Multi-state occupancy modeling and optimal allocation of survey resources for Common Loons in Washington State

Hannah Sipe

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2019

Reading Committee:

Sarah J. Converse, Chair

Beth Gardner

Aaron Wirsing

Program Authorized to Offer Degree:

Quantitative Ecology and Resource Management

University of Washington

Abstract

Multi-state occupancy modeling and optimal allocation of survey resources for Common Loons in Washington State

Hannah Sipe

Chair of the Supervisory Committee:
Sarah J. Converse
Unit Leader, USGS Washington Cooperative Fish and Wildlife Research Unit
Associate Professor, School of Environmental and Forest Sciences (SEFS) & School of Aquatic and Fishery Sciences (SAFS), University of Washington

Common Loons (*Gavia immer*) are a state listed sensitive species in Washington State; however, little is known about the distribution of the Common Loon or the habitat associations of this species. This is complicated by the limited resources available for management to monitor Common Loons in Washington. In Chapter 1, I develop a novel multi-state occupancy model to integrate citizen science eBird data with monitoring data. This framework was then applied to Common Loons in Washington State. Occupancy probabilities were influenced by level of human disturbance and physical lake characteristics. There was temporal autocorrelation in

reproduction, such that reproduction at a site in a given year was positively associated with reproduction in the previous year. For eBird observers, detection of Common Loons at sites with reproduction was negatively related to site area and distance travelled during an observation bout, and positively related to time spent surveying. Washington Department of Fish and Wildlife observers were more likely to detect a Common Loon at sites with and without reproduction than were eBird observers. My results provide a better understanding of the distribution and breeding habitat requirements of Common Loons.

In Chapter 2, I develop a framework to identify a study design that will optimally allocate limited survey resources to maximize the information gained from occupancy analyses. Placing study design within the broader framework of optimal decision making provides a structured and transparent approach to optimal survey design, while Bayesian analytical methods provide the opportunity to leverage Bayesian updating in the evaluation of candidate survey designs. The Common Loon is monitored by the Washington Department of Fish and Wildlife, but there are limited personnel hours available to conduct surveys each season. I formulated optimal survey design for the Common Loon in Washington as a resource allocation problem. Alternative designs were built through application of alternative decision rules wherein sites were selected based on various estimates from an initial occupancy analysis. The decision rule that minimized the predicted state-wide mean uncertainty in occupancy probability selected sites based on the site-specific uncertainty in covariate relationships.

Together, the frameworks developed provide methods for utilizing multiple data sources and identifying a decision rule that can be used for survey design planning. I also demonstrate that a framework for including citizen science data with traditional monitoring data can lead to an expanded scope of inference in understanding the ecology and conservation needs of species.

This optimal design framework is applicable to occupancy models generally, and provides a quantitative assessment of the outcome of potential survey designs with respect to a given monitoring objective.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my funding sources, including the QERM Program at UW, the Washington Department of Fish and Wildlife, and the Washington Cooperative Fish and Wildlife Research Unit. The support of these organizations was vital to me pursuing and completing my thesis, to which I am grateful. Moreover, I would like to express my gratitude to Gretchen Blatz, Steve Desimone, Ilai Keren, Scott Pearson, and the entire Washington Department of Fish and Wildlife personnel who played a huge part in this thesis, providing invaluable knowledge, feedback, and encouragement. In addition, I would like to thank the Washington Department of Fish and Wildlife biologists who collected the data used in this project. I would like to thank the entire QERM program for the constant sense of community and motivation, especially Tim Essington and Erica Owens, along with my cohort John Best, Robert Emmet, Lillian McGill, and Elizabeth Ng. A special thanks to the Quantitative Conservation Lab members and Verna Blackhurst, all of who are among the best humans on the planet, providing me with inspiration and a warm working environment throughout this thesis. I would like to express my gratitude to my committee, Beth Gardner and Aaron Wirsing, who provided superb guidance for analysis, feedback on my thesis, and acted as allies throughout this project. Finally, to my advisor, Sarah J. Converse, thank you for your mentorship, leadership, support, and always offering an exemplary example of how to be a scientist. I have grown tremendously throughout this process and I attribute this growth to your teaching and guidance.

Chapter 1: Multi-state occupancy modeling using citizen science and survey data

1.1 Abstract

Monitoring a species to better understand its status, ecology, or management needs may be expensive or difficult. Citizen science data have the potential to expand our understanding for minimal cost but require development of appropriate analytical frameworks. I expanded on existing approaches for integrating data from the eBird citizen science platform with monitoring data. I developed a novel multi-state occupancy model to account for the structural differences in eBird data and agency-collected monitoring data. This framework was applied to Common Loons (*Gavia immer*) in Washington State. Occupancy probabilities were influenced by level of human disturbance and physical lake characteristics. There was temporal autocorrelation in reproduction, such that reproduction at a site in a given year was positively associated with reproduction in the previous year. For eBird observers, detection of Common Loons at sites with reproduction was negatively related to site area and distance travelled during an observation bout, and positively related to time spent surveying. Washington Department of Fish and Wildlife observers were more likely to detect a Common Loon at sites with and without reproduction than were eBird observers. My results provide a better understanding of the distribution and breeding habitat requirements of Common Loons. I also demonstrate that a framework for including citizen science data with traditional monitoring data can lead to an expanded scope of inference in understanding the ecology and conservation needs of species.

1.2 Introduction

Management agencies monitor species of interest to better understand their status, ecology, and management needs (Nichols and Williams, 2006). In many situations species can be prohibitively expensive or difficult to monitor, resulting in a lack of statistical power in the data (Field et al., 2005). When increasing monitoring effort is not an option due to cost or availability of personnel, auxiliary data sources can contribute to achieving monitoring goals (Hochachka et al., 2012). New data streams are available, such as citizen science data, that can substantially expand the information available to improve our understanding of species ecology and approaches to conservation (Hochachka et al., 2012; Sullivan et al., 2009). However, utilization of citizen science data requires development of appropriate frameworks in order to deal with the challenges inherent in these data (Hochachka et al., 2012; Kosmala et al., 2016). Developing novel frameworks that account for the difficulties and uncertainties in citizen science data has the potential to help conservation agencies make more efficient use of scarce monitoring resources.

Citizen science programs seek to engage the public in the scientific process and to enlist the public in collecting data that can be used by the scientific community, with varied objectives and protocols depending on the program (Bonney et al., 2014). Involving the public in data collection is not new; for example, the Breeding Bird Survey (BBS, established in 1966) and the Christmas Bird Count (CBC, established in 1900) continue to engage the public in data collection for birds (Dickinson et al., 2010). Increasing technological advances have allowed the public to participate in data collection on a wider scale (Dickinson et al., 2010; Kosmala et al., 2016). The variety of systems on which citizen science programs collect information is broad (Pocock et al., 2017). Citizen science programs range from structured (with defined locations,

objectives and instruction by a professional) to unstructured (observations are opportunistic and decided by the user) (Brown and Williams, 2019; Tulloch et al., 2013a).

By far the most attractive feature of citizen science data is the ability to capture information across spatial and temporal scales that would be nearly impossible by traditional means (Bonney et al., 2014; Dickinson et al., 2010; Kelling et al., 2015). In ecology, volunteers have contributed to datasets covering vast geographic and temporal scales and large numbers of species, allowing scientists to answer a wide range of ecological questions (Pocock et al., 2017). Monitoring data collected over expansive spatiotemporal scales can provide information about threats or changes in the system that would not be apparent from local-scale studies. Programs with online portals (e.g. eBird, iNaturalist, and eMammal) allow volunteers to submit data in real time, allowing for rapid transfer of data from collector to user (Dickinson et al., 2010). Citizen science programs also provide benefits associated with volunteers learning about and becoming invested in the study system. For species or systems of conservation concern, citizen science platforms allow the public to take a stake in conservation management (Dickinson et al., 2012). Whereas the vast quantities of data and public engagement in ecological research make citizen science attractive, citizen science data do have challenges in practice when applied to answering ecological questions.

One of the dominant citizen science programs in ecology today is eBird (ebird.org). In 2002, the Cornell Lab of Ornithology and the National Audubon Society launched this public observation platform for bird observations. Standardized protocols are used by eBirders to record observations of birds, and observations are subjected to quality control measures before inclusion in a database that can be accessed for research purposes. Users have the option to report their observations as complete checklists, indicating that all the species observed were recorded.

Complete checklists implicitly contribute ‘non-detection’ data that can be exploited along with submitted effort information to model the observation process. eBird data will often provide a larger spatial and temporal scale than survey data collected by researchers or management agencies because of the large number of observers using eBird. However, challenges associated with eBird data include low and variable detectability due to variation in observer experience, heterogeneous distribution of effort across landscapes, and species mis-identification (Sullivan et al., 2014, 2009). In order to take advantage of this data source, it is important to employ appropriate analytical techniques. Understanding the process used to collect the data is important in determining which statistical methods to use in analyzing the data and understanding the inferential strength of results (Altwegg and Nichols, 2019).

Occupancy modeling uses detection/non-detection data to estimate the proportion of sites in an area that is occupied by a species of interest (MacKenzie et al., 2002). Occupancy models rely on repeated surveys within a period during which the occupancy state of a site can be assumed static; i.e., the population is assumed not to be undergoing demographic or geographic change. The flexible nature of occupancy models as initially formulated, as well as subsequent extensions, allow for multiple seasons with dynamics occurring between seasons (MacKenzie et al., 2003; Royle and Kéry, 2007), missing observations, variability in detection over space and time, use of multiple data sources (Clare et al., 2017; Pacifici et al., 2017), and multiple occupancy states (MacKenzie et al., 2009; Nichols et al., 2007; Royle and Link, 2005). Multi-state occupancy models allow for finer resolution on the state of a site. For example, multi-state occupancy models could be used to distinguish whether a site is occupied by breeding versus non-breeding individuals (Kéry and Schaub, 2012). The flexibility and relative ease of data

collection for occupancy modeling have led to its wide adoption for monitoring changes in species distribution or trends over time (MacKenzie and Royle, 2005).

Citizen science data have potential to contribute substantially to large-scale occupancy analyses, but accounting for potential sources of bias is crucial. Isaac et al. (2014) simulated various forms of sampling bias to determine which types of trend estimation models are robust to biases in opportunistic data; they found that occupancy models are relatively robust because they explicitly model the detection process. Occupancy-based trends appear to be robust to missing zeros (unrecorded non-detections) in many cases, but missing zeros do yield overestimates of both detection and occupancy probabilities. The variation in effort found in opportunistic data can be accounted for by including relevant covariates on detection parameters (Kéry et al., 2010). Similarly, van Strien et al. (2013) compared trends between opportunistic citizen science data and data from designed surveys using occupancy models, and showed that through use of appropriate covariates to account for variation in detection in opportunistic data, the results were similar to results based on designed surveys (van Strien et al., 2013). Hence, through use of appropriate analytical techniques to accurately represent the data collection process, robust occupancy estimates can be obtained from citizen science data.

While citizen science data present challenges, more traditional (e.g., agency-led) monitoring programs may also present difficulties for analysts. First, monitoring design is often a trade-off between cost and statistical power (Field et al., 2005) and so may be hindered by a shortage of resources. Similarly, agency-led monitoring programs may also be opportunistic, and when there is no formal survey design, these programs will fall victim to similar biases as citizen science data (Yoccoz et al., 2001). Spatial variability may arise from the fact that it is often not feasible for agency personnel to survey entire areas due to cost, resulting in data that present

complications when the goal is to make inference over a large area (Nichols and Williams, 2006). Integrating citizen science data with more traditional types of monitoring data can provide benefits by increasing the spatial and temporal range of data. When the goal is to make inference to a large area, citizen science data may be particularly useful (Dickinson et al., 2010). Pacifici et al. (2017) merged citizen science and designed survey data in an occupancy modeling framework. They considered various parameterizations of occupancy models with both datasets included and compared them to a model without citizen science data. They found that by considering the types of bias that influence citizen science data and the observation processes that produce these data, the two sources can be usefully integrated, such that models including citizen data sources may perform better than models exploiting only designed survey data.

The Common Loon (*Gavia immer*) is a state listed Sensitive species in Washington. Common Loons are waterbody-obligate birds, found primarily in North America. Breeding adults spend summers on freshwater territories, either occupying an entire lake or sharing larger waterbodies (Evers, 2004) in the Northern United States, Canada, Greenland, and Iceland. Winters are spent in marine waters along the coast of North America. Common Loons reach reproductive maturity at approximately 5 years of age, but may not breed until 7 to 11 years of age (McIntyre, 1998). In Washington state, the Washington Department of Fish and Wildlife (WDFW) has been monitoring known summer breeding sites of Common Loons for approximately 40 years. These efforts are limited by available resources and the secluded breeding habitat of Common Loons. Although monitoring has been conducted, the distribution of breeding and non-breeding Common Loons in Washington state during the breeding season is not well understood and has not been formally analyzed (Richardson et al., 2000).

Single state, dynamic occupancy models have been used to understand Common Loon breeding distribution in northern Michigan (Field and Gehring, 2015) and northwestern Montana (Hammond et al., 2012). Factors influencing Common Loon occupancy were different in each analysis. In both Montana and Michigan, physical lake characteristics, intraspecific variables, and human disturbance were considered. In Montana, intraspecific covariates were the best predictors of occupancy (Hammond et al., 2012). Physical lake attributes, specifically lake area and presence of an island, were found to best explain breeding season occupancy in Michigan (Field and Gehring, 2015). Nest success models were also applied to Common Loons in Michigan, where physical lake features were found to be dominant predictors of nest success (Field and Gehring, 2015). Selection of breeding sites by Common Loons was analyzed in New Hampshire using habitat suitability models (Kuhn et al., 2011). Results suggested that Common Loons selected breeding sites based on physical lake features, intraspecific factors, and human disturbance (Kuhn et al., 2011).

Here, I expand on previous approaches for integrating citizen science data with monitoring data in an occupancy modeling framework to estimate the distribution of breeding and non-breeding Common Loons in Washington State. I formulated a multistate occupancy model that accounts for the structural differences in detection/non-detection eBird data and multistate occupancy data collected by the WDFW. I applied this framework to investigate Common Loon occupancy, reproduction, and habitat associations in potential breeding sites throughout Washington.

1.3 METHODS

1.3.1 DEFINITION OF SITE AND SEASON

Consultation with WDFW biologists and a review of the literature on habitat associations of Common Loons were used to define sites for Common Loons, and to determine the months when they could be assumed available on breeding territories. I defined sites (for the purpose of use in an occupancy model) as all fresh waterbodies in Washington with an area greater than 15 acres and below 5000 ft in elevation. Rivers were divided into sections based on county lines. This resulted in 2324 sites. I assumed that Common Loons would be available for detection on their breeding territories in June, July, and August, and that sites would be closed to geographic and demographic changes within these months (see Evers, 2007; Kuhn et al., 2011; McIntyre, 1998; Richardson et al., 2000).

1.3.2 DATA

1.3.2.1 STATE SURVEY DATA

Common Loons are monitored during the summer months by WDFW district biologists. Surveys provide information on detection/non-detection of adults and whether evidence of reproduction is observed if detection occurs. Evidence of reproduction is gleaned from more detailed observation data based on the detection of adults, young, and nest condition (i.e., empty, occupied, in disrepair). The sites that the WDFW surveys are secluded lakes or reservoirs (not free-flowing rivers), some of which are on access-restricted land. Of the available WDFW data, only data collected in 2017 were used in the occupancy model. Unlike other years, these data were collected following a designed survey, wherein surveyed sites were chosen randomly from a set of candidate sites. A total of 53 sites were surveyed by WDFW biologists in 2017.

