x)
  }
)
```

Because the nimbleFunction `dOcc` is a valid and compilable probability distribution, it can be used directly in a model to define the probability of data arising from an occupancy distribution given parameters p (function argument "`probDetect`") and ψ (function argument "`probOcc`"), thus eliminating the need for a node representing the latent state z in the graphical model. The NIMBLE code in section 2.1 could be rewritten to use this marginalized distribution as

```
nimbleCode({
    x[1:J] ~ dOcc(psi, p, J)

    # Priors
    psi ~ dunif(0, 1)
    p ~ dunif(0, 1)
})
```

2.3 The nimbleEcology R package

Installing and using nimbleEcology

nimbleEcology is available as an R package on CRAN at `https://cran.r-project.org/web/packages/nimbleEcology/index.html` [54]. When loaded, the package initializes and registers custom distributions for six model families. The six model families provided are: the hidden Markov model, dynamic hidden Markov model, occupancy model, dynamic occupancy model, Cormack-Jolly-Seber capture-recapture model, and N-mixture model. Once the package is loaded, all custom distributions representing available parameterizations and variants of these six model families become usable in the R environment and in NIMBLE code. Documentation is provided for each distribution and can be accessed using R's "help()" or "`?`" syntax preceding either the name of the model family or specific distribution (so, `?dDynOcc` and `?dDynOcc_ssv` both bring up the documentation for the dynamic occupancy model family).

Source code for nimbleEcology is available at `https://github.com/nimble-dev/nimbleEcology`, which may be useful to users wishing to develop new custom distributions or modify those provided.

nimbleEcology parameterizations and syntax

Each marginal distribution family represented in nimbleEcology contains a set of functions implementing different parameterizations of that distribution. To handle compilation of arbitrary functions, NIMBLE requires that argument types are fixed at the time of model specification. This includes whether a node or an argument in a custom function is a scalar (length = 1) or vector (length > 1). This is different from base R, which lets the user provide a scalar or a vector equivalently for many cases. Since the usefulness of one parameterization over another is context-dependent, nimbleEcology provides multiple implementations of each distribution, many of which vary only in whether certain parameters are vector or scalar, in an attempt to cover all the parameterizations users may want.

Different parameterizations of a nimbleEcology custom distribution are indicated with suffixes appended to function names. Suffixes refer to whether each parameter with multiple possible types is provided as a scalar (s), vector (v), or matrix (m). For example, the function `dOcc_s` takes a scalar (s for scalar; visit-invariant) detection probability, while the function `dOcc_v` takes a vector (v for vector; visit-specific) detection probability. The order of suffixes corresponds to the order in which the relevant parameters are provided to the function. In this chapter, I will often replace prefixes with asterisks, as in `dOcc_*`, to refer to all variants of a model family. Multiple asterisks are used to replace multiply suffixed functions, such as `dDynOcc_***`. When referring to a certain subset of these functions I may only replace certain suffixes with asterisks, as in `dDynOcc_**v`, which would refer to the four dynamic occupancy distribution functions that take vector detection probabilities.

2.4 Distributions provided by nimbleEcology

In this section, I present the six families of custom distributions provided by nimbleEcology. I provide a brief definition of each model family, discuss some of its applications, and then present the relevant nimbleEcology function and its usage. The six model families are the hidden Markov model, dynamic hidden Markov model, occupancy model, dynamic occupancy model, Cormack-Jolly-Seber model, and N-mixture model.

Hidden Markov model

Model definition

The hidden Markov model (HMM) is a statistical model for sequential data [94, 52] used to estimate transition probabilities between discrete true states of a system when those states are observed imperfectly. It can represent a wide variety of ecological systems dynamics, from estimating behavioral states of individuals to population dynamics at large scales [47, 132]. Other common ecological models, including the capture-recapture survival model and dynamic occupancy model, are special cases of the HMM [94].

The HMM gives the probability of a vector of data, $\mathbf{x}$, where x_t is the observed state of an individual at time step $t = 1...T$. The latent state s_t represents the "true" unobserved state of the individual at time $t = 1...T$. Then, in the latent state formulation,

$$P(s_1 = k|\boldsymbol{\pi}) = \pi_k$$
$$P(s_t = k|\boldsymbol{R}, s_{t-1}) = r_{s_{t-1},k} \text{ for } t = 2...T$$
$$P(x_t = l|\boldsymbol{O}, s_t) = o_{s_t,l} \text{ for } t = 1...T$$

where $\boldsymbol{R}$ is a $(S \times S)$ transition probability matrix where element $r_{i,j}$ gives the probability of transitioning from state i to state j, $\boldsymbol{O}$ is a $(S \times N)$ observation probability matrix where element $o_{i,j}$ gives the probability of observing state j given that the individual is in state i, $\boldsymbol{\pi}$ is a vector of initial probabilities where element π_i gives the probability of the individual being in state i at time $t = 1$, S is the total number of states the individual can take and N is the total number of observable states.

Estimating hidden Markov models is possible via MCMC sampling of latent states in any flexible model specification environment such as JAGS or NIMBLE [101, 133]. The R package momentuHMM also supports estimating HMMs via MLE [93]. nimbleEcology provides a marginalized implementation of the HMM likelihood.

A marginalized likelihood function for a single HMM detection history can be calculated by considering the probability of data arising from transitions and observation probabilities sequentially in time. A full description of the process of marginalizing the HMM is given by Turek et al. and reproduced briefly here [130]. The likelihood is calculated by iteratively computing the following quantities over the detection history:

$$\boldsymbol{p}_1 = \boldsymbol{\pi}$$
$$\boldsymbol{p}_t = \boldsymbol{R}_t \boldsymbol{q}_{t-1} \text{ for } t = 2...T$$
$$\boldsymbol{q}_t = \boldsymbol{o}'_{x_t,1:N} * \boldsymbol{p}_t / L_t \text{ for } t = 1...T$$
$$L_t = \boldsymbol{o}_{x_t,1:N} \boldsymbol{p}_t \text{ for } t = 1...T$$

where $\boldsymbol{R}_t$ is the transition probability matrix as above; $o_{i,1:N}$ gives the ith row in the observation matrix O as a vector; and $*$ and $/$ denote element-wise multiplication and scalar division, respectively. $\boldsymbol{p}_t$ and $\boldsymbol{q}_t$ are column vectors used for intermediate calculation. The likelihood of the parameters given a single detection history is given by

$$L(\boldsymbol{R}, \boldsymbol{O}, \pi | \mathbf{x}) = \prod_t L_t$$

Four of the six distribution families provided by nimbleEcology related to or specific cases of the hidden Markov model: the hidden Markov model itself, the dynamic hidden Markov model, the dynamic occupancy model, and the Cormack-Jolly-Seber model. Implementations of these four distribution families use this approach for marginalization.

nimbleEcology distributions: `dHMM` and `dHMMo`

nimbleEcology provides two custom distribution parameterizations for the marginalized hidden Markov model: `dHMM` and `dHMMo`. When used in a NIMBLE model, the distribution `dHMM` takes the following arguments: `init`, the initial probability vector ($\boldsymbol{\pi}$); `probObs`, an observation probability matrix ($\boldsymbol{O}$); and `probTrans`, a transition probability matrix ($\boldsymbol{R}$). `dHMM` also has an optional argument `checkRowSums` which takes a Boolean value. If `checkRowSums` $= 1$ (the default), the likelihood function internally checks that rows in each matrix sum to 1 and are therefore valid sets of probabilities. Setting `checkRowSums` to 0 will skip this check for efficiency if the user is confident that both matrices are defined such that rows sum to 1.

A second variation of the hidden Markov model, `dHMMo`, is provided, which accommodates temporally variable observation probabilities. `dHMMo` is identical to `dHMM`, except that the observation matrix `probObs` is parameterized as a 3-dimensional array with an additional dimension of length T, where element $o_{i,j,t}$ gives the probability of observing an individual in state i at time t in observation state j. Thus the third equation in the preceding formula is replaced with

$$P(x_t = l | \boldsymbol{O}, s_t) = o_{s_t,l,t} \text{ for } t = 1...T$$

Example code

The following code chunk defines a NIMBLE model where observed data follow a hidden Markov model.

```
codeHMMo <- nimbleCode({
  for (i in 1:numHists) {
    observedStates[i, 1:Tt] ~ dHMMo(
      init = initStates[1:S],
      probObs = observationProbs[1:S, 1:N, 1:Tt],
      probTrans = transitionProbs[1:S, 1:S],
      len = Tt, checkRowSums = 1)
  }
})
```

Dynamic hidden Markov model

Model definition

The dynamic hidden Markov model (DHMM) is an extension of the hidden Markov model, with the important difference that transition probabilities are allowed to vary with time. Rather than a $(S \times S)$ matrix, the transition probabilities R are provided in a $(S \times S \times (T-1))$ three-dimensional array, and the second equation in the HMM definition is replaced in the DHMM with

$$P(s_t = k|\boldsymbol{R}) = r_{s_{t-1},k,t-1} \text{ for } t = 2...T$$

DHMMs may be used in ecology in similar contexts to HMMs, when the probability of transitioning from one state to the next may vary with time.

Marginalization in the DHMM nimbleEcology functions is achieved following the method used for HMMs.

nimbleEcology distributions: `dDHMM` and `dDHMMo`

Two distributions of `dDHMM` are provided, `dDHMM` and `dDHMMo`. The former is parameterized with time-invariant observation probabilties and the latter is parameterized with time-variant observation probabilities, exactly as in the `dHMM*` family (see above for more details).

Example code

The following code chunk defines a NIMBLE model where observed data follow a dynamic hidden Markov model.

```
codeDHMMo <- nimbleCode({
    for (i in 1:numHists) {
      observedStates[i, 1:Tt] ~ dDHMMo(
        init = initStates[1:S],
        probObs = observationProbs[1:S, 1:N, 1:Tt],
        probTrans = transitionProbs[1:S, 1:S, 1:(Tt-1)],
        len = Tt, checkRowSums = 1)
    }
  })
```

Occupancy model

Model definition

The hierarchical occupancy model, sometimes called the site-occupancy model, is a common ecological model used to estimate the occurrence probability of a species at a number of sites from replicate detection-nondetection surveys at each site [87, 88, 79]. For a single site, the occupancy model is defined as

$$z|\psi \sim \text{Bernoulli}(\psi)$$

$$x_j|\psi, p \sim \text{Bernoulli}(zp) \text{ for } j = 1...J$$

where x_j is 1 if the species was detected on the jth sampling occasion and 0 otherwise, ψ is probability that the species truly occupies the site, p is the probability of detecting the species at the site given that it is present, and z is a latent state taking 1 or 0 representing whether or not the species occupies the site.

Occupancy models are used in a variety of ecological contexts to infer a species' patterns of presence and absence from data sampled with imperfect detection [88]. The single-site definition given above can be embedded in a variety of occupancy model extensions of varying complexity, for example a multispecies hierarchical occupancy model [67, 37].

Marginalization for the occupancy model is presented in section 2.2. To briefly recap, the occupancy model's marginalized implementation is achieved by summing together the probability of the data given that the site is occupied times the occupancy probability with the probability of the data given that it is not times one minus the occupancy probability.

nimbleEcology distributions: `dOcc_v` and `dOcc_s`

Two parameterizations of the occupancy distribution are provided: `dOcc_v` and `dOcc_s`. The suffixes distinguish between whether the detection probability p is a visit-invariant scalar (`dOcc_s`) or a potentially visit-variant vector (`dOcc_v`).

Example code

The following code defines a nimbleModel where a series of replicate observations at each of N sites follow an occupancy distribution.

```
codeOccV <- nimbleCode({
    for (i in 1:N) {
        obs[i, 1:J] ~ dOcc_v(
            probOcc = psi[i],
            probDetect = p[i, 1:J],
            len = J
        )
    }
})
```

Dynamic occupancy model

Model definition

The dynamic occupancy model represents a series of replicate surveys at a site in consecutive seasons, where the species of interest can undergo dynamics of local colonization and extinction, and where detection is imperfect [89, 88]. Dynamic occupancy models are used in multi-season studies of species occupancy where rates of colonization and persistence at sites between seasons are of interest. The model is defined as

$$x_{j,t} \sim \text{Bernoulli}(z_t p)$$

$$z_1 \sim \text{Bernoulli}(\psi_1)$$

$$z_t \sim \text{Bernoulli}(\pi z_{t-1} + \gamma(1 - z_{t-1})) \text{ for } t = 2...T$$

where $x_{j,t}$ is a datum indicating whether the species was detected during the jth survey in season (discrete time period) t, z_t is the latent occupancy state at time t, ψ_1 is the probability that the site is occupied in time $t = 1$, π is the persistence probability (the probability that the site is occupied in time $t > 1$ given that it was occupied in time $t - 1$), and γ is the colonization probability (the probability that the site is occupied in time $t > 1$ given that it was *not* occupied in time $t - 1$).

The dynamic occupancy model is a special case of the hidden Markov model with replicate observations of a given unit in one time period necessary for identifiability. The model has two observed states (detected and not detected) and two true states (unoccupied and occupied). Colonization and extinction probabilities contribute to a 2×2 transition matrix and the detection probability p contributes to a 2×2 observation matrix. As such, marginalization in the dynamic occupancy model follows the method used in the generic hidden Markov model described above.

nimbleEcology distributions: `dDynOcc_***`

Twelve parameterizations of the dynamic occupancy model are provided to accomodate all combinations of three parameters that each can be time-variant. The probabilities of persistence, π, and colonization, γ, can be temporally constant (scalar) or vary between seasons (vector). The probability of detection, p can be time-invariant (scalar), vary across seasons (vector), or vary across seasons and between visits within a season (matrix). Parameterizations are differentiated by suffixes `s`, `v`, and `m` in the order that parameters are called, such that `dDynOcc_svm` requires a scalar persistence probability, a vector colonization probability, and a matrix detection probability.

Because there are only $T-1$ transitions between seasons, the persistence and colonization probabilities are provided as vectors of length $T - 1$ when time-variant.

Example code

The following code defines a nimbleModel where observations follow a dynamic occupancy model at each of `N` sites over `Tt` seasons.

```
codeDynOccVVM <- nimbleCode({
  for (i in 1:N) {
    obs[i, 1:J, 1:Tt] ~ dDynOcc_vvm(
      init = init[i],
      probPersist = pi[i, 1:(Tt-1)],
      probColonize = gamma[i, 1:(Tt-1)],
      p = detProb[i, 1:J, 1:Tt],
      start = start_vec[1:Tt],
      end = end_vec[1:Tt]
    )
  }
})
```

Cormack-Jolly-Seber capture-recapture model

Model definition

The Cormack-Jolly-Seber model is used to estimate survival probabilities for individuals that are marked at some initial time and then either recaptured or not at each of several subsequent sampling points [84, 25, 72, 116]. The probability that an individual is observed on replicate surveys is defined as

$$
\begin{gathered}
x_1 = 1 \\
z_1 = 1 \\
x_t | z_t, p_t \sim \text{Bernoulli}(z_t p_t) \text{ for } t = 2...T \\
z_t | \boldsymbol{\phi}, z_{t-1} \sim \text{Bernoulli}(\phi_{t-1} z_{t-1}) \text{ for } t = 2...T
\end{gathered}
$$

where x_t is a datum taking 1 if the individual was recaptured at time t and 0 otherwise, p_t is the probability that the species is recaptured during a resampling event at time t given that it is alive at that time, ϕ_t is the probability that the species survives from time t to time $t+1$, and z_t is a latent variable representing whether the individual is alive ($z_t = 1$) or dead ($z_t = 2$) at time t. Time $t = 1$ corresponds to the first capture of the individual. Since individuals always begin in the alive state with a successful capture, the datum x_1 does not contribute to the likelihood of the parameters.

The Cormack-Jolly-Seber model is a special case of the hidden Markov model with two observed states, “captured” or “not captured,” two hidden states, “alive” or “dead,” and with the probability of transitioning from the “dead” to “alive” state fixed at 0. Marginalization is achieved following the method used for hidden Markov models (see above for details).

nimbleEcology distributions: `dCJS_**`

nimbleEcology provides four parameterizations of the Cormack-Jolly-Seber model. The survival and detection probabilities can each be time-invariant (scalar) or time-variant (vector); combining these possibilities gives four combinations, `dCJS_ss`, `dCJS_sv`, `dCJS_vs`, and `dCJS_vv`.

nimbleEcology expects the initial capture to be included in the data, so `x[1]` should always equal 1. In the time-variant parameterizations, the length of the detection vector `probDetection` should be the same as the length of `x`, although its first value is ignored. The length of the survival probability vector `probSurvive` should be one less than the length of `x`, with `probSurvive[t]` giving the probability of surviving from time t to $t+1$.

Example code

The following code chunk defines a NIMBLE model where a detection history for an individual is associated with the Cormack-Jolly-Seber model.

```
codeCJS <- nimbleCode({
  for (i in 1:N) {
    obs[i, 1:Tt] ~ dCJS_vv(
      probSurvive = phi[1:(Tt-1)],
      probCapture = p[1:Tt],
      len = Tt)
  }
})
```

N-mixture model

Model definition

The N-mixture model is a hierarchical model for estimating the abundance of a species from count data when detection is imperfect [114]. Under this model, a site has a true fixed number of individuals, represented by a latent state N drawn from a Poisson distribution whose mean represents the mean expected abundance at the site. Then, on each of J replicate visits to the site, an observed count is drawn from a binomial distribution with size N and a detection probability representing the chance that each individual is detected. We write this as

$$x_j|N, p_j \sim \text{Binom}(N, p_j) \text{ for } j = 1...J$$

$$N|\lambda \sim \text{Poisson}(\lambda)$$

N-mixture models are commonly used in ecology for to estimate the abundance or relative abundance of a species from counts produced by repeated site visits [53, 14, 43].

Unlike for preceding distributions, estimating a marginalized probability for an N-mixture model is nontrivial. Because the latent space of N is infinite, marginalizing over its possible values requires approximating an infinite sum. nimbleEcology follows the standard approach to approximating this infinite sum used in other marginalized implementations of the N-mixture model, such as those used in MLE packages like `unmarked` [46]. The user is required to specify a parameter `Nmax` (called K in other packages), the upper bound at which the sum is truncated, such that only values of N from 0...`Nmax` contribute to the likelihood. The tradeoff in this case is between accuracy of the estimate and computational efficiency; as `Nmax` increases, the calculated likelihood more closely approximates the full infinite sum, but computational time increases. Typically, one aims to choose a value of `Nmax` large enough such that the contribution of values of N greater than `Nmax` to the infinite sum are a negligible fraction of the total sum. This can be difficult to identify in advance, and users are advised to pay careful attention to this choice, especially in light of a known tendency for the N-mixture model to estimate near-infinite abundance [78, 53].

nimbleEcology uses a fast algorithm for calculating the truncated infinite sum developed by Meehan et al. which takes advantage of recursive properties of the Poisson and binomial likelihoods [96]. The Meehan algorithm, covering the standard Poisson-binomial case, was extend to the beta-binomial and negative binomial cases by Goldstein and de Valpine (see Chapter 3) [53]. This fast algorithm dramatically improves computation time especially as

`Nmax` becomes large in the beta-binomial, as it requires only a single call to the underlying beta function that can be slow to calculate rather than `Nmax` calls.

nimbleEcology distributions: `dNmixture_⋆⋆`

nimbleEcology provides eight versions of N-mixture distribution functions. These are best thought of as four model variants of the N-mixture distribution, each of which has two parameterizations. Unlike preceding model families, the four N-mixture variants differ in their underlying definitions.

In the standard N-mixture model defined above, the latent state is drawn from a Poisson distribution and the observed data are drawn from a binomial distribution. The Poisson and binomial may be substituted for a negative binomial and beta-binomial distribution, respectively, to accommodate overdispersed data [114, 92]. In each case, an extra parameter (s or θ) is added that determines the variance of the distribution. nimbleEcology offers all combinations of these distributions, yielding four forms of N-mixture model: `dNmixture_⋆`, the standard binonmial-Poisson; `dNmixture_BNB_⋆`, the binomial-negative binomial; `dNmixture_BBP`, the beta-binomial-Poisson; and `dNmixture_BBNB_⋆`, the beta-binomial-negative binomial.

Each of these four variants may be suffixed with either `_v` or `_s` to indicate whether detection probability p is provided as a vector (potentially different for each replicate) or scalar (constant), respectively, giving eight total distribution functions.

For the two Poisson variants of the N-mixture model, nimbleEcology's distribution offers the option to dynamically select a value of `Nmax`, the upper bound of the truncated infinite sum, by setting the parameters `Nmin` and `Nmax` to -1. Note that this option should only be used during MCMC model estimation and not during MLE estimation, as dynamic calculations interfere with optimization.

Example code

The following code chunk defines a NIMBLE model using an N-mixture distribution.

```
codeNmixBBNBV <- nimbleCode({
  for (i in 1:N) {
    obs[i, 1:J] ~ dNmixture_BBNB_v(
      lambda = lambda[i],
      prob = p[i, 1:J],
      theta = theta,
      s = s,
      Nmin = 0,
      Nmax = 1000,
      len = J
    )
  }
})
```

2.5 Worked example: estimating a dynamic occupancy model with maximum likelihood

In this section, I provide an example estimating a dynamic occupancy model with MLE using simulated data.

Data context

In this example, the goal of modeling is to estimate rates of colonization and persistence in the occupancy of a bird species at various sites [89]. We also want to understand how colonization and persistence vary with two environmental covariates of interest, forest cover and water availability, and we want to take into account variation in detection driven by wind conditions. Examples of dynamic occupancy models applied to observations of birds to understand their site-level colonization and extinction dynamics include Yackulic et al. [142] and Donaldson et al. [39].

I simulated data collected at 100 sites throughout a study region. At each site, four replicate detection-nondetection observations of the species are made in each of six consecutive years, giving a total of 24 detection or nondetection data at each site. During each field survey, we measured wind speed, which we suspect impacts bird detectability as high winds make it harder to identify bird calls. We also measured percent forest cover and water availability at each site, which we think influences colonization and persistence probabilities.

We wish to estimate the model

$$
\begin{gathered}
x_{i,j,t} \sim \text{Bernoulli}(z_{i,t} p_{i,j,t}) \\
z_{i,1} \sim \text{Bernoulli}(\psi) \\
z_{i,t} \sim \text{Bernoulli}(\pi_i z_{i,t-1} + \gamma_i (1 - z_{i,t-1})) \text{ for } t = 2...T \\
\text{logit}(\pi_i) = \text{logit}(\pi_0) + \beta_{\pi 1}\text{forest}_i + \beta_{\pi 2}\text{water}_i \\
\text{logit}(\gamma_i) = \text{logit}(\gamma_0) + \beta_{\gamma 1}\text{forest}_i + \beta_{\gamma 2}\text{water}_i \\
\text{logit}(p_{i,j,t}) = \text{logit}(p_0) + \beta_{p1}\text{wind}_{i,j,t}
\end{gathered}
$$

which has nine parameters: the initial occupancy probability at a site (ψ), persistence intercept and covariate coefficients (π_0, $\beta_{\pi 1}$, $\beta_{\pi 2}$), colonization intercept and covariate coefficients (γ_0, $\beta_{\gamma 1}$, $\beta_{\gamma 2}$), and detection intercept and covariate coefficients (p_0, β_{p1}). For more detail on the structure of the dynamic occupancy model, see section 4.4.

Defining, building, and compiling the NIMBLE model

First, nimbleEcology must be installed and loaded.

```
install.packages("nimbleEcology")
library(nimbleEcology)
```

Once we've loaded the nimbleEcology package, the first step in the process of estimating any model is to define the model in NIMBLE code. For the sake of simplicity, basic familiarity with BUGS-like syntax is assumed; for more information on this, see the NIMBLE User Manual at `https://r-nimble.org/manuals/NimbleUserManual.pdf`.

In our model, colonization and extinction probabilities are constant at a site across all seasons (since forest cover and water availability don't change between seasons), while detection probability varies from visit to visit based on wind speed. In this case, we will use the nimbleEcology custom distribution `dDynOcc_ssm`, where the suffix `_ssm` indicates that persistence and colonization probabilities are time-invariant scalars, while the detection probability varies across seasons and surveys.

Using the marginalized distribution `dDynOcc_ssm`, we avoid the need to define a latent state z in model code. This is a necessary step for MLE. Note that since we intend to estimate the model via maximum likelihood, we do not include priors in the model code.

The full model code can be written as:

```
# NIMBLE models are defined using the function nimbleCode
dynocc_model_code <- nimbleCode({

  # Loop over sites
  for (i in 1:nsite) {

    # Calculate logit-linked persistence and
    # colonization by site
    logit(pers[i]) <- logit(per.int) +
      per.beta.forest * forest[i] +
      per.beta.water * water[i]
    logit(coln[i]) <- logit(col.int) +
      col.beta.forest * forest[i] +
      col.beta.water * water[i]

    # Rather than defining a latent state node,
    # we use dDynOcc_ssm to directly
    # link observed data to model parameters
    y[i, 1:nyear, 1:nrep] ~ dDynOcc_ssm(
        init = init,
        probPersist = pers[i],
        probColonize = coln[i],
        p = det[i, 1:nyear, 1:nrep],
        start = start[1:6],
        end = end[1:6]
    )

    # Loop over year and visit,
    # calculate logit-linked deteciton prob.
    for (j in 1:nyear) {
      for (k in 1:nrep) {
        logit(det[i,j,k]) <- logit(det.int) +
            det.beta.wind * wind[i,j,k]
      }
    }
  }
})
```

Having defined the model, we use the functions `nimbleModel` and `compileNimble` to build and compile the model. Though not necessary in this case, it is good practice to provide a list of initial values for all parameters when the model is built to confirm that the model likelihood can be calculated without error.