1.3.2.2 eBird Filtering

All complete checklists within Washington state from June-August during 2000-2017, both with and without Common Loon detections, were obtained from the eBird website (eBird Basic Dataset, 2017; eBird Sampling Dataset, 2017). These data include a georeferenced point location, distance travelled (for travelling protocol types only), area covered (for area protocol types only), date and time of observation, elapsed time of observation, and whether the observer recorded all species observed (i.e., whether the list constitutes a complete checklist). Any lists that did not follow the protocol types 'travelling', 'area', or 'stationary' were removed. eBird data allow users to share checklists among groups and identify them with a 'group identifier'. One of the group checklists (the list that was submitted first) was retained and the others removed, along with any checklists that had exactly matching information. I used the R package auk ('auk: eBird Data Extraction and Processing with AWK', Strimas-Mackey et al., 2018) to facilitate filtering, merging, and zero-filling the non-detection events into a useable format for analysis.

To remove those observations that did not take place near a site, the georeferenced point location associated with a complete checklist observation was mapped in ArcMap (ESRI, 2011). Shape files for waterbodies in Washington State were downloaded from the National Hydrography Database (NHD, U.S. Geological Survey, n.d.). Each observation point was buffered by a circle with radius equal to the distance travelled or equal to the area covered information from the eBird dataset. The stationary observations were given a 20-m buffer; i.e., we assumed that any stationary observer within 20 m of a lake would have a non-zero probability of detecting a Common Loon if it was present. Lists with buffers that intersected marine waters were removed. Any observation with a buffer that did not intersect a fresh waterbody was

removed. The dataset was further filtered based on the criteria for a site, as described above, using the elevation shapefile (U.S. Geological Survey, n.d.) and the area of each waterbody from the NHD shapefile. eBird observations that intersected a site were assigned to that site, and if an observation intersected multiple sites, the observation was assigned to the site that was closest to the georeferenced location of the observation. I assumed that, if a Common Loon occupied a given site, there would be a non-zero probability of detection on an eBird list assigned to that site.

Of the 2324 sites that met the minimum criteria for a Common Loon site, 740 sites had eBird observations. Combining the sites with eBird data and the sites with WDFW data resulted in 766 total sites; 713 sites had only eBird observations, 26 sites had only WDFW observations, and 27 sites had both. All observations were assumed to be independent between and within datasets. WDFW observations can provide information on the breeding state of a site (i.e., occupied by breeding loons, occupied by non-breeding loons), whereas eBird observers only record detection/non-detection data with no information on the reproductive state of the site.

1.3.3 Model

1.3.3.1 General Framework of Multi-state Occupancy Models

Sites, described above, were the sampling unit. There was a total of N sites, and $i = 1, \ldots, N$ indexes the sites. The surveys at each site are indexed by $j = 1, \ldots, M$ surveys per year (where M varies by site and year) over a total of $t = 1, \ldots, T$ years. Following the multi-state Bayesian hierarchical structure of Royle and Link (2005), I introduce a latent random variable $z_{i,t}$ describing the true state for site i in year t, which is partially observed and takes on values of $s = 0,1,2$. The possible states s are interpreted as 0: unoccupied, 1: occupied by non-breeding

Common Loons, and 2: occupied by breeding Common Loons. The true latent occupancy state is modeled as a random variable drawn from a categorical distribution, $z_{i,t} \sim categorical(\Phi_{i,t}^{s})$. The multinomial probability of being in each state can be expressed through a state probability vector for all sites, *i*, and years, *t*:

$$\Phi_{i,t} = \begin{bmatrix} (1 - \Psi_{i,t}) & \Psi_{i,t}(1 - R_{i,t}) & \Psi_{i,t}R_{i,t} \end{bmatrix} \quad 1.1$$

where $\Psi_{i,t}$ is the probability that site *i* is occupied in year *t,* and $R_{i,t}$ is the probability that reproduction occurred at site i in year *t*, given that it was occupied. The parameters Ψ and R can be modeled as functions of site-specific covariates to evaluate the influence of covariates on occupancy and reproduction probabilities (MacKenzie et al., 2009). Note that different years at a given site are modeled as repeated measures; i.e., with a random effect over sites. I also include annual random effects to account for variation over years; i.e.,

$$logit(\Psi_{i,t}) = \alpha_{\Psi} + \boldsymbol{X}\beta_{\Psi} + \varepsilon_{\Psi,i} + \varepsilon_{\Psi,t} \quad 1.2$$

$$with \ \varepsilon_{\Psi,i} \sim Normal(0, \sigma^{\Psi,i})$$

$$and \ \varepsilon_{\Psi,t} \sim Normal(0, \sigma^{\Psi,t})$$

where α_{Ψ} is the intercept for $\Psi_{i,t}$, $\boldsymbol{X}$ is a matrix of site-specific covariates, β_{Ψ} is a vector of coefficient parameters for $\Psi_{i,t}$, $\varepsilon_{\Psi,i}$ is the random site-effect with variance of $\sigma^{\Psi,i}$, and $\varepsilon_{\Psi,t}$ is the random year effect with variance of $\sigma^{\Psi,t}$. An equivalent parameterization was developed for $R_{i,t}$ (see Eq. 1.9). Whereas it is possible to model dynamics in the occupancy process to account for variation over years (see MacKenzie et al., 2009), adequate data to properly employ the dynamic form were not available, so instead I included the random site effect to account for the repeated measures.

The observation data are represented by the random variable, $y_{i,t,j}$ where the observation for site i in year t on survey j can take values of $y = 0,1,2$ for the WDFW data and values of $y = 0,1$ for the eBird data (i.e., information on reproductive state is not captured by eBird). The observations, $y_{i,t,j}$ can be modeled using a categorical distribution conditioned on the true state; i.e. $y_{i,t,j}|z_{i,t} \sim categorical(\Upsilon^{[s,r]}_{i,t,j,k})$. The observation process is represented with the observation matrix, $\Upsilon^{[s,r]}_{i,j,t,k}$, containing the probabilities, given the true state is *s*, of recording observation $y = r$ at site i in year *t* on survey *j* for data type k. The index *k* takes values of *k* = 1 or 2 (eBird data or WDFW data, respectively). The observation matrix varies by data type *k*. For eBird observations (*k* = 1):

$$\Upsilon^{[s,r]}_{i,t,j,1} = \textit{True State} \quad \begin{array}{c} \\ \textbf{0} \\ \textbf{1} \\ \textbf{2} \end{array} \begin{array}{c} \textit{Observation} \\ \textbf{0} \quad\quad\quad \textbf{1} \\ 1 \quad\quad\quad 0 \\ (1 - p^1_{i,t,j,1}) \quad p^1_{i,t,j,1} \\ (1 - p^2_{i,t,j,1}) \quad p^2_{i,t,j,1} \end{array} \tag{1.3}$$

where row *s* represents the true state (unoccupied, occupied by a nonbreeding loon, or occupied by a breeding loon) and column *r* represents the observations (loon observed, or loon not observed). The cells in the matrix represent the probabilities for the respective combination of true state and observation. There are separate detection probabilities for sites with non-breeding loon(s) and sites with breeding loons; i.e., I modeled the probability of an eBird observer, at site *i* in year *t* on visit *j*, observing a Common Loon at a site in state *s*, given by $p^s_{i,t,j,1}$. Note that for eBird data, visit *j* can also be thought of as list *j*, where there are *M* total lists that may have included observations at that site in that year.

For WDFW data, the observation matrix is:

$$\Upsilon_{i,t,j,2}^{[s,r]} = \textit{True State} \quad \begin{array}{c} \\ \textbf{0} \\ \textbf{1} \\ \textbf{2} \end{array} \begin{array}{c} \textit{Observation} \\ \textbf{0} \quad\quad\quad \textbf{1} \quad\quad\quad \textbf{2} \\ 1 \quad\quad\quad 0 \quad\quad\quad 0 \\ (1 - p_{i,t,j,2}^{1}) \quad\quad p_{i,t,j,2}^{1} \quad\quad 0 \\ (1 - p_{i,t,j,2}^{2}) \quad\quad p_{i,t,j,2}^{2}(1-\delta) \quad\quad p_{i,t,j,2}^{2}\delta \end{array} \tag{1.4}$$

where, similar to Eq. 1.3, $p_{i,t,j,2}^{s}$ is the probability of a WDFW biologist, at site *i* in year *t* on visit *j,* observing a Common Loon in state *s*. For WDFW data, there is also a classification probability. This is the probability, given by δ, of observing evidence of reproduction, and therefore being able to definitively classify a site as being in state *s = 2*, given that there was a Common Loon observed at a site in state *s* = 2. This probability is modeled as a constant in this application but could be modeled as a function of survey or site-specific covariates.

The conditional detection probabilities (i.e., $p_{i,t,j,1}^{s}$ and $p_{i,t,j,2}^{s}$) for WDFW and eBird were modeled as functions of survey- or site-specific covariates. For state *s* = 1 or *s* = 2, the probabilities were constrained to have the same intercept and included an additional term for WDFW data type. For example, the detection probabilities of observing a Common Loon in state *s* are:

$$logit\left(p_{i,t,j,1}^{s}\right) = \alpha_{p1s} + \boldsymbol{Y}\beta_{ps1} \tag{1.5}$$

$$logit\left(p_{i,t,j,2}^{s}\right) = \alpha_{p1s} + \lambda_{s} + \boldsymbol{Z}\beta_{ps2} \tag{1.6}$$

where α_{ps} is the intercept for $p_{i,t,j,1}^{s}$ and $p_{i,t,j,2}^{s}$, λ_{s} is the term for WDFW data type, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are the data-type specific covariates for each data type, β_{ps1} and β_{ps2} are the coefficient vectors for each data type.

1.3.3.2 TEMPORAL AUTOCORRELATION

While I did not model occupancy dynamics, I recognized that the probability of occupancy in one year may be related to the probability in previous years, even after habitat relationships are accounted for. This process can be modelled through temporal autocorrelation in $\Psi_{i,t}$ and $\mathrm{R}_{i,t}$ by allowing the current year's probability of occupancy or reproduction to be dependent on the previous year's occupancy state (i.e. $z_{i,t-1}{\sim}categorical(\Phi_{i,t-1})$ as in MacKenzie et al. (2018). For example, $\Psi_{i,t}$ can be extended from Eq. 2 to include a temporal autologistic term for $t \geq 2$:

$$logit(\Psi_{i,t}) = \alpha_{\Psi} + \boldsymbol{X}\beta_{\Psi} + \varepsilon_{\Psi,i} + \varepsilon_{\Psi,t} + \beta_{\Psi,\mathrm{m}} * m_{i,t} \quad 1.7$$

$$m_{i,t} = \begin{cases} 0 \; if \; z_{l,t-1} = 0 \\ 1 \; if \; z_{l,t-1} \geq 1 \end{cases} \quad 1.8$$

where the term $m_{i,t}$ is the occupancy state of the focal site in the previous year. Similarly, for $t \geq 2$, $\mathrm{R}_{i,t}$ can be modeled as a function of site-specific covariates and a temporal autocorrelation effect:

$$logit(\mathrm{R}_{i,t}) = \alpha_{R} + \boldsymbol{X}\beta_{R} + \varepsilon_{\mathrm{R},i} + \varepsilon_{\mathrm{R},t} + \beta_{R,g} * g_{i,t} \quad 1.9$$

$$g_{i,t} = \begin{cases} 0 \; if \; z_{l,t-1} = 0 \\ 1 \; if \; z_{l,t-1} \geq 2 \end{cases} \quad 1.10$$

α_{R} is the intercept for $\mathrm{R}_{i,t}$, $\boldsymbol{X}$ is a matrix of site-specific covariates, β_{R} is a vector of coefficient parameters for $\mathrm{R}_{i,t}$, $\varepsilon_{\mathrm{R},i}$ is the random site-effect with variance of $\sigma^{\mathrm{R},i}$, $\varepsilon_{\mathrm{R},t}$ is the random year effect with variance of $\sigma^{\mathrm{R},t}$, and the term $g_{i,t}$ is the reproductive state of the focal site in the previous year.

1.3.4 Covariate Information

The site-specific covariates considered for modeling occupancy and reproduction probabilities were waterbody area (km2; NHD dataset), waterbody perimeter (km; NHD dataset), perimeter complexity ($\sqrt{\text{area}}/\text{perimeter}$), elevation (m), Human Influence Index (HII; Wildlife Conservation Society, 2005), and tree canopy cover (Yang et al., 2018). Land cover type was included as a categorical variable, with 4 categories: (1) open water and wetlands (including woody and emergent herbaceous wetlands), (2) cultivated and developed land (including hay, pasture, crops, low, medium and high development), (3) forest (including deciduous, evergreen, and mixed), and (4) an 'other' category (including barren land, scrub/shrub land, and herbaceous land, Yang et al., 2018).

The survey-specific detection covariates for eBird observations included start time of list, duration of list, and distance travelled/area covered (km). The WDFW observations included survey-specific records of the duration of time spent surveying and the start time of the observation.

1.3.5 Model Implementation

Models were written using NIMBLE, a Bayesian hierarchical interface that compiles and runs models through C++ (de Valpine et al., 2017). Data were processed and NIMBLE was accessed through R (R Core Team, 2018). I employed penalized complexity priors on all model coefficients which allowed me to make inference about the magnitude of each covariate's effect (see Simpson et al., 2017). The full model was fit with priors on coefficients; i.e. $\beta_j \sim Normal(0, \sigma_j)$ where $\sigma_j \sim Exponential(1)$. Priors on intercepts (i.e. α_{Ψ}, α_{R}, α_{p1}, α_{p2}) were specified as $Normal(0{,}10)$. The parameter δ was given a prior of $Beta(1{,}1)$, the priors for the

random effects of $\Psi_{i,t}$ (i.e. $\sigma^{\Psi,\mathrm{i}}$ and $\sigma^{\Psi,t}$) and $\mathrm{R}_{i,t}$ (i.e. $\sigma^{\mathrm{R,i}}$ and $\sigma^{\mathrm{R},t}$) were $Uniform(0,10)$. The temporal autocorrelation structure described in Eqs. 1.9 and 1.11 were used to model $\Psi_{i,t}$ and $\mathrm{R}_{i,t}$.

The full model was fit in NIMBLE with 4 chains, 125,000 iterations and 50,000 burn-in per chain. I thinned the samples at a rate of 10 to reduce file size. The Gelman-Rubin convergence diagnostic was used to ensure chains had converged. The code for the model is provided in Appendix 1. I interpreted covariates as influential when their coefficients had 95% credible intervals (CIs) that did not overlap 0.

1.4 Results

The number of sites with detections of Common Loons by eBirders for all years (2000-2017) was 227. The number of sites with detections differed (and generally increased) across years, from a minimum of 0 sites in 2001 to a maximum of 33 sites in 2017. In 2017, WDFW detected Common Loons without evidence of breeding at 12 sites and with evidence of breeding at 14 sites, of 53 sites surveyed. The number of sites with a Common Loon detection (with or without breeding) by both eBird and WDFW for all years was 245. The model estimated that the mean occupancy (Ψ) and reproduction (R) probabilities for all sites included in the model (n = 766) over all years was 0.390 (95% CI = 0.129, 0.922) and 0.105 (95% CI = 0.011, 0.333) respectively (Figure 1.1). Occupancy was predicted to be highest at sites located in northeastern Washington, with high occupancy also along rivers east of the Cascade Mountains (Figure 1.2). Reproduction was predicted to be highest at sites located in northeastern Washington (Figure 1.3).

Inference can be made to all sites (as defined above, n = 2324) using site-specific covariate information and model parameter estimates. This results in a state-wide occupancy probability estimate of mean = 0.365 (95% CI = 0.163, 0.839) and a state-wide reproduction probability estimate of mean 0.129 (95% CI = 0.007, 0.410; Figure 1.4).

All coefficients reported are on the logit scale. The fixed effect of HII had a negative effect on the probability of occupancy ($\Psi_{i,t}$) with a mean posterior estimate of -1.002 (95% CI = -1.734, -0.336), as did elevation with a mean posterior estimate of -1.171 (95% CI = -2.344, -0.231). All other covariates were estimated to have a minimal effect on occupancy probability (Table 1.1, Figure 1.5, Figure 1.6). The mean estimates of SD for the random year and site effects were $\sigma^{\Psi,t}$ = 0.358 (95% CI = 0.013, 1.367) and $\sigma^{\Psi,i}$ = 3.465 (95% CI = 2.212, 4.943; Table 1.1).

The estimated temporal autocorrelation effect on reproduction was positive and strong, with a mean of 25.539 (95% CI = 7.761, 87.913). All other covariates were estimated to have a minimal effect on reproduction probability (Table 1.2, Figure 1.7, Figure 1.8). The mean estimates of SD for the random year and site effects were $\sigma^{R,t}$ = 8.120 (95% CI = 4.926, 9.933) and $\sigma^{R,i}$ = 1.099 (95% CI = 0.179, 2.871; Table 1.2).

Detection by eBird observers (i.e., $p^{s}_{i,t,j,1}$) was negatively influenced by the area of a site for sites with reproduction (mean = -1.916, 95% CI = -3.743, -0.539). Detection by eBird observers at sites with reproduction was positively influenced by the duration of the survey (mean = 0.666, 95% CI = 0.041, 1.472). Other covariates were estimated to have a minimal effect on detection at sites with or without reproduction (Table 1.3, Table 1.4, Figure 1.9).

WDFW observers (i.e., $p^{s}_{i,t,j,2}$) were more likely than eBirders to detect Common Loons at sites without reproduction (mean = 3.111, 95% CI = 0.527, 6.583) and at sites with reproduction (mean = 3.166, 95% CI = 0.704, 5.313). The probability of a WDFW observer detecting evidence of reproduction at a site with reproduction, δ, was relatively high (mean = 0.766; 95% CI = 0.636, 0.875; Table 1.3, Table 1.4, and Figure 1.10).

1.5 Discussion

I developed a new model to combine eBird data with standardized survey data in order to gain insights into the breeding distribution and ecology of the Common Loon, a species of interest in the state of Washington. The distribution of breeding and non-breeding Common Loons in Washington was previously not well understood. Through using eBird data, I was able to include more years and sites, while maintaining greater ecological specificity by using agency-collected data (i.e., by differentiating sites that are occupied without breeding from those that are occupied with breeding). An understanding of where Common Loons are absent, or present with or without breeding, is key to making informed conservation decisions for this sensitive species. My results show that integrating citizen science data provided valuable information on the distribution of breeding and non-breeding Common Loons. However, there are a number of challenges that must be addressed when integrating structured and unstructured data sources. These challenges will vary depending on the type of citizen science data and the objectives of the analysis (Tulloch et al., 2013). While I focused on confronting challenges associated with integrating eBird data with agency-collected data in a multi-state occupancy framework, these challenges are similar to those that arise in many other situations using similar data sources.

The model identified site features that influenced Common Loon occupancy and reproduction probabilities. This result would not have been possible without eBird data, given

the small sample size of WDFW surveys. My results indicate that elevation negatively influences occupancy. Sites at lower elevation (i.e., near the coast or rivers) may be associated with higher occupancy of non-breeding Common Loons as they are generally closer to winter grounds (i.e., sites near the coast or on rivers), and require a shorter migration than sites at higher elevations. Common Loons' sensitivity to human disturbance was borne out in my analysis. Common Loons occupancy increased with decreasing human impact index, which likely indicates a combination of habitat degradation and human disturbance. Breeding Common Loons are known to have high between-year site fidelity (Evers, 2004; Richardson et al., 2000) and the model estimated that reproduction was much more likely at sites where reproduction occurred in the previous year. However, temporal autocorrelation was not supported for occupancy. My results are similar to those from other regions using single state occupancy models; specifically, I found that human disturbance, physical lake characteristics, and intraspecific factors were important for predicting locations occupied by Common Loons. One factor that may be quite influential is the presence or density of fishes in the lake, given that fish are the preferred food of Common Loons (Evers, 2004; McIntyre, 1998; Richardson et al., 2000). However, data on presence and density of fish are not available for Washington, and fish stocking activities result in a constantly changing landscape of fish-occupied lakes and reservoirs. Data on fish population size and characteristics have the potential to substantially improve the predictive capacity for Common Loon occupancy models.

While including eBird data increased the information available to model Common Loon occupancy in Washington, caution is still warranted when using these data. Pitfalls to watch for include lack of independence between observers, false positives, or unexplainable heterogeneity in detection probabilities (Altwegg and Nichols, 2019). Lack of independence will occur when

multiple observers in a group submit lists. eBird allows users to share information between members of a group, but records this sharing with a group identifier. During the data filtering process, I only included one list (the list submitted first) from groups, and also deleted one of two lists if the two lists were exactly identical, even if not coded as group lists. I assumed that the Common Loons were never falsely detected when absent. I was reasonably confident that eBird users could correctly identify the species. Common Loons are large, conspicuous birds with striking plumage and a distinct call. False positives are a greater concern when the focal species is less recognizable, but there is a range of potential solutions to this problem (McClintock et al., 2017; Miller et al., 2011; Ruiz-Gutierrez et al., 2016). The assumption that detection probabilities have no unmodeled heterogeneity (MacKenzie et al., 2018) is likely violated when using citizen science data (Altwegg and Nichols, 2019). Heterogeneity in citizen science data arises due to unequal sampling effort across sites and among users, variable expertise among users, and bias towards particular species (van Strien et al., 2013). Through only using complete eBird checklist observations, I sought to regulate the data for preferences toward particular species. Because effort information is included with eBird observations, I was able to include these measures as covariates on eBird detection parameters and capture the variability in observation effort among users. Kelling et al. (2015) described an approach for quantifying variability in expertise of eBirders. Johnston et al. (2018) used expertise scores as covariates on detection probability for an occupancy analysis using eBird data and found that expertise measures improved model fit. Including an expertise metric as a covariate would be possible in my framework as well.

eBird observations do not have exact location information (given that eBirders can be travelling during the submission of a checklist), resulting in uncertainty in where observations

were made. I assumed that an eBirder had a non-zero probability of detecting a Common Loon at the site closest to their recorded observation location, if that site was within the buffered distance defined by distance traveled or area searched information. This assumption may introduce heterogeneity into the occupancy model's observation process, as some of those observations may have not taken place near enough to a site to observe a Common Loon. However, through adapting the occupancy model's structure, specifically the observation process, it would be possible to account for this uncertainty. A process could be built to model the probability that an eBird observation truly occurred at a site, constituting a 'list inclusion probability.' For species such as Common Loons, with strong associations to discrete habitat features, this stochastic inclusion process could be informed by other species on the eBird checklist associated with those same habitat features (e.g., other waterbirds), thereby reducing the heterogeneity in detection brought about by including lists when there is uncertainty in the location surveyed. This approach would account for the uncertainty in the survey location. Building this feature into the occupancy model structure is currently in progress.

There are various methods for combining multiple data sources, and the method that is appropriate in a given application depends upon the goals of analysis, analysis type, and available data sources (Miller et al., 2019). Fletcher et al. (2019) identified common approaches for combining data sources, including (1) pooling data together, (2) combining independent models (i.e., independent models for each data type are built and predictions are combined or compared), (3) using one data type as auxiliary data or to create informative priors, and (4) data integration (models for each data type are combined to estimate model parameters). Methods for combining data vary in the degree to which they can account for sampling biases and uncertainties that are dataset-specific. For example, data pooling makes it more difficult to

account for unique issues with each data type, whereas data integration explicitly models each data collection process (Fletcher et al., 2019). I chose to formally integrate the data, developing distinct observation models for each data type, with the same underlying latent state process. My approach allowed me to explicitly account for the differences in data structure, collection protocols, and sampling issues.

The eBird survey detection parameter estimates, for detection at sites with reproduction, were positively influenced by the duration of the observation, indicating that detection was a function of increased time spent surveying. The detection of a Common Loon by eBirders was also higher at breeding sites with smaller areas. Sites with smaller areas will generally be easier to survey, as there is less area to cover. Detection by WDFW personnel, at sites with and without reproduction, was slightly higher during morning hours. Detection of Common Loons at sites with reproduction was marginally positively influenced by the duration of the observation for WDFW observers. Not surprisingly, the WDFW observers were better at detecting Common Loons overall. My findings about the observation process could inform future survey efforts; for example, future surveys would be most effective if conducted in the morning.

Unequal sampling effort across sites arises in eBird data, as users tend to visit sites with easy access, particular species of interest, or an overall high-quality birding experience (Hochachka and Fink, 2012). Roads, land cover, and human density have been shown to be important explanatory variables in resolving spatial bias in unstructured citizen science data (Geldmann et al., 2016) and I included these types of variables in the analysis. Johnston et al. (2019) recommends proper data filtering in conjunction with effort covariates when using eBird data in occupancy models. If the species of interest is rarely detected by eBirders, alternate filtering techniques may be considered to obtain balanced data and unbiased parameter estimates.

One technique is to under-sample non-detections while retaining all detections (Robinson et al., 2018). The benefit of this approach could be tested in this application by filtering eBird data for balance, refitting, and comparing parameter estimates. However, I expect that the generally broad coverage of data and the use of covariates have resulted in reasonably robust inference in the application to Common Loons in Washington.

While their quality can bring challenges, citizen science efforts do provide a lot of useful information for researchers and managers. As such, citizen science data can allow ecologists to explore spatial and temporal patterns that would otherwise be infeasible to investigate. Through the application to Common Loons in Washington state, I realized the potential for citizen science to expand the scope of ecological inference to gain a better understanding of species-habitat relationships across a large area. The exact methods employed here are specific to the Common Loon in Washington State, but the underlying ideas I present can be adapted for other applications. When traditional survey data are limited, citizen science data can lead to large gains in information. Citizen science data are not restricted to occupancy analysis but could be useful when applied to other monitoring targets (e.g., abundance), if appropriate analytical techniques are employed. The popularity of citizen science data has led to a large amount of data that should not be overlooked when information is needed to inform management or ecological understanding.

1.6 References

Altwegg, R., Nichols, J.D., 2019. Occupancy models for citizen-science data. Methods Ecol. Evol. 10, 8–21. https://doi.org/10.1111/2041-210X.13090

Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J., Parrish, J.K., 2014. Next steps for citizen science. Science (80-.). 343, 1436–1437. https://doi.org/10.1126/science.1251554

Brown, E.D., Williams, B.K., 2019. The potential for citizen science to produce reliable and useful information in ecology. Conserv. Biol. 33, 561–569. https://doi.org/10.1111/cobi.13223

Clare, J., McKinney, S.T., Depue, J.E., Loftin, C.S., 2017. Pairing field methods to improve inference in wildlife surveys while accommodating detection covariance: Ecol. Appl. 27, 2031–2047. https://doi.org/10.1002/eap.1587

de Valpine, P., Turek, D., Paciorek, C.J., Lang, D.T., Bodik, R., 2017. Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. J. Comput. Graph. Stat. 26(2), 403–413.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T., Purcell, K., 2012. The current state of citizen science as a tool for ecological research and public engagement. Front. Ecol. Environ. 10, 291–297. https://doi.org/10.1890/110236

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

eBird Basic Dataset, 2017. EBD_US-WA_relNov-2017. Cornell Lab Ornithol. Ithaca, New York.

eBird Sampling Dataset, 2017. EBD_sampling_relNov-2017. Cornell Lab Ornithol. Ithaca, New York.

ESRI, 2011. ArcGIS Desktop.

Evers, D.C., 2007. Evers, D. C. 2007. Status assessment and conservation plan for the Common Loon (Gavia immer) in North America: 2007. BRI Report 2007-20. U.S. Fish and Wildlife Service, Hadley, MA. BRI Rep. 2007-20. U.S. Fish Wildl. Serv. Hadley, MA.

Evers, D.C., 2004. Status assessment and conservation plan for the Common loon (Gavia immer) in North America. U.S. Fish Wildl. Serv. 1–87.

Field, M., Gehring, T.M., 2015. Physical, human disturbance, and regional social factors influencing Common Loon occupancy and reproductive success. Condor 117, 589–597. https://doi.org/10.1650/condor-14-195.1

Field, S.A., Tyre, A.J., Possingham, H.P., 2005. Optimizing Allocation of Monitoring Effort Under Economic and Observational Constraints. J. Wildl. Manage. 69, 473–482. https://doi.org/10.2193/0022-541x(2005)069[0473:oaomeu]2.0.co;2

Fletcher, R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A., Dorazio, R.M., 2019. A practical guide for combining data to model species distributions. Ecology 100, 0–

3. https://doi.org/10.1002/ecy.2710

Geldmann, J., Heilmann-Clausen, J., Holm, T.E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., Tøttrup, A.P., 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers. Distrib. 22, 1139–1149. https://doi.org/10.1111/ddi.12477

Hammond, C.A.M., Mitchell, M.S., Bissell, G.N., 2012. Territory occupancy by common loons in response to disturbance, habitat, and intraspecific relationships. J. Wildl. Manage. 76, 645–651. https://doi.org/10.1002/jwmg.298

Hochachka, W., Fink, D., 2012. Broad-scale citizen science data from checklists: prospects and challenges for macroecology. Front. Biogeogr. 4. https://doi.org/10.21425/f5fbg15350

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.K., Kelling, S., 2012. Data-intensive science applied to broad-scale citizen science. Trends Ecol. Evol. 27, 130–137. https://doi.org/10.1016/j.tree.2011.11.006

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

Johnston, A., Fink, D., Hochachka, W.M., Kelling, S., 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 9, 88–97. https://doi.org/10.1111/2041-210X.12838

Kelling, S., Fink, D., La Sorte, F.A., Johnston, A., Bruns, N.E., Hochachka, W.M., 2015. Taking a 'Big Data' approach to data quality in a citizen science project. Ambio 44, 601–611. https://doi.org/10.1007/s13280-015-0710-4

Kéry, M., Gardner, B., Monnerat, C., 2010. Predicting species distributions from checklist data using site-occupancy models. J. Biogeogr. 37, 1851–1862. https://doi.org/10.1111/j.1365-2699.2010.02345.x

Kéry, M., Schaub, M., 2012. Bayesian population analysis using WinBUGS: a hierarchical perspective, 1st ed. Boston: Academic Press.

Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Front. Ecol. Environ. 14, 551–560. https://doi.org/10.1002/fee.1436

Kuhn, A., Copeland, J., Cooley, J., Vogel, H., Taylor, K., Nacci, D., August, P., 2011. Modeling habitat associations for the Common Loon (Gavia immer) at multiple scales in Northeastern North America. Avian Conserv. Ecol. 6. https://doi.org/10.5751/ACE-00451-060104

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, A.B., 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84, 2200–2207. https://doi.org/10.1890/02-3090

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.A., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83, 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., Hines, J.E., 2018. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence,

2nd ed. Academic Press.