```
init_list <- list(
  det.int = 0.5,
  det.beta.wind = 0,
  col.int = 0.5,
  col.beta.water = 0,
  col.beta.forest = 0,
  per.int = 0.5,
  per.beta.water = 0,
  per.beta.forest = 0,
  init = 0.5
)

dynocc_model <- nimbleModel(dynocc_model_code,
    constants = list(
      nyear = nyear, nsite = nsite, nrep = nrep,
      wind = wind_data, forest = site_data$forest,
      water = site_data$water,
      start = rep(1, 6), end = rep(4, 6)
    ),
    data = list(
      y = dnd_data
    ),
    inits = init_list)

Cdynocc <- compileNimble(dynocc_model)
```

We now have a compiled model object whose likelihood can be calculated with the `Cdynocc$calculate()` method.

Maximum likelihood estimation

The principle behind MLE with NIMBLE models is to treat the model object log likelihood calculation as an objective function for use with an optimizer. In this case, the objective function takes a vector of model parameter values, assigns them to the relevant nodes in the model object, and calculates and returns the model's log likelihood. An optimizer like R's `optim()` can use a variety of algorithms to identify the maximum log likelihood value over the parameter space.

We can use a NIMBLE custom function to create the necessary objective function. Since this function will set parameter values and calculate a log likelihood, I name it `setAndCalculate`. We define this as a custom nimbleFunction using a slightly different syntax than that used to define custom distributions. The `setAndCalculate` function will have a "setup" as well as a "run" function. The "setup" function will take as arguments the set of node names corresponding to the parameters over which we will optimize and the model object itself. It will return a "set-and-calculate" object with a "run" function that sets the values of the specified model nodes, then calculates and returns the log-likelihood of the model with those values. This object can itself be compiled. Note that this `setAndCalculate` function can be used as written for an arbitrary NIMBLE model.

```
setAndCalculate <- nimbleFunction(
  name = 'setAndCalculate',
  setup = function(model, targetNodes) {
    targetNodesAsScalar <-
```

```
      model$expandNodeNames(targetNodes,
                            returnScalarComponents = TRUE)
    calcNodes <- model$getDependencies(targetNodes)
  },
  run = function(targetValues = double(1)) {
    values(model, targetNodesAsScalar) <<- targetValues
    lp <- calculate(model, calcNodes)

    returnType(double())
    return(lp)
  }
)

target_nodes <- c("det.int", "det.beta.wind",
  "col.int", "col.beta.water", "col.beta.forest",
  "per.int", "per.beta.water", "per.beta.forest",
  "init")

set_and_calc_obj <- setAndCalculate(dynocc_model,
                                    target_nodes)

Csco <- compileNimble(set_and_calc_obj)
```

Now that the `setAndCalculate` function has been defined, instantiated, and compiled, the likelihood maximum can be identified using R's `optim`. We provide a vector of initial values, the objective function (the `run` method of our compiled `setAndCalculate` object), and other control arguments. Since the log likelihood function is to be maximized but `optim` finds a minimum by default, the control argument `fnscale=-1` must be included, which multiplies the objective function by -1 before optimizing and therefore finds the maximum of the original function.

```
optim_result <- optim(par = unlist(init_list),
                      fn = Csco$run,
                      method = "BFGS",
                      control = list(fnscale = -1,
                                     maxit = 10000),
                      hessian = TRUE)
```

Following maximum likelihood theory, we estimate the variances of the parameter estimates as the diagonal of the inverse of the Hessian matrix evaluated at the likelihood maximum. Since we multiplied the function by -1 for optim, we need to undo the operation before taking the inverse of the Hessian matrix.

This procedure yields an estimated value for each parameter with an accompanied standard error. We can visualize these estimates and compare them to the true values used during simulation (Figure 2.1). In this case, the parameters used for simulation fall within the 95% confidence intervals of each estimated parameter (approximated as the point estimate ±1.96 times the estimated standard error).

```
optim_SE <- sqrt(diag(solve(-optim_result$hessian)))

parameter_estimates <- data.frame(
  param = target_nodes,
  est = optim_result$par,
  se = optim_SE,
  type = "MLE"
)
```

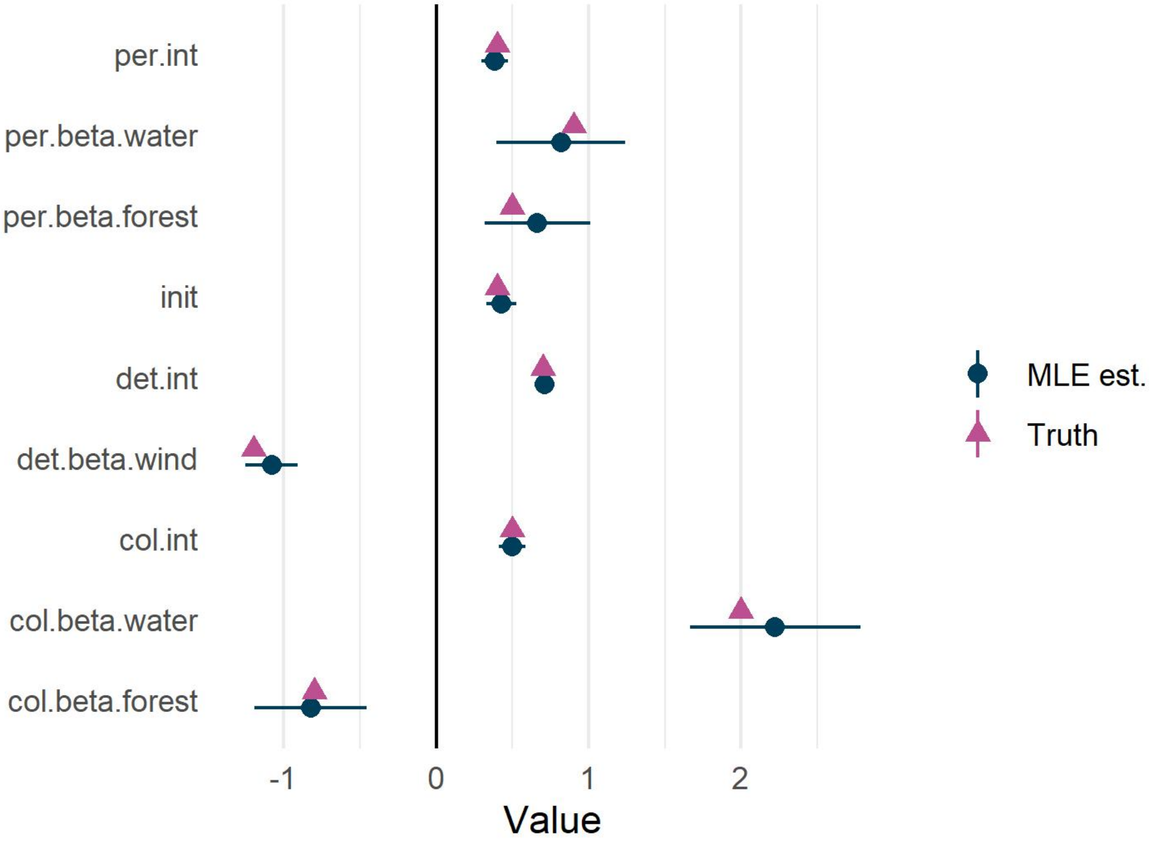

Figure 2.1: Comparing maximum likelihood estimates of dynamic occupancy model parameters to true parameters used in simulation. Blue points show point estimates with approximate 95% confidence intervals (point estimate ±1.96 times the estimated standard error). Purple triangles show the values used for simulation. In this case, all true values fell within their corresponding estimated 95% confidence intervals.

2.6 Worked example: estimating a dynamic hidden Markov model with MCMC

In this example, I show how a nimbleModel with a nimbleEcology custom distribution can be used to simulate such data and illustrate the process of estimating a model with MCMC. As a case example, I focus on the dynamic hidden Markov model (DHMM). The DHMM is sometimes used in wildlife studies to associate movement data with behavioral states. Recent published applications of HMMs to characterize behavioral states from movement data include estimating the impacts of sound disturbance on blue whale feeding behavior [36], classifying albatross flight types from accelerometer data [24], and identifying patterns of foraging in macaroni penguins [60]. In this example, we will consider a simplified version of the data used in these studies to illustrate the use of nimbleEcology distributions for data simulation and MCMC estimation.

Specifying the model

Rather than starting with a simulated dataset and writing code to suit it, in this example we will first define and construct a nimbleModel, then use it to simulate data. We will specify the model in BUGS code, set values for parameters, and simulate observed data following that model. We will then be able to use the same model object to estimate the parameters based on the simulated data.

Consider a study using accelerometers to monitor species behavior. We collect observation data for 20 individuals of a species. Every hour for three days, we categorize each individual's movement pattern into one of three observable states: (O1) not moving, (O2) moving slowly, or (O3) moving quickly. These observable states are meant to correspond to three true behavioral states: (S1) sleeping, (S2) foraging, or (S3) traveling. However, there is a possibility of misclassification, so the true state is unobserved. We have 192 observations per individual, representing one observed behavioral state in each hour over an eight-day period.

For a given true state, an observation probability matrix will give the probability that the species is detected in each category. We think that when the species is sleeping, we will always correctly categorize it as not moving, so the first row in the observation matrix is $(1, 0, 0)$ representing that perfect correspondence. We think that foraging individuals may be miscategorized as sleeping, and moving individuals may be miscategorized as foraging, so those rows in the observation matrix will have multiple nonzero elements. Specifically, the second row will have $(1 - p_f, p_f, 0)$ and the third row will have $(0, 1 - p_m, p_m)$, where p_m and p_f give the probabilities of accurately categorizing a moving and foraging individual, respectively.

For a given true state, an array of transition probabilities gives the probability of transitioning from each state into each state between time steps. For the sake of simplicity in this example, we think an individual will only transition from the sleeping state to the foraging state, from the foraging state to the moving state, and from the moving state to the sleeping state. The probability of each of these transitions will vary with time of day. We parameterize these as logit-linked quadratic relationships with hour of day, plus a random effect of individual. We will add additive random effects of individual to these probabilities representing behavioral variation between individuals. (See Figure 2.4 for the true transition probabilities by time of day.)

Our model will ultimately have 17 parameters to estimate: a length-three vector of initial probabilities (constrained to sum to one, these are really two parameters, although from an MCMC estimation perspective three nodes will be sampled); the probability of correctly observing a foraging animal as foraging; the probability of correctly observing a moving animal as moving; nine parameters linking transition probabilities to quadratic functions with time of day; and the standard deviations of three normally distributed random effects of individual on transition probability.

To begin, we specify the model in NIMBLE code. We manually specify elements of the observation matrix in terms of parameters, and define the relationships between transition parameters and the transition matrix. For the sake of monitoring mean transition probabilities across species, we separately define nodes for those means on the logit scale (such as `logit_pwake`) and for individual transition probabilities (with random effects added; `ind_pwake`).

```
dDHMM_movement_model_code <- nimbleCode({

  # Observation probability matrix
  probObs[1,1] <- 1
  probObs[1,2] <- 0
  probObs[1,3] <- 0
  probObs[2,1] <- 1 - obs22
  probObs[2,2] <- obs22
  probObs[2,3] <- 0
  probObs[3,1] <- 0
  probObs[3,2] <- 1 - obs33
  probObs[3,3] <- obs33

  for (t in 1:ntime) {
    logit_pwake[t]  <- pw.int + pw.b1 * time[t] +
                       pw.b2 * time[t]^2
    logit_pmove[t]  <- pm.int + pm.b1 * time[t] +
                       pm.b2 * time[t]^2
    logit_psleep[t] <- ps.int + ps.b1 * time[t] +
                       ps.b2 * time[t]^2
  }

  # Time-variant transition probabilities
  for (k in 1:n_individuals) {

    # Define random effects as normally distributed
    ranef_wake[k] ~  dnorm(0, sd = sd_wake)
    ranef_move[k] ~  dnorm(0, sd = sd_move)
    ranef_sleep[k] ~ dnorm(0, sd = sd_sleep)

    for (t in 1:(ntime-1)) {
      logit(ind_pwake[t,k])  <- logit_pwake[t] + ranef_wake[k]
      logit(ind_pmove[t,k])  <- logit_pmove[t] + ranef_move[k]
      logit(ind_psleep[t,k]) <- logit_psleep[t] +
                                    ranef_sleep[k]

      probTrans[1,1,t,k] <- 1 - ind_pwake[t,k]
      probTrans[1,2,t,k] <- ind_pwake[t,k]
      probTrans[1,3,t,k] <- 0

      probTrans[2,1,t,k] <- 0
      probTrans[2,2,t,k] <- 1 - ind_pmove[t,k]
      probTrans[2,3,t,k] <- ind_pmove[t,k]

      probTrans[3,1,t,k] <- ind_psleep[t,k]
      probTrans[3,2,t,k] <- 0
      probTrans[3,3,t,k] <- 1 - ind_psleep[t,k]
    }

  ### dDHMM likelihood
    y[k, 1:ntime] ~ dDHMM(
        probObs = probObs[1:3, 1:3],
        probTrans = probTrans[1:3, 1:3, 1:(ntime-1), k],
        init = inits[1:3],
        len = ntime,
        checkRowSums = 0
    )
  }


  #### Priors
  inits[1:3] ~ ddirch(alpha = init_alphas[1:3])

  # Detection probs.
  obs22 ~ dunif(0, 1)
  obs33 ~ dunif(0, 1)

  # Transition parameters
  pw.int ~ dnorm(0, sd = 50)
```

```
  pm.int ~ dnorm(0, sd = 50)
  ps.int ~ dnorm(0, sd = 50)
  pw.b1  ~ dnorm(0, sd = 50)
  pm.b1  ~ dnorm(0, sd = 50)
  ps.b1  ~ dnorm(0, sd = 50)
  pw.b2  ~ dnorm(0, sd = 50)
  pm.b2  ~ dnorm(0, sd = 50)
  ps.b2  ~ dnorm(0, sd = 50)

  # Random effect standard deviations
  sd_wake  ~ dunif(0, 10)
  sd_move  ~ dunif(0, 10)
  sd_sleep ~ dunif(0, 10)
})
```

When simulating data, we need to specify constants when the nimbleModel is built. Constants are values that define data dimensions and structure; they are "baked in" to the model graph and cannot be changed once the model is built. We will provide our covariate vector, time, as a constant as well. We must set initial values for our parameters that we want to simulate from. We do not provide individual detection histories, since we want to simulate them. This step instantiates an NIMBLE model object that has initial values and constants and a full graph but `NA` values for all observations.

```
DHMM_model <- nimbleModel(
  code = dDHMM_movement_model_code,
  constants = list(
    time = rep(0:23, 8),
    ntime = length(rep(0:23, 8)),
    n_individuals = 20,
    init_alphas = c(1,1,1)
  ),
  inits = list(
    inits   = c(0.8, 0.19, 0.01),
    obs22   = 0.8,
    obs33   = 0.7,
    pw.int  = -15.5,
    pw.b1   = 2.05,
    pw.b2   = -.06,
    pm.int  = 0,
    pm.b1   = -0.3,
    pm.b2   = 0.015,
    ps.int  = -3,
    ps.b1   = 3,
    ps.b2   = -0.3,
    sd_wake = 0.1,
    sd_sleep = 0.2,
    sd_move = 0.1
  ),
  calculate = FALSE
)
```

Note that attempting to calculate the model likelihood (a default behavior of `nimbleModel`) without data in this case prints an error message, which is expected behavior. By setting `calculate = FALSE` when building the model we avoid errors related to the fact that data are not yet valid.

Simulating data

Now that we have a model object without data, we can use it to simulate data. First, we need to populate the observation and transition arrays. Since some of these elements don't depend on any parameters (i.e. those that are set to 1 and 0) they will be NA until model calculation occurs. We can do this by using the `calculate` function and providing nodes to calculate.

```
nodes_to_calc <- c("probObs", "probTrans")
DHMM_model$calculate(nodes_to_calc)
```

To make sure to simulate all nodes downstream from our model parameters, including random effects as well as observations, we start by listing the parameter nodes in the model, then ask the model to identify every node that depends on those. We then pass that list of nodes to the model to simulate. Deterministic nodes downstream of parameters, such as elements of the probability arrays, will be calculated, while stochastic data nodes will be simulated.

```
target_params <- c("inits", "obs22", "obs33",
                   "pw.int", "pw.b1", "pw.b2",
                   "pm.int", "pm.b1", "pm.b2",
                   "ps.int", "ps.b1", "ps.b2",
                   "sd_move", "sd_wake", "sd_sleep")

nodes_to_sim <- DHMM_model$getDependencies(
    target_params, self = F, downstream = T
  )

DHMM_model$simulate(nodes_to_sim)
```

For each individual, we've simulated response data from the model as specified with the initial parameter values. Note that since we used the marginalized distribution dDHMM, we did not simulate true latent states.

We can visualize the individuals' simulated observation histories as timeseries (Figure 2.2).

Before MCMC estimation, it is important to indicate to the model via the `setData()` method that the observations we simulated are now "data" in the sense that we do not want the `y` nodes to be sampled.

```
DHMM_model$setData("y")
```

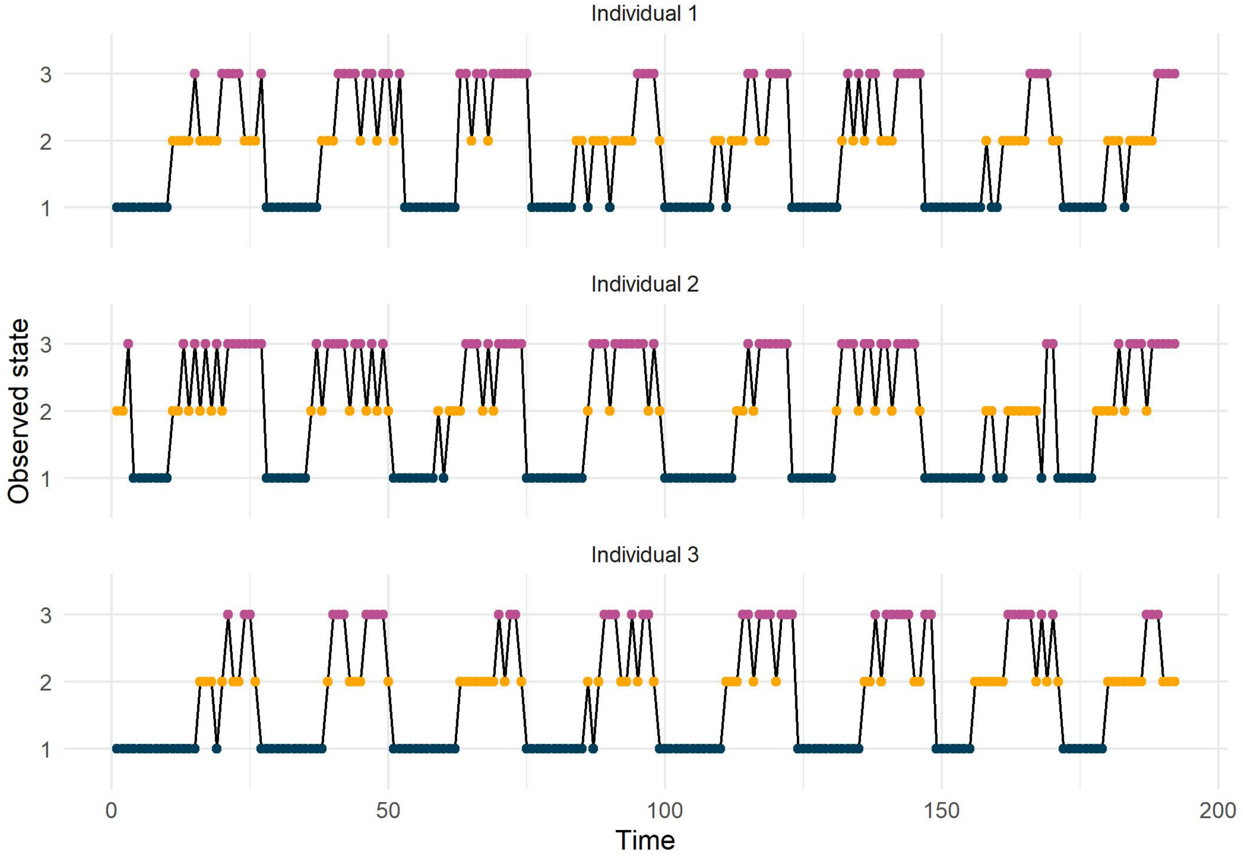

Figure 2.2: Simulated observations over eight days from a dynamic hidden Markov model. Transitions between three true states are functions of time of day with individual-level random variation, and states are observed imperfectly.

Estimating model parameters with MCMC

Now we are ready to estimate model parameters using MCMC.

First, we need to build an MCMC object. NIMBLE will automatically identify the stochastic nodes in the model, including parameters and random effects, and assign default samplers. In this case, I want to use custom samplers. For efficient mixing, it makes sense for each set of three parameters describing how transition probabilities vary over time to be sampled together. One appropriate choice is the adaptive factor slice sampler (AFSS). To set these samplers, we create an MCMC configuration object using `configureMCMC()`, call `removeSampler()` to drop the default random walk samplers for those nodes, then use the `addSampler` with argument `type = "AF_slice"` for each set of three parameters.

```
DHMM_MCMC_conf <- configureMCMC(DHMM_model)

# Remove default RW samplers for transition prob. parameters
DHMM_MCMC_conf$removeSampler(c("pm.b1", "pm.b2", "pm.int",
                               "pw.b1", "pw.b2", "pw.int",
                               "ps.b1", "ps.b2", "ps.int"))

# Add AF slice samplers for each group of 3 parameters
DHMM_MCMC_conf$addSampler(
  target = c("pm.b1", "pm.b2", "pm.int"), type = "AF_slice")
DHMM_MCMC_conf$addSampler(
  target = c("pw.b1", "pw.b2", "pw.int"), type = "AF_slice")
DHMM_MCMC_conf$addSampler(
  target = c("ps.b1", "ps.b2", "ps.int"), type = "AF_slice")

DHMM_MCMC_conf
```

Since we are interested in how transition probabilities change over time, we also add monitors on the vectors `pwake`, `psleep`, and `pmove` before we finally build the MCMC object. Adding these monitors will store the value of these derived quantities at each MCMC iteration, yielding a posterior distribution of each value. We can examine posterior predictive distributions of these nodes.

```
DHMM_MCMC_conf$addMonitors("logit_pwake",
      "logit_psleep", "logit_pmove")

DHMM_MCMC <- buildMCMC(DHMM_MCMC_conf)
```

Having built a model and an MCMC object, we can compile them. If we provide both to `compileNimble()` in a single call, the function will return a named list whose elements are the compiled components.

```
compiled_objs <- compileNimble(DHMM_model, DHMM_MCMC)
```

Now we use `runMCMC()` to execute the MCMC algorithm. I use two chains of 5000 iterations each, with a burn-in period of 1000 samples. I also specify new initial values in order not to begin sampling at the true values used for simulation.