MacKenzie, D.I., Nichols, J.D., Seamans, M.E., Gutiérrez, R.J., Ecology, S., Mar, N., James, D., 2009. Modeling Species Occurrence Dynamics with Multiple States and Imperfect Detection. Ecology 90, 823–835.

MacKenzie, D.I., Royle, J.A., 2005. Designing occupancy studies: General advice and allocating survey effort. J. Appl. Ecol. 42, 1105–1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x

McClintock, B.T., Bailey, L.L., Pollock, K.H., Theodore, R., Mcclintock, B.T., Bailey, L.L., Pollock, K.H., Simons, T.R., 2017. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections Published by : Wiley on behalf of the Ecological Society of America Stable URL : http://www.jstor.org/stable/27860809 REFERENCES Link 91, 2446–2454.

McIntyre, J.W., 1998. The Common Loon: Spirit of Northern Lakes. Minneapolis: University of Minnesota Press.

Miller, D.A., Nicholas, J.D., T., M.B., Campbell Grant, E.H., Bailey, L.L., 2011. error occur : non-detection and species misidentification R eports R eports. Ecology 92, 1422–1428.

Miller, D.A.W., Pacifici, K., Sanderlin, J.S., Reich, B.J., 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods Ecol. Evol. 10, 22–37. https://doi.org/10.1111/2041-210X.13110

Nichols, J.D., Hines, J.E., Mackenzie, D.I., Seamans, M.E., Gutiérrez, R.J., 2007. Occupancy estimation and modeling with multiple states and state uncertainty. Ecology 88, 1395–1400. https://doi.org/10.1890/06-1474

Nichols, J.D., Williams, B.K., 2006. Monitoring for conservation. Trends Ecol. Evol. 21, 668–673. https://doi.org/10.1016/j.tree.2006.08.007

Pacifici, K., Reich, B.J., Miller, D.A.W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., Collazo, J.A., 2017. Integrating multiple data sources in species distribution modeling: A framework for data fusion. Ecology 98, 840–850. https://doi.org/10.1002/ecy.1710

Pocock, M.J.O., Tweddle, J.C., Savage, J., Robinson, L.D., Roy, H.E., 2017. The diversity and evolution of ecological and environmental citizen science. PLoS One 12, 1–17. https://doi.org/10.1371/journal.pone.0172579

R Core Team, 2018. R: A language and environment for statistical computing.

Richardson, S., Hays, D., Spencer, R., Stofel, J., 2000. Washington state status report for the common loon. Washingt. Dep. Fish Wildlife, Olympia. 53.

Robinson, O.J., Ruiz-Gutierrez, V., Fink, D., 2018. Correcting for bias in distribution modelling for rare species using citizen science data. Divers. Distrib. 24, 460–472. https://doi.org/10.1111/ddi.12698

Royle, J.A., Kéry, M., 2007. A Bayesian state-space formulation of dynamic occupancy models. Ecology 88, 1813–1823. https://doi.org/10.1890/06-0669.1

Royle, J.A., Link, W.A., 2005. A general class of multinomial mixture models for anuran calling survey data. Ecology 86, 2505–2512. https://doi.org/10.1890/04-1802

Ruiz-Gutierrez, V., Hooten, M.B., Campbell Grant, E.H., 2016. Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias. Methods Ecol. Evol. 7, 900–909. https://doi.org/10.1111/2041-210X.12542

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Stat. Sci. 32, 1–28. https://doi.org/10.1214/16-STS576

Strimas-Mackey, M., Miller, E., Hochachka, W., 2018. auk: eBird Data Extraction and Processing with AWK R package.

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K. V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W.K., Wood, C.L., Yu, J., Kelling, S., 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biol. Conserv. 169, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J., Martin, T.G., 2013a. Realising the full potential of citizen science monitoring programs. Biol. Conserv. 165, 128–138. https://doi.org/10.1016/j.biocon.2013.05.025

Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J., Martin, T.G., 2013b. Realising the full potential of citizen science monitoring programs. Biol. Conserv. 165, 128–138. https://doi.org/10.1016/j.biocon.2013.05.025

U.S. Geological Survey, n.d. National Hydrography Dataset (NHD).

U.S. Geological Survey, n.d. National Elevation Dataset (NED) accessed:

van Strien, A.J., van Swaay, C.A.M., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J. Appl. Ecol. 50, 1450–1458. https://doi.org/10.1111/1365-2664.12158

Wildlife Conservation Society, 2005. Global Human Influence Index (HII) Dataset. Last Wild Proj. Version 2. https://doi.org/https://doi.org/10.7927/H4BP00QC

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S.M., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Granneman, B., Liknes, G.C., Rigge, M., Xian, G., 2018. A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. ISPRS J. Photogramm. Remote Sens. 146, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time. Trends Ecol. Evol. 16, 446–453. https://doi.org/10.1016/S0169-5347(01)02205-4

1.7 FIGURES AND TABLES

Table 1.1 Posterior estimates of model coefficients and random effects parameters ($\sigma_{site,}$ σ_{year}) on occupancy probability of Common Loons in Washington, including mean, standard deviation (SD), and quantiles (2.5%, 25%, 50%, 75%, and 97.5%).

	Mean	SD	2.50%	25%	50%	75%	97.50%
Intercept[1]	-0.446	0.287	-1.006	-0.639	-0.448	-0.259	0.131
Focal[2]	-0.916	0.745	-2.506	-1.378	-0.852	-0.378	0.347
Area[3]	1.981	1.583	-0.188	0.820	1.645	2.820	5.779
Canopy[4]	-0.014	0.366	-0.726	-0.255	-0.017	0.216	0.740
Elevation[5]	-1.171	0.576	-2.344	-1.566	-1.158	-0.741	-0.131
HII[6]	-1.002	0.348	-1.734	-1.225	-0.995	-0.767	-0.336
Perimeter Complexity[7]	0.781	0.527	-0.230	0.431	0.762	1.113	1.890
Perimeter[8]	-0.408	0.921	-2.334	-0.975	-0.368	0.229	1.310
LC: Forest[9]	-0.967	1.137	-3.572	-1.593	-0.831	-0.211	0.948
LC: Developed[10]	-0.463	1.077	-2.707	-0.990	-0.386	0.181	1.401
LC: Other[11]	-0.937	0.852	-2.804	-1.450	-0.870	-0.359	0.583
σ_{year}	0.358	0.358	0.013	0.130	0.256	0.458	1.367
σ_{site}	3.465	0.715	2.212	2.955	3.416	3.931	4.943

1. The intercept term in the model for the occupancy parameter. This is the reference group for the categorical land cover data types, i.e. open water and wetland (includes woody and emergent herbaceous wetlands) land types.
2. Focal is the coefficient for the temporal autocorrelation effect on occupancy.
3. Area is the coefficient for the area of a site.
4. Canopy is the coefficient for the percent tree canopy cover at a site.
5. Elevation is the coefficient for the elevation of a site.
6. HII is the coefficient for the Human Influence Index at a site.
7. Perimeter complexity is the coefficient for the complexity of a site's perimeter.
8. Perimeter is the coefficient for the length of the perimeter of a site.
9. Forest is the coefficient for the land cover type forest, which includes evergreen, deciduous, and mixed forests. The model intercept estimates open water/wetland sites, such that the Forest coefficient is the difference between open water/wetland sites and forest.
10. Developed is the coefficient parameter for the land cover types of developed (including low, medium, and high levels of development) and cultivated (including hay, pasture, and crops). The model intercept estimates open water/wetland sites, such that the Developed coefficient is the difference between open water/wetland sites and developed/cultivated areas.
11. Other is the coefficient parameter for the land cover types barren land, scrub/shrub land, and herbaceous. The model intercept estimates open water/wetland sites, such that the Other coefficient is the difference between open water/wetland sites and these land types.

Table 1.2 Posterior estimates of model coefficients and random effects parameters ($\sigma_{site,}$ σ_{year}) on reproduction probability of Common Loons in Washington, including mean, standard deviation (SD), and quantiles (2.5%, 25%, 50%, 75%, and 97.5%).

	Mean	SD	2.50%	25%	50%	75%	97.50%
Intercept[1]	-0.206	0.312	-0.810	-0.416	-0.208	0.006	0.399
Focal[2]	25.539	19.938	7.761	15.234	19.792	27.750	87.913
Area[3]	0.210	1.043	-1.770	-0.381	0.168	0.750	2.415
Canopy[4]	0.555	0.637	-0.532	0.127	0.488	0.917	1.962
Elevation[5]	0.824	0.828	-0.572	0.269	0.749	1.285	2.758
HII[6]	-1.309	0.866	-3.290	-1.767	-1.199	-0.720	0.077
Perimeter Complexity[7]	0.096	0.678	-1.265	-0.329	0.097	0.528	1.471
Perimeter[8]	-1.197	1.658	-5.177	-1.858	-0.927	-0.161	1.135
LC: Forest[9]	-1.670	2.426	-8.267	-2.443	-1.033	-0.175	1.203
LC: Developed[10]	0.223	1.301	-2.190	-0.447	0.206	0.897	2.780
LC: Other[11]	-0.500	1.073	-2.880	-1.072	-0.417	0.159	1.411
σ_{year}	8.195	1.375	4.926	7.340	8.496	9.310	9.933
σ_{site}	1.099	0.749	0.179	0.510	0.899	1.541	2.871

1. The intercept term in the model for the reproduction parameter. This is the reference group for the categorical land cover data types, i.e. open water and wetland (includes woody and emergent herbaceous wetlands) land types.
2. Focal is the coefficient for the temporal autocorrelation effect on reproduction.
3. Area is the coefficient for the area of a site.
4. Canopy is the coefficient for the percent tree canopy cover at a site.
5. Elevation is the coefficient for the elevation of a site.
6. HII is the coefficient for the Human Influence Index at a site.
7. Perimeter complexity is the coefficient for the complexity of a site's perimeter.
8. Perimeter is the coefficient for the length of the perimeter of a site.
9. Forest is the coefficient for the land cover type forest, which includes evergreen, deciduous, and mixed forests. The model intercept estimates open water/wetland sites, such that the Forest coefficient is the difference between open water/wetland sites and forest.
10. Developed is the coefficient parameter for the land cover types of developed (including low, medium, and high levels of development) and cultivated (including hay, pasture, and crops). The model intercept estimates open water/wetland sites, such that the Developed coefficient is the difference between open water/wetland sites and developed/cultivated areas.
11. Other is the coefficient parameter for the land cover types barren land, scrub/shrub land, and herbaceous. The model intercept estimates open water/wetland sites, such that the Other coefficient is the difference between open water/wetland sites and these land types.

Table 1.3 Posterior estimates of model coefficients on detection parameters – at sites without reproduction – for eBird and WDFW data types of Common Loons in Washington, including mean, standard deviation (SD), and quantiles (2.5%, 25%, 50%, 75%, and 97.5%).

	Mean	SD	2.50%	25%	50%	75%	97.50%
Intercept[1]	-2.552	0.130	-2.812	-2.637	-2.553	-2.465	-2.298
Survey[2]	3.111	2.007	0.527	2.077	2.910	3.809	6.583
eBird Area[3]	-0.325	0.203	-0.649	-0.458	-0.362	-0.236	0.154
eBird Distance[4]	-0.550	0.612	-1.988	-0.892	-0.465	-0.129	0.427
eBird Duration[5]	0.103	0.080	-0.062	0.052	0.106	0.155	0.258
eBird Start[6]	-0.142	0.090	-0.323	-0.203	-0.141	-0.082	0.033
WDFW Area[7]	-0.031	1.975	-3.962	-0.841	-0.012	0.792	3.722
WDFW Duration[8]	0.540	1.239	-1.333	-0.172	0.370	1.035	3.455
WDFW Start[9]	-4.183	6.093	-21.920	-4.512	-2.118	-1.096	0.025

1. The intercept term for detection at sites without reproduction, shared between data types.
2. Survey is the term for WDFW detection at sites without reproduction.
3. eBird area is the coefficient for the area of a site.
4. eBird distance is the coefficient for the distance travelled or area covered during an eBird observation.
5. eBird duration is the coefficient for the duration of an eBird survey.
6. eBird start is the coefficient for the time an eBird observation started.
7. WDFW area is the coefficient for the area of a site.
8. WDFW duration is the coefficient for the duration of a WDFW survey.

Table 1.4 Posterior estimates of model coefficients on detection parameters – at sites with reproduction – for eBird and WDFW data types of Common Loons in Washington, including mean, standard deviation (SD), and quantiles (2.5%, 25%, 50%, 75%, and 97.5%).

	Mean	SD	2.50%	25%	50%	75%	97.50%
Intercept[1]	-0.236	0.216	-0.652	-0.382	-0.236	-0.098	0.197
Survey[2]	3.166	1.121	0.704	2.538	3.163	3.827	5.313
eBird Area[3]	-1.916	0.808	-3.743	-2.357	-1.754	-1.441	-0.539
eBird Distance[4]	-2.850	2.583	-9.057	-4.326	-2.204	-0.880	0.577
eBird Duration[5]	0.666	0.373	0.040	0.389	0.653	0.897	1.472
eBird Start[6]	-0.038	0.165	-0.351	-0.150	-0.039	0.070	0.298
WDFW Area[7]	-1.015	3.379	-10.720	-1.480	-0.349	0.470	2.765
WDFW Duration[8]	0.676	0.783	-0.414	0.127	0.540	1.066	2.552
WDFW Start[9]	-0.546	0.542	-1.782	-0.857	-0.492	-0.173	0.369
Delta[10]	0.766	0.061	0.636	0.727	0.770	0.809	0.875

1. The intercept term for detection at sites with reproduction, shared between data types.
2. Survey is the term for WDFW detection at sites with reproduction.
3. eBird area is the coefficient for the area of a site.
4. eBird distance is the coefficient for the distance travelled or area covered during an eBird observation.
5. eBird duration is the coefficient for the duration of an eBird survey.
6. eBird start is the coefficient for the time an eBird observation started.
7. WDFW area is the coefficient for the area of a site.
8. WDFW duration is the coefficient for the duration of a WDFW survey.
9. WDFW start is the coefficient for the time a WDFW observation started.
10. Delta is the probability of correctly classifying an observation as with reproduction.

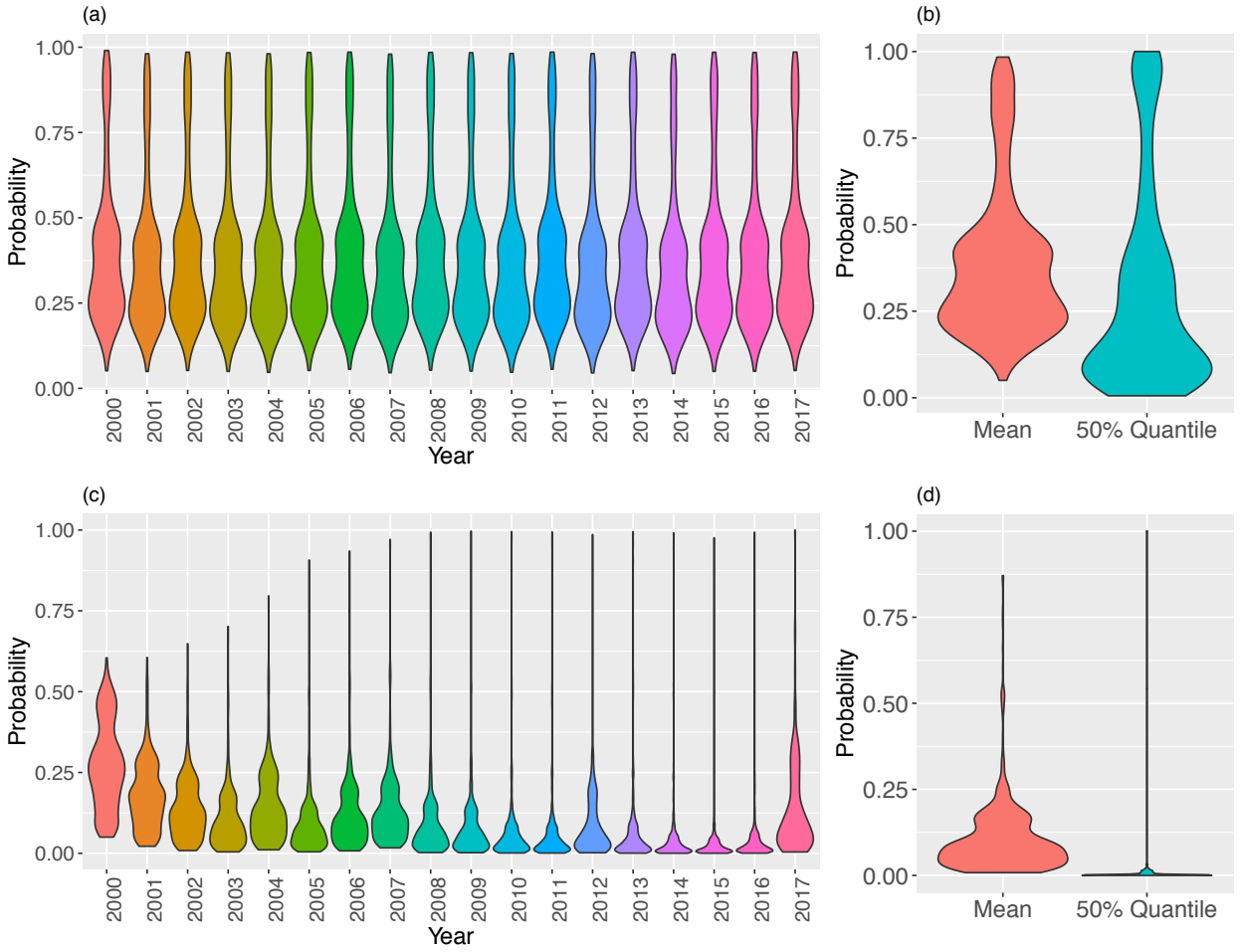

Figure 1.1 Estimated occupancy and reproduction probabilities for sites included in the multi-state occupancy model (*n* = 766). (a) is the mean occupancy probability estimates for all sites in each year (2000-2017). **(b)** is the mean and 50% quantile (median) of occupancy probability for all sites and all years. **(c)** is the mean reproduction probability estimates for all sites in each year (2000-2017). **(d)** is the mean reproduction and 50% quantile (median) of the reproduction probability for all sites and all years.

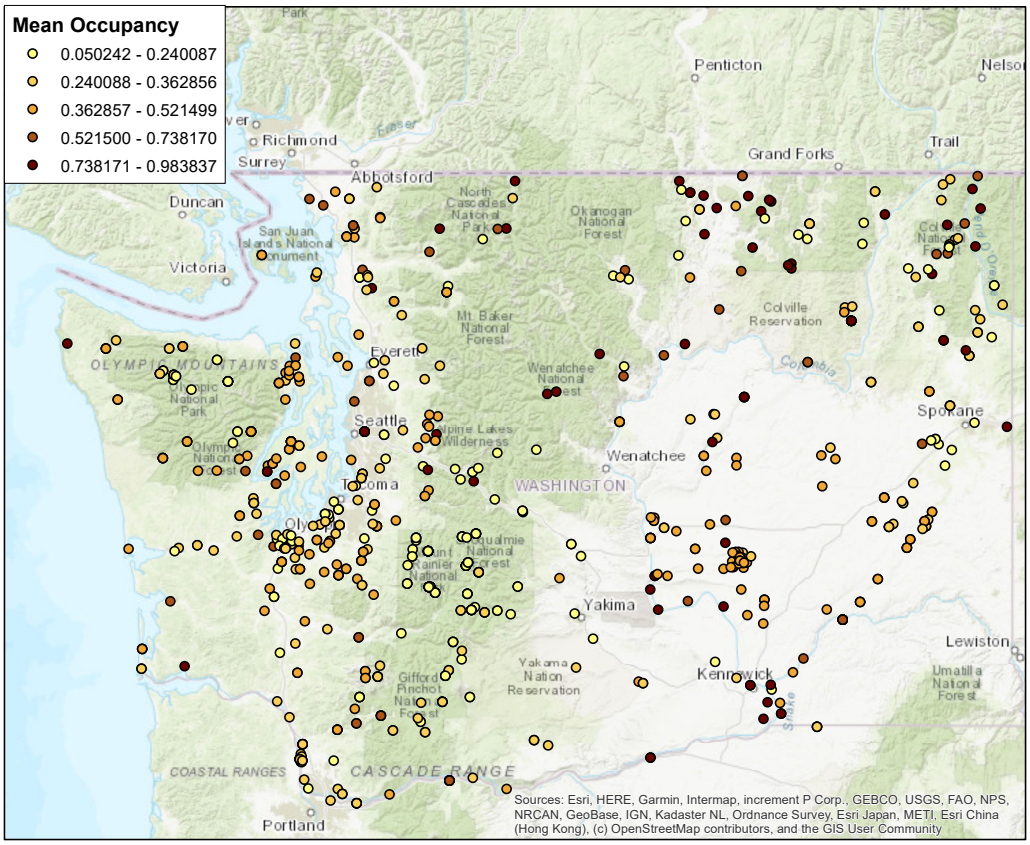

Figure 1.2 Mean estimated occupancy probabilities across years, for sites included in the multi-state occupancy model (*n* = 766). Legend shows the corresponding occupancy probability ranges, from 0.050 to 0.984.

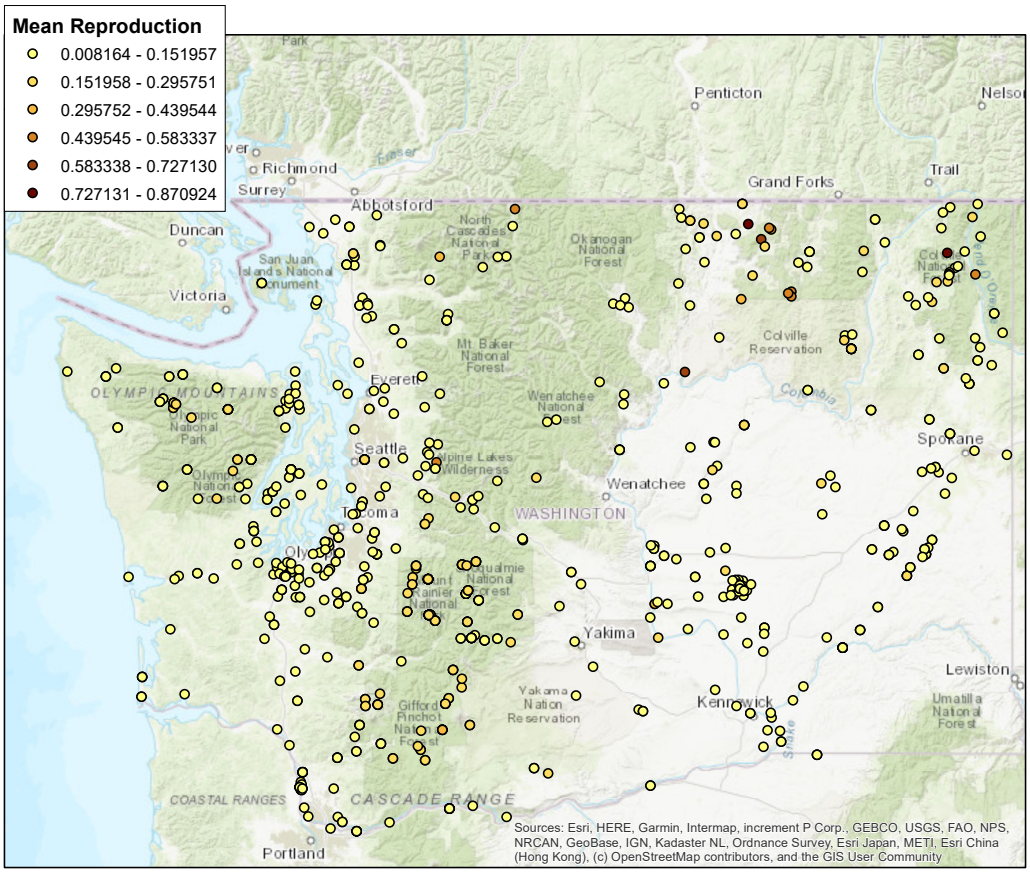

Figure 1.3 Mean estimated reproduction probabilities across years, for sites included in the multi-state occupancy model (*n* = 766). Legend shows the corresponding reproduction probability ranges, from 0.008 to 0.871.

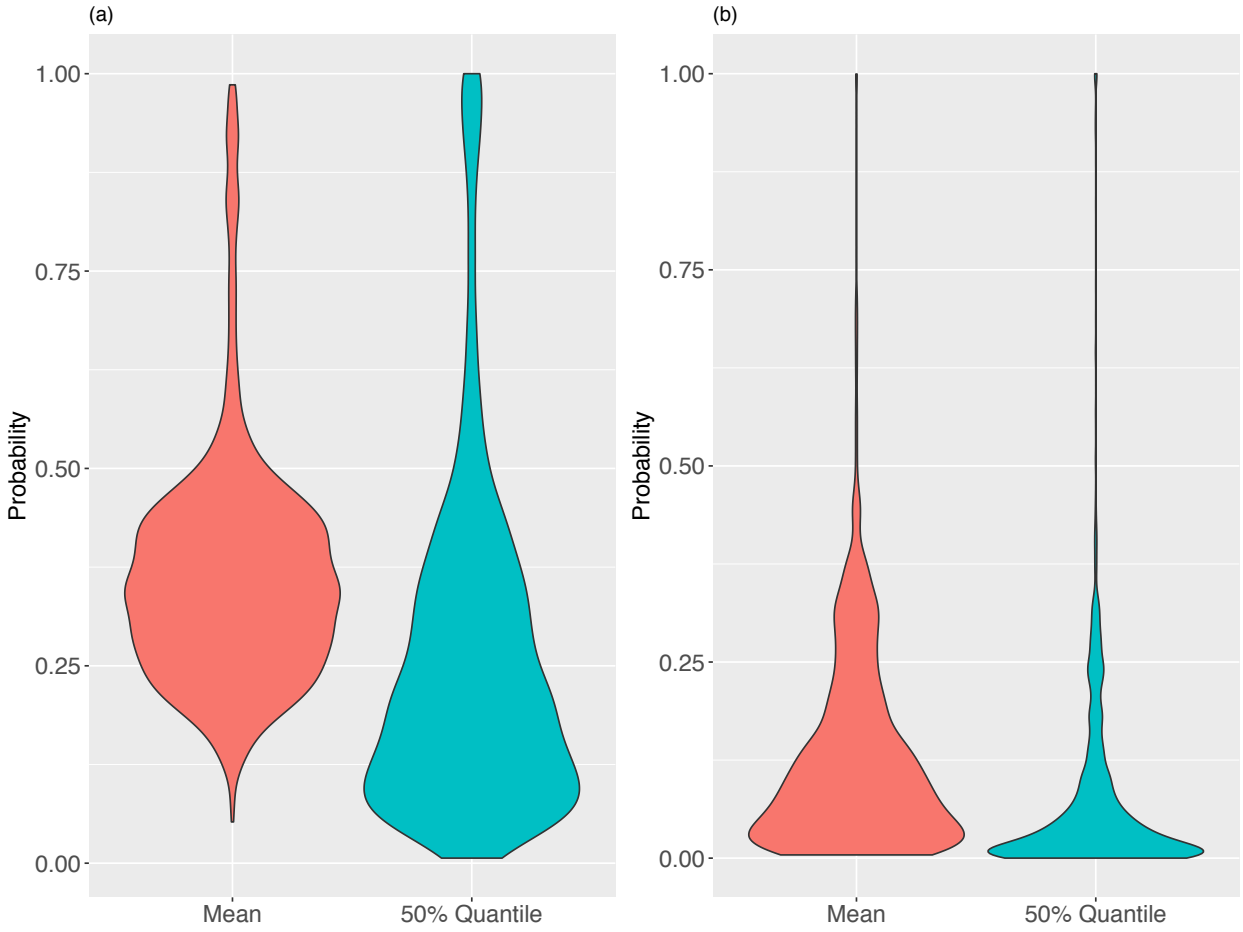

Figure 1.4 Occupancy and reproduction probabilities across Washington State (*n* = 2324 sites), predicted using occupancy and reproduction parameter estimates and site-specific covariate information. Mean and 50% quantile (median) predictions of **(a)** occupancy and **(b)** reproduction for all sites. The mean estimated state-wide occupancy across sites is 0.365 (95% CI = 0.163, 0.839) and the mean estimated state-wide reproduction is 0.129 (95% CI = 0.007, 0.410).

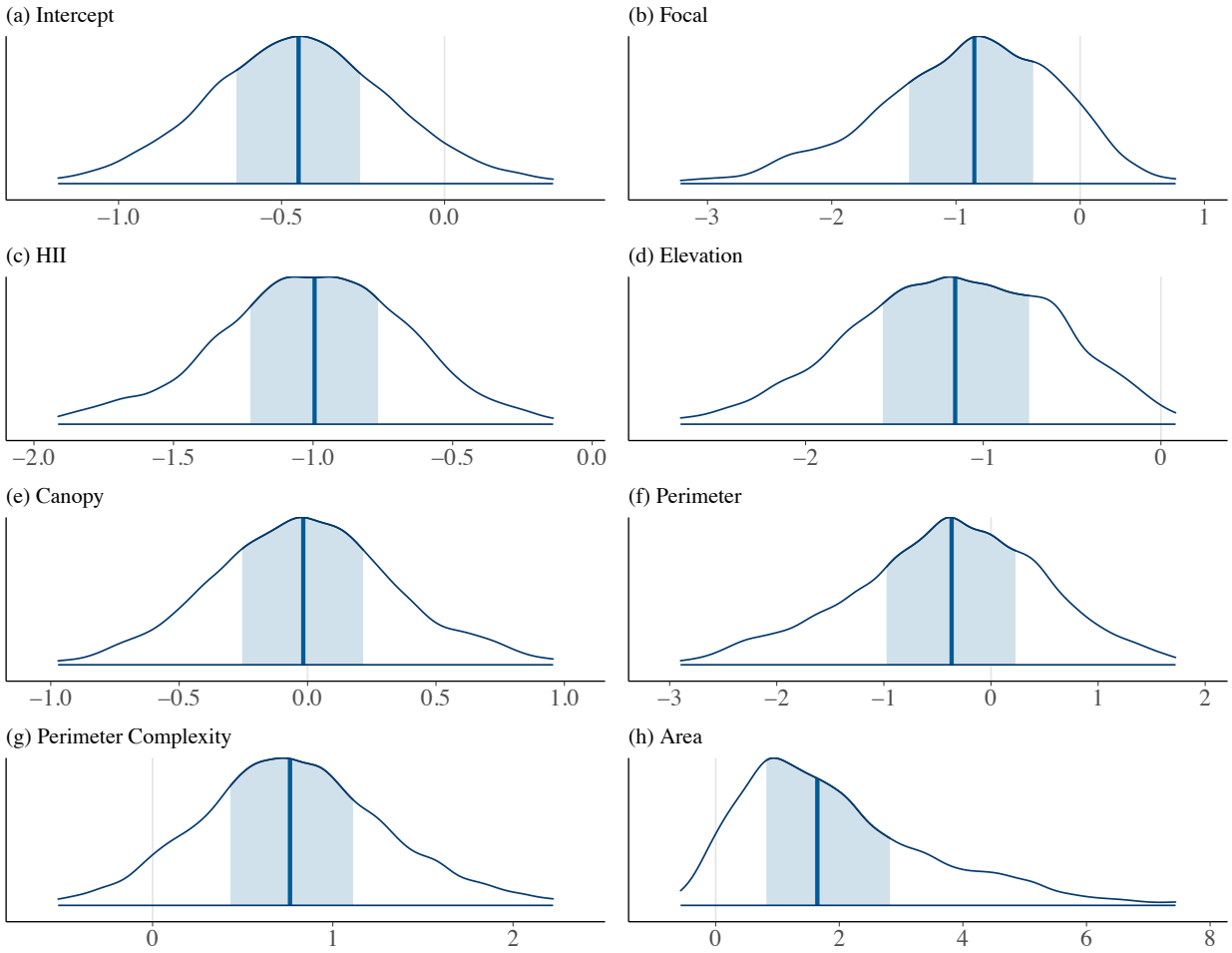

Figure 1.5 Occupancy parameter posterior distributions for continuous parameters, estimates are shown with medians and 50% credible intervals shaded. (a) is the intercept term for the occupancy parameter, **(b)** is the temporal autocorrelation coefficient where the previous year's occupancy is used to model the current year's occupancy. **(c)** is the Human Influence Index coefficient, **(d)** is the coefficient for elevation of the site (m). **(e)** is the coefficient for percent tree canopy cover. **(f-h)** are the coefficients for perimeter (km), perimeter complexity, and area (km^2) of a site.