```
samples <- runMCMC(compiled_objs$DHMM_MCMC, ni = 5000, nc = 2,
                  nburnin = 1000,
                  samplesAsCodaMCMC = TRUE, inits = list(
                      inits = c(0.33, 0.34, 0.33),
                      obs22 = 0.5, obs33 = 0.5,
                      pw.int = 0, pw.b1  = 0, pw.b2  = 0,
                      pm.int = 0, pm.b1  = 0, pm.b2  = 0,
                      ps.int = 0, ps.b1  = 0, ps.b2  = 0,
                      sd_wake = 0.2, sd_move = 0.2,
                      sd_sleep = 0.2
                  ))
```

We examine trace plots of some parameters (Figure 2.3). The mixing for these parameters looks good.

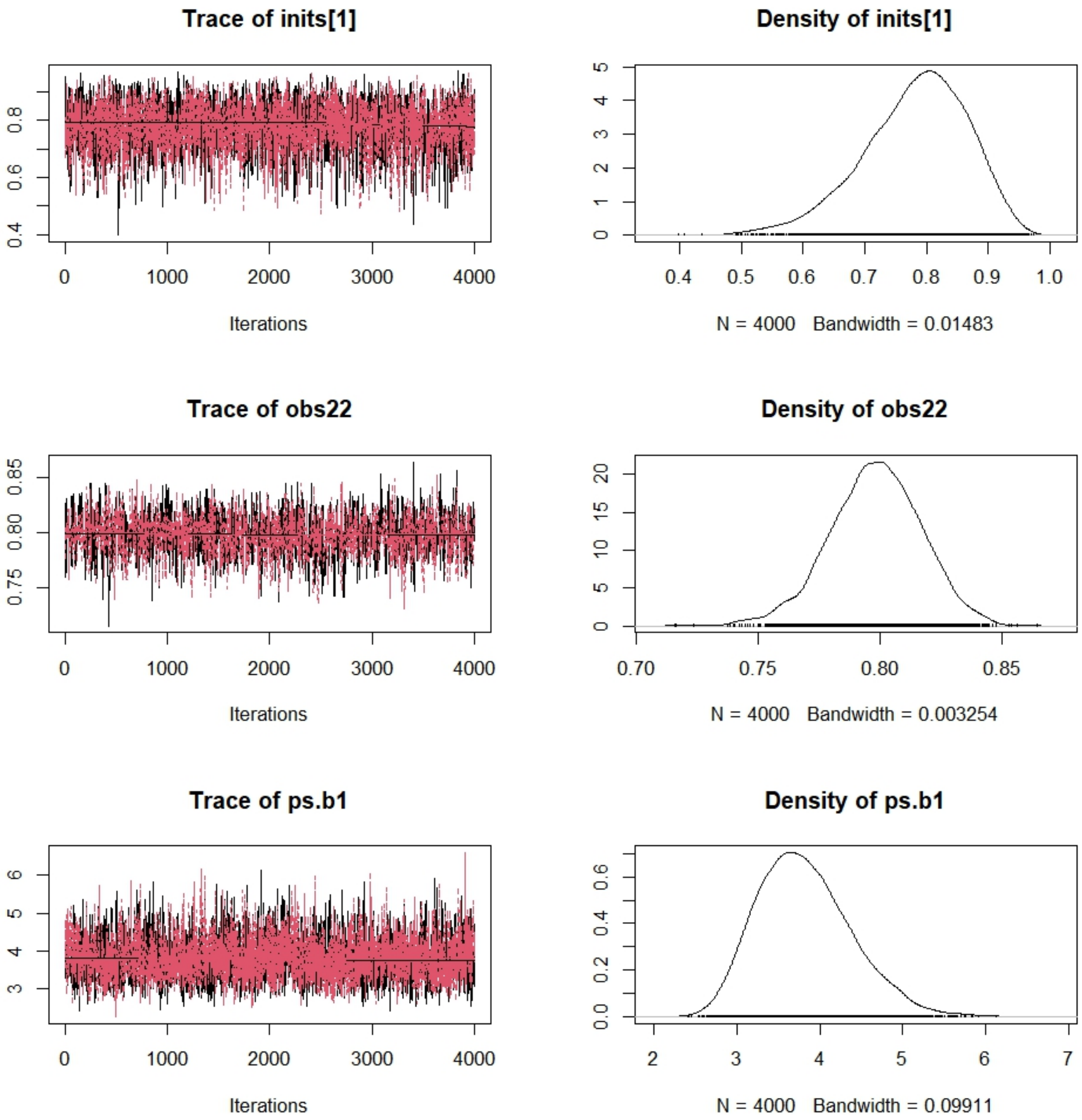

Figure 2.3: Trace plots for three estimated parameters. “Grassy” behavior and similar means and variances between chains indicates good mixing.

We can then visualize the estimated relationships between time and each transition probability using posterior predictions for each transition probability vector. With some manipulation, we can use the summaries of these parameters to visualize how we predict transition probabilities to vary through time, and compare these to the true values used in simulation (Figure 2.4).

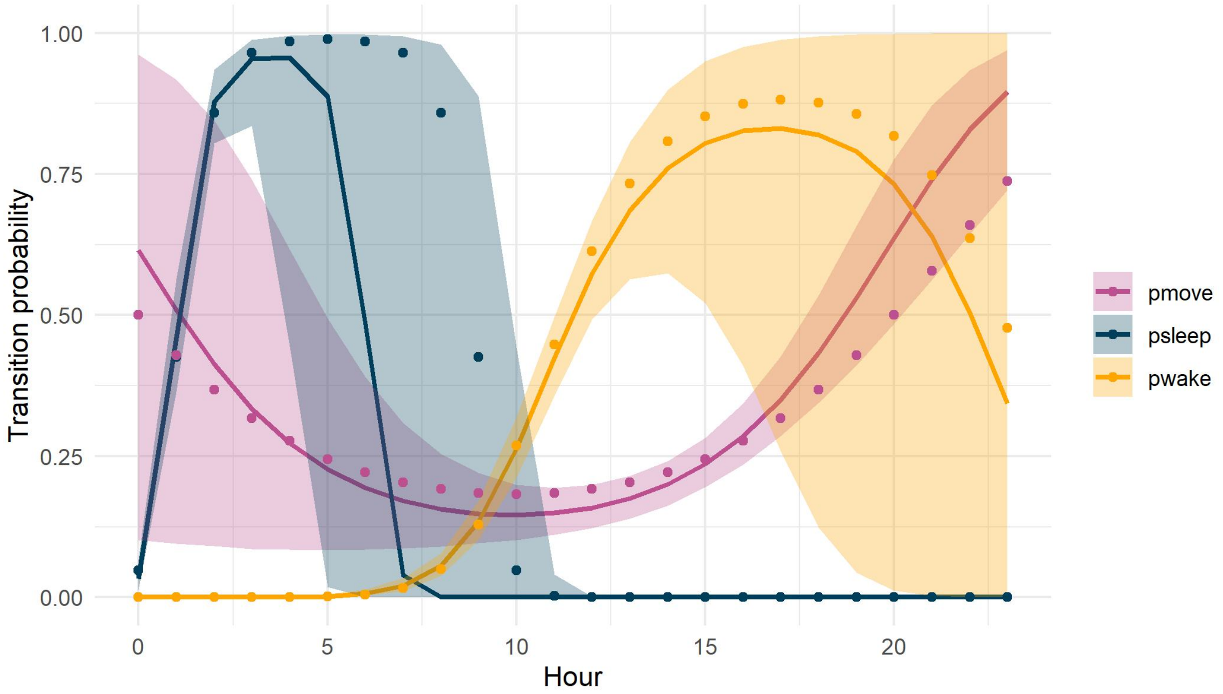

Figure 2.4: Comparing 95% credible regions of three transition probabilities (shaded regions; lines indicate median estimates) to the true values used in simulation (points). True values mostly fall within 95% credible intervals.

2.7 Worked example: choosing between a marginalized and a latent state formulation for a Cormack-Jolly-Seber model

When there is no *a priori* reason to prefer marginalization over sampling latent states, modelers using MCMC may choose between these two methods based on their efficiency. Improvements in MCMC estimation efficiency due to marginalization are highly context-dependent and may depend on data dimensions and the presence of additional hierarchical structure as well as on the type of model under consideration [103]. For this reason, it is recommended that the modeler evaluate the efficiency gains of marginalization on a case-by-case basis.

The goal of this worked example is to demonstrate how to compare the estimation efficiency of a marginalized model to that of a latent state formulation of the same model. The workflow we will use is simple and generic. First, we will create two separate model definitions in nimbleCode specifying the same model using marginalized and latent state formulations. We will be careful to make sure that the two models are statistically identical. We will build each model and generate MCMC samples. Then we will compare the effective sample size and computational time for both models to determine which is more efficient.

Data context

In this example, I consider a simple mark-recapture model based on Stewart et al. [120]. The authors collected marked and re-sighted Yaqui catfish (*Ictalurus pricei*) in San Bernardino National Wildlife Refuge, Arizona, USA. They fit a Cormack-Jolly-Seber mark-recapture model to these data with the goal of estimating survival rates and capture probabilities of the species. The authors included season-specific intercepts for capture probabilities and survival probabilities, such that the model has 32 parameters to be estimated (16 season-specific capture probability intercepts and 16 season-specific survival probability intercepts). The latent state formulation of the model to be estimated is given by

$$y_{i,j} | z_{i,j}, p_j \sim \text{Bernoulli}(z_{i,j} p_j)$$

$$z_{i,j} | z_{i,j-1} \sim \text{Bernoulli}(z_{i,j-1} S_j)$$

where the length-16 parameter vectors $\boldsymbol{S}$ and $\boldsymbol{p}$ will be estimated. We put a uniform prior from 0 to 1 on each of these parameters.

I simulated data approximating the data dimensions and parameters estimated by Stewart et al. In this example, capture-recapture data are collected for each of 350 individuals over a 16-year period. In the first year, 50 fish are captured for the first time, with an additional 50 new fish being added in each of the following 6 years. We begin the exercise with capture histories of each fish in each season (Figure 2.5).

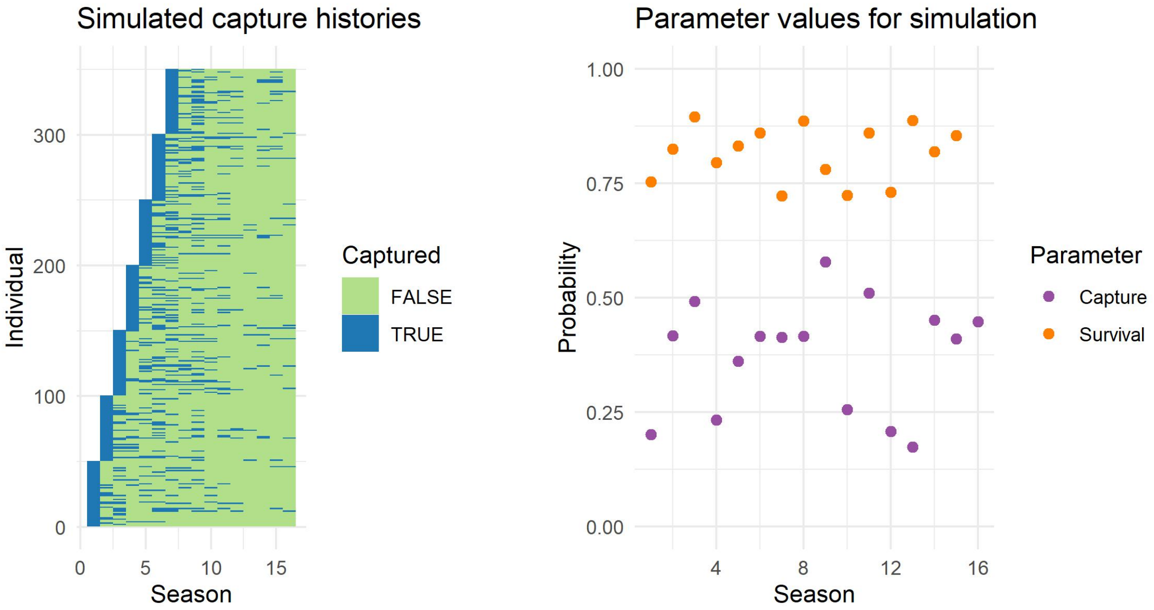

Figure 2.5: (Left) Simulated capture histories for 350 Yaqui catfish. 50 individuals are added each season. (Right) Season-specific probabilities of survival and capture used to generate capture histories.

Defining the marginalized and latent state model formulations

To begin, we write NIMBLE code for each model. Each version of the model code has two components. First I define parameter vectors for survival probability and capture probability, which I call `pcap` and `psurv`, as following uniform priors. This component is the same in both models. Then, I relate these probabilities to the data. In the marginalized version, I loop over individuals and simply link each individual's detection history to the CJS model using the nimbleEcology function `dCJS_vv`, which gives the probability distribution for time-variant survival and capture probabilities. In the latent state formulation, I loop over individuals and seasons, linking the data to parameters via two Bernoulli distributions and a latent state, `z`, representing whether each individual is alive at each time step.

```
marginalized_CJS <- nimbleCode({
  #### Set priors for parameters (same in both models) ####
  # Uniform prior on season-specific survival prob. and
  # mean capture prob.
  for (j in 1:(nseason - 1)) {
    psurv[j] ~ dunif(0, 1)
    pcap[j + 1] ~ dunif(0, 1)
  }
  pcap[1] <- 1 # Capture at t=1 is ignored,
               # so we don’t need to sample it

  #### Relate data to parameters ####
  # Loop over individuals and associate each capture history
  # with the CJS model using a single line of code
  for (i in 1:nind) {
    y[i, start[i]:nseason] ~ dCJS_vv(
      probSurvive = psurv[start[i]:(nseason - 1)],
      probCapture = pcap[start[i]:nseason],
      len = nseason - start[i] + 1
    )
  }
})

latentstate_CJS <- nimbleCode({
  #### Set priors for parameters (same in both models) ####
  # Uniform prior on season-specific survival prob.
  # and mean capture prob.
  for (j in 1:(nseason-1)) {
    psurv[j] ~ dunif(0, 1)
    pcap[j + 1] ~ dunif(0, 1)
  }
  pcap[1] <- 1 # Capture at t=1 is ignored,
               # so we don’t need to sample it

  #### Relate data to parameters ####
  # Loop over individuals and define the relationship between
  # data (y), latent states (is_alive), and
  # probabilities of survival and capture (psurv and pcap)
  for (i in 1:nind) {

    is_alive[i, start[i]] <- 1

    for (j in (start[i] + 1):nseason) {
      is_alive[i, j] ~ dbern(psurv[j - 1] *
                             is_alive[i, j - 1])

      y[i, j] ~ dbern(pcap[j] * is_alive[i, j])
    }
  }

})
```

Comparing model efficiency

Having specified both models in NIMBLE code, we need to build and generate MCMC samples for both. We start by creating model objects and providing initial values for all model parameters.

```
latentstateMod <- nimbleModel(
  code = latentstate_CJS,
  constants = list(
    nind = nind,
    nseason = nseason,
    start = start_vec),
  data = list(y = y),
  inits = list(
    psurv = rep(0.5, nseason - 1),
    pcap = rep(0.5, nseason),
    is_alive = matrix(1, nrow(y), ncol(y))
  )
)

marginalizedMod <- nimbleModel(
  code = marginalized_CJS,
  constants = list(
    nind = nind,
    nseason = nseason,
    start = start_vec
  ),
  data = list(y = y),
  inits = list(
    psurv = rep(0.5, nseason - 1),
    pcap = rep(0.5, nseason)
  )
)

marginalizedMod$calculate()
latentstateMod$calculate()
```

Note that, at model build time, NIMBLE provides two warnings relating to the fact that the capture histories contain NA elements (for seasons before an individual was observed) and due to the fact that survival probability and capture probability vectors are different lengths. These warnings are defensive and can be safely ignored in this case.

We use the standard NIMBLE workflow to build MCMC objects for both models and compile the models and MCMC objects.

```
# Build MCMC for both models, then compile everything
latentMCMC   <- buildMCMC(latentstateMod)
marginalMCMC <- buildMCMC(marginalizedMod)

Clatent   <- compileNimble(latentstateMod, latentMCMC)
Cmarginal <- compileNimble(marginalizedMod, marginalMCMC)
```

We now execute MCMC estimation of each model. To evaluate efficiency, it is not necessary that our MCMC chains are of sufficient length for final results. We only need to run enough samples to reliably estimate the effective sample size of the posterior and thus get a clear sense of the rate at which the model is generating effectively independent samples. While in this example, the MCMCs run quickly, for models that are quite slow to estimate it is useful to evaluate estimation efficiency without a full run.

I run each model for 2 chains of 1000 iterations each, with a burn-in period of 0 samples. I use the command `system.time()` to monitor runtime for each model.

```
latent_time <- system.time(
  samples_latent <- runMCMC(Clatent$latentMCMC,
                            niter = 1000,
                            nchains = 2, nburnin = 0,
                            samplesAsCodaMCMC = TRUE)
)

marginal_time <- system.time(
  samples_marginal <- runMCMC(Cmarginal$marginalMCMC,
                              niter = 1000,
                              nchains = 2, nburnin = 0,
                              samplesAsCodaMCMC = TRUE)
)

latent_summary   <- MCMCvis::MCMCsummary(samples_latent)
marginal_summary <- MCMCvis::MCMCsummary(samples_marginal)
```

In this example, we want to compare two features of sampling: time taken per MCMC iteration and effective sample size [103]. For the latter, we're specifically interested in the minimum ESS across all sampled nodes. We can query these features with the following commands.

```
marginal_time[3]
latent_time[3]

min(marginal_summary$n.eff[rownames(marginal_summary) !=
    "pcap[1]"])
min(latent_summary$n.eff[rownames(marginal_summary) !=
    "pcap[1]"])
```

We can define MCMC efficiency as the number of effectively independent samples generated per unit time [103]. We can calculate the number of independent samples generated per second by dividing minimum ESS by time taken for each model.

```
# Marginalized model efficiency
min(marginal_summary$n.eff[rownames(marginal_summary) !=
    "pcap[1]"]) / marginal_time[3]