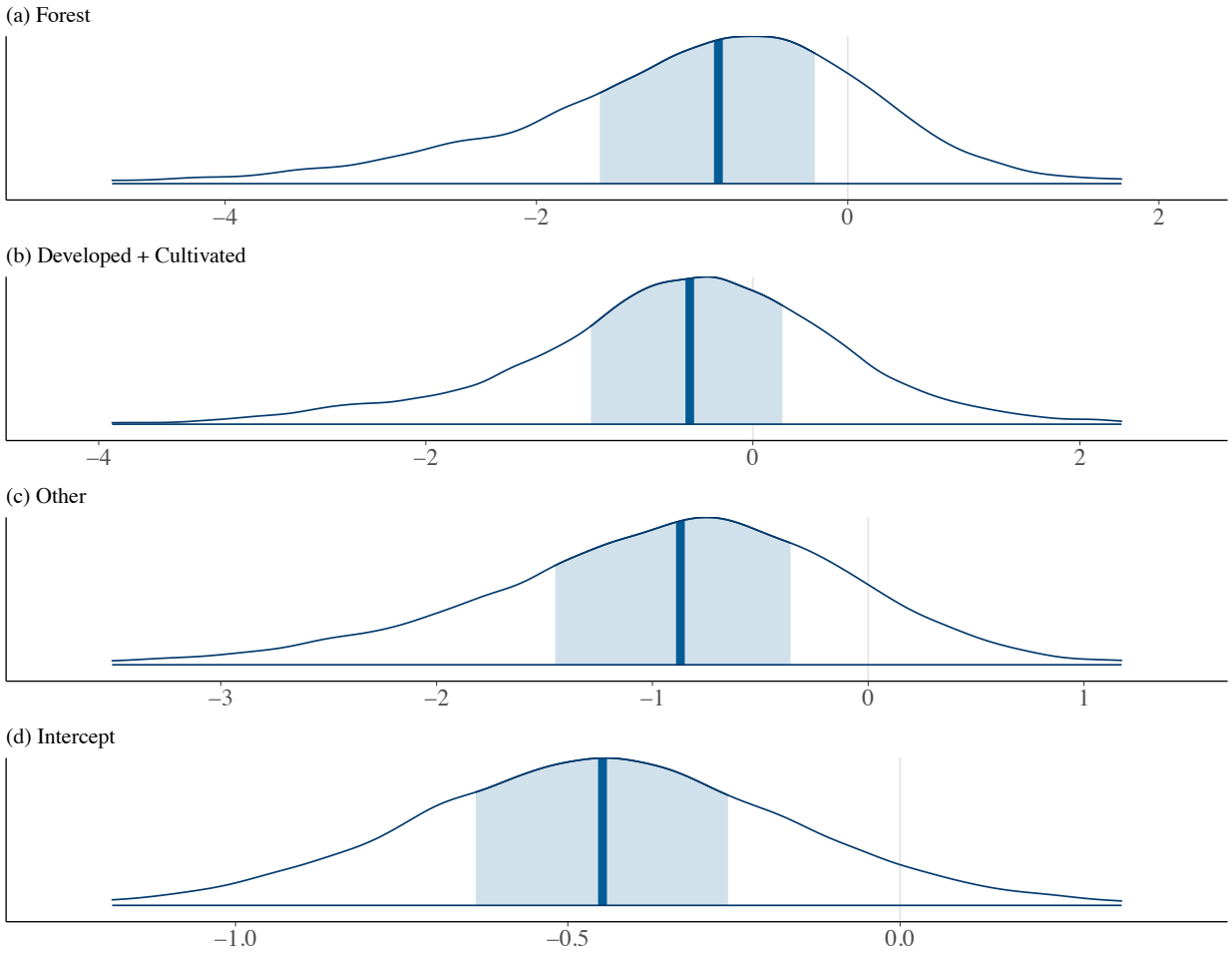

Figure 1.6 Occupancy parameter posterior distributions for categorical parameters, estimates are shown with medians and 50% credible intervals shaded. (a-d) are coefficient estimates for National Land Cover Database categorical land types. **(a)** the posterior estimate for the land type of forest (including evergreen, deciduous, and mixed forests). **(b)** the posterior estimate for the land types of developed (including low, medium, and high levels of development) and cultivated (including hay, pasture, and crops). **(c)** the posterior for the land type of other, which includes land types barren land, shrub/scrub types, and herbaceous land. **(d)** is the intercept term for the model, which acts as a reference group for the final land types of open water and wetlands (including woody and herbaceous wetlands).

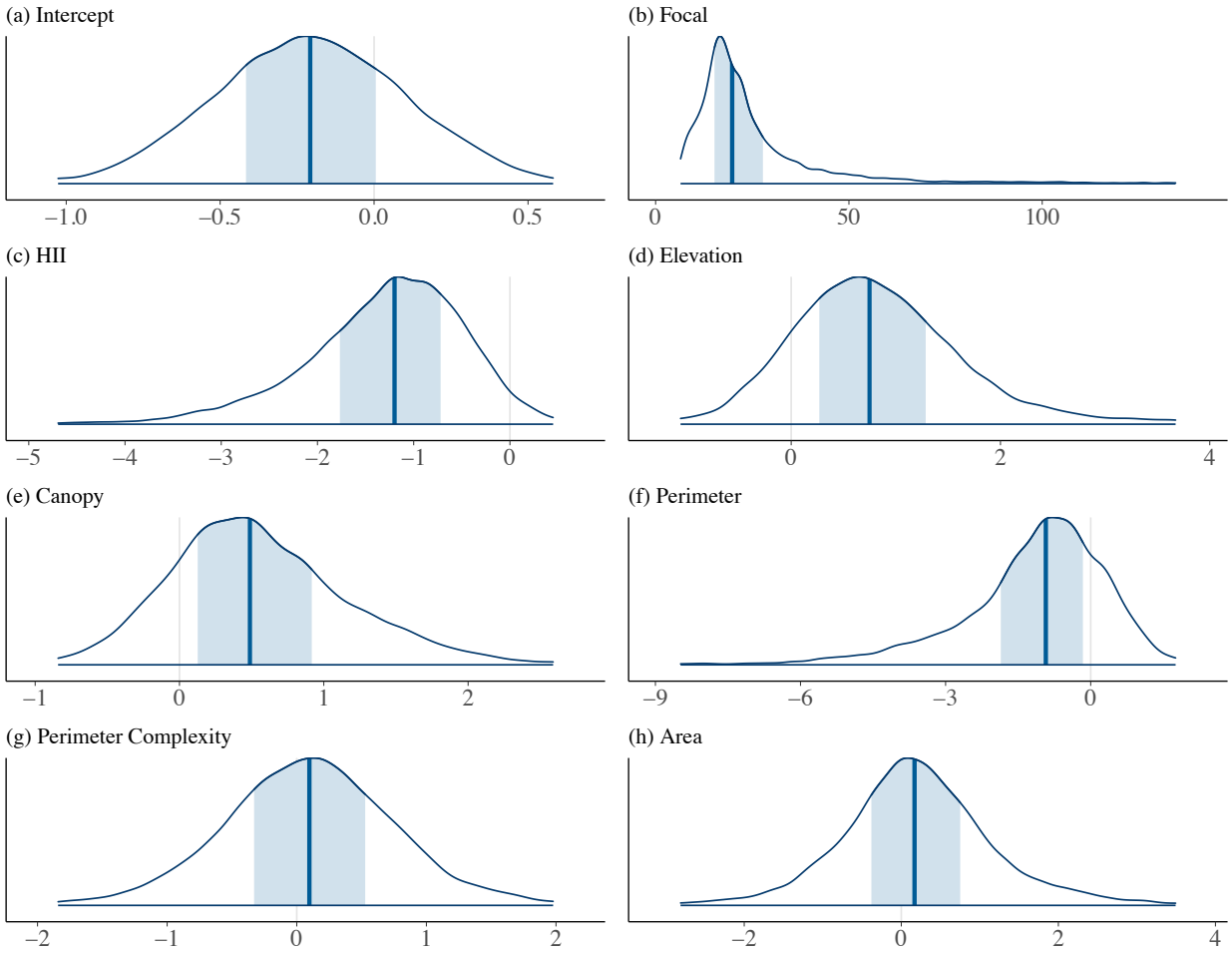

Figure 1.7 Reproduction parameter posterior distributions for continuous parameters, estimates are shown with medians and 50% credible intervals shaded. (a) is the intercept term for the reproduction parameter. **(b)** is the temporal autocorrelation coefficient where the previous year's reproduction probability is used to model the current year's reproduction probability. **(c)** is the Human Influence Index coefficient. **(d)** is the coefficient for elevation (m) of the site. (e) is the coefficient for percent tree canopy cover. **(f-h)** are the coefficients for perimeter (km), perimeter complexity, and area (km^2) of a site.

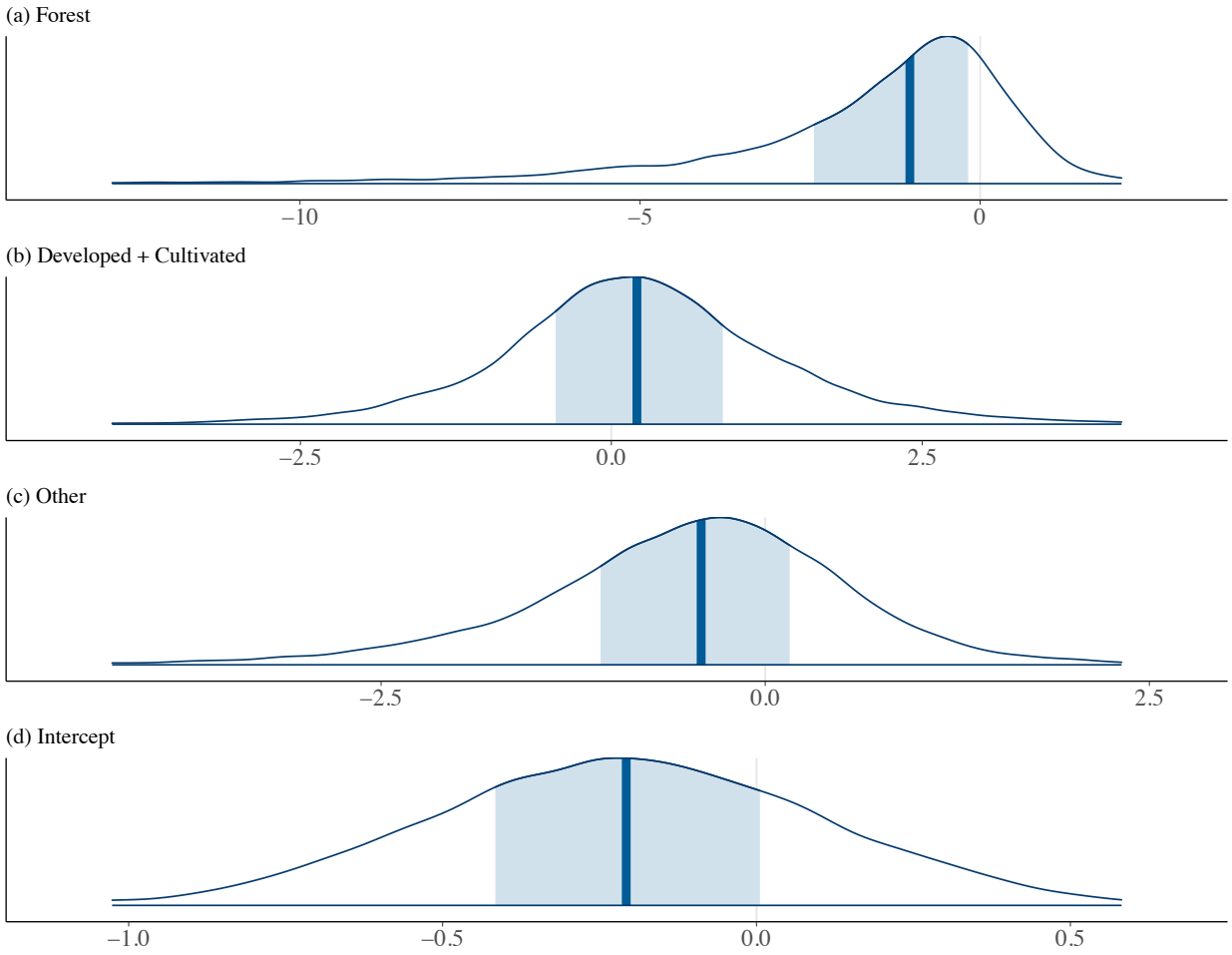

Figure 1.8 Reproduction parameter posterior distributions for categorical parameters, estimates are shown with medians and 50% credible intervals shaded. (a-d) are coefficient estimates for National Land Cover Database categorical land type. **(a)** the posterior estimate for the land type of forest (including evergreen, deciduous, and mixed forests). **(b)** the posterior estimate for the land types of developed (including low, medium, and high levels of development) and cultivated (including hay, pasture, and crops). **(c)** the posterior for the land type of other, which includes land types barren land, shrub/scrub types, and herbaceous land. **(d)** is the intercept term for the model, which acts as a reference group for the final land types of open water and wetlands (including woody and herbaceous wetlands).