min(latent_summary$n.eff[rownames(marginal_summary) !=
    "pcap[1]"]) / latent_time[3]
```

In this case, the marginalized model generated a minimum of 11.7 independent samples per second, while the latent state formulation generated 3.1 independent samples per second. We can conclude that estimation via the marginalized implementation of this model is more efficient.

2.8 Conclusion

nimbleEcology was created to make implementing many common ecological models in NIMBLE easier. Predefined custom distributions for ecological models grant flexibility in model implementation while reducing the incidence of coding error and saving time, giving users a middle ground between specifying a model completely from scratch and giving up control over model specification. Moreover, marginalized distributions open the door for more efficient model estimation and maximum likelihood estimation of hierarchical models.

The nimbleEcology R package is always growing and actively considering additional model distributions and package functionality. We welcome ideas for additional distributions, and especially appreciate the submission of code defining distributions for inclusion in the package.

For a comprehensive user guide to the NIMBLE modeling tool, see the NIMBLE user manual at `https://r-nimble.org/documentation-2`. For further information on nimbleEcology, see the package's official documentation on CRAN along with the package vignette at `https://cran.r-project.org/web/packages/nimbleEcology/index.html`. For additional context on the models described in this chapter, we recommend the two-volume textbook *Applied Hierarchical Modeling in Ecology* by Kéry and Royle [79].

Chapter 3

Comparing N-mixture models and GLMMs for relative abundance estimation with a citizen science dataset

This chapter was previously published as Goldstein and de Valpine, 2022 [53] and is included here with permission from the co-author.

In Chapter 2, I implemented a number of common ecological models using marginalization, enabling the use of maximum likelihood estimation. In this chapter, I apply these software advances to conduct a methodological comparison of two ecological models for relative abundance estimation in a maximum likelihood framework, specifically via the use of information criteria based in maximum likelihood theory. My findings, which may guide ecologists in choosing ecological models for their applications, depend on the maximum likelihood framework made possible in the previous chapter.

3.1 Background

Understanding how species' abundances are associated with covariates of interest is a primary goal of species distribution modeling [42]. Often the best one can hope to estimate are patterns of *relative abundance*, sometimes referred to as an index of abundance. Relative abundance values are equal to absolute abundance – the number of individuals in a known area – multiplied by an unknown constant such as a detection rate and/or effective area sampled. When the unknown constant can be assumed to be the same between two areas, the ratio of relative abundances is the same as the ratio of absolute abundances, allowing comparison of sites relative to each other. Relative abundance can sometimes be estimated from data when absolute abundance cannot, but doing so can be challenging when data are collected with heterogeneous sampling protocols because variation in the data collection process can obscure or confound those patterns. Large, heterogeneous datasets are becoming more prominent in ecology, such as those produced by citizen science [21, 118], autonomous recorders [48, 74], and camera traps [119]. Analyzing these data requires statistical models that fit the data well, account for details of study design and sampling if possible, and can be estimated efficiently.

Two existing model types can satisfy this requirement: generalized linear mixed models (GLMMs) and N-mixture models. The GLMM is a linear model extension that models relationships between non-Gaussian (e.g. count) response data while allowing hierarchical structure via random effects. To estimate relative abundance, count data are modeled as Poisson-distributed with an expected value defined by a log-linked linear combination of important covariate data, and a link-scale random effect is added to account for relatedness between replicate observations at a site [12]. The GLMM is a heuristic model developed to explain patterns but does not correspond to a data generating process for repeated counts of unknown, finite numbers of individuals.

In contrast to the GLMM, the N-mixture model's development was motivated by process-based thinking, and its architecture corresponds to an idealized data generating process [114]. In the N-mixture model, a latent state N represents the absolute abundance of a species at a sampling site. N varies between sites according to a Poisson distribution with expected value log-linked to linear covariates and is assumed to be constant across replicate observations at a site (the "closure" assumption). The observed data are binomial distributed with size N and probability p, representing the detection process, such that variation in counts within a site is due only to observer error [114, 115, 79]. While all within-site variation is considered observation error and all between-site variation stems from the underlying abundance process, in practice heterogeneous observation error and movement of animals can lead to detection-driven variation between sites and abundance-driven variation within sites.

Though mathematically distinct, both GLMMs and N-mixture models partition variation into between- and within-site components using hierarchical relationships. Between sites, the N-mixture has a Poisson random latent abundance at each site, while the GLMM has a log-scale normal random effect at each site. Within sites, the N-mixture model uses binomial counts, while the GLMM uses Poisson counts. Each model has variants that accommodate overdispersion in counts: in the GLMM, a negative binomial distribution may replace the Poisson, while in the N-mixture model both the Poisson and binomial may be replaced with a negative binomial or beta-binomial distribution, respectively, to account for overdispersion in the within- and/or between-site submodel [92, 114]. However, in the N-mixture model, the latter variants can be highly computationally demanding, so one contribution here is a set of more efficient ways to calculate likelihoods for these variants. More detailed descriptions of both models are presented in the "Model implementation" section of the Methods.

These models exist within a shared heuristic framework in that both models have parameters that predict "number of individuals observed" at a standardized location and primarily differ in their assumptions of between- and within-site variation in observed counts. It is important to distinguish between estimation of parameters and estimation of latent states, such as the latent state N in the N-mixture model. Parameter estimation and model selection arise from making a model do as well as possible as a probability distribution for observed data, not latent states. In other words, both models predict “how many individuals will I see.” Users of N-mixture models are often interested in the follow-on question of “how many individuals are there, adjusted for imperfect detection,” but that is an interpretation of parameters and latent states outside of fitting criteria. Hence, the models can be compared based on their fit to observed data. Although “all models are wrong,” knowing which model fits the data best is useful for characterizing patterns.

Within this heuristic framework, both the GLMM and N-mixture model come with advantages and disadvantages. The main advantage in using a GLMM is that it is robust to unmodeled heterogeneity in data generation due to its relatively simple structure. Robustness to unmodeled variation is a useful quality when analyzing count data with unobserved heterogeneity in sampling protocol or skill, unknown and heterogeneous sampling areas, or complicated non-independence between observations. The primary drawback to using the GLMM is that this model does not explicitly account for the detection process. In fact, GLMMs only estimate relative abundance under the assumption that the pattern of interest is not confounded with detection after accounting for other modeled variables [34]. When this assumption is unreasonable, the GLMM may not be useful for understanding biological patterns, even when it fits the data best.

The N-mixture model, in contrast, explicitly separates abundance from imperfect detection, but at the cost of strong sensitivity to violations in model assumptions, especially those assumptions related to an absence of unmodeled heterogeneity. In recent years, simulation studies have characterized the degree to which N-mixture models produce biased or nonsensical estimates of abundance in the presence of unmodeled heterogeneity [6, 41, 80, 92, 85, 97]. Additionally, Kéry (2018) identified estimation instability, likely attributable to a likelihood maximum in the limit of zero detection [35], but recommended the use of N-mixture models when estimation instability does not occur [78]. Several studies have found that the N-mixture model produces estimates of absolute abundance that agree with more rigorous sampling methods [14, 26, 43, 22] though this finding is not universal [27]. Still, established sensitivities and computational pathologies are grounds for caution in recommending N-mixture models when data do not strictly conform to modeling assumptions.

In this chapter, we provide guidance for choosing between the GLMM and N-mixture model for relative abundance estimation in an empirical context. We ask which of a set of GLMM and N-mixture model variants best fits single-species subsets of eBird point-count data on a small spatial scale. eBird, the largest and most systematic citizen science data repository of its kind, is increasingly used to estimate bird species’ spatiotemporal distributions [124, 125]. Because eBird data inherently contain unobserved heterogeneity, they present an interesting challenge for both GLMMs and N-mixture models.

eBird citizen scientists report their observations in the form of “checklists", lists of detected species associated with sampling metadata. Almost 90% of eBird checklists are “complete checklists”, which imply zero counts for all unreported species. While much statistical modeling of eBird data has addressed estimation over large spatial extents [45, 65, 71], a simpler yet still challenging goal is to estimate local abundance patterns using data from regions with concentrated replicate observations. We select 396 species-subregion (SSR) subsets of eBird across gradients of space, checklist density, and species abundance. Within each SSR, we use only eBird checklists from stationary sampling locations (i.e. checklists obtained from a single spatial point) to have the highest chance of satisfying N-mixture model assumptions.

We consider four variants of the N-mixture model and two of the GLMM. Each N-mixture variant is defined by the two distributions in the within- and between-site submodels, and we consider four variants: the classic binomial-Poisson (B-P), binomial-negative binomial (B-NB), beta-binomial-Poisson (BB-P) and beta-binomial-negative binomial (BB-NB). In the GLMM, we consider the traditional Poisson distribution for counts alongside the negative binomial to allow for overdispersion. We fit each model variant to each dataset with step-wise variable selection. To characterize relative fit across models, we use the Akaike information criterion (AIC) because it is derived to select the model with lowest out-of-sample prediction error [18]. We explore patterns in selection across levels of abundance and sampling intensity and apply a suite of goodness-of-fit and estimation checks to characterize known issues with both the N-mixture and GLMM. We investigate the special issue of estimation instability with a reparameterization of the N-mixture abundance and detection intercepts. We also implement new, fast algorithms for calculating N-mixture likelihoods for variants accommodating overdispersion.

3.2 Methods

eBird and covariate data

eBird data are structured as follows. Birders submit observations as species checklists with counts of each species they identify. They report associated metadata, such as location, date and time, duration of the observation period, number of observers, and sampling protocol [124, 125, 122]. The birder indicates whether their checklist is “complete”; complete checklists yield inferred zeroes for all species not reported on a checklist.

We retrieved the eBird Basic Dataset containing all eBird observations and sampling metadata. We extracted all complete checklists that occurred within the U.S. state of California between April 1 and June 30, 2019. Four survey-level covariates were retrieved from eBird checklist metadata as detection covariates: number of observers, checklist duration, date of year, and time of day; any checklist that failed to report one or more of these variables was dropped. Corresponding to best practices for use of eBird data, we filtered the data for quality according to the following criteria: we discarded checklists other than those following the "Stationary" survey protocol (observations made at a single spatial location) with duration shorter than 4 hours and at most 10 observers in the group [70, 122].

We selected twenty circular regions of high sampling intensity with 10 km radii across California. These spanned the state’s many habitats including coastal, agricultural, wetland, and mountain areas, and contained active birding areas such as parks and human population centers. In each subregion, we selected 10 species with the highest reporting rate (proportion of checklists including that species) and 10 representing an intermediate reporting rate. An additional 10 species were selected that were detected in many regions to enable cross-region comparisons, yielding 407 species-subregion (SSR) datasets. Across 20 subregions, we accepted 6,094 eBird checklists for analysis, each with an associated count (potentially zero) for each species. Observations were aggregated to sampling sites defined by a 50 m spatial grid. The 50 m grid was chosen to conservatively identify related surveys and was not motivated by biological processes, nor does it represent the sampling area of each survey. In this context, the concept of "closure" in the latent state is already suspect due to the fact that eBird checklist sampling areas are inconsistent. Data were processed in R using the 'auk' package [106, 121].

An elevation surface for the state of California was retrieved from WorldClim at $8.3x10^{-3}$ decimal degrees resolution using the R package raster [44, 63]. This commonly used covariate was included as a baseline spatial covariate to enable comparison of estimation properties across sites, but its biological relevance to abundance is not crucial to our analysis [122]. Land cover data were retrieved from the LandFire GIS database's Existing Vegetation Type layer [83]. For each unique survey location, a 500 m buffer was calculated around the reported location, and the percent of the buffer which was water, tree cover, agriculture or other vegetation (shrub or grassland) was calculated. We used the following five site-level covariates: elevation, and percent of the landscape within a 500 m buffer of the site that was water, trees, agricultural land, or other vegetation. We included six checklist-level covariates: duration, number of observers, time of day, time of day squared, Julian date, and Julian date squared. Covariates were dropped in datasets where only a single unique value was observed for that covariate.

Model implementation and selection

We considered four variants of the N-mixture model and two variants of the GLMM comprising a total of 6 distinct models, defined by the distributions used in the model or sub-model.

The GLMM for count data that we considered is defined as

$$y_{ij} \sim D(\mu_{ij}, [\theta])$$

$$\log(\mu_{ij}) = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \alpha_i$$

$$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha)$$

where y_{ij} is th jth observation at site i, D is a probability distribution (which may contain an extra parameter θ to account for overdispersion), μ_{ij} represents the mean expected count and is a logit-linear combination of observed site- and observation-level covariates x_{ij}, β are coefficients representing the effect of those covariates, β_0 is a log-scale intercept corresponding to the expected log count at the mean site (i.e. with all centered covariates set to 0), and α_i is the random effect of site i following a normal distribution. Due to the right skew of $\exp(y_{ij})$, by log-normal distribution theory the log of the expected count at the mean site is $\beta_0 + 0.5\sigma_\alpha^2$. We considered two forms of this model, where D was either a Poisson or a negative binomial distribution, in the latter case with the extra parameter θ.

The N-mixture model is defined as

$$y_{ij} \sim D_w(N_i, p_{ij}, [\theta_w])$$

$$N_i \sim D_b(\lambda_i, [\theta_b])$$

$$\text{logit}(p_{ij}) = \text{logit}(p_0) + \mathbf{x}_{ij(w)}\boldsymbol{\beta}_w$$

$$\log(\lambda_i) = \log(\lambda_0) + \mathbf{x}_{i(b)}\boldsymbol{\beta}_b$$

$$p_0 = e^{\frac{\phi_1+\phi_2}{2}}$$

$$\lambda_0 = e^{\frac{\phi_1-\phi_2}{2}}$$

where D_b and D_w are probability distributions representing between- and within-site variation, respectively; N_i is a site-level latent variable normally representing the “true" abundance at site i; p_{ij} is the detection probability of each individual on the jth observation event at site i; λ_i is the mean abundance at site i; and $x_{(w)}$ and $x_{(b)}$ are covariate vectors for detection and abundance, respectively, with corresponding coefficients β_w and β_b. For reasons described more below, we reparameterize the intercept parameters of the N-mixture submodels, $\log(\lambda_0)$ and $\text{logit}(p_0)$, in terms of two orthogonal parameters $\phi_1 = \log(\lambda_0 p_0)$ and $\phi_2 = \log(p_0/\lambda_0)$. Now ϕ_1 and ϕ_2 represent the expected log count and the contrast between detection and abundance, respectively, at the mean site. This parameterization allows us to investigate stability of parameter estimation. The log-scale expected count of the N-mixture model is $\phi_1 = \log(\lambda_0 p_0)$, analogous to $\beta_0 + 0.5\sigma_\alpha^2$ in the GLMM. Each submodel distribution D could include or not include an overdispersion parameter (θ_w and θ_b), yielding four possible N-mixture model variants: binomial-Poisson (B-P), binomial-negative binomial (B-NB), beta-binomial-Poisson (BB-P), and beta-binomial-negative binomial (BB-NB) [92, 114].

We chose to fit models with maximum likelihood estimation (MLE) for computational feasibility and because key diagnostic tools, such as AIC and methods for checking goodness of fit and autocorrelation, were best suited to MLE estimation [80]. We fit N-mixture models with the nimble and nimbleEcology R packages starting with a conservatively large choice of K, the truncation value of the infinite sum in the N-mixture likelihood calculation [133, 54]. We fit GLMMs with the R package glmmTMB [15]. We applied forward AIC selection to choose the best covariates for each model with each dataset (illustrated in Figure S1). One spatial covariate (elevation) and two checklist metadata covariates (duration and number of observers) were treated as *a priori* important and were included in all models. In the N-mixture model, checklist-specific sampling metadata were only allowed in the detection submodel, while land cover covariates and the interactions between them were allowed in both the detection and abundance submodels. Interactions were dropped in datasets when interaction values showed a correlation of >0.8 with one of their first-order terms. In N-mixture models, additions to both submodels were considered simultaneously during forward AIC selection.

For comparisons between models, we selected a heuristic threshold of $\Delta\text{AIC} > 2$ to say that one model is supported over another [18].

Fit, estimation, and computation

Goodness-of-fit

We used the Kolmogorov-Smirnov (KS) test, a p-value based metric, to evaluate goodness-of-fit on each selected model. For GLMMs, residuals were obtained using the DHARMa R package's 'simulateResiduals' and the KS test was applied using the 'testUniformity' function [61]. For N-mixture models, we considered the site-sum randomized quantile (SSRQ) residuals described by Knape et al. [80], computing these for each N-mixture model and running a KS test against the normal CDF. We assumed that covariate effects did not vary by space within subregions and chose not to use spatially explicit models [70, 122]. To test this assumption, we applied Moran's I test to the SSRQ or DHARMa-generated residuals for each site or observation.

Parameter estimation

We compared two abundance parameters of interest across models: coefficients for elevation and log expected count at a standard site (in the GLMM, $\beta_0 + 0.5\sigma^2_\alpha$; in the N-mixture model, $\log(\lambda_0 p_0)$). We examined absolute differences in point estimates and the log-scale ratios between their standard errors.

Stability of estimated parameters

Attempting to decompose the expected value of observed data into within- and between-site components can lead to ridged likelihood surfaces with difficult-to-estimate optima. Kéry found that instability of model estimates with increasing K occurred when there was a likelihood tradeoff between detection and abundance, resulting in a tendency in abundance toward positive infinity restrained only by K [79]. Dennis et al. showed that N-mixture models could in fact yield estimates of absolute abundance at infinity [35]. We interpreted this as a case of a boundary parameter estimate rather than non-identifiability and explored it by reparametrizing as follows. We estimated the intercepts for detection and abundance with two orthogonal parameters (rotated in log space) $\phi_1 = \log(\lambda_0 p_0)$ and $\phi_2 = \log(p_0/\lambda_0)$, where λ_0 and p_0 are real-scale abundance and detection probability at the mean site. We hypothesized that in unstable cases, ϕ_1, log expected count, is well-informed by the data, but ϕ_2, the contrast between abundance and detection, is not well-informed, corresponding to a likelihood ridge as $\phi_2 \to -\infty$ due to detection probability approaching 0 and abundance approaching infinity. This reparameterization isolates the likelihood ridge to one parameter direction, similar to a boundary estimate as $\exp(\phi_2) \to 0$. Boundary estimates occur in many models and are distinct from non-identifiability in that they result from particular datasets. Confidence regions extending from a boundary estimate may include reasonable parameters, reflecting that there is information in the data. We defined a practical lower bound for ϕ_2. When ϕ_2 was estimated very near that bound, we conditioned on that boundary for ϕ_2 when estimating confidence regions for other parameters.

In the N-mixture case, diagnosing a boundary estimate for ϕ_2 is made more difficult by the need to increase K for large negative ϕ_2 to calculate the likelihood accurately. We used an approach like that of Dennis et al. (2015) to numerically diagnose unstable cases. For each N-mixture variant in each SSR, the final model was refitted twice, using values of K 2000 and 4000 greater than the initial choice. Estimates were considered unstable if the absolute value of the difference in AIC between these two large-K refits was above a tolerance of 0.1. We monitored whether MLE estimates of ϕ_1 and ϕ_2 also varied with increasing K.

Evaluating the fast N-mixture calculation

We extended previous work by Meehan et al. to drastically improve the efficiency of N-mixture models using negative binomial or beta-binomial distributions in submodels [96].

We ran benchmarks of this likelihood calculation for a single site against the traditional algorithm, which involves iterating over values of N to compute a truncated infinite sum. We calculated the N-mixture likelihood at 5,000 sites and compared the computation time between the two methods for all four N-mixture model variations. We ran benchmarks along gradients of length(y_i) (number of replicate observations at the simulated site) and K (the upper bound of the truncated infinite sum) for each variant.

All figures were produced using the R package ggplot2 v3.3.5 [137].

3.3 Results

Species-subregion (SSR) datasets

Our procedure for aggregating "subregions" yielded 20 circular subregions of 10 km radii containing the highest density of eBird activity in California during the 2019 breeding season. Subregions corresponded to human population centers with access to natural habitat such as large parks or coastal areas. Subregions contained between 140-1000 high quality checklists distributed across 12 to 140 unique locations. In each subregion, we selected 10 species with high detection rates and 10 species with intermediate detection rates. We further selected 10 species overall with sufficient data in the most total subregions to compare model selection for individual species across space. In all, we compared models of interest across 396 species-subregion (SSR) subsets of eBird.

Model selection

We fit each of six N-mixture and GLMM variations to each SSR dataset. Across SSRs, the six models we considered were chosen by AIC at the following rates: Poisson GLMMs were selected in 88 datasets (22%), negative binomial GLMMs in 61 datasets (15%), BB-NB N-mixture models in 79 datasets (20%), BB-P N-mixture models in 74 datasets (19%), B-NB N-mixtures in 46 datasets (12%) and B-P N-mixtures in 48 datasets (12%) (Figure 3.1). AIC clearly selected (ΔAIC > 2) the best GLMM over the best N-mixture model in 102 (26%) datasets, while the best N-mixture model was selected in 199 (50%). AIC rankings indicated overall support for models incorporating overdispersion.

Negative binomial GLMMs outperformed Poisson GLMMs in 308 of 396 datasets. However, they were largely superseded in those cases by N-mixture models, especially those that included a beta-binomial-distributed detection submodel (BB-P and BB-NB N-mixtures). Among N-mixture variants, the best N-mixture model by AIC used a beta-binomial distribution in 223 datasets (56%), and these cases largely corresponded to cases where GLMMs also outperformed binomial-submodel N-mixture models (B-P and B-NB). When beta-binomial N-mixtures were excluded from the analysis, GLMMs were dominant, being selected clearly by AIC (ΔAIC > 2) in 237 (60%) datasets.

Subregions with more checklists were associated with higher rates of selection of the most complicated model (by number of overdispersion parameters), the BB-NB N-mixture model (Figure 3.2). We attribute this to the phenomenon that more data contains more information to be explained by additional model structure.

We did not detect patterns in model selection related to whether a species was overall highly detected or detected at intermediate rates (Figure 3.1). We did not identify patterns in model selection varying by species identity among widespread species.

Eleven SSR datasets were removed from the analysis after modeling due to computational issues with GLMM estimation.

Fit, estimation, and computation

Goodness-of-fit

We tested goodness-of-fit (GOF) from residuals for each model and assessed systematic patterns of fit by examining the distributions of GOF p-values for each model type. Among N-mixture models, distributions of goodness-of-fit p-values did not deviate from the uniform, meaning that most N-mixture models selected by AIC fit well for this metric. Both sets of selected GLMMs showed deviation from a uniform distribution of p-values, indicating that these models' residuals deviated meaningfully from modeling assumptions for some datasets (Figure 3.3). GLMMs were selected for several datasets where GOF checks for GLMMs failed, despite those same datasets passing goodness-of-fit checks for N-mixture models, indicating that AIC model selection did not correspond exactly to goodness-of-fit metrics.

None of the models considered incorporated spatial autocorrelation. We tested that assumption using Moran's I tests on residuals for all models. Nine percent of model-dataset combinations had p-values less than 0.05 from Moran's I test, indicating more cases of non-random spatial structure in residuals than expected by chance. This suggested the presence

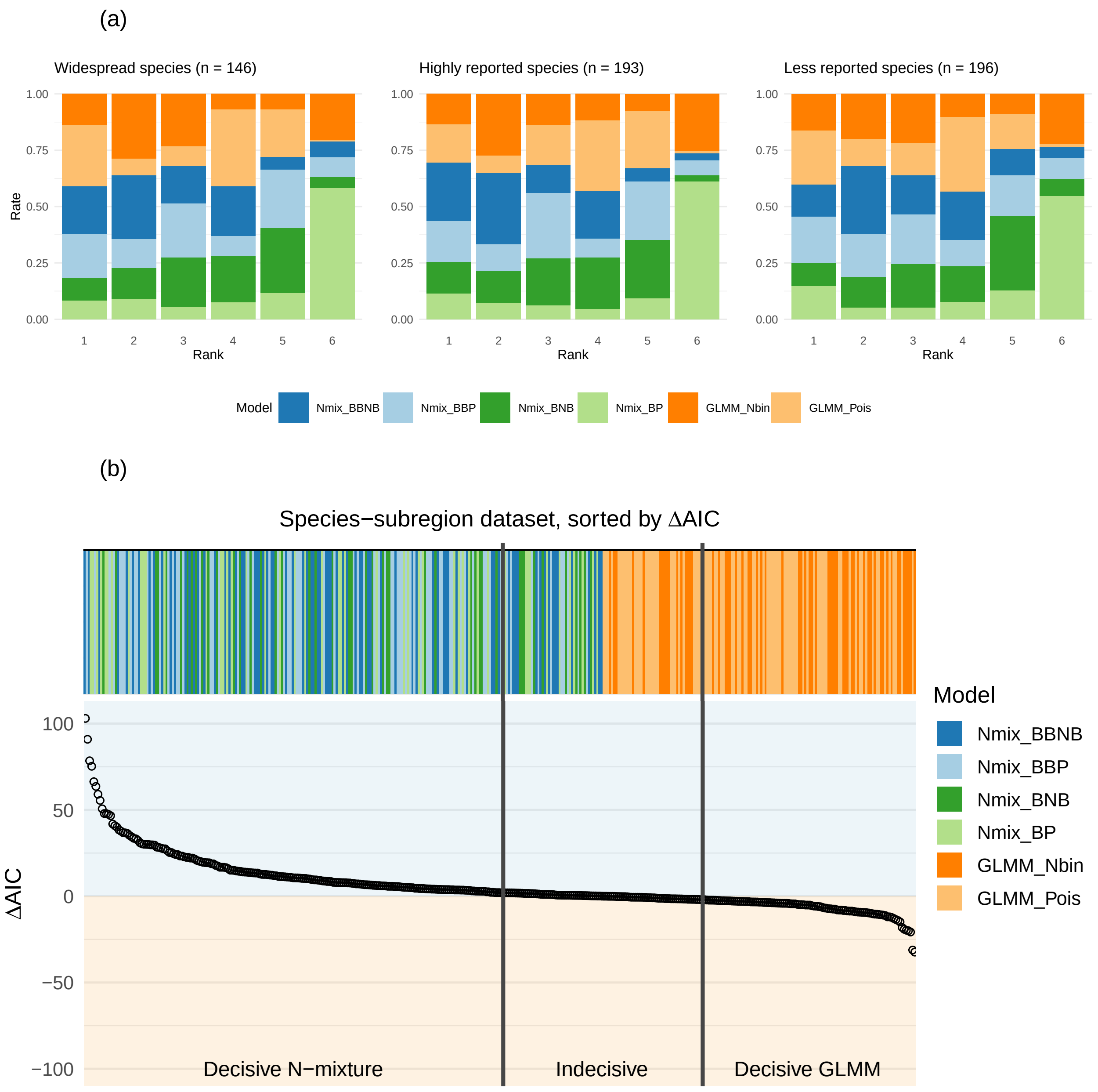

Figure 3.1: (a) Model rankings by species reporting rate category (1 = best). (b) Magnitude of AIC difference between the best N-mixture model and the best GLMM for each dataset. Each stripe-point pair represents one SSR dataset. Bar color indicates AIC model choice; point position on the y-axis indicates the difference in AIC (ΔAIC) between the best N-mixture model and the best GLMM.

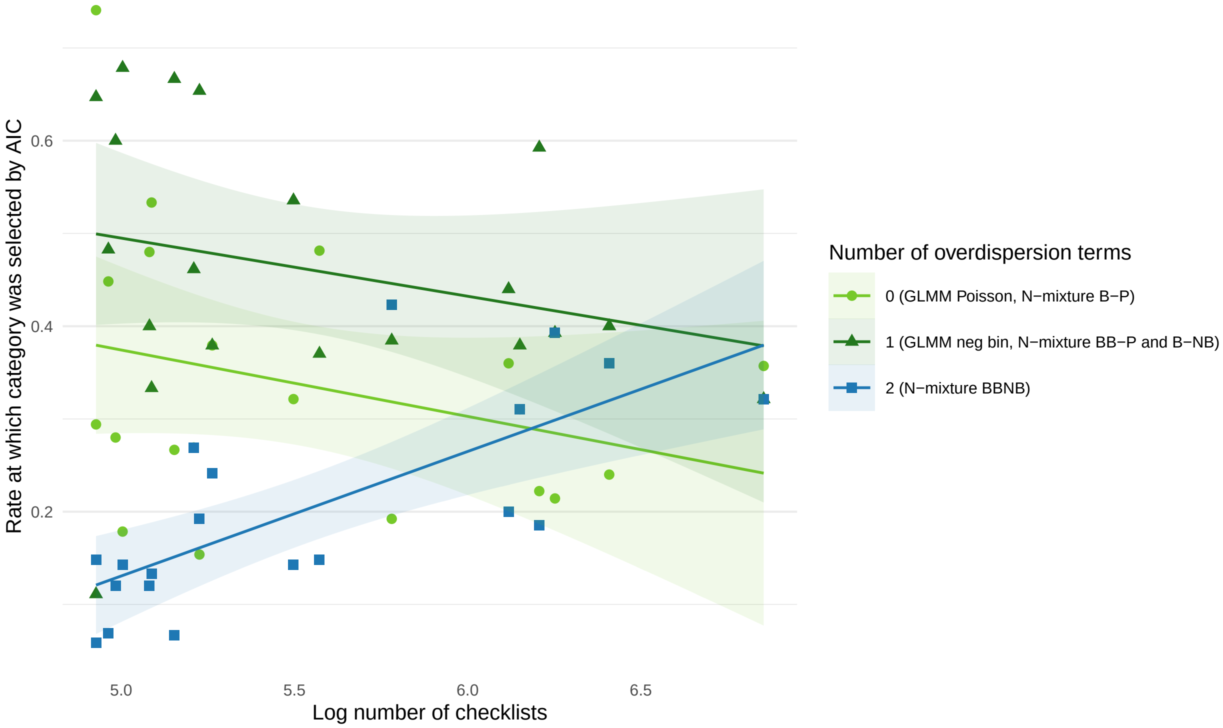

Figure 3.2: For each subregion, we plot the proportion of each model group selected for each of three model flexibility categories against the number of checklists in that subregion. Solid lines show best fit lines for each category. As the number of checklists in a subregion is increased, the most flexible BB-NB N-mixture model is supported at an increasing rate.

of spatial autocorrelation in some datasets, but at a low rate we considered acceptable for this study's conclusions.

Parameter estimation

Point estimates of the log expected count at the mean site (i.e. with all centered covariates set to 0), adjusted for the log-normal random effects of the GLMM, were similar for GLMMs and N-mixture models (Figure 3.4). Within model type, N-mixture models agreed closely with one another, as did GLMMs. Differences in a point estimate of a coefficient, site elevation, were centered around zero for all model combinations, indicating no systematic differences between or within model types.

Estimated standard errors of covariates were systematically different between models (Figure 3.5). Both GLMMs estimated standard errors systematically larger than all N-

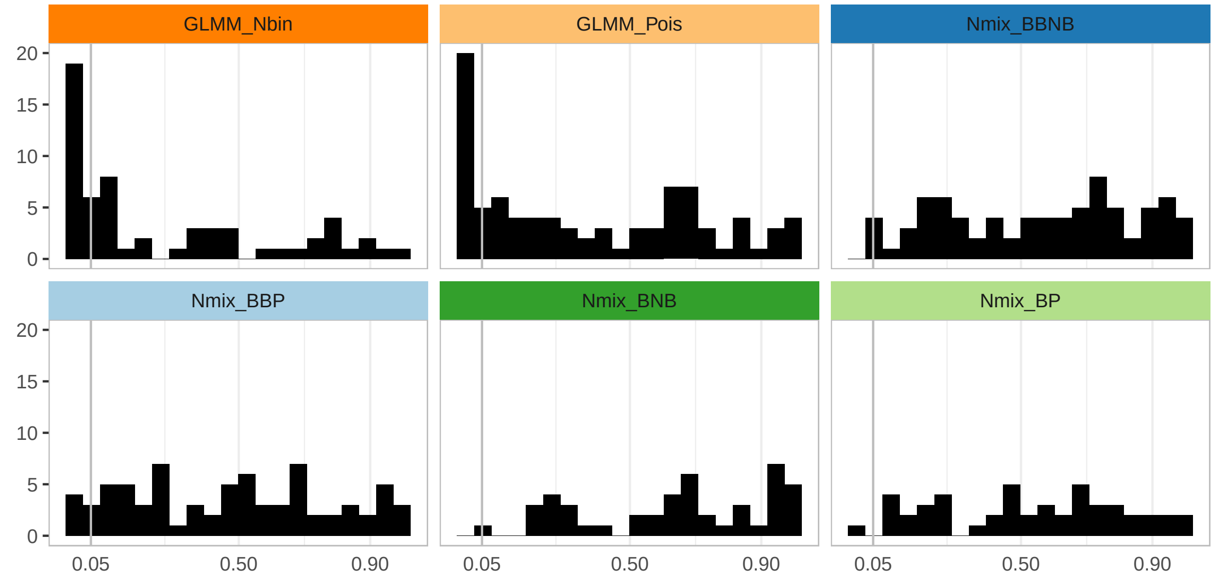

Figure 3.3: Goodness-of-fit p-values for N-mixture models selected by AIC had near-uniform distributions, while a subset of selected GLMMs showed goodness-of-fit failures.

mixture models, while more complex N-mixture models estimated larger standard errors (as expected within a model family).

Stability of parameter estimates

We investigated rates of instability in N-mixture estimation by monitoring whether estimated AIC and two intercept parameters changed as the upper bound of the truncated infinite sum, K, was increased [78]. Across 396 species-region datasets, 6% of B-P N-mixture models, 7% of BB-P models, 37% of B-NB models, and 11% of BB-NB models were found to be unstable in AIC for a tolerance of $\Delta\text{AIC} = 0.1$. While the B-NB showed by far the highest rate of instability, agreeing with previous findings by Kéry (2018), some instability was detected across all N-mixture models. To illustrate that instability could be attributed to indeterminate tradeoff between detection and abundance, we reparameterized the two intercepts with one parameter for (intercept of) log observed count (abundance × detection) and another for (intercept of) log ratio between detection and abundance (abundance / detection). The log observed count intercept was stable in 93-98% of datasets in each of the four N-mixture models. The log ratio intercept, representing tradeoff between detection and abundance, was unstable in patterns mirroring those of AIC. This suggested that instability with increasing K was due to ridged likelihoods in a single parameter direction that was a

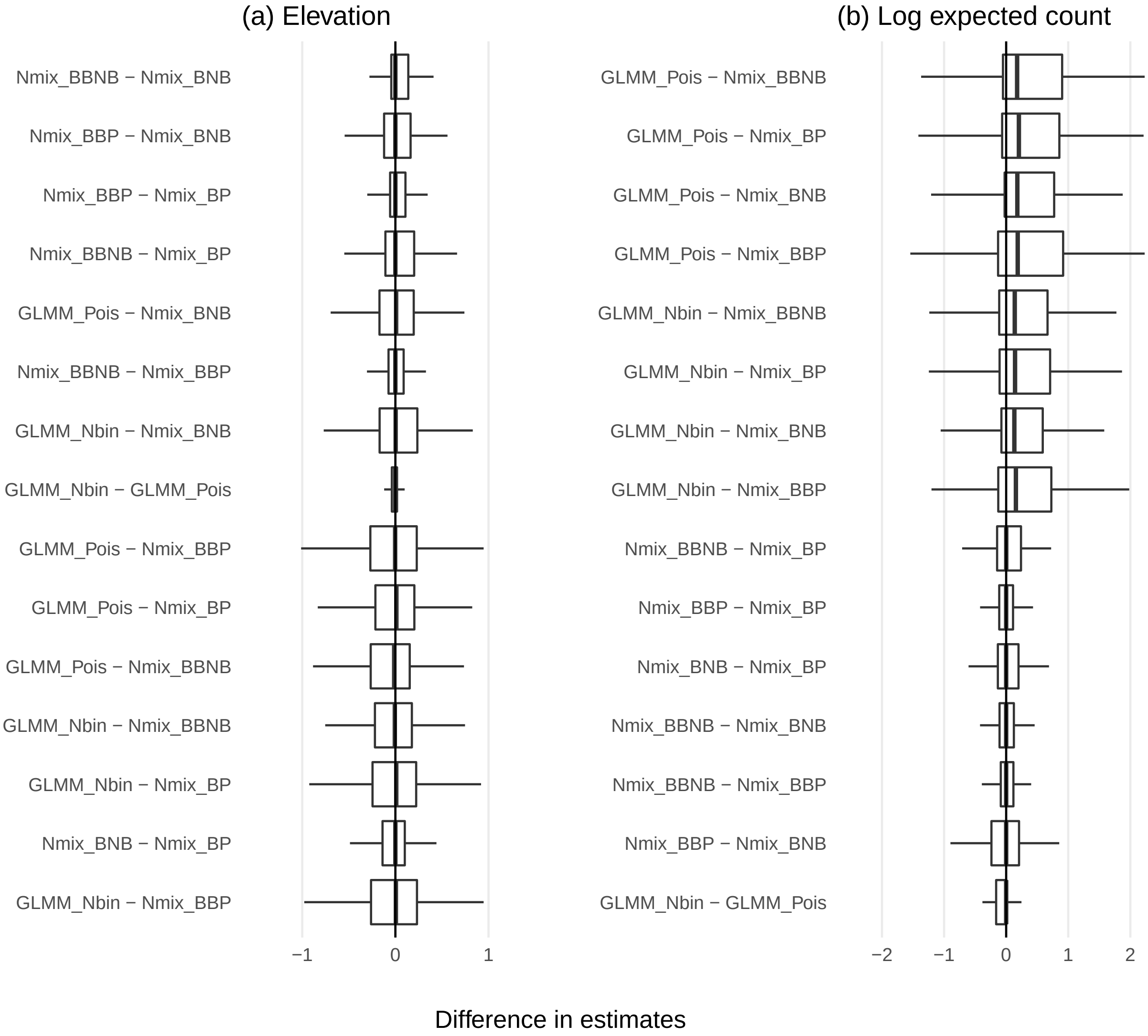

Figure 3.4: Distributions of the absolute differences between models in (a) log expected count and (b) log-scale effect of elevation. Model pairs are ordered by median difference. Outliers are excluded for legibility. Log expected count is the log-scale intercept plus 0.5 times the random effects variance for the GLMMs and simply the log-scale expected count intercept for N-mixture models. The latter is defined in our parameterization as a combination of abundance and detection intercepts.

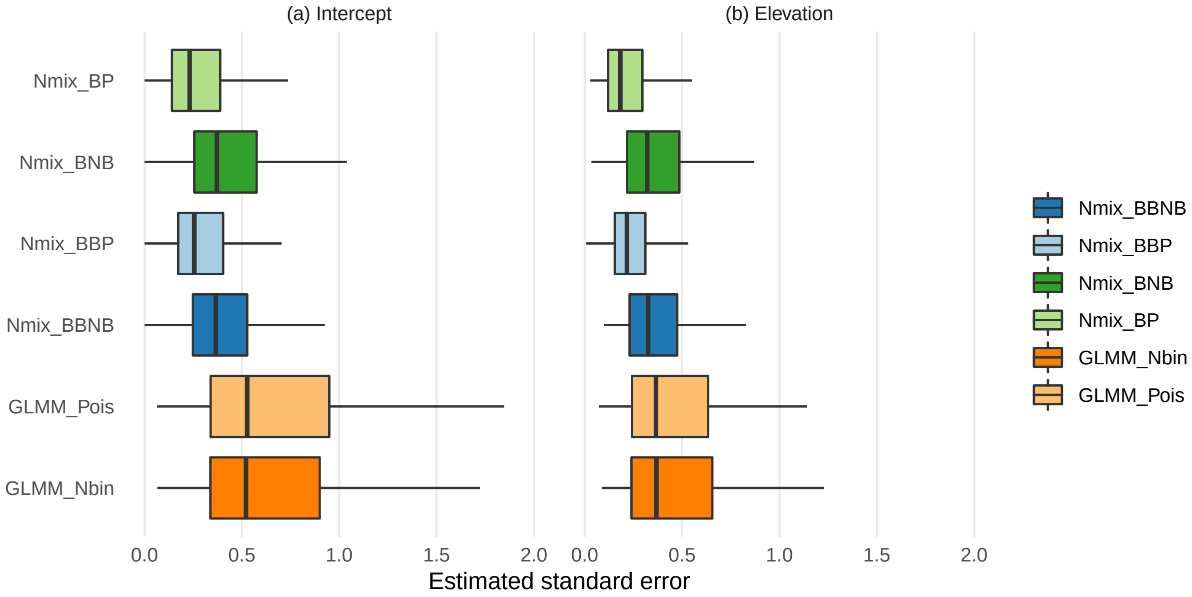

Figure 3.5: Comparing standard errors of estimates of the data intercept (defined as expected log count at a standard site) and a target relative abundance driver, elevation. Outliers are excluded for legibility.

combination of the original parameters.

Performance of the fast N-mixture calculation

We extended previous work by Meehan et al. (2020) to implement an algorithm for fast N-mixture likelihood calculations. Fast N-mixture calculations led to 20- to 125-fold improvements in computation time compared to the naive method for calculating the truncated infinite sum for $100 \leq K \leq 2000$ and $5 \leq \text{length}(y_i) \leq 20$, where y_i was the vector of detection-nondetection observations at a site. These gains increased with $\text{length}(y_i)$, such that the contribution of the fast algorithm was important when more replicate visits were made to a simulated site. In eBird, sites are clustered such that a small number of locations contain many replicate observations, making this improvement per replicate particularly relevant. Improvements in computation time were largest for the beta-binomial and negative binomial N-mixture variants, which take orders of magnitude longer than the traditional B-P N-mixture due to their higher computational costs.

Discussion

When estimating relative abundance from count data, ecologists can consider both N-mixture models and GLMMs alongside one another in a model selection context. Both N-mixture models and GLMMs act as predictors of counts that vary between and within sites and can be compared via standard model comparison tools. By adopting a heuristic perspective towards models based on predictive fit, we let the data speak about which model is better rather than guessing *a priori* based on assumptions about the data-generating process. Put another way, we accept that all models under consideration are “wrong" in that assumptions are bound not to be perfectly satisfied. By selecting among the candidate models with AIC, we learn which model approximately minimizes out-of-sample prediction error.

Across 396 species-subregion data subsets of eBird, we found that it was usually possible to distinguish between N-mixture and GLMM fit and that N-mixture models outperformed GLMMs somewhat more often than the reverse. This pattern was contingent on the inclusion of beta-binomial N-mixtures, which greatly outperformed binomial N-mixture variants. All four N-mixture model variants also more consistently passed goodness-of-fit checks. Consideration of only one model type across all these data would produce worse overall fit than dataset-by-dataset selection.

We identified no patterns in model selection across species characteristics or site identity, indicating that relative fit was highly context-dependent. We observed one pattern in model selection: datasets containing more checklists (and therefore more information) were more likely to select more complex models. This trend does not suggest that more complex models like the BB-NB N-mixture model were “true" for our data, and that it was therefore wrong to use a simpler model for low-information datasets, since in fact simpler models minimize out-of-sample prediction error in lower-information contexts. The overall lack of patterning suggests that ecologists should not assume that a particular model will fit better for a given dataset.

Point estimates of a parameter of interest, effect of elevation, agreed between N-mixture models and GLMMs, but GLMMs estimated systematically larger standard errors than N-mixture models. In general, lower standard errors are not reason for recommending a model, because lower standard errors can result from either correctly characterized improvements in precision, or overconfidence (and increased Type I error); without access to true data generating parameters these are indistinguishable without additional fit metrics. If both models fit reasonably well, then agreement between parameters may indicate support for the assumption that elevation was not confounded with detection. While most selected N-mixture models passed goodness-of-fit checks, a substantial portion of GLMMs chosen by AIC failed them, indicating that GLMMs fit the data somewhat more poorly overall. A simulation study targeting goodness-of-fit and model selection patterns across known data-generating conditions could clarify whether this relatively poorer fit is due to actual better correspondence between N-mixture models and the eBird data generating process, or whether N-mixture models are more flexible when data are heterogeneous in general. Comparing intercepts between the models (expected log count at a “typical” site) is more complicated due to their different structures. Specifically, the expected count from the GLMM's log-normal intercept distribution needs adjustment by the random effects variance to be comparable to the corresponding N-mixture parameter. After this adjustment, we saw that log expected counts are not systematically higher or lower for one kind of model.

Estimation instabilities in N-mixture models were largely attributable to a single dimension representing the decomposition between abundance and detection, while parameter estimates of relative abundance drivers were estimated stably. These instabilities correspond to a boundary estimate of detection probability being nearly zero. Although this is implausible in the mechanistic motivation of N-mixture models, from a statistical perspective a boundary estimate is not necessarily a pathological estimation outcome, so these unstable cases can still be compared alongside other models. We found a high rate of instability in the B-NB N-mixture model in agreement with Kéry (2018), but found that this high rate did not extend to the BB-NB and BB-P N-mixture model variants, which showed rates of instability comparable to the traditional B-P N-mixture. Still, instability was present at low rates in all four N-mixture variants. Future theoretical work may clarify how to interpret this phenomenon when it occurs.

We analyzed datasets from a single important database (eBird), during a single year, and at modest spatial scales, so the particular patterns we observed in model selection are limited. We also chose an aggressive data filtering approach that brought the data closer into alignment with N-mixture modeling assumptions. It is possible that N-mixture models would be less successful on less aggressively filtered data. For feasibility, we chose to only consider two model families of interest. One relevant linear model extension, the generalized additive mixed model (GAMM), was excluded from this analysis. GAMMs can be used to introduce flexibility into parameters which are allowed to vary over space and/or time [122, 23]. Because of the potential for estimating spatially variable covariate effects, ecologists working at larger spatial scales may want to consider GAMMs alongside N-mixture models and GLMMs where appropriate. In this application, we expected that this flexibility would not be relevant at the spatial scale considered. Neither GLMMs nor N-mixture models overwhelmingly outperformed one another by AIC, but it is conceivable that new models could be developed to extend or bridge these approaches.

When choosing between GLMMs and N-mixture models for relative abundance estimation, practitioners may weigh trade-offs beyond the models' different abilities to explain the data (as characterized by an information criterion and goodness-of-fit checks). When an abundance covariate of interest is confounded with detection, using an N-mixture model may disentangle these two components, while using a GLMM will not. If estimating absolute abundance is of interest, the GLMM will similarly not satisfy this need, while the N-mixture model could if effective area sampled is known. Practitioners interested in either of these two approaches should collect their data in line with N-mixture modeling assumptions and should consider modeling assumptions during study design. On the other hand, the GLMM may be preferred when computational efficiency is important, such as when the number of observations is very large. The marginalized likelihoods for N-mixture variants presented in this chapter reduce the computational costs of N-mixture model fitting. Fast implementations of the N-mixture likelihood calculation reduced computation times 20- to 90-fold. These implementations are available in the R package nimbleEcology [54].

We found that observed instability with increasing K [78, 41] was not prohibitive for interpreting relative abundance estimates. We used a parameterization that suggests such instability arises from a parameter estimate on the boundary of the parameter space, which is a manageable and not uncommon problem in other kinds of models. Relative abundance point estimates agreed between N-mixture models and equivalently constructed GLMMs. This suggests that practitioners not reject the N-mixture model in the presence of this form of estimation instability, and can consider both N-mixture models and GLMMs when estimating relative abundance with count data. Without access to known truth, we could not investigate whether the estimates produced reflect true relationships between covariates and species abundance, so we cannot say whether either N-mixture models or GLMMs estimated parameters accurately in that sense. We suggest that practitioners interested in modeling relative abundance from count data consider the assumptions of both models, including known features of N-mixture model robustness in practice and theory [41, 79, 35, 6, 92, 85, 97, 14, 26, 43, 22], along with other practical trade-offs. If both GLMMs and N-mixture models are potentially appropriate, we recommend the model selection and goodness-of-fit checking procedure outlined in this chapter for choosing the most parsimonious model with adequate fit.

3.4 Discussion

To estimate species' relative abundance patterns from increasingly common heterogeneous datasets, ecologists will find both N-mixture models and GLMMs useful on a context-dependent basis. Despite their distinct origins in heuristic and process-based thinking, these two models are structurally analogous tools for predicting counts that vary both across and within sampling units. We have presented results to indicate that selecting between these models for a particular dataset is possible on a context-dependent basis. We encourage ecologists to adopt a holistic approach to model selection, considering different models alongside one another and bearing in mind that statistical models, whether or not they are process-motivated by design, are ultimately tools for fitting data.

Chapter 4

Drought influences habitat associations and abundances of birds in California's Central Valley

In Chapters 2 and 3, I investigated the methodology behind relative abundance estimation with participatory science data. I illustrated a marginalized approach to implementing N-mixture models and gave an interpretive framework for understanding abundance patterns with these models. In an investigation into eBird data, I showed that N-mixture models and GLMMs performed comparably for estimating relative abundance, but that models with the ability to accommodate more overdispersion fit the data better as the size of the dataset grew. In this chapter, I apply these methodological findings to answer a pressing ecological question: how does severe drought influence the abundance and distribution of birds?

4.1 Introduction

The frequency of hot, dry periods constituting ecological drought is increasing in many parts of the globe [135, 29, 28, 38, 32]. Conservation management during drought requires careful ecological study, as ecological impacts are multifaceted and species- and habitat-specific. Reduced precipitation during drought can impact wildlife directly through water stress-induced mortality or indirectly by altering the availability of food resources [20]. When high temperatures and dry periods occur simultaneously, synergistic impacts occur, especially increased cooling costs in the form of higher water or food requirements [20, 75, 110, 109, 91]. Ultimately, a species' vulnerability to drought is a combination of its level of exposure (the degree to which regional climatic change is experienced by individuals of the species) and its sensitivity (the degree to which a species' abundance changes per unit change in exposure to drought) [140].

Previous studies have established that extreme drought can have a major impact on bird abundance [86, 104, 99, 11, 117, 2, 3]. A variety of mechanisms explaining which species decline, and where, have been investigated in isolation and in tandem. Certain (especially natural) habitats promote species resilience via increased availability of food or microhabitat [109, 99, 11, 51, 68]. Birds' cooling costs correlate with body mass, so larger-bodied birds may be more sensitive to water deficits [110]; on the other hand, smaller-bodied birds experience greater relative evaporative water loss, so extreme heat might more strongly impact smaller species [4]. Trophic niche may also determine sensitivity to the indirect effects of drought, as the availability of food in response to drought changes differently in different habitats for herbivores and carnivores [104]. Resource pressure may lead to density-dependent relationships between drought and bird abundance [104, 19]. Behavioral plasticity and mobility may play a role in structuring sensitivity; for example, migratory species may have greater spatial flexibility in choosing breeding sites but may be more phenologically restricted [95, 49]. Species with more suitable habitat may also be better equipped to seek out new territory during drought.

Habitat composition and structure both affect the extent to which birds are impacted by regional climatic extremes. Diversity in habitat structure can allow animals to moderate exposure to extreme heat by using microclimates. The availability of resources, especially water, also varies differently during drought in different habitats within a region. The role of human-modified habitat in moderating species' exposure to drought is especially difficult to predict. Human modification and temperature increases are likely to interact in their impacts on species, as evidenced by the fact that human-modified habitat tends to support species that are further from their thermal limits [138]. Agricultural lands support smaller populations than natural lands, especially during hot periods, and especially for less drought-tolerant species [62, 138, 139]. Species more sensitive to drought may also be more vulnerable to habitat degradation as the availability of natural lands decreases [129]. On the other hand, human activity may buffer smaller-scale climate impacts on heavily used lands as water is artificially redirected for agricultural and domestic use. Understanding how species respond to drought in heavily modified landscapes requires a holistic modeling framework that considers interacting effects of habitat type and drought.

In this paper, we investigate the impact of dry periods on birds' abundances and their relative use of different habitats in the Central Valley ecoregion in California, USA. The Central Valley comprises several large metropolitan areas and is an important agricultural region dominated by fruit, vegetable, and nut production as well as working pasture. The area also includes substantial remnant natural habitat including riparian zones around perennial streams and grasslands, though many of the region's grasslands are working pasture. Like much of western North America, this region has experienced a number of severe droughts in recent decades [38]. This region has seen an increase in bird conservation efforts in recent years as its importance as a migratory corridor and breeding habitat for birds is recognized [33, 55, 107]. Conservation planning in this region must contend with the expectation that extreme droughts will continue to occur frequently, so understanding how species respond in these conditions is of fundamental importance.

We modeled counts of birds submitted to the participatory science platform eBird. We characterized changes in reported counts of 66 common Central Valley birds from 2010-

2019 in this region, taking advantage of the density of eBird activity to achieve a high degree of spatiotemporal resolution. We applied stacked single-species N-mixture models with overdispersion and random effects structures for conservative estimation of parameters driving variation in counts. We asked three questions to better understand how counts of birds change during drought. First, we asked whether each of 66 species' overall abundance changed with drought severity, and whether those overall changes were related to species traits. Second, we asked whether the effect of drought on each species varied meaningfully between habitat types. Finally, we asked which individual climate or environmental variables were responsible for overall species abundance changes during drought. Since drought is a multivariate environmental process, it may or may not be possible to attribute a drought-abundance association to solely one dimension, such as change in temperature.

4.2 Methodology

Study region

Comprising the Sacramento Valley in the north and the San Joaquin valley in the south, the Central Valley is a predominantly developed and agricultural system with remnant riparian and natural grassland habitat. Despite the relative lack of unmodified habitat, the Valley serves as an important breeding habitat for many birds and a migratory habitat for more [33]. Climate change is expected to increase the frequency of extreme droughts in California as the probability of co-occurring dry and warm periods increases [38], so understanding birds' responses to drought in this area is crucial for their conservation. This region has a high density of eBird sampling activity spread across a variety of habitats and climatic conditions. It has experienced multiple periods of severe drought and non-drought in recent years (4.1). This combination of features makes the ecoregion an ideal study area for resolving the effect of drought on species in different habitats and via different mechanisms. To delineate the study area we used the "Central California Valley" ecoregion as defined by the USGS [56].

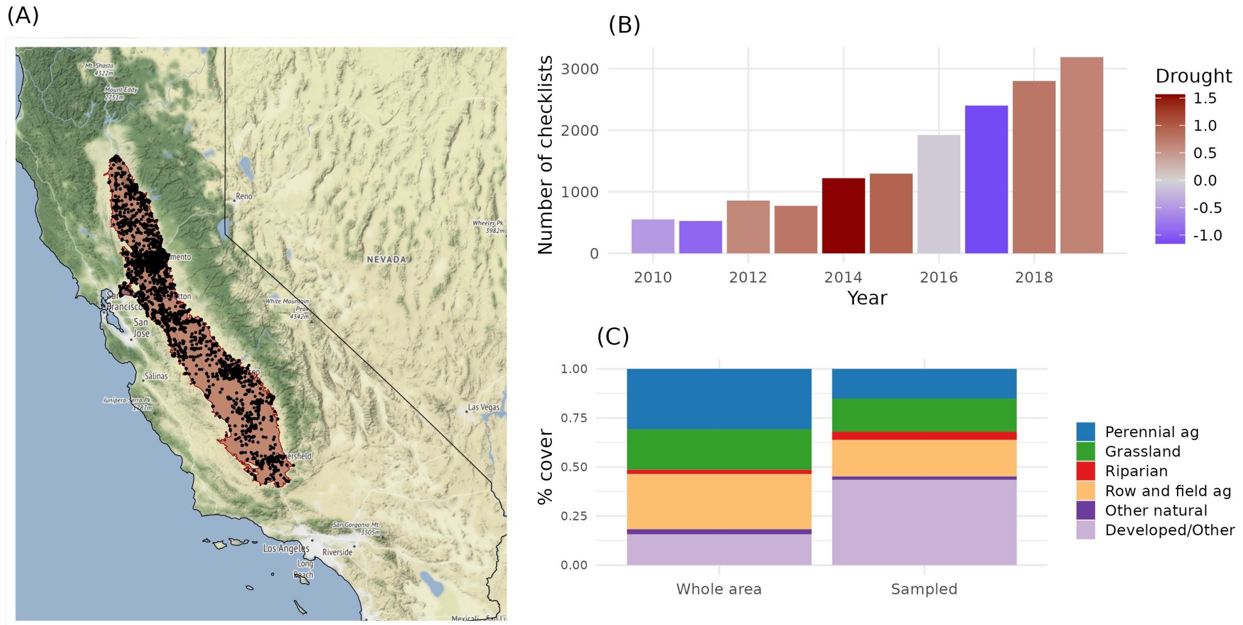

Figure 4.1: (A) Locations of eBird checklists (black points) in the Central Valley ecoregion (red polygon) of California. Checklists were densest near developed areas but distributed throughout the region, with some gaps in the southern Central Valley. (B) Due to increases in eBird activity over time, more recent years were better represented in the dataset. Bar colors indicate annual average drought as defined by the standard precipitation-evapotranspiration index (SPEI); our study period comprised two relatively wet and two relatively dry periods. (C) Land cover proportions of the entire study region (left bar) compared to the areas sampled in the data (right bar). The developed/other category was overrepresented in the data, comprising 16% of the Central Valley region but 43% of sampled area.

Bird count and trait data

eBird is a participatory science birding data repository where volunteer observers report checklists comprising counts of bird species detected on discrete sampling occasions [124, 125, 122]. Reported counts from eBird have been used in the past in the Central Valley to inform conservation planning [107, 55] and assess species status [113]. eBird data are semi-structured, meaning that most checklists are associated with sampling metadata describing effort. Checklists may also be tagged as "complete" (meaning all detected birds were reported, and any non-reported bird was not observed). These two features make it possible to model eBird counts while partially controlling for sampling variation. However, eBird data also contain heterogeneity, especially heterogeneity in area sampled, that precludes inferring absolute abundance from these data [53]. We therefore focus our interpretations on relative abundance while accounting for variation in detection, rather than the absolute abundance of birds.

We analyzed eBird checklists documenting point count surveys within the Central Valley region in California, USA (Figure 4.1A). We extracted all eBird complete checklist data for observations that took place in California's Central Valley in the years 2010-2019 during the months of April, May, and June, selected as a rough approximation of the breeding season in this region. From each checklist, we retrieved the following metadata: date of year, time of day, duration of sampling event, and number of observers in the observation group. Any checklist missing one or more of these metadata was excluded. We also excluded checklists other than those that followed the "stationary sampling" protocol, which specifies that all birds were detected from a single point in space, and excluded any checklists with a recorded duration of more than 3 hours or more than 8 observers to minimize unmodeled heterogeneity and spatial error in the data [70]. Of 73,853 eBird checklists conducted in the Central Valley during the study period, 15,522 eBird checklists met all quality criteria and were admitted to analysis. Because each checklist included was "complete," each was associated with either a count or a nondetection (count of 0) for all species, meaning that all single-species models were fit to observations from all checklists. We chose to model 66 species that were detected on at least 500 (or 3%) of admitted eBird checklists.

One aim of this study was to identify whether species functional traits were associated with drought responses. For each of 66 modeled species, we retrieved the species' trophic niche and whether or not the species is migratory [128]. We also retrieved each species' taxonomic order and grouped species into two groups for comparison (Passerines and non-Passerines).

Landscape covariate data

Observations were assigned to spatiotemporal cell-year units according to a grid of 1 km × 1 km × 1 year covering the Central Valley ecoregion over the period 2010-2019. A spatial resolution of 1 km was chosen to maximize resolution while allowing eBird checklists conducted very nearby one another to be associated. In the N-mixture model, observations in a cell-year were modeled as sharing a latent abundance state. Checklists were distributed across a total of 4,821 cell-years each containing between 1 and 190 checklists (mean = 3.2 checklists; median = 1 checklist; 90th quantile = 5 checklists; 95th quantile = 10 checklists.).

For each cell, habitat covariates were produced representing the proportion of each cell covered by each habitat type, as defined by land cover and crop type data. Crop types were retrieved from the California Statewide Crop Mapping dataset for 2018. At the time this work was initiated, only cropping data for 2016 and 2018 were available, so we opted to use the 2018 dataset only under the expectation that aggregated land use categories at the 1km scale were constant during the study period [136]. Gaps in the Statewide Crop Mapping dataset, which covers only agricultural land, were filled in with land cover data from the Functional Vegetation LANDFIRE dataset [83]. Land cover classes were aggregated into 6 categories: row and field crops, perennial crops, grassland and pasture, natural riparian habitat, other natural and semi-natural habitat, and developed/other. The "developed/other" category comprised 91.4% urban habitat.