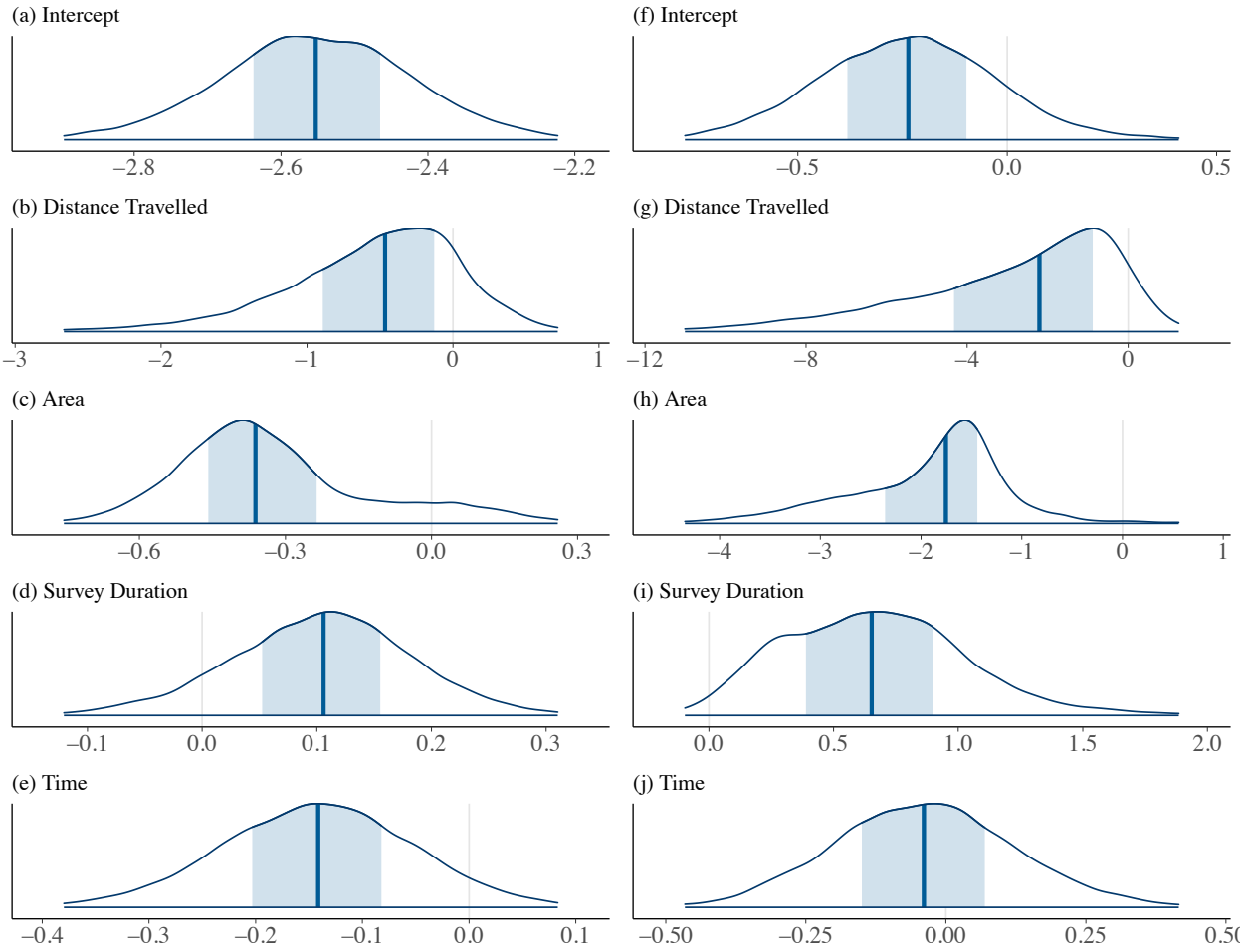

Figure 1.9 Detection parameter posterior distributions for eBird observations, estimates are shown with medians and 50% credible intervals shaded. (a-e) Coefficient parameter estimates for eBird detection at sites without reproduction. **(f-j)** Coefficient parameter estimates for eBird detection at sites with reproduction. **(a)** and **(f)** are the intercept terms for detection at sites without reproduction and with reproduction, respectively.

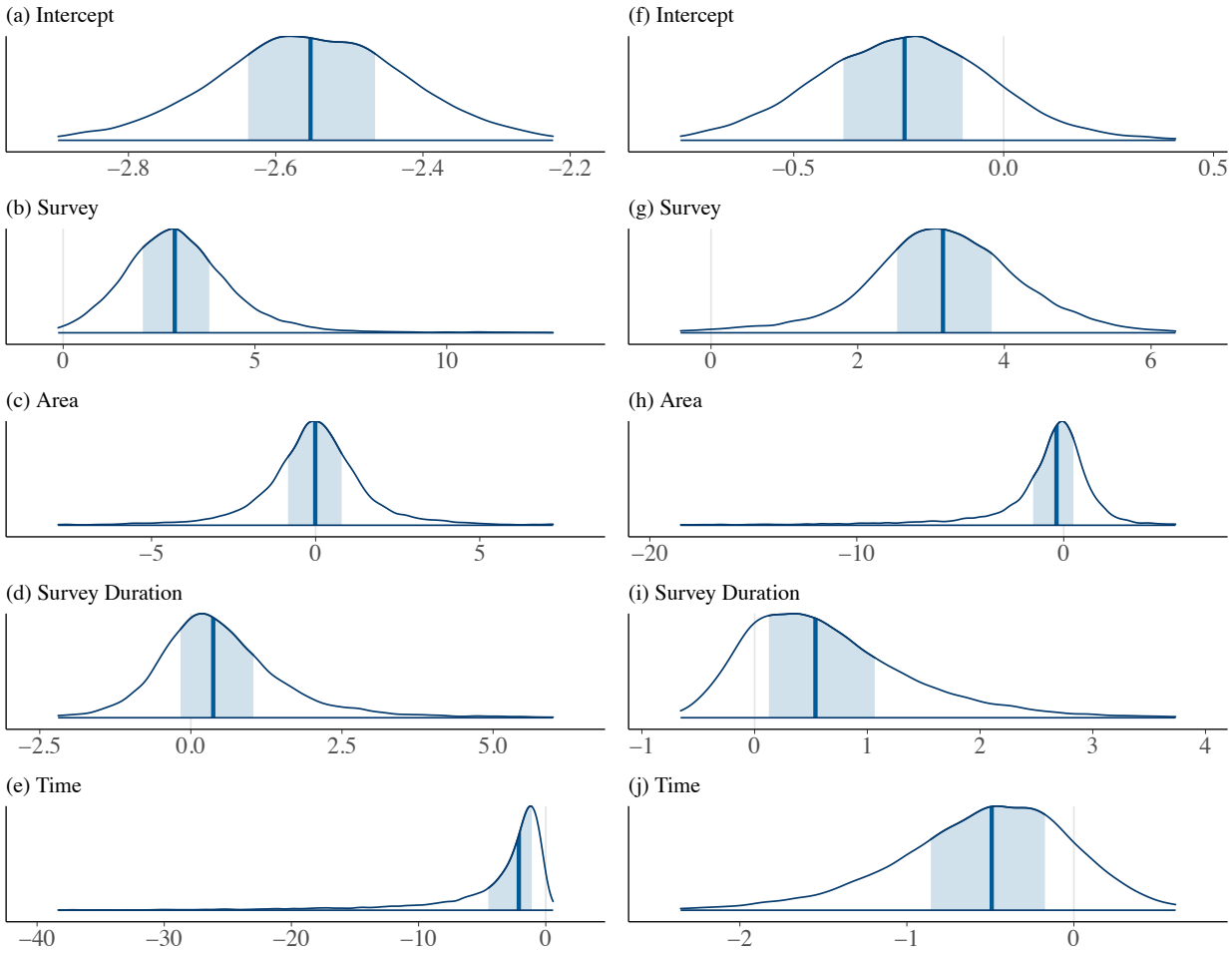

Figure 1.10 Detection parameter posterior distributions for WDFW observations, estimates are shown with medians and 50% credible intervals shaded. (a-e) Coefficient parameter estimates for WDFW detection at sites without reproduction. **(f-j)** Coefficient parameter estimates for WDFW detection at sites with reproduction. **(a)** and **(f)** are the intercept terms for detection at sites without reproduction and with reproduction, respectively. **(b)** and **(g)** are the WDFW specific survey effects.

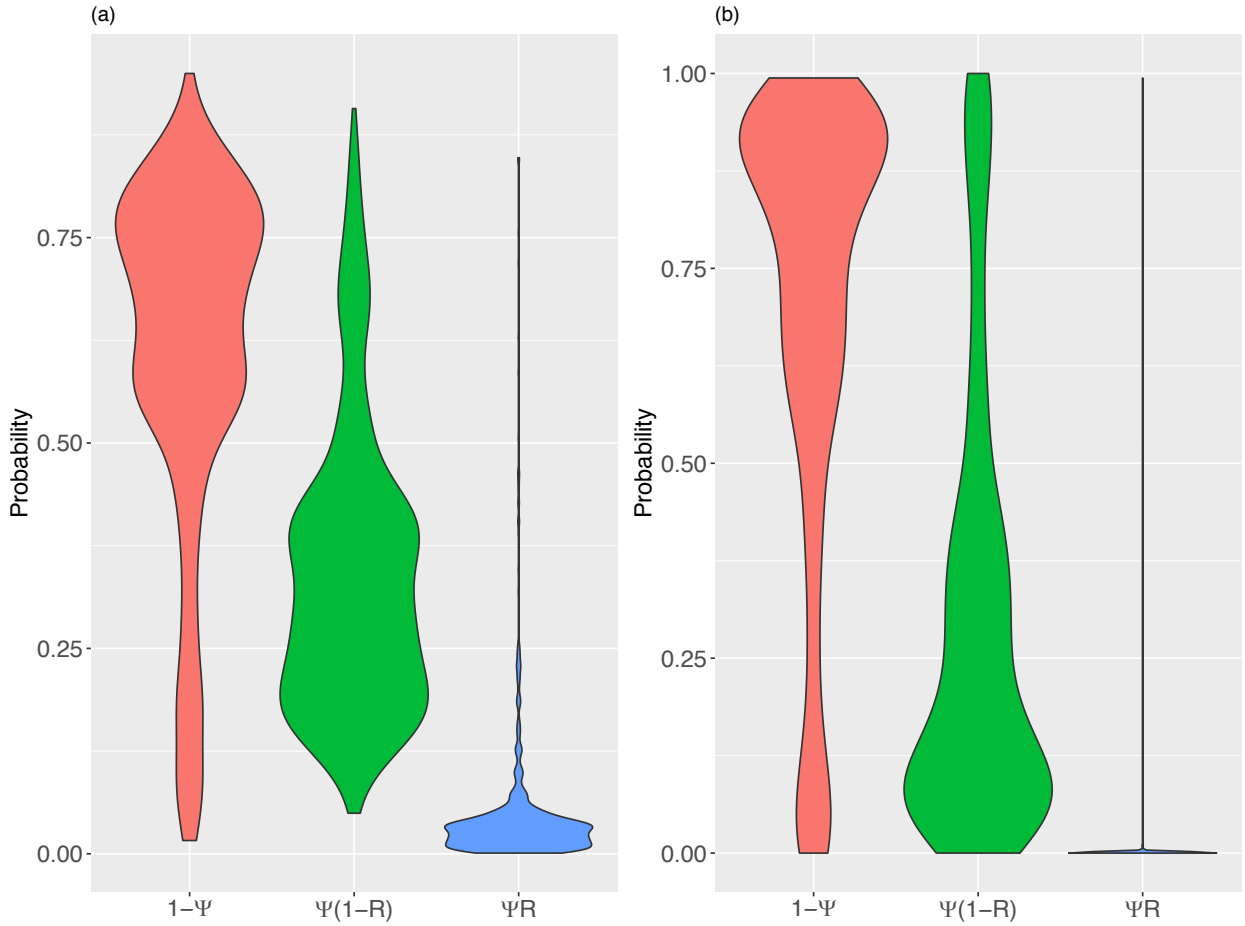

Figure 1.11 Occupancy states by site for all years, for sites included in the model (*n* = 766), $(1 - \Psi)$ is the probability of a site being unoccupied, $\Psi * (1 - R)$ is the probability that a site is occupied without reproduction, and ΨR is the probability that a site is occupied with reproduction. (a) are the mean estimated occupancy state probabilities and **(b)** are the 50% quantile (median) of the estimated occupancy state probabilities.

1.8 APPENDIX 1

Full multi-state occupancy model incorporating eBird citizen science data and state monitoring data. The table of covariates provides how covariates were represented within the NIMBLE model code found below.

Table of Covariates

Covariates for Occupancy (psi) and Reproduction (R)

hii	elev	area	pc	canopy	focal	Land Cover types
Human Influence Index	Elevation in meters	Area in km^2	Perimeter complexity	NLCD Percent tree cover	Temporal autocorrelation	11:Forest 2:Developed 31:Other

Covariates for detection parameters

timeobs	dur	dist
Time of the observation	Duration of time spent	Distance covered or area travelled