We retrieved four environmental variables of interest in each grid cell: temperature, precipitation, the normalized differential water index (NDWI), and the enhanced vegetation index (EVI). We retrieved daily precipitation and temperature data from the PRISM climate group [59]. We computed the average maximum daily temperature in the sampling period April-June in each cell-year and calculated the amount of precipitation in the preceding year (July of the previous year through June of the current year, including the April-June sampling period). We also included two remotely sensed variables: EVI, an index of vegetative productivity [73, 134], and NDWI, a measure of the amount of standing water in an area [131]. We retrieved EVI data at 500 m daily resolution and NDWI data at 30 m bi-weekly resolution. For both EVI and NDWI, we computed averages within each cell-year during the months April-June.

We used the standardized precipitation evapotranspiration index (SPEI) as a continuous measure of drought severity. SPEI incorporates information from several climate measures such as water availability and temperature. SPEI has been used in previous studies of birds to quantify drought [66, 19]. SPEI is informed by lagged impacts of climate, building in some potential to explain lagged climate effects. However, it does not account for potential biological delays in bird responses to drought, such as the possibility that ecological drought leads to worse breeding outcomes and reduced populations several years later. Some evidence exists from related systems that short-term water availability is most important in determining bird responses [104], although there is also evidence that the timing of rain events in relation to the breeding season is important, which would suggest that abundance declines would appear on a time lag [111]. We investigate lagged effects of drought with a separate model (see below).

We obtained monthly measures of SPEI from the global SPEI database at 1° resolution [7]. For each cell-year, we extracted SPEI April 1, the beginning of the sampling period for that year. Because SPEI data were available at a resolution of 1° (roughly 85 km in central California or 111 km at the equator), they are interpreted as a regional measure of drought.

For comparisons of drought vs non-drought conditions using estimated models, we selected two levels of SPEI to represent a typical wet year and a typical extremely dry year in this system based on the lowest and highest median annual SPEI (in 2014 and 2017, respectively). SPEI is parameterized such that lower SPEI indicates drier conditions, so a positive effect of SPEI on occupancy means a negative effect of drought on occupancy.

Abundance models and counterfactual approach

Overview of multi-model framework

Our primary objective was to understand how the abundance of 66 Central Valley bird species changed between non-drought and drought conditions, which we accomplished with a novel multi-model framework (Figure 4.2). We fit single-species N-mixture models to estimate how eBird counts varied with a set of covariates including drought-related environmental variables [34]. We also accounted for variation in detection, nonindependence between checklists, and overdispersion. To differentiate between the roles of collinear environmental variables representing different aspects of drought while still estimating an overall drought effect, we use a second model to predict how environmental variables changed with drought. Together, the two models allow us to use posterior predictive methodology to estimate (a) sets of environmental variables representing drought conditions and (b) distributions of predicted abundance based on those environmental variables.

Details on the model and on the specific posterior predictive methods used to answer each main research question are presented below.

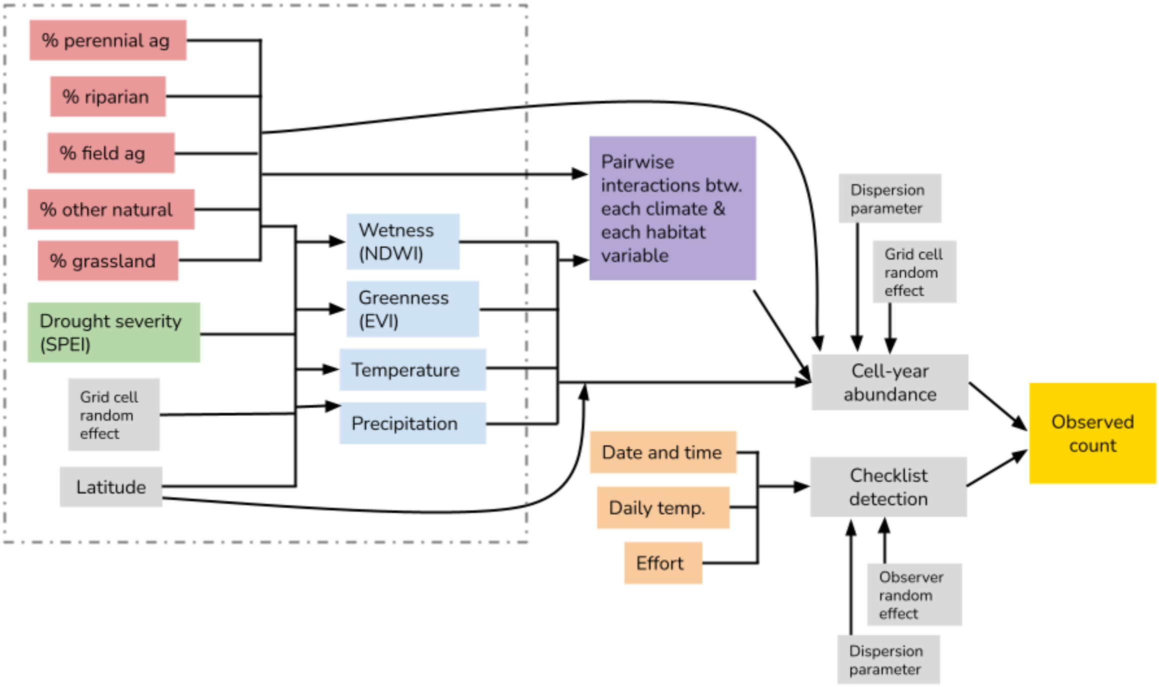

Figure 4.2: A conceptual diagram illustrating relationships between variables in the model. Linear mixed models (LMMs), delineated by the dotted gray box, were used to explore how four environmental variables (NDWI, EVI, maximum temperature, annual precipitation) changed with change in a drought index (SPEI) differently in different habitat types. Posterior predictions of the four environmental variables under drought and non-drought conditions were generated with linear mixed models. N-mixture models were used to estimate the effect of covariates on eBird counts, and included two random effects and two layers of possible overdispersion. Posterior predictions of environmental variables from LMMs were then used as input data to predict posterior distributions of bird counts under drought and non-drought conditions.

N-mixture models for bird counts

We analyzed eBird reported counts using single-species N-mixture models[114]. Since N-mixture models are somewhat sensitive to unmodeled variation in counts [85], and since eBird data are sampled heterogeneously, we parameterized our models with maximal flexibility, accounting for two layers of potential overdispersion with a beta-binomial detection submodel and a negative binomial abundance submodel. We also include two types of random effects. We fit single-species rather than community N-mixture models because we did not want to rely on the assumption that covariate effects are normally distributed. Single-species models allow us to base our conclusions only on species with enough data to be informative. The single-species N-mixture models we estimated were defined as

$$\begin{aligned}
y_{ijt}|N_{it} &\sim \text{BetaBinomial}(N_{it}, p_{ijt}, \theta_1) \\
N_{it} &\sim \text{NegativeBinomial}(\lambda_{it}, \theta_2) \\
\text{logit}(p_{ijt}) &= \mathbf{x}_{ijt}\boldsymbol{\beta} + \alpha_{o(ijt)} \\
\log(\lambda_{it}) &= \boldsymbol{w}_{it}\boldsymbol{\gamma} + \alpha_i \\
\alpha_{o(ijt)} &\sim \mathcal{N}(0, \sigma_{\alpha_o}) \\
\alpha_i &\sim \mathcal{N}(0, \sigma_{\alpha_i})
\end{aligned}$$

The datum y_{ijt} is the observed count of the species for the jth checklist submitted in grid cell i and year t. These counts follow a beta-binomial distribution with size N_{it}, a cell-year-level latent variable representing the expected count under perfect detection; probability p_{ijt}, the detection probability of each individual on the jth observation event at grid cell i, year t; and overdispersion parameter θ_1 to account for extra-binomial variation in counts within a cell-year due to unobserved heterogeneity. N_{it} follows a negative binomial distribution with expected value λ_{it} representing the mean abundance in grid cell i in year t and overdispersion parameter θ_2 representing extra-Poisson variation in underlying counts across cell-years. The values λ_{it} and p_{ijt} are log- and logit-linear functions of checklist-level and cell-year-level covariates $\mathbf{x}_ijt$ and $\boldsymbol{w_it}$, respectively, with corresponding coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We include a random effect of grid cell, α_i, on abundance to account for potential nonindependence between counts in each grid cell in different years. A random effect of observer (grouping checklists submitted by the same eBird user), $\alpha_{o(ijt)}$, on detection is also included, where $o(ijt)$ gives the observer ID for the jth checklist submitted in grid cell i and year t. This helps accommodate potential differences in skill between users, which if ignored could lead to biased abundance estimates as the observer pool turns over [69]. Each of the random effects is normally distributed with standard deviation parameters σ_{α_i} and σ_{α_o}.

Estimating the absolute abundance of species with these data was impossible for two reasons. First, effective area sampled varies by eBird checklist depending on microsite characteristics and observer choices. Second, in aggregating observations to a spatiotemporal grid, we expect to have violated the N-mixture closure assumption. Instead we restrict our interpretation to the effect of covariates on eBird counts, rather than underlying abundance. We do not interpret the N-mixture model as a one-to-one representation of the data-generating process but rather a heuristic model with the flexibility to partition variation in reported counts into within- and between-site components [53].

The following checklist-level covariates were included in the detection submodel (as $\mathbf{x}_{ijt}$) to account for variation in effort and detectability: sampling duration, time of day, time of day squared, day of year, day of year squared, and number of observers in group. Maximum daily temperature as retrieved from PRISM was also included as a detection covariate to account for the fact that birds vocalize differently depending on temperature [95].

Ten cell-year-level covariates were included in the abundance submodel (as $\boldsymbol{w}_{it}$): latitude; categorical effect of year (nine levels); habitat type percentages (0-1 values for each of perennial agriculture, row and field agriculture, grassland, riparian, and other natural habitat); and four continuous environmental variables (EVI, NDWI, average daily temperature, and annual precipitation). We also included twenty pairwise interactions between each habitat variable and each climate variable and an interaction between wetness (NDWI) and average daily maximum temperature. Interactions allowed us to test whether the effect of drought on observed counts varied by habitat type through any or all of the four environmental variables. Because habitat types summed to 1 for all cells, the "developed/other" habitat type was excluded from the model (i.e. when all land cover classes are 0 on the real scale, that cell is 100% "developed/other"). All covariates in both submodels were centered and scaled. There was substantial collinearity between EVI and NDWI (correlation of 0.74) and between some EVI and NDWI interaction terms (3 of 5 pairs EVI and NDWI interaction terms showed correlation > 0.7). Rather than discarding these terms from the analysis, we proceeded with modeling. Collinearity may make it difficult to disentangle the effects of these two variables, diminishing our power to answer our third question; however, in the context of the Bayesian posterior predictive method used throughout, we judged that collinearity between variables would not be an issue in predicting changes in abundance (see section on "Counterfactual drought prediction and linear mixed models").

We expected that drought impacts on abundance would coincide with the drought events themselves. Since it is possible that ecological drought leads to worse breeding outcomes and reduced populations after the drought event has concluded, the impact of drought on abundance may occur on a delay. To test whether predicted changes in abundance were robust to the choice to use year-of environmental variables, we replicated the N-mixture modeling and posterior predictive steps for all species with one year lags for the drought index and four drought-related environmental variables.

We implemented single-species N-mixture models in NIMBLE v0.12.2 [133]. We chose all priors to be minimally informative on their relevant scales [100]. For the detection intercept β_0, we used a logistic prior. For all other coefficients in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we used normal priors with mean 0 and standard deviation 2.25. For both random effect standard deviation priors, we used uniform distributions from 0.001 to 10. For priors on beta-binomial and negative binomial overdisperion parameters, we used uniform distributions from 0.0001 to 25. Models were estimated with MCMC using custom sampler assignments for improved mixing. For each species, we ran three chains of 15,000 iterations with 5,000 iterations of burn-in and a thinning interval of 10. We checked whether each model had a minimum effective sample size of 100 for all stochastic parameters, and ran additional chains for each species until this condition was met.

We opted not to use multispecies abundance models for two reasons. First, as we were primarily interested in species-level abundance estimates and in identifying potential differences in drought response between species, we thought that single-species estimation was more appropriate than a model that assumed that drought responses across species followed a shared normal distribution. Second, model estimation was computationally prohibited when using multispecies models with random effects for each species. Since these random effects and overdispersion terms were critical to robust inference with eBird data, we chose to prioritize model flexibility over multispecies inference.

Linear mixed models of environmental covariates

SPEI, an index of drought level, is derived from remotely sensed climate data. In this application, we invert this logic and use SPEI to predict observed environmental variables. This enables us to generate posterior predictions of environmental variables for a given location in the study area under a reference level of SPEI representing a typical dry or wet year.

We used explanatory linear mixed models (LMMs) to estimate how each of the four climate covariates varied with SPEI. For a given climate variable C, we fit LMMs defined by the equations

$$
\begin{gathered}
\mathrm{C}_{it} \sim \mathcal{N}(\mu_{it}, \sigma_{\epsilon}) \\
\mu_{it} = \beta_0 + x_{it}\boldsymbol{\beta}_c + \alpha_i \\
\alpha_i \sim \mathcal{N}(0, \sigma_{\alpha})
\end{gathered}
$$

where the value of the climate variable at cell i in year t, C_{it}, was normally distributed with mean μ_{it} and residual variation σ_{ϵ}. The mean climate variable at each cell-year μ_{it} was a linear combination of covariates x_{it} with coefficients $\boldsymbol{\beta}_c$. Covariates included were SPEI, five habitat types (as in N-mixture models), and interactions between SPEI and each habitat type. We included a normally distributed additive random effect of grid cell, α_i. All covariates were centered and scaled. Data for the years 2010-2019 for all 2,566 grid cells containing eBird data were included.

Linear mixed models were estimated with the R package "brms" [17]. We used normal priors with mean 0 and standard deviation 5 for all β covariates, and half-Cauchy priors with scale parameter 2 for the prior on σ_{α}. We ran three chains of 15,000 iterations with 5,000 iterations of burn-in. We then obtained posterior predictions of each climate variable at each cell-year in the study area using actual habitat values and the two reference levels of SPEI, representing a distribution of potential climate conditions under drought and non-drought scenarios.

Question 1: do species' overall counts change with drought?

We combined the N-mixture models and LMMs in a joint posterior predictive framework to estimate changes in bird counts during drought (Figure 4.2). Because we modeled the effect of environmental variables on birds' counts, we needed posterior predictions generated by LMMs to predict how those environmental variable change between drought and non-drought conditions. Using predicted levels of those variables from LMMs and posterior samples of abundance coefficients from single-species N-mixture models, we computed an expected abundance λ_i for each species at each cell-year in each MCMC iteration. We drew random negative binomial counts using these expected abundances and draws of the overdispersion parameter, θ_2, ultimately yielding posterior predictive distributions of underlying counts of each species in both drought and non-drought conditions.

To evaluate whether each species declined or increased during drought, we took the difference in count summed across cell-years between drought and non-drought conditions of SPEI at each iteration for each species. If the 95% credible interval of this distribution did not overlap zero, we interpreted this as evidence that the species had either a positive or negative association between drought and reported counts. We used chi-squared tests to ask whether changes in species counts were associated with trophic niche, whether or not a species is migratory, and taxonomic group (comparing Passerines and non-Passerines). We corrected p-values obtained from chi-squared tests across both Question 1 and Question 2 by controlling the false discovery rate [10].

Question 2: do species' habitat associations change during drought?

Whether species' habitat associations—their predicted relative abundance on each habitat type—varied between drought and non-drought conditions was a major question of this study. In a linear modeling context, the question "does the effect of covariate 1 on the response variable change with the level of covariate 2?" can be represented by including an interaction term and testing whether that term is different from zero. In the multi-model framework presented above, we estimate an interaction effect of drought and habitat on the four climate variables and an interaction effect of those climate variables with habitat on bird counts. This structure creates multiple "pathways" through the model by which both habitat and drought can influence abundance and multiple opportunities for interaction effects to occur (Figure 4.2). This means that a simple interaction term is not estimated. However, we can use partial derivatives to analytically derive the quantity the interaction term represents—the rate of change of the effect of covariate 1 on the response variable with respect to covariate 2—in our multi-model framework.

We estimate an interaction between drought and each habitat type term for each species, $\frac{dY}{dDdh}$, representing the degree to which each species' abundance in each habitat varied differently with drought. To test whether interactions were statistically meaningful, we computed posterior predictive distributions of the derived interaction terms for each species-habitat type combination. If the 95% CI of the posterior distribution of the interaction term between drought and one or more habitat variable did not overlap zero, we interpreted that as evidence that the species shifted its overall use of habitat types during drought. We used one-dimensional credible intervals of each interaction rather than a credible region on all interactions for a species because the latter is difficult to characterize in higher dimensions.

To characterize the magnitude of habitat shifts on a biologically relevant scale, we derived posterior distributions of the abundance of each species in each of 6 habitat types during drought and non-drought conditions. We used the posterior predictive approach in Question 1 to draw abundances for each grid cell for each species, then allowed each grid cell to contribute to the species' abundance on each habitat in proportion to the habitat makeup of that grid cell. We then derived the fraction of the species in each habitat type and each drought condition, which we use to visualize and interpret habitat shifts. Among species whose overall distributions were found to shift, we used chi-squared tests to ask whether species were more likely to decrease or increase their use of each habitat type.

We also used chi-squared tests to ask whether habitat shifts were associated with trophic niche, whether or not the species is migratory, and taxonomic order (Passerines vs. non-Passerines).

Question 3: are species' changes with drought attributable to changes in environmental variables?

To estimate the effect of each climate variable's change during drought on the count of each species, we adapted the counterfactual count generation workflow. In generating counts for Question 1 we produced two sets of counts, one using values of all four environmental variables drawn from predictions of drought conditions and the other with all four environmental variables predicted in non-drought conditions. To understand the impact of each climate variable in isolation, we instead predicted counterfactual counts with only one climate variable drawn from predictions in drought conditions, while the others were predicted in non-drought conditions. By comparing these new count distributions with counts under non-drought conditions, we were able to identify the amount of change in each species' abundance attributable to change in each climate variable. We refer to these tests as "one-variable counterfactual scenarios."

If the 95% credible interval of the difference in predicted count between each one-variable counterfactual scenario and the non-drought baseline scenario did not overlap zero, we interpreted this as evidence of that variable's importance in driving the species' overall abundance during drought.

Figures 4.1, 4.3, and 4.4 were created using the R packages ggplot2 v3.3.6, ggtern, and urbnmapr [137, 58, 123]. MCMC samples were processed using the package MCMCvis [143].

4.3 Results

Summary of eBird coverage

More recent years are represented by more checklists in the data than years further into the past, mirroring trends in eBird usage overall [125] (Figure 4.1). This weakens our ability to identify trends in time relative to early (low-information) years, but recent years contain both wet and dry conditions, so inference on drought effects should be robust to this pattern. Coverage of habitat types was comparable to the distribution of habitat types across the landscape, with the major exception that developed habitats were strongly overrepresented. This oversampling should have no bearing on the interpretation of the results, except that the effect of developed habitat types on species abundance may be better informed than that of other habitat types.

Linear mixed model results

In all six habitat groups (row and field agriculture, perennial agriculture, grassland and pasture, riparian, other natural habitat, and developed/other) temperature was positively associated with drought. Riparian habitat was associated with the highest magnitude of temperature increase (95% CI 0.88-1.19 degree C increase per 1 unit decrease in SPEI) and other natural habitat was associated with the lowest magnitude of increase (95% CI 0.67-0.99 degree C increase per 1 unit decrease in SPEI).

Similarly, in all six habitat groups precipitation was negatively associated with drought. Riparian habitat was associated with the highest magnitude of precipitation decrease (95% CI 106.22-133.25 mm decrease per 1 unit decrease in SPEI) and other natural habitat was associated with the lowest magnitude of decrease (95% CI 68.81-97.17 mm decrease per 1 unit decrease in SPEI).

While change in NDWI and EVI, the remotely sensed measures of landscape wetness and greenness, were credibly associated with drought in some habitats, the magnitudes of these relationships were so low as to not be biologically relevant. NDWI decreased with drought in other natural habitat at a very low magnitude (95% CI 0.01-0.02 unit decrease per 1 unit decrease in SPEI). We estimated that NDWI actually increased with SPEI in three habitats (other/developed, grassland, and row and field agriculture), though at similarly low magnitudes (the highest change was in grassland, 95% CI 0.00-0.02 unit decrease per 1 unit decrease in SPEI). For comparison, the estimated standard error of the random effect of grid cell was 0.06, and the estimated residual standard error was 0.06. Similarly, EVI changed little with drought severity. EVI increased with drought in four habitats (both agriculture groups, other/developed, and riparian) and decreased in one habitat (grassland) though at very small magnitudes; in no habitat did the magnitude of the relationship exceed a 0.01 unit change in EVI per unit change in SPEI. For comparison, the estimated standard error of the random effect of cell was 0.05. Altogether these model results indicate that variation EVI and NDWI in this system are mostly not attributable to drought.

Do species' overall counts vary during drought?

We estimated posterior distributions of the change in overall abundance of each of 66 bird species between drought and non-drought conditions. Of these, 21 species had 95% credible intervals of change in abundance that did not overlap zero. We infer that counts of 5 species increased during drought, while counts of 16 species decreased (Figure 4.3).

Chi-squared tests indicated that there was no evidence of associations between species traits or taxonomic group (Passerines vs. non-Passerines) and whether the species' overall abundance changed during drought.

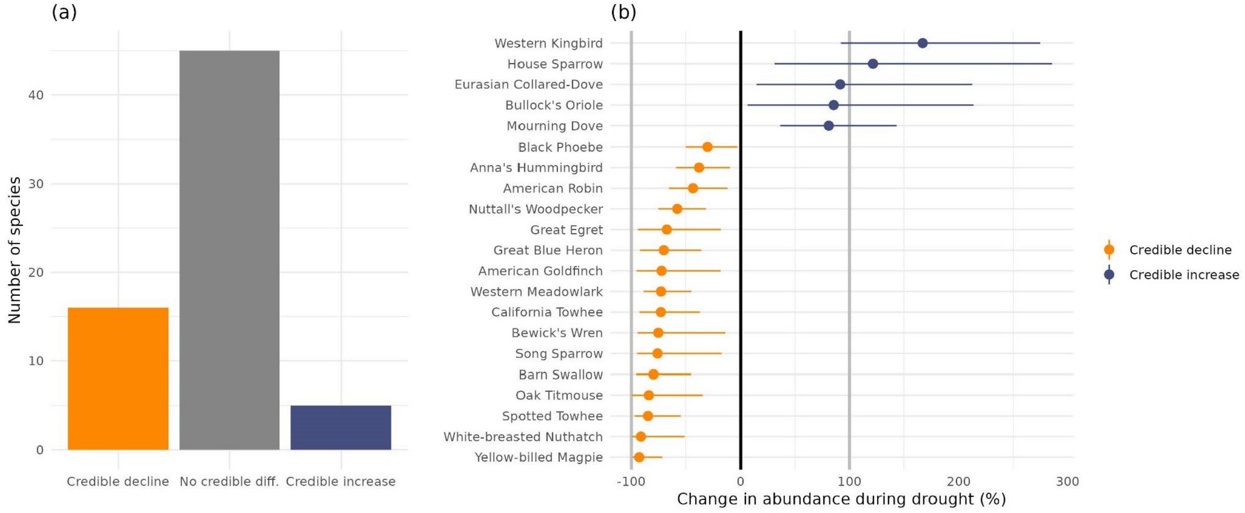

Figure 4.3: Summary of shifts in overall abundance by species. (a) Counts indicating the rate at which species decreased, increased, or showed no relationship between overall abundance and drought. 16 species declined in abundance with drought, 5 increased, and the remaining 45 species showed no relationship between overall abundance and drought. (b) Percent change in abundance with 95% credible intervals for 21 species with credibly nonzero relationships between overall abundance and drought.

To test whether year-of drought effects were appropriate, we replicated the models including drought variables on a one-year lag. Results of these models were nearly identical to results produced by the main year-of models. Under these models, fifteen species decreased and four increased their overall abundance, with all but 3 species effects being the same direction as in the primary models. This suggested that the choice to use year-of models did not obfuscate a lagged effect of drought. We proceed with interpreting year-of model results.

Do species' habitat associations change during drought?

We estimated posterior distributions of interaction terms representing how abundance changed differently for each species in each of 5 habitats during drought. Across 66 species, 27 species had one or more credibly nonzero interaction terms. We infer that associations between counts of those species and habitat type changed with drought level, suggesting that species used habitat types in different proportion during drought. We did not identify patterns in habitat shifts with taxonomy or any functional traits.

To characterize multispecies patterns in habitat-drought relationships, we visualize how the proportion of each species in each habitat shifts between drought and non-drought conditions (Figure 4.4). Across the 27 species with habitat shifts, chi-squared tests indicated that birds were more likely to increase than decline in developed habitat (25/27 species increased use; adjusted p-value < 0.001) and in perennial agricultural habitat (21/27 species increased use; adjusted p-value < 0.05). Relative increases in use of perennial agriculture and developed habitats were offset by decreases in the other four habitat types. More species declined than increased use of riparian habitat (19/27 decreased) and other natural habitat (17/27 decreased), though chi-squared tests did not indicate that these patterns were statistically different from an even pattern of increases and decreases after controlling for false detection.

Chi-squared tests indicated that there was no evidence of associations between species traits or taxonomic group (Passerines vs. non-Passerines) and whether the species changed its relative use of habitat types during drought.

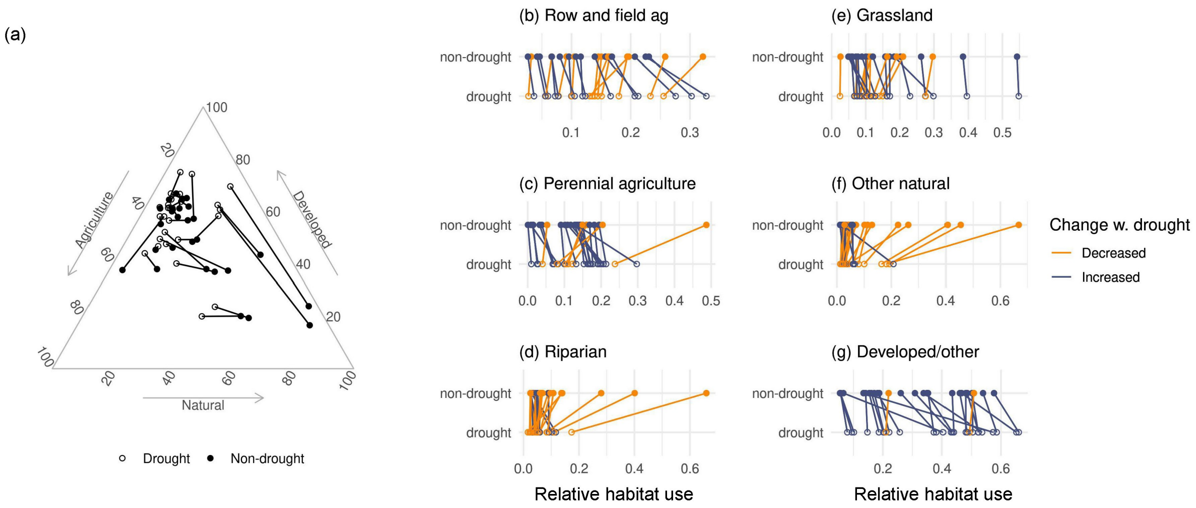

Figure 4.4: Visualizing how 27 species shifted their habitat associations during drought. The 39 species that did not change their relative use of habitat with drought are excluded. (a) A ternary plot shows how species shift in three-dimensional habitat space. Each point pair represents one species for which a habitat shift was estimated. Habitats have been aggregated into three categories: agriculture (combining perennial and row/field agriculture), natural (riparian, grassland, and other natural) and developed. Species overall show shifts away from natural habitat and toward developed and agricultural habitat during drought (moving from filled to empty circles, species largely shift up and to the right). (b-g) The shift in use of each species in each habitat. Colors indicate whether each species' median posterior predicted proportional use increased or declined in the drought condition. We identify a pattern of increase agriculture and developed habitat, while species declined at the greatest rate in riparian and other natural habitats. Note that both plots visualize median posterior predicted proportional habitat use and that credible intervals have been omitted for legibility.

Are species' changes with drought attributable to changes in environmental variables?

Among 16 species with overall decline, 7 declined in a temperature-only one-variable counterfactual scenario, and 2 species declined in a precipitation-only counterfactual scenario. Among 5 species with overall increase, 3 species increased in a temperature-only and 3 species increased in a precipitation-only drought counterfactual; one species with an overall increase declined in the precipitation-only scenario. Additionally, we find very low rates of marginal counterfactual difference among species with no overall change (4 species of 45). No species declined in EVI-only or NDWI-only counterfactual scenarios, which is explained by the finding in the LMM phase that EVI and NDWI did not vary with drought to a relevant degree.

We chose *a posteriori* to predict abundance changes under an additional counterfactual scenario where precipitation and temperature were both allowed to vary with drought, but EVI and NDWI were not. Under this scenario, 15 species declined in abundance and 5 increased, indicating that the combination of change in temperature and precipitation were jointly responsible for nearly all abundance changes predicted by the model.

4.4 Discussion

Birds' responses to drought depend on habitat type. Using a novel multi-model framework to analyze eBird data, we provide the strongest evidence to date that changes in the relative importance of habitats may be a more common and immediate consequence of drought than changes in overall abundance. Nearly half of a set of common Central Valley species, including many species whose overall abundances did not change during drought, changed their use of habitat types depending on drought level. We also found that the region-wide abundance of a moderate number of species declined meaningfully during severe drought, a pattern comparable to that identified in previous research [86, 99, 117, 104]. However, the rate of abundance declines was lower across species than the rate of habitat shifts.

In California's Central Valley, birds used human-modified habitat more during periods of extreme drought compared to non-drought periods. This pattern contrasts with research in other systems indicating that natural habitat supports greater biodiversity and promotes resilience during drought [99]. This discrepancy might be explained by the intensity and character of human activity in the region. The Central Valley is dominated by irrigated agricultural land, and the distribution of water between "environmental" applications (including river flow, wildlife habitat maintenance, and scenic waterways) is highly regulated and varies dramatically between drought and non-drought periods. As an illustration, in one characteristic dry year, 2014, the allocation of water for environmental use was cut to a quarter of its allocation compared to a characteristic wet year, 2006, while the allocation for agriculture actually increased slightly to offset precipitation deficits [98]. In addition to driving short-term changes in relative abundance, changes in water availability were a major driver of avian community composition change in the Central Valley over the last 100 years, supporting the idea that water is major factor determining species distributions [90]. However, the effect of water availability on bird abundance may be dwarfed by effects of climate change and land use change over long timescales [9]. Agricultural and developed habitats, which may be less preferable for species in normal climate conditions, experience less change in water availability due to human intervention compared with natural landscapes, which may dry entirely as intermittent streams stop flowing.

Of the 16 species whose overall abundance declined during drought, the yellow-billed magpie experienced the greatest declines. The yellow-billed magpie is a species of conservation concern whose range is restricted almost entirely to the Central Valley. This finding implicates frequent extreme drought as a compounding factor driving this species' recent observed population declines [31].

While we identified greater rates of drought-related abundance decline among long-distance migrants, this pattern was not statistically significant. Migratory species may be phenologically restricted and thus less able to adapt to short-term changes in temperature during the breeding season, but further study is needed to test this hypothesis in the Central Valley [3, 2, 95].

Predicted abundance declines during drought were similar when considering effects of drought on a one-year lag, suggesting that the year-of model was appropriate. Drought responses were observed during drought events, rather than after them. The rapid response by birds is more consistent with the hypothesis that observed habitat shifts are driven by individuals moving across the landscape as opposed to by habitat-dependent mortality gradients, which would be more evident over longer timescales. However, our model ultimately cannot differentiate between animal movement and mortality gradients. Monitoring and conserving birds during droughts will require considering that birds' use of habitat is climate-dependent, and that shifting relative use of habitat and drought-driven movement across the landscape may influence bird demographics, persistence, and exposure to other threats, especially along a gradient of human modification.

We found that nearly half of overall drought-related species declines in this system would occur under only the influence of extremely high temperatures such as those that occur during drought years, while all changes were attributable to a combination of high temperature and low precipitation. Birds' sensitivity to drought is in large part driven by heat stress [110, 109], and we predicted few species declines when temperatures were normal. The importance of temperature is consistent with the fact that species increased their use of perennial agriculture—fruits, nuts, and vineyards that provide year-round microhabitats in the form of vegetative structure—but not row and field crops during drought. As temperatures in California are expected to continue to increase [38], water availability may be insufficient to promote resilience. Conservation must take into account the necessity of shifts in range and habitat use by species to buffer exposure to warming by seeking out habitats with available microrefugia.

Conserving birds in the Central Valley requires balancing the needs of wildlife with the reality of extensive human modification of the landscape. As extremely high temperatures synergize with water deficits to produce abundance declines among birds, habitats with stable sources of water and sufficient microrefugia may support the persistence of sensitive species. Our results, which show that species' relative use of developed and perennial agricultural habitat is greater during periods of drought, indicate that birds are likely already buffering some effects of anthropogenic climate change by tracking human-induced gradients in water availability across suitable habitats. Conservation managers can work with this trend by placing a stronger emphasis on conservation on working landscapes during drought [81]. Agricultural and developmental practices that promote biodiversity in the context of human modification, such as crop diversity and remnant natural habitat, could have a greater proportional effect on birds during drought when modified habitats are of greater relative importance [108, 8, 50]. However, a conservation paradigm that ties the persistence of birds during extremely hot, dry periods to agricultural and developed land poses potential problems. Increasing human-wildlife interaction can expose birds to additional stressors such as disturbance, noise, and pollution, which could constitute an ecological trap in which species prefer human-modified habitat despite having worse demographic outcomes there [112]. Negative impacts on human systems must also be considered, such as increased consumption of crops by birds, although birds may also predate pests and provide other ecosystem services. A conservation plan that emphasizes working lands in this system should focus on mitigating the impacts of human disturbance on birds and promoting biodiversity on human-dominated habitat during drought.

While this study represents a major step toward a comprehensive picture of drought impacts on birds, tailored conservation decision-making will require careful observational study of individual systems of interest to clarify the extent to which demographic processes and species movement separately contribute to changes in relative habitat use during drought. We suggest that ecologists emphasize interactions between habitat type and drought in future studies and experimental interventions. eBird data likely contain observer variation, cell variation, and overdispersion, all of which we accounted for in the model but which potentially limited our power to detect changes in abundance. Higher statistical power may be achieved via more targeted sampling in future studies. Because this study was restricted to the 66 most commonly detected birds in the Central Valley, our ability to identify impacts on rare species was limited. eBird data may be insufficient for understanding how rare species respond to drought, so ecologists may wish to prioritize targeted monitoring of rare species. We note also that while the N-mixture approach is an effective way to account for between- and within-site variation such as that generated by detection heterogeneity [114], it is possible that additional unmeasured variation in the detection process beyond that accounted for with covariates and random effects can introduce bias in parameter estimates or lead to misattribution of variation. For instance, if eBird observer behavior differed systematically during periods of extreme temperature beyond what was accounted for by effort covariates and observer-level random effects, we may infer biological relationships from detection-driven variation.

Our ability to identify changes in bird abundance during drought, and isolate those changes to particular environmental variables and habitat types, depended on the new multi-model framework presented in this manuscript. By hierarchically structuring the impacts of drought and habitat on abundance, we were able to estimate parameters across a complex set of ecological relationships for a large number of species. We propose that joint posterior predictive methodology will be a valuable tool for ecologists and environmental scientists seeking to leverage high-volume datasets to understand such systems.

Shifting habitat associations, more than abundance declines, define birds' responses to drought. Patterns in ten years of eBird data suggest that species respond rapidly to severe drought, and that individuals are likely able to track gradients of habitat suitability to meet temperature and water needs. This pattern is part of a global trend of increased human-wildlife interaction driven by climate change [1]. When human-induced resource gradients lead species onto agricultural and developed land, conservation managers must be prepared to follow. Conservation planning for such species should adopt a working lands approach that considers species' habitat associations not as fixed properties but as dynamic and climate-dependent [81].

Chapter 5

Conclusion

Participatory science data are a powerful resource for ecological inference. As the volume and spatiotemporal coverage of these data grow, ecologists will find increasing value in their use for wildlife monitoring. This trend creates a pressing need for methodological guidance. Ecologists must be equipped with clear, robust, and accessible modeling tools to use these data effectively.

In this dissertation, I presented methodological and applied work illustrating the use of participatory science data in gaining insight into the distribution of wildlife. Over the course of three chapters, several key insights emerged. First, statistical models that accommodate overdispersion and variation in the sampling process are essential when using these data. I found in a methodological comparison that, as the size of the participatory dataset increases, more flexible models fit the data better more often. I also showed that inference was dependent on being able to reasonably claim that a model's estimated relationship between observed counts and a potential abundance driver was not confounded with variation in the detection process. When dealing with participatory science data, models are only useful when a reasonable attempt has been made to model extra variation in detection, such that that variation is not incorrectly attributed to the variables of interest. Careful thought must always be given to the data context before the ecologist can assert that unmodeled variation in detection has not been misattributed to the process of interest.

I further found that participatory data were in practice useful for revealing novel aspects of ecological systems. In an applied investigation into drought impacts in the Central Valley, I attempted to decompose the effects of a complex set of correlated environmental processes on each of a number of species. The size and coverage of the eBird data I used were critical in allowing me to develop and estimate a sufficiently complex model. While this analysis posed significant computational and statistical challenges, they were not insurmountable. Rather, these observations constituted a unique resource for understanding the study system in question. I anticipate that this dynamic will continue to play out across the field of wildlife ecology. As computational and statistical advances make it easier for ecologists to estimate more complex models with larger datasets, the use of participatory data in ecological research will become increasingly mainstream. In concert with targeted and standardized surveys, participatory data will play a key role in the future of wildlife science and monitoring.

References

[1] Briana Abrahms et al. “Climate change as a global amplifier of human–wildlife conflict”. en. In: *Nature Climate Change* (Feb. 2023). Publisher: Nature Publishing Group, pp. 1–11. ISSN: 1758-6798. DOI: 10.1038/s41558-023-01608-5. URL: https://www.nature.com/articles/s41558-023-01608-5 (visited on 03/03/2023).

[2] Thomas P. Albright et al. “Combined effects of heat waves and droughts on avian communities across the conterminous United States”. en. In: *Ecosphere* 1.5 (2010). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/ES10-00057.1, art12. ISSN: 2150-8925. DOI: 10.1890/ES10-00057.1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1890/ES10-00057.1 (visited on 04/04/2023).

[3] Thomas P. Albright et al. “Effects of drought on avian community structure”. en. In: *Global Change Biology* 16.8 (2010). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2009.02120.x, pp. 2158–2170. ISSN: 1365-2486. DOI: 10.1111/j.1365-2486.2009.02120.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2009.02120.x (visited on 04/04/2023).

[4] Thomas P. Albright et al. “Mapping evaporative water loss in desert passerines reveals an expanding threat of lethal dehydration”. en. In: *Proceedings of the National Academy of Sciences* 114.9 (Feb. 2017), pp. 2283–2288. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1613625114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1613625114 (visited on 01/31/2020).

[5] Thomas Auer et al. *EOD – eBird Observation Dataset*. en. 2022. DOI: 10.15468/AOMFNB. URL: https://www.gbif.org/dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e (visited on 03/16/2023).

[6] Richard J. Barker et al. “On the reliability of N-mixture models for count data”. en. In: *Biometrics* 74.1 (Mar. 2018), pp. 369–377. ISSN: 0006341X. DOI: 10.1111/biom.12734. URL: http://doi.wiley.com/10.1111/biom.12734 (visited on 11/10/2019).

[7] Santiago Beguería. *sbegueria/SPEIbase: Version 2.7*. Jan. 2022. DOI: 10.5281/zenodo.5864391. URL: https://doi.org/10.5281/zenodo.5864391.

[8] Damien Beillouin et al. "Positive but variable effects of crop diversification on biodiversity and ecosystem services". en. In: *Global Change Biology* 27.19 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.15747, pp. 4697–4710. ISSN: 1365-2486. DOI: 10.1111/gcb.15747. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15747 (visited on 02/01/2023).

[9] Steven R. Beissinger et al. "Concordant and opposing effects of climate and land-use change on avian assemblages in California's most transformed landscapes". In: *Science Advances* 9.8 (Feb. 2023). Publisher: American Association for the Advancement of Science, eabn0250. DOI: 10.1126/sciadv.abn0250. URL: https://www.science.org/doi/10.1126/sciadv.abn0250 (visited on 04/05/2023).

[10] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x, pp. 289–300. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x (visited on 04/17/2023).

[11] Joanne M. Bennett et al. "Climate drying amplifies the effects of land-use change and interspecific interactions on birds". en. In: *Landscape Ecology* 30.10 (Dec. 2015), pp. 2031–2043. ISSN: 0921-2973, 1572-9761. DOI: 10.1007/s10980-015-0229-x. URL: http://link.springer.com/10.1007/s10980-015-0229-x (visited on 05/04/2022).

[12] Benjamin M. Bolker et al. "Generalized linear mixed models: a practical guide for ecology and evolution". en. In: *Trends in Ecology & Evolution* 24.3 (Mar. 2009), pp. 127–135. ISSN: 01695347. DOI: 10.1016/j.tree.2008.10.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534709000196 (visited on 09/09/2021).

[13] Benjamin M. Bolker et al. "Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS". en. In: *Methods in Ecology and Evolution* 4.6 (2013). _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12044, pp. 501–512. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12044. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12044 (visited on 03/06/2023).

[14] Yves Bötsch, Lukas Jenni, and Marc Kéry. "Field evaluation of abundance estimates under binomial and multinomial N -mixture models". en. In: *Ibis* 162.3 (July 2020), pp. 902–910. ISSN: 0019-1019, 1474-919X. DOI: 10.1111/ibi.12802. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ibi.12802 (visited on 01/11/2021).

[15] Mollie E. Brooks et al. “glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling”. In: *The R Journal* 9.2 (2017), pp. 378–400. URL: `https://journal.r-project.org/archive/2017/RJ-2017-066/index.html`.

[16] Steve Brooks, ed. *Handbook for Markov chain Monte Carlo*. Boca Raton: Taylor & Francis, 2011. ISBN: 978-1-4200-7941-8.

[17] Paul-Christian Bürkner. “brms: An R Package for Bayesian Multilevel Models Using Stan”. In: *Journal of Statistical Software* 80.1 (2017), pp. 1–28. DOI: `10.18637/jss.v080.i01`.

[18] Kenneth P. Burnham and David Raymond Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. en. 2nd ed. OCLC: ocm48557578. New York: Springer, 2002. ISBN: 978-0-387-95364-9.

[19] Samantha M. Cady et al. “Species-specific and temporal scale-dependent responses of birds to drought”. en. In: *Global Change Biology* 25.8 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.14668, pp. 2691–2702. ISSN: 1365-2486. DOI: `10.1111/gcb.14668`. URL: `http://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14668` (visited on 06/10/2022).

[20] Abigail E. Cahill et al. “How does climate change cause extinction?” en. In: *Proceedings of the Royal Society B: Biological Sciences* 280.1750 (Jan. 2013), p. 20121890. ISSN: 0962-8452, 1471-2954. DOI: `10.1098/rspb.2012.1890`. URL: `https://royalsocietypublishing.org/doi/10.1098/rspb.2012.1890` (visited on 01/30/2020).

[21] Mark Chandler et al. “Contribution of citizen science towards international biodiversity monitoring”. en. In: *Biological Conservation* 213 (Sept. 2017), pp. 280–294. ISSN: 00063207. DOI: `10.1016/j.biocon.2016.09.004`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S0006320716303639` (visited on 11/17/2020).

[22] Sonja A. Christensen, Matthew T. Farr, and David M. Williams. “Assessment and novel application of N -mixture models for aerial surveys of wildlife”. en. In: *Ecosphere* 12.8 (Aug. 2021). ISSN: 2150-8925, 2150-8925. DOI: `10.1002/ecs2.3725`. URL: `https://onlinelibrary.wiley.com/doi/10.1002/ecs2.3725` (visited on 09/10/2021).

[23] Jeremy M. Cohen, Daniel Fink, and Benjamin Zuckerberg. “Avian responses to extreme weather across functional traits and temporal scales”. en. In: *Global Change Biology* 26.8 (Aug. 2020), pp. 4240–4250. ISSN: 1354-1013, 1365-2486. DOI: `10.1111/gcb.15133`. URL: `https://onlinelibrary.wiley.com/doi/10.1111/gcb.15133` (visited on 09/09/2021).

[24] Melinda G. Conners et al. “Hidden Markov models identify major movement modes in accelerometer and magnetometer data from four albatross species”. In: *Movement Ecology* 9.1 (Feb. 2021), p. 7. ISSN: 2051-3933. DOI: `10.1186/s40462-021-00243-z`. URL: `https://doi.org/10.1186/s40462-021-00243-z` (visited on 03/07/2023).

[25] R. M. Cormack. “Estimates of Survival from the Sighting of Marked Animals”. In: *Biometrika* 51.3/4 (1964). Publisher: [Oxford University Press, Biometrika Trust], pp. 429–438. ISSN: 0006-3444. DOI: `10.2307/2334149`. URL: `https://www.jstor.org/stable/2334149` (visited on 01/27/2023).

[26] Andrea Costa, Antonio Romano, and Sebastiano Salvidio. “Reliability of multinomial N-mixture models for estimating abundance of small terrestrial vertebrates”. en. In: *Biodiversity and Conservation* 29.9-10 (Aug. 2020), pp. 2951–2965. ISSN: 0960-3115, 1572-9710. DOI: `10.1007/s10531-020-02006-5`. URL: `http://link.springer.com/10.1007/s10531-020-02006-5` (visited on 01/11/2021).

[27] Thibaut Couturier et al. “Estimating abundance and population trends when detection is low and highly variable: A comparison of three methods for the Hermann’s tortoise: Three Methods for Estimating *T. hermanni* Abundance”. en. In: *The Journal of Wildlife Management* 77.3 (Apr. 2013), pp. 454–462. ISSN: 0022541X. DOI: `10.1002/jwmg.499`. URL: `https://onlinelibrary.wiley.com/doi/10.1002/jwmg.499` (visited on 01/20/2022).

[28] Shelley D. Crausbay et al. “Defining Ecological Drought for the Twenty-First Century”. en. In: *Bulletin of the American Meteorological Society* 98.12 (Dec. 2017), pp. 2543–2550. ISSN: 0003-0007, 1520-0477. DOI: `10.1175/BAMS-D-16-0292.1`. URL: `https://journals.ametsoc.org/doi/10.1175/BAMS-D-16-0292.1` (visited on 06/10/2022).

[29] Shelley D. Crausbay et al. “Unfamiliar Territory: Emerging Themes for Ecological Drought Research and Management”. en. In: *One Earth* 3.3 (Sept. 2020), pp. 337–353. ISSN: 25903322. DOI: `10.1016/j.oneear.2020.08.019`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S2590332220304280` (visited on 06/08/2022).

[30] Noel Cressie et al. “Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling”. en. In: *Ecological Applications* 19.3 (Apr. 2009), pp. 553–570. ISSN: 1051-0761. DOI: `10.1890/07-0744.1`. URL: `http://doi.wiley.com/10.1890/07-0744.1` (visited on 07/11/2019).

[31] Scott P. Crosbie et al. “Early Impact of West Nile Virus on the Yellow-Billed Magpie (Pica Nuttalli)”. In: *The Auk* 125.3 (July 2008), pp. 542–550. ISSN: 1938-4254. DOI: `10.1525/auk.2008.07040`. URL: `https://doi.org/10.1525/auk.2008.07040` (visited on 04/05/2023).

[32] Aiguo Dai. “Increasing drought under global warming in observations and models”. en. In: *Nature Climate Change* 3.1 (Jan. 2013). Number: 1 Publisher: Nature Publishing Group, pp. 52–58. ISSN: 1758-6798. DOI: 10.1038/nclimate1633. URL: https://www.nature.com/articles/nclimate1633 (visited on 06/10/2022).

[33] William V DeLuca et al. “The Colorado River Delta and California’s Central Valley are critical regions for many migrating North American landbirds”. en. In: *Ornithological Applications* 123.1 (Mar. 2021), duaa064. ISSN: 0010-5422, 2732-4621. DOI: 10.1093/ornithapp/duaa064. URL: https://academic.oup.com/condor/article/doi/10.1093/ornithapp/duaa064/6119082 (visited on 02/11/2022).

[34] Francisco V. Dénes, Luís Fábio Silveira, and Steven R. Beissinger. “Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation”. en. In: *Methods in Ecology and Evolution* 6.5 (May 2015). Ed. by Nick Isaac, pp. 543–556. ISSN: 2041-210X, 2041-210X. DOI: 10.1111/2041-210X.12333. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12333 (visited on 12/12/2020).

[35] Emily B. Dennis, Byron J.T. Morgan, and Martin S. Ridout. “Computational aspects of N-mixture models: Computational Aspects of N-Mixture Models”. en. In: *Biometrics* 71.1 (Mar. 2015), pp. 237–246. ISSN: 0006341X. DOI: 10.1111/biom.12246. URL: http://doi.wiley.com/10.1111/biom.12246 (visited on 11/19/2019).

[36] Stacy L. DeRuiter et al. “A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure”. In: *The Annals of Applied Statistics* 11.1 (Mar. 2017). Publisher: Institute of Mathematical Statistics, pp. 362–392. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/16-AOAS1008. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-11/issue-1/A-multivariate-mixed-hidden-Markov-model-for-blue-whale-behaviour/10.1214/16-AOAS1008.full (visited on 03/06/2023).

[37] Kadambari Devarajan, Toni Lyn Morelli, and Simone Tenan. “Multi-species occupancy models: review, roadmap, and recommendations”. en. In: *Ecography* (Feb. 2020), ecog.04957. ISSN: 0906-7590, 1600-0587. DOI: 10.1111/ecog.04957. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.04957 (visited on 08/10/2020).

[38] Noah S. Diffenbaugh, Daniel L. Swain, and Danielle Touma. “Anthropogenic warming has increased drought risk in California”. en. In: *Proceedings of the National Academy of Sciences* 112.13 (Mar. 2015), pp. 3931–3936. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1422385112. URL: https://pnas.org/doi/full/10.1073/pnas.1422385112 (visited on 05/06/2022).

[39] Lynda Donaldson et al. "Quantifying resistance and resilience to local extinction for conservation prioritization". en. In: *Ecological Applications* 29.8 (2019). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/eap.1989, e01989. ISSN: 1939-5582. DOI: 10.1002/eap.1989. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/eap.1989 (visited on 03/09/2023).

[40] Jeffrey W. Doser, Andrew O. Finley, and Sudipto Banerjee. "Joint species distribution models with imperfect detection for high-dimensional spatial data". In: *arXiv preprint arXiv:2204.02707* (2022). URL: https://arxiv.org/abs/2204.02707.

[41] Adam Duarte, Michael J. Adams, and James T. Peterson. "Fitting N-mixture models to count data with unmodeled heterogeneity: Bias, diagnostics, and alternative approaches". en. In: *Ecological Modelling* 374 (Apr. 2018), pp. 51–59. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2018.02.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304380018300607 (visited on 11/19/2019).

[42] Jane Elith and John R. Leathwick. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time". en. In: *Annual Review of Ecology, Evolution, and Systematics* 40.1 (Dec. 2009), pp. 677–697. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev.ecolsys.110308.120159. URL: http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.110308.120159 (visited on 08/30/2019).

[43] Gentile Francesco Ficetola et al. "N-mixture models reliably estimate the abundance of small vertebrates". en. In: *Scientific Reports* 8.1 (Dec. 2018), p. 10357. ISSN: 2045-2322. DOI: 10.1038/s41598-018-28432-8. URL: http://www.nature.com/articles/s41598-018-28432-8 (visited on 11/18/2019).