Full model written in NIMBLE code

```
mod<-nimbleCode({
  delta ~ dbeta(1,1)
  l.psi~ dnorm(0,10)
  sig_b.area~dexp(1)
  b.area~dnorm(0, sig_b.area)
  sig_b.peri~dexp(1)
  b.peri~dnorm(0, sig_b.peri)
  sig_b.elev~dexp(1)
  b.elev~dnorm(0, sig_b.elev)
  sig_b.hii~dexp(1)
  b.hii~dnorm(0,sig_b.hii)
  sig_b.pc~dexp(1)
  b.pc~dnorm(0, sig_b.pc)
  sig_b.canopy~dexp(1)
  b.canopy~dnorm(0,sig_b.canopy)
  sig_b.11~dexp(1)
  b.11~dnorm(0,sig_b.11)
  sig_b.2~dexp(1)
  b.2~dnorm(0,sig_b.2)
  sig_b.31~dexp(1)
  b.31~dnorm(0,sig_b.31)
  #R
  l.R ~ dnorm(0,10)
```

```
sig_r.area~dexp(1)
r.area~dnorm(0,sig_r.area)
sig_r.hii~dexp(1)
r.hii~dnorm(0,sig_r.hii)
sig_r.elev~dexp(1)
r.elev~dnorm(0,sig_r.elev)
sig_r.peri~dexp(1)
r.peri~dnorm(0,sig_r.peri)
sig_r.pc~dexp(1)
r.pc~dnorm(0, sig_r.pc)
sig_r.focal~dexp(1)
r.focal~dnorm(0,sig_r.focal)
sig_r.canopy~dexp(1)
r.canopy~dnorm(0, sig_r.canopy)
sig_r.11~dexp(1)
r.11~dnorm(0,sig_r.11)
sig_r.2~dexp(1)
r.2~dnorm(0,sig_r.2)
sig_r.31~dexp(1)
r.31~dnorm(0,sig_r.31)
#detection
l.po1 ~ dnorm(0,10)
l.po2 ~ dnorm(0,10)
sig_b.survey1~dexp(1)
b.survey1~dnorm(0,sig_b.survey1)
sig_p11.area~dexp(1)
p11.area~dnorm(0, sig_p11.area)
sig_p12.area~dexp(1)
p12.area~dnorm(0, sig_p12.area)
sig_p21.area~dexp(1)
p21.area~dnorm(0, sig_p21.area)
sig_p22.area~dexp(1)
p22.area~dnorm(0, sig_p22.area)
sig_b.survey2~dexp(1)
b.survey2~dnorm(0, sig_b.survey2)
sig_p11.dur~dexp(1)
p11.dur~dnorm(0, sig_p11.dur)
sig_p12.dur~dexp(1)
p12.dur~dnorm(0, sig_p12.dur)
sig_p21.dur~dexp(1)
p21.dur~dnorm(0, sig_p21.dur)
sig_p22.dur~dexp(1)
p22.dur~dnorm(0, sig_p22.dur)
sig_p11.start~dexp(1)
p11.start~dnorm(0, sig_p11.start)
sig_p12.start~dexp(1)
```

```
p12.start~dnorm(0, sig_p12.start)
sig_p21.start~dexp(1)
p21.start~dnorm(0, sig_p21.start)
sig_p22.start~dexp(1)
p22.start~dnorm(0, sig_p22.start)
sig_p11.dist~dexp(1)
p11.dist~dnorm(0, sig_p11.dist)
sig_p21.dist~dexp(1)
p21.dist~dnorm(0, sig_p21.dist)
sd_alpha_psi~dunif(0,10)
tau_alpha_psi<-pow(sd_alpha_psi, -2)
sd_alpha_psiN~dunif(0,10)
tau_alpha_psiN<-pow(sd_alpha_psiN, -2)

for(t in 1:T){
  alpha_psi[t]~dnorm(0, tau_alpha_psi)
}
for(i in 1:N){
  alpha_psiN[i]~dnorm(0, tau_alpha_psiN)
}

for(i in 1:N){
  logit(psi[i,1])<-l.psi +alpha_psiN[i]+alpha_psi[1]+b.elev*elev[i,1]+b.pc*pc[i,1]+
    b.hii*hii[i,1]+b.area*area[i,1]+b.peri*peri[i,1]+b.canopy*canopy[i,1]+
    b.11*x11[i,1]+b.2*x2[i,1]+b.31*x31[i,1]
  logit(R[i,1]) <- l.R +r.area*area[i,1]+r.hii*hii[i,1]+r.elev*elev[i,1]+
    r.peri*peri[i,1]+r.pc*pc[i,1]+ r.canopy*canopy[i,1]+
    r.11*x11[i,1]+r.2*x2[i,1]+r.31*x31[i,1]
}
for(i in 1:N){
  for(t in 2:T){
    f[i,t]<-getFocalOcc(z[i,t-1])
    fr[i,t]<-getFocalR(z[i,t-1])
    logit(psi[i,t])<-l.psi +alpha_psi[t]+alpha_psiN[i]+
      b.focal*f[i,t]+
      b.hii*hii[i,t]+b.elev*elev[i,t]+b.pc*pc[i,t]+
      b.area*area[i,t]+b.peri*peri[i,t]+b.canopy*canopy[i,t]+
      b.11*x11[i,t]+b.2*x2[i,t]+b.31*x31[i,t]
    logit(R[i,t]) <- l.R +r.area*area[i,t]+r.hii*hii[i,t]+r.elev*elev[i,t]+
      r.focal*fr[i,t]+
      r.peri*peri[i,t]+r.pc*pc[i,t]+ r.canopy*canopy[i,t]+
      r.11*x11[i,t]+r.2*x2[i,t]+r.31*x31[i,t]
  }
}
for(i in 1:N){
  for(t in 1:T){
```

```
    z[i,t] ~ dcat(phi[i,t,1:3])
    phi[i,t,1]<-1-psi[i,t]
    phi[i,t,2]<-psi[i,t]*(1-R[i,t])
    phi[i,t,3]<-psi[i,t]*R[i,t]
   }
  }
  for(i in 1:N){
   for(t in 1:T){
    for(j in 1:Jm[i,t]){
     logit(po1[i,t,j,1]) <- l.po1 + b.survey1*(ind[i,t,j] - 1)
+p11.area*area[i,t]+p11.dur*dur[i,t,j]+p11.start*timeobs[i,t,j]+p11.dist*dist[i,t,j]
     logit(po1[i,t,j,2]) <- l.po1 + b.survey1*(ind[i,t,j] - 1)
+p12.area*area[i,t]+p12.dur*dur[i,t,j]+p12.start*timeobs[i,t,j]
     logit(po2[i,t,j,1]) <- l.po2 + b.survey2*(ind[i,t,j] - 1)
+p21.area*area[i,t]+p21.dur*dur[i,t,j]+p21.start*timeobs[i,t,j]+p21.dist*dist[i,t,j]
     logit(po2[i,t,j,2]) <- l.po2 + b.survey2*(ind[i,t,j] - 1) +
p22.area*area[i,t]+p22.dur*dur[i,t,j]+p22.start*timeobs[i,t,j]
     p1[i,t,j,1] <- po1[i,t,j,1]
     p1[i,t,j,2] <- po1[i,t,j,2]
     p2[i,t,j,1] <- po2[i,t,j,1]
     p2[i,t,j,2] <- po2[i,t,j,2]

     pvec[i,t,j,1:3]<-obsFun(ind=ind[i,t,j], p11=p1[i,t,j,1], p12=p1[i,t,j,2], p21=p2[i,t,j,1],
p22=p2[i,t,j,2],z=z[i,t],delta=delta)
     y[i,t,j] ~ dcat(pvec[i,t,j,1:3])
    }
   }
  }
})
#function for the detection probability matrix depending on which data type the observation is
obsFun<-nimbleFunction(
 run=function(ind=integer(), p11=double(),p12=double(),
          p21=double(), p22=double(), z=integer(), delta=double()){
  p<-matrix(nrow=3, ncol=3)
  if(ind==1){
   p[1,1:3]<-c(1,0,0)
   p[2,1:3]<-c((1-p11),p11, 0)
   p[3,1:3]<-c((1-p21), p21, 0)
  } else{
   if(ind==2){
    p[1,1:3]<-c(1,0,0)
    p[2,1:3]<-c((1-p12), p12,0)
    p[3,1:3]<-c((1-p22), p22*(1-delta), p22*delta)
   }else{
    p[1,1:3]<-c(1,0,0)
    p[2,1:3]<-c((1-p12), p12,0)
```

```
      p[3,1:3]<-c((1-p22), p22*(1-delta), p22*delta)
    }
   }
   pvec<-p[z,1:3]
   return(pvec[1:3])
   returnType(double(1))
  })
#Functions for the temporal effect on occupancy and reproduction
getFocalOcc<-nimbleFunction(
  run=function(z=double()){
   if(z>1) return(1)
   else return(0)
   returnType(integer())
  })
getFocalR<-nimbleFunction(
  run=function(z=double()){
   if(z>2) return(1)
   else return(0)
   returnType(integer())
  })
```

Chapter 2: Optimally allocating survey resources for occupancy analysis of Washington State Common Loons

2.1 Abstract

Collecting robust data on a species for conservation monitoring purposes can be challenging when the area of interest is large or difficult to access, the species requires intensive effort to survey due to low incidence or low detection probability, or there are limited resources available for conducting surveys. Data collection for use in an occupancy analysis further requires that sites are repeatedly visited throughout the study season. Identifying a study design that will optimally allocate limited resources is key to maximizing information gain from occupancy analyses. Placing study design within the broader framework of optimal decision making provides a structured and transparent approach to optimal survey design, while Bayesian analytical methods provide the opportunity to leverage Bayesian updating in the evaluation of candidate survey designs. The Common Loon (*Gavia immer*) is a state-listed sensitive species in Washington State, but little is known about its distribution or habitat associations. The species is monitored by the Washington Department of Fish and Wildlife, but there are limited personnel hours available to conduct surveys each season. I formulated optimal survey design for the Common Loon in Washington as a resource allocation problem, with a monitoring objective of minimizing the uncertainty of Common Loon occupancy parameter estimates over time given a constraint on survey effort. Alternative designs were built through application of alternative decision rules wherein sites were selected based on various estimates from an initial occupancy analysis; e.g., site-specific occupancy probability, site-specific uncertainty in occupancy probability, or site-specific uncertainty in covariate relationships. Simulated data from each alternative design were fit to an occupancy model with priors derived from the initial analysis

(i.e., using Bayesian updating). The decision rule that minimized the predicted state-wide mean uncertainty in occupancy probability selected sites based on the site-specific uncertainty in covariate relationships. This optimal design framework is applicable to occupancy models generally, and provides a quantitative assessment of the outcome of potential survey designs with respect to a given monitoring objective.

2.2 Introduction

Species monitoring is the practice of collecting information on parameters of interest through time, thereby providing quantitative measures to assess temporal trends or to predict responses of the species to management actions (Yoccoz et al., 2001). Identifying the initial state of the species being monitored provides a baseline for comparison (Pollock et al., 2002; Yoccoz et al., 2001). Effective monitoring seeks to represent the population of interest, and requires spatially and temporally representative sampling, adequate sample size, and methods to ensure adequate detection of the species (Buckland and Johnston, 2017). However, collecting robust data can be challenging due to the variability intrinsic in ecological systems and inability to completely observe system dynamics, both of which increase uncertainty in estimated trends or responses to management actions (Williams, 2011). Additionally, limited resources to conduct surveys will tend to result in low precision and accuracy (Field et al., 2007). Nonetheless, survey design is often overlooked or underspecified in species monitoring programs (Legg and Nagy, 2006). An inadequate survey design can lead to an inability to detect the response of an implemented action, suboptimal decisions, and suboptimal use of scarce resources (Lindenmayer and Likens, 2010). Through a structured survey design process, strategies can be evaluated based on their ability to achieve monitoring objectives prior to implementation (Reynolds et al., 2011).

Challenges in survey design arise when the area to be surveyed is large, the species is found in secluded locations, or the species is difficult to detect, characteristics that can result in difficulty identifying optimal spatial and temporal sampling schemes. For monitoring programs that seek to survey a large geographical area, sampled locations should provide a representative sample of the entire area, thereby allowing inference to the entire area (Pollock et al., 2002). General guidelines for robust survey design include accounting for spatial variability through randomly selecting sampling locations and including information to estimate the detectability of the species (Pollock et al., 2002; Yoccoz et al., 2001). However, for species that are cryptic, rare, and non-uniformly distributed across a landscape or strata, random sampling may result in few detections of the species (Guisan et al., 2006). Consequently, when a species is not perfectly detected, as is the case for most species, increased sampling effort is required (Field et al., 2005).

Occupancy modeling is a commonly used technique for modeling species distributions across a geographic area, and involves explicitly modeling the detection process (MacKenzie et al., 2003). Repeated surveys are taken within a period of time during which the population can be assumed demographically and geographically closed, thus allowing the detection process to be separated from the ecological process of interest. General study design recommendations for occupancy modeling are available, including advice for conducting sampling depending on the suspected or estimated detection and occupancy of a given species. For example, if detection is high and occupancy is low, i.e., the species is rare but conspicuous, sampling a greater number of sites with fewer repeat visits is advised. Conversely, species with low detectability and high occupancy (i.e. common but cryptic species), require increased sampling at fewer sites (MacKenzie et al., 2018; MacKenzie and Royle, 2005). However, further consideration is often required based on the unique characteristics of the species and landscape. Occupancy studies

have been designed using a power analysis approach (Bailey et al., 2007; Barata et al., 2017; Field et al., 2005; Guillera-Arroita et al., 2010; Guillera-Arroita and Lahoz-Monfort, 2012; Sewell et al., 2012). Power analysis allows for identification of the sample size required to detect an effect of a specified magnitude with a specified level of statistical significance (Di Stefano, 2003; Field et al., 2004), which has limitations (Wade, 2000). For example, uncertainty in parameter estimates is not easily integrated into power analysis (Di Stefano, 2003) and the sample size required to detect a specific effect with a given significance level may exceed the available resources (Ellis et al., 2014).

An alternative approach is to consider survey design through a Bayesian framework. In Bayesian analysis, parameter estimates are described as probability distributions and naturally include uncertainty (Ascough et al., 2008; Ellison, 2004; Wade, 2000). In comparison to typical power analyses, Bayesian methods can be used to describe the expected full probability distribution for a given effect under a candidate survey design, thereby providing richer information on which to base decisions (Wade, 2000). Additional information can be included as it becomes available using Bayesian updating (Ellison, 2004). Consequently, Bayesian methodology has tremendous utility in decision making (Ascough et al., 2008; Wade, 2000), including decision making about optimal survey design.

Decision analysis provides an avenue for making decisions in a structured way, using available information and considering uncertainties (Clemen, 1996). A decision-analytic framework is composed of four fundamental elements: objectives, alternative actions, predicted responses to the alternative actions, and a solution or optimization method (Clemen, 1996; Lyons et al., 2008). For management decisions, the objective is defined by managers based on their desired outcomes. The management alternatives are explicitly defined, and management

constraints can be imposed by limiting the set of feasible alternatives. A model of the system is used to predict system responses under each alternative and an optimization method is used to identify the optimal alternative(s) given all other components of the analysis (Possingham et al., 2001). Uncertainties are recognized throughout the decision-making process, providing management with a complete picture of modelled outcomes (Lyons et al., 2008).

Often, ecological decisions are guided by multiple management objectives and these objectives may be conflicting (e.g., increase conservation returns and decrease cost). Such decisions can be evaluated using principles of multi-objective decision making (Williams and Kendall, 2017), which include the subfields of multi-criteria decision analysis (Mendoza and Martins, 2006) and multi-objective programming (Ehrgott, 2005)). One approach to multi-objective decision making is to use scalarization techniques, which re-formulate the problem as a single-objective optimization problem (Marler and Arora, 2004). One such technique translates all but one objective to constraints (e.g., acceptable above a threshold, unacceptable below a threshold) and optimizes the single objective subject to a range of constraint values; this is formally known as the epsilon constraint method (Haimes et al., 1971). When decision makers can provide a preferred range of values for the constraints, the problem can be solved using the boundary objective function method (Marler and Arora, 2004).

In application to survey design for species monitoring programs, the two primary management objectives are typically to minimize the cost of surveying while maximizing the precision of resulting parameter estimates. Such a problem can be represented mathematically, and optimal solutions can be found, through scalarizing the problem. Critical to solving the problem is input from decision makers on the exact form of the objectives and constraints.

The formulation of a survey design problem in this manner is similar to an optimal resource allocation problem, in context and construction. Resource allocation problems require identification of the optimal allocation of a fixed resource. Optimal resource allocation techniques have been used to find solutions to difficult conservation problems, including the optimal allocation of funding across species management actions (Gerber et al., 2018; Martin et al., 2007), land parcels (Moilanen et al., 2006), regions (Wilson et al., 2006), and surveys (Gerber et al., 2005), and the optimal allocation of staff time across conservation projects (Converse et al., 2011). The common thread among all examples is that each seeks to optimize objectives within constraints. Optimal resource allocation approaches can accommodate uncertainty, presenting managers with the information necessary to make informed decisions (McCarthy, 2014; McCarthy et al., 2010).

In Washington State, the Common Loon (*Gavia immer*) is listed as a Sensitive species, but the distribution and trends of this species are not well understood. Common Loons defend breeding territories during the summer months on oligotrophic lakes with little human disturbance (Richardson et al., 2000). Historically, breeding Common Loons were thought to be abundant in Washington. However, with increasing habitat degradation (i.e., through shoreline development, disturbance, and pollutants on fresh waterbodies), it is hypothesized that breeding Common Loons have declined, and may continue to, in Washington. The Washington Department of Fish and Wildlife (WDFW) has been monitoring known breeding lakes across the state for approximately 40 years. Because Common Loons are found in secluded and relatively inaccessible areas, it is difficult for biologists to survey all lakes that are potentially occupied by breeding Common Loons. Survey efforts are divided by WDFW districts, where often a single biologist is responsible for surveying an entire district for breeding Common Loons in the

summer months. Consequently, time and budgetary restrictions limit the number of sites that can be visited by biologists during a survey season. In most years, survey sites have not been selected with respect to any defined sampling design, complicating state-wide inference. Hence, there is a need for a state-wide survey design with the ability to provide information on state-wide occupancy and trends.

I approached the development of a survey design for Common