[44] S.E. Fick and Robert J. Hijmans. "WorldClim 2: new 1km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12 (2017), pp. 4302–4315.

[45] Daniel Fink et al. "Spatiotemporal exploratory models for broad-scale survey data". en. In: *Ecological Applications* 20.8 (Dec. 2010), pp. 2131–2147. ISSN: 1051-0761. DOI: 10.1890/09-1340.1. URL: http://doi.wiley.com/10.1890/09-1340.1 (visited on 08/07/2018).

[46] Ian Fiske and Richard Chandler. "unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance". In: *Journal of Statistical Software* 43.10 (2011), pp. 1–23. URL: http://www.jstatsoft.org/v43/i10/.

[47] Alastair Franke, Terry Caelli, and Robert J Hudson. "Analysis of movements and behavior of caribou (Rangifer tarandus) using hidden Markov models". en. In: *Ecological Modelling* 173.2 (Apr. 2004), pp. 259–270. ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2003.06.004. URL: https://www.sciencedirect.com/science/article/pii/S0304380003003983 (visited on 01/27/2023).

[48] Brett J. Furnas and Richard L. Callas. “Using automated recorders and occupancy models to monitor common forest birds across a large geographic region: Automated Recorders Monitoring Common Birds”. en. In: *The Journal of Wildlife Management* 79.2 (Feb. 2015), pp. 325–337. ISSN: 0022541X. DOI: 10.1002/jwmg.821. URL: http://doi.wiley.com/10.1002/jwmg.821 (visited on 08/11/2020).

[49] Brett J. Furnas and Michael C. McGrann. “Using occupancy modeling to monitor dates of peak vocal activity for passerines in California”. en. In: *The Condor* 120.1 (Feb. 2018), pp. 188–200. ISSN: 0010-5422, 1938-5129. DOI: 10.1650/CONDOR-17-165.1. URL: https://academic.oup.com/condor/article/120/1/188-200/5152992 (visited on 08/03/2020).

[50] Lucas A. Garibaldi et al. “Working landscapes need at least 20% native habitat”. en. In: *Conservation Letters* 14.2 (2021). _eprint: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12773, e12773. ISSN: 1755-263X. DOI: 10.1111/conl.12773. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/conl.12773 (visited on 02/01/2023).

[51] T. Luke George et al. “Impacts of a Severe Drought on Grassland Birds in Western North Dakota”. en. In: *Ecological Applications* 2.3 (1992). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1941861, pp. 275–284. ISSN: 1939-5582. DOI: 10.2307/1941861. URL: http://onlinelibrary.wiley.com/doi/abs/10.2307/1941861 (visited on 06/10/2022).

[52] Richard Glennie et al. “Hidden Markov models: Pitfalls and opportunities in ecology”. en. In: *Methods in Ecology and Evolution* 14.1 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13801, pp. 43–56. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13801. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13801 (visited on 03/06/2023).

[53] Benjamin R. Goldstein and Perry de Valpine. “Comparing N-mixture models and GLMMs for relative abundance estimation in a citizen science dataset”. en. In: *Scientific Reports* 12.1 (July 2022). Number: 1 Publisher: Nature Publishing Group, p. 12276. ISSN: 2045-2322. DOI: 10.1038/s41598-022-16368-z. URL: http://www.nature.com/articles/s41598-022-16368-z (visited on 08/25/2022).

[54] Benjamin R. Goldstein et al. *nimbleEcology: Distributions for Ecological Models in nimble.* 2020. URL: https://cran.r-project.org/package=nimbleEcology.

[55] Gregory H. Golet et al. “Using ricelands to provide temporary shorebird habitat during migration”. en. In: *Ecological Applications* 28.2 (Mar. 2018), pp. 409–426. ISSN: 10510761. DOI: 10.1002/eap.1658. URL: http://doi.wiley.com/10.1002/eap.1658 (visited on 01/24/2019).

[56] G.E. Griffith et al. *Ecoregions of California (poster): U.S. Geological Survey Open-File Report.* en. Open-File Report. Series: Open-File Report. 2016.

[57] Joseph Grinnell. “The Niche-Relationships of the California Thrasher”. In: *The Auk* 34.4 (1917). Publisher: American Ornithological Society, pp. 427–433. ISSN: 0004-8038. DOI: 10.2307/4072271. URL: https://www.jstor.org/stable/4072271 (visited on 03/15/2023).

[58] Nicholas E. Hamilton and Michael Ferry. “ggtern: Ternary Diagrams Using ggplot2”. In: *Journal of Statistical Software, Code Snippets* 87.3 (2018), pp. 1–17. DOI: 10.18637/jss.v087.c03.

[59] Edmund M. Hart and Kendon Bell. *prism: Download data from the Oregon prism project.* 2015. DOI: 10.5281/zenodo.33663. URL: https://github.com/ropensci/prism.

[60] Tom Hart et al. “Behavioural switching in a central place forager: patterns of diving behaviour in the macaroni penguin (Eudyptes chrysolophus)”. en. In: *Marine Biology* 157.7 (July 2010), pp. 1543–1553. ISSN: 1432-1793. DOI: 10.1007/s00227-010-1428-2. URL: https://doi.org/10.1007/s00227-010-1428-2 (visited on 03/06/2023).

[61] Florian Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models.* 2022. URL: https://CRAN.R-project.org/package=DHARMa.

[62] J. Nicholas Hendershot et al. “Intensive farming drives long-term shifts in avian community composition”. en. In: *Nature* 579.7799 (Mar. 2020). Number: 7799 Publisher: Nature Publishing Group, pp. 393–396. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2090-6. URL: https://www.nature.com/articles/s41586-020-2090-6 (visited on 01/24/2023).

[63] Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling.* 2020. URL: https://CRAN.R-project.org/package=raster.

[64] E. Hines. *PRESENCE - Software to estimate patch occupancy and related parameters.* 2006. URL: http://www.mbr-pwrc.usgs.gov/software/presence.html.

[65] Wesley M. Hochachka et al. “Data-intensive science applied to broad-scale citizen science”. en. In: *Trends in Ecology & Evolution* 27.2 (Feb. 2012), pp. 130–137. ISSN: 01695347. DOI: 10.1016/j.tree.2011.11.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0169534711003296 (visited on 08/07/2018).

[66] Kelly J. Iknayan and Steven R. Beissinger. “Collapse of a desert bird community over the past century driven by climate change”. en. In: *Proceedings of the National Academy of Sciences* 115.34 (Aug. 2018), pp. 8597–8602. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1805123115. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1805123115 (visited on 01/30/2020).

[67] Kelly J. Iknayan et al. “Detecting diversity: emerging methods to estimate species diversity”. en. In: *Trends in Ecology & Evolution* 29.2 (Feb. 2014), pp. 97–106. ISSN: 01695347. DOI: 10.1016/j.tree.2013.10.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534713002619 (visited on 11/06/2019).

[68] Michelle C. Jackson et al. “Net effects of multiple stressors in freshwater ecosystems: a meta-analysis”. en. In: *Global Change Biology* 22.1 (Jan. 2016), pp. 180–189. ISSN: 13541013. DOI: 10.1111/gcb.13028. URL: http://doi.wiley.com/10.1111/gcb.13028 (visited on 02/20/2020).

[69] Alison Johnston, Eleni Matechou, and Emily B. Dennis. “Outstanding challenges and future directions for biodiversity monitoring using citizen science data”. en. In: *Methods in Ecology and Evolution* (Mar. 2022), pp. 2041–210X.13834. ISSN: 2041-210X, 2041-210X. DOI: 10.1111/2041-210X.13834. URL: https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13834 (visited on 04/21/2022).

[70] Alison Johnston et al. “Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions”. en. In: *Diversity and Distributions* 27.7 (July 2021). Ed. by Yoan Fourcade, pp. 1265–1277. ISSN: 1366-9516, 1472-4642. DOI: 10.1111/ddi.13271. URL: https://onlinelibrary.wiley.com/doi/10.1111/ddi.13271 (visited on 08/24/2021).

[71] Alison Johnston et al. “Estimating species distributions from spatially biased citizen science data”. en. In: *Ecological Modelling* 422 (Apr. 2020), p. 108927. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2019.108927. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304380019304351 (visited on 08/07/2020).

[72] G. M. Jolly. “Explicit Estimates from Capture-Recapture Data with Both Death and Immigration-Stochastic Model”. In: *Biometrika* 52.1/2 (1965). Publisher: [Oxford University Press, Biometrika Trust], pp. 225–247. ISSN: 0006-3444. DOI: 10.2307/2333826. URL: https://www.jstor.org/stable/2333826 (visited on 01/27/2023).

[73] C.O. Justice et al. “The Moderate Resolution Imaging Spectroradiometer (MODIS): land remote sensing for global change research”. In: *IEEE Transactions on Geoscience and Remote Sensing* 36.4 (July 1998). Conference Name: IEEE Transactions on Geoscience and Remote Sensing, pp. 1228–1249. ISSN: 1558-0644. DOI: 10.1109/36.701075.

[74] Stefan Kahl et al. “BirdNET: A deep learning solution for avian diversity monitoring”. en. In: *Ecological Informatics* 61 (Mar. 2021), p. 101236. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2021.101236. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574954121000273 (visited on 09/09/2021).

[75] Michael Kearney, Richard Shine, and Warren P. Porter. "The potential for behavioral thermoregulation to buffer "cold-blooded" animals against climate warming". In: *Proceedings of the National Academy of Sciences* 106.10 (Mar. 2009). Publisher: Proceedings of the National Academy of Sciences, pp. 3835–3840. DOI: 10.1073/pnas.0808913106. URL: https://www.pnas.org/doi/full/10.1073/pnas.0808913106 (visited on 06/10/2022).

[76] Kenneth F. Kellner and Robert K. Swihart. "Accounting for Imperfect Detection in Ecology: A Quantitative Review". en. In: *PLOS ONE* 9.10 (Oct. 2014). Publisher: Public Library of Science, e111436. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0111436. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0111436 (visited on 03/09/2023).

[77] Kenneth F. Kellner et al. "ubms: An R package for fitting hierarchical occupancy and N-mixture abundance models in a Bayesian framework". In: *Methods in Ecology and Evolution* 13 (2021), pp. 577–584. URL: https://doi.org/10.1111/2041-210X.13777.

[78] Marc Kéry. "Identifiability in *N* -mixture models: a large-scale screening test with bird data". en. In: *Ecology* 99.2 (Feb. 2018), pp. 281–288. ISSN: 00129658. DOI: 10.1002/ecy.2093. URL: http://doi.wiley.com/10.1002/ecy.2093 (visited on 11/18/2019).

[79] Marc Kéry and J. Andrew Royle. *Applied hierarchical modeling in ecology*. Amsterdam: Elsevier/AP, 2016.

[80] Jonas Knape et al. "Sensitivity of binomial N-mixture models to overdispersion: The importance of assessing model fit". en. In: *Methods in Ecology and Evolution* 9.10 (Oct. 2018). Ed. by Nick Isaac, pp. 2102–2114. ISSN: 2041210X. DOI: 10.1111/2041-210X.13062. URL: http://doi.wiley.com/10.1111/2041-210X.13062 (visited on 06/13/2019).

[81] C. Kremen and A. M. Merenlender. "Landscapes that work for biodiversity and people". In: *Science* 362.6412 (Oct. 2018). Publisher: American Association for the Advancement of Science, eaau6020. DOI: 10.1126/science.aau6020. URL: https://www.science.org/doi/10.1126/science.aau6020 (visited on 01/31/2023).

[82] J. L. Laake. *RMark: An R Interface for Analysis of Capture-Recapture Data with MARK*. AFSC Processed Rep. 2013-01. Seattle, WA: Alaska Fish. Sci. Cent., NOAA, Natl. Mar. Fish. Serv., 2013, p. 25. URL: https://apps-afsc.fisheries.noaa.gov/Publications/ProcRpt/PR2013-01.pdf.

[83] LANDFIRE. *LANDFIRE Remap 2016 Existing Vegetation Type (EVT) CONUS*. Tech. rep. Earth Resources Observation and Science Center (EROS), U.S. Geological Survey, 2020. (Visited on 07/21/2020).

[84] Jean-Dominique Lebreton et al. “Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach with Case Studies”. en. In: *Ecological Monographs* 62.1 (1992). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/2937171, pp. 67–118. ISSN: 1557-7015. DOI: 10.2307/2937171. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/2937171 (visited on 01/27/2023).

[85] William A. Link et al. “On the robustness of N-mixture models”. en. In: *Ecology* 99.7 (July 2018), pp. 1547–1551. ISSN: 00129658. DOI: 10.1002/ecy.2362. URL: http://doi.wiley.com/10.1002/ecy.2362 (visited on 11/18/2019).

[86] Ralph Mac Nally et al. “Collapse of an avifauna: climate change appears to exacerbate habitat loss and degradation”. en. In: *Diversity and Distributions* 15.4 (July 2009), pp. 720–730. ISSN: 13669516, 14724642. DOI: 10.1111/j.1472-4642.2009.00578.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2009.00578.x (visited on 05/04/2022).

[87] Darryl I MacKenzie et al. “Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One”. en. In: (Aug. 2002), p. 9.

[88] Darryl I. MacKenzie, ed. *Occupancy estimation and modeling: inferring patterns and dynamics of species*. Amsterdam ; Boston: Elsevier, 2006. ISBN: 978-0-12-088766-8.

[89] Darryl I. MacKenzie et al. “ESTIMATING SITE OCCUPANCY, COLONIZATION, AND LOCAL EXTINCTION WHEN A SPECIES IS DETECTED IMPERFECTLY”. en. In: *Ecology* 84.8 (Aug. 2003), pp. 2200–2207. ISSN: 0012-9658. DOI: 10.1890/02-3090. URL: http://doi.wiley.com/10.1890/02-3090 (visited on 10/03/2019).

[90] Sarah A. MacLean et al. “A century of climate and land-use change cause species turnover without loss of beta diversity in California’s Central Valley”. In: *Global Change Biology* 24 (2018), pp. 5882–5894. DOI: 10.1111/gcb.14458.

[91] Chrystal S. Mantyka-pringle, Tara G. Martin, and Jonathan R. Rhodes. “Interactions between climate and habitat loss effects on biodiversity: a systematic review and meta-analysis”. en. In: *Global Change Biology* 18.4 (2012). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2011.02593.x, pp. 1239–1252. ISSN: 1365-2486. DOI: 10.1111/j.1365-2486.2011.02593.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2011.02593.x (visited on 01/25/2023).

[92] Julien Martin et al. “Accounting for non-independent detection when estimating abundance of organisms with a Bayesian approach: Correlated behaviour and abundance”. en. In: *Methods in Ecology and Evolution* 2.6 (Dec. 2011), pp. 595–601. ISSN: 2041210X. DOI: 10.1111/j.2041-210X.2011.00113.x. URL: http://doi.wiley.com/10.1111/j.2041-210X.2011.00113.x (visited on 07/30/2019).

[93] Brett T. McClintock and Théo Michelot. "momentuHMM: R package for generalized hidden Markov models of animal movement". In: *Methods in Ecology and Evolution* 9.6 (2018), pp. 1518–1530. DOI: 10.1111/2041-210X.12995. URL: http://dx.doi.org/10.1111/2041-210X.12995.

[94] Brett T. McClintock et al. "Uncovering ecological state dynamics with hidden Markov models". en. In: *Ecology Letters* 23.12 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13610, pp. 1878–1903. ISSN: 1461-0248. DOI: 10.1111/ele.13610. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13610 (visited on 01/27/2023).

[95] Michael C. McGrann and Brett J. Furnas. "Divergent species richness and vocal behavior in avian migratory guilds along an elevational gradient". en. In: *Ecosphere* 7.8 (2016). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecs2.1419, e01419. ISSN: 2150-8925. DOI: 10.1002/ecs2.1419. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ecs2.1419 (visited on 01/23/2023).

[96] Timothy D. Meehan, Nicole L. Michel, and Håvard Rue. "Estimating Animal Abundance with N-Mixture Models Using the *R* - **INLA** Package for *R*". en. In: *Journal of Statistical Software* 95.2 (2020). ISSN: 1548-7660. DOI: 10.18637/jss.v095.i02. URL: http://www.jstatsoft.org/v95/i02/ (visited on 10/07/2021).

[97] Adrian P. Monroe et al. "The importance of simulation assumptions when evaluating detectability in population models". en. In: *Ecosphere* 10.7 (July 2019). ISSN: 2150-8925, 2150-8925. DOI: 10.1002/ecs2.2791. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ecs2.2791 (visited on 01/11/2021).

[98] J Mount and E Hanak. *Water Use in California.* Tech. rep. PPIC Water Policy Center, July 2016.

[99] Dale G. Nimmo et al. "Riparian tree cover enhances the resistance and stability of woodland bird communities during an extreme climatic event". en. In: *Journal of Applied Ecology* 53.2 (Apr. 2016). Ed. by Jeremy James, pp. 449–458. ISSN: 00218901. DOI: 10.1111/1365-2664.12535. URL: https://onlinelibrary.wiley.com/doi/10.1111/1365-2664.12535 (visited on 05/04/2022).

[100] Joseph M. Northrup and Brian D. Gerber. "A comment on priors for Bayesian occupancy models". en. In: *PLOS ONE* 13.2 (Feb. 2018). Ed. by Yong Deng, e0192819. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0192819. URL: https://dx.plos.org/10.1371/journal.pone.0192819 (visited on 03/31/2021).

[101] Martyn Plummer. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Ed. by K Hornik, F Leisch, and A Zeileis. Technische Universität Wien, Vienna, Austria, 2003.

[102] Martyn Plummer. "Simulation-Based Bayesian Analysis". en. In: *Annual Review of Statistics and Its Application* 10.1 (Mar. 2023), annurev–statistics–122121–040905. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-122121-040905. URL: https://www.annualreviews.org/doi/10.1146/annurev-statistics-122121-040905 (visited on 02/28/2023).

[103] Lauren C. Ponisio et al. "One size does not fit all: Customizing MCMC methods for hierarchical models using NIMBLE". en. In: *Ecology and Evolution* 10.5 (Mar. 2020), pp. 2385–2416. ISSN: 2045-7758, 2045-7758. DOI: 10.1002/ece3.6053. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.6053 (visited on 04/29/2020).

[104] Laura R. Prugh et al. "Ecological winners and losers of extreme drought in California". en. In: *Nature Climate Change* 8.9 (Sept. 2018), pp. 819–824. ISSN: 1758-678X, 1758-6798. DOI: 10.1038/s41558-018-0255-1. URL: http://www.nature.com/articles/s41558-018-0255-1 (visited on 04/17/2019).

[105] *Publications - eBird Science.* en. URL: https://science.ebird.org/en/research-and-conservation/publications (visited on 03/15/2023).

[106] R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2020. URL: https://www.R-project.org/.

[107] Mark D. Reynolds et al. "Dynamic conservation for migratory species". en. In: *Science Advances* 3.8 (Aug. 2017), e1700707. ISSN: 2375-2548. DOI: 10.1126/sciadv.1700707. URL: http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1700707 (visited on 01/24/2019).

[108] Lindsey N. Rich et al. *An evaluation of avifaunal diversity in California's Great Valley.* Tech. rep. Department of Environmental Science, Policy & Management, University of California, 2017.

[109] E. A. Riddell et al. "Exposure to climate change drives stability or collapse of desert mammal and bird communities". en. In: *Science* 371.6529 (Feb. 2021), pp. 633–636. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abd4605. URL: https://www.science.org/doi/10.1126/science.abd4605 (visited on 04/12/2022).

[110] Eric A. Riddell et al. "Cooling requirements fueled the collapse of a desert bird community from climate change". en. In: *Proceedings of the National Academy of Sciences* (Sept. 2019), p. 201908791. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1908791116. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1908791116 (visited on 10/09/2019).

[111] Jason Riggio et al. "Long-term monitoring reveals the impact of changing climate and habitat on the fitness of cavity-nesting songbirds". en. In: *Biological Conservation* 278 (Feb. 2023), p. 109885. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2022.109885. URL: https://www.sciencedirect.com/science/article/pii/S0006320722004384 (visited on 01/31/2023).

[112] Bruce A. Robertson and Richard L. Hutto. "A Framework for Understanding Ecological Traps and an Evaluation of Existing Evidence". en. In: *Ecology* 87.5 (2006). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/0012-9658%282006%2987%5B1075%3AAFFUET%5D2.0.CO%3B2, pp. 1075–1085. ISSN: 1939-9170. DOI: 10.1890/0012-9658(2006)87[1075:AFFUET]2.0.CO;2. URL: https://onlinelibrary.wiley.com/doi/abs/10.1890/0012-9658%282006%2987%5B1075%3AAFFUET%5D2.0.CO%3B2 (visited on 02/06/2023).

[113] Orin J. Robinson et al. "Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models". en. In: *Diversity and Distributions* 26.8 (Aug. 2020). Ed. by Luigi Maiorano, pp. 976–986. ISSN: 1366-9516, 1472-4642. DOI: 10.1111/ddi.13068. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13068 (visited on 08/10/2020).

[114] J. Andrew Royle. "*N* -Mixture Models for Estimating Population Size from Spatially Replicated Counts". en. In: *Biometrics* 60.1 (Mar. 2004), pp. 108–115. ISSN: 0006-341X, 1541-0420. DOI: 10.1111/j.0006-341X.2004.00142.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2004.00142.x (visited on 10/09/2019).

[115] J. Andrew Royle and Robert M. Dorazio. "Hierarchical models of animal abundance and occurrence". en. In: *Journal of Agricultural, Biological, and Environmental Statistics* 11.3 (Sept. 2006), pp. 249–263. ISSN: 1085-7117, 1537-2693. DOI: 10.1198/108571106X129153. URL: http://link.springer.com/10.1198/108571106X129153 (visited on 09/02/2021).

[116] G. A. F. Seber. "A Note on the Multiple-Recapture Census". In: *Biometrika* 52.1/2 (1965). Publisher: [Oxford University Press, Biometrika Trust], pp. 249–259. ISSN: 0006-3444. DOI: 10.2307/2333827. URL: https://www.jstor.org/stable/2333827 (visited on 01/27/2023).

[117] Katherine E. Selwood et al. "High-productivity vegetation is important for lessening bird declines during prolonged drought". en. In: *Journal of Applied Ecology* 55.2 (Mar. 2018). Ed. by Steve Willis, pp. 641–650. ISSN: 0021-8901, 1365-2664. DOI: 10.1111/1365-2664.13052. URL: https://onlinelibrary.wiley.com/doi/10.1111/1365-2664.13052 (visited on 10/20/2022).

[118] Jonathan Silvertown. "A new dawn for citizen science". en. In: *Trends in Ecology & Evolution* 24.9 (Sept. 2009), pp. 467–471. ISSN: 01695347. DOI: 10.1016/j.tree.2009.03.017. URL: http://linkinghub.elsevier.com/retrieve/pii/S016953470900175X (visited on 08/21/2018).

[119] Robin Steenweg et al. "Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors". en. In: *Frontiers in Ecology and the Environment* 15.1 (Feb. 2017), pp. 26–34. ISSN: 15409295. DOI: 10.1002/fee.1448. URL: https://onlinelibrary.wiley.com/doi/10.1002/fee.1448 (visited on 09/10/2021).

[120] David R. Stewart et al. "Mark-recapture models identify imminent extinction of Yaqui catfish Ictalurus pricei in the United States". en. In: *Biological Conservation* 209 (May 2017), pp. 45–53. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2017.02.010. URL: https://www.sciencedirect.com/science/article/pii/S0006320716310084 (visited on 03/06/2023).

[121] Matthew Strimas-Mackey, Eliot Miller, and Wesley Hochachka. *auk: eBird Data Extraction and Processing with AWK*. 2018. URL: https://cornelllabofornithology.github.io/auk/.

[122] Matthew Strimas-Mackey et al. *Best Practices for Using eBird Data. Version 1.0*. Ithaca, New York: Cornell Lab of Ornithology, 2020. URL: https://cornelllabofornithology.github.io/ebird-best-practices/.

[123] Sarah Strochak, Kyle Ueyama, and Aaron Williams. *urbnmapr: State and county shapefiles in sf and tibble format*. 2022. URL: https://github.com/UrbanInstitute/urbnmapr.

[124] Brian L. Sullivan et al. "eBird: A citizen-based bird observation network in the biological sciences". en. In: *Biological Conservation* 142.10 (Oct. 2009), pp. 2282–2292. ISSN: 00063207. DOI: 10.1016/j.biocon.2009.05.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S000632070900216X (visited on 08/11/2020).

[125] Brian L. Sullivan et al. "The eBird enterprise: An integrated approach to development and application of citizen science". en. In: *Biological Conservation* 169 (Jan. 2014), pp. 31–40. ISSN: 00063207. DOI: 10.1016/j.biocon.2013.11.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006320713003820 (visited on 11/30/2020).

[126] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual.* 2023. URL: https://mc-stan.org.

[127] E.J. Theobald et al. "Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research". en. In: *Biological Conservation* 181 (Jan. 2015), pp. 236–244. ISSN: 00063207. DOI: 10.1016/j.biocon.2014.10.021. URL: http://linkinghub.elsevier.com/retrieve/pii/S0006320714004029 (visited on 08/07/2018).

[128] Joseph A. Tobias et al. "AVONET: morphological, ecological and geographical data for all birds". en. In: *Ecology Letters* 25.3 (Mar. 2022). Ed. by Tim Coulson, pp. 581–597. ISSN: 1461-023X, 1461-0248. DOI: 10.1111/ele.13898. URL: https://onlinelibrary.wiley.com/doi/10.1111/ele.13898 (visited on 05/06/2022).

[129] J. M. J. Travis. "Climate change and habitat destruction: a deadly anthropogenic cocktail". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.1514 (Mar. 2003). Publisher: Royal Society, pp. 467–473. DOI: 10.1098/rspb.2002.2246. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2002.2246 (visited on 01/25/2023).

[130] Daniel Turek, Perry de Valpine, and Christopher J. Paciorek. "Efficient Markov chain Monte Carlo sampling for hierarchical hidden Markov models". en. In: *Environmental and Ecological Statistics* 23.4 (Dec. 2016), pp. 549–564. ISSN: 1352-8505, 1573-3009. DOI: 10.1007/s10651-016-0353-z. URL: http://link.springer.com/10.1007/s10651-016-0353-z (visited on 09/24/2019).

[131] USGS. *USGS Landsat 7 Collection 1 Tier 1 and Real-Time data Raw Scenes [Data set].* 2022. URL: https://www.usgs.gov/landsat-missions/landsat-collection-1.

[132] Perry de Valpine and Alan Hastings. "FITTING POPULATION MODELS INCORPORATING PROCESS NOISE AND OBSERVATION ERROR". en. In: *Ecological Monographs* 72.1 (2002), p. 20.

[133] Perry de Valpine et al. "Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE". en. In: *Journal of Computational and Graphical Statistics* 26.2 (Apr. 2017), pp. 403–413. ISSN: 1061-8600, 1537-2715. DOI: 10.1080/10618600.2016.1172487. URL: https://www.tandfonline.com/doi/full/10.1080/10618600.2016.1172487 (visited on 02/13/2019).

[134] E. Vermote and R. Wolfe. *MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1kmand 500m SIN Grid V006 [Data set].* 2015. URL: https://doi.org/10.5067/MODIS/MOD09GA.006 (visited on 01/19/2023).

[135] Sergio M. Vicente-Serrano et al. "A review of environmental droughts: Increased risk under global warming?" en. In: *Earth-Science Reviews* 201 (Feb. 2020), p. 102953. ISSN: 00128252. DOI: 10.1016/j.earscirev.2019.102953. URL: https://linkinghub.elsevier.com/retrieve/pii/S0012825218306421 (visited on 09/20/2021).

[136] California Department of Water Resources. *2018 Statewide Crop Mapping GIS Map Service*. 2020. URL: https://gis.water.ca.gov/arcgis/rest/services/Planning/i15_Crop_Mapping_2018/MapServer.

[137] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[138] Jessica J. Williams and Tim Newbold. "Local climatic changes affect biodiversity responses to land use: A review". en. In: *Diversity and Distributions* 26.1 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.12999, pp. 76–92. ISSN: 1472-4642. DOI: 10.1111/ddi.12999. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.12999 (visited on 01/24/2023).

[139] Jessica J. Williams and Tim Newbold. "Vertebrate responses to human land use are influenced by their proximity to climatic tolerance limits". en. In: *Diversity and Distributions* 27.7 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.13282, pp. 1308–1323. ISSN: 1472-4642. DOI: 10.1111/ddi.13282. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13282 (visited on 01/24/2023).

[140] Stephen E Williams et al. "Towards an Integrated Framework for Assessing the Vulnerability of Species to Climate Change". en. In: *PLoS Biology* 6.12 (Dec. 2008). Ed. by Craig Moritz, e325. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0060325. URL: https://dx.plos.org/10.1371/journal.pbio.0060325 (visited on 01/23/2020).

[141] Charles B. Yackulic et al. "A need for speed in Bayesian population models: a practical guide to marginalizing and recovering discrete latent states". en. In: *Ecological Applications* 30.5 (2020). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/eap.2112, e02112. ISSN: 1939-5582. DOI: 10.1002/eap.2112. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/eap.2112 (visited on 02/27/2023).

[142] Charles B. Yackulic et al. "To predict the niche, model colonization and extinction". en. In: *Ecology* 96.1 (2015). _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/14-1361.1, pp. 16–23. ISSN: 1939-9170. DOI: 10.1890/14-1361.1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1890/14-1361.1 (visited on 03/09/2023).

[143] Casey Youngflesh. “MCMCvis: Tools to visualize, manipulate, and summarize MCMC output”. In: *Journal of Open Source Software* 3.24 (2018), p. 640. DOI: `10.21105/joss.00640`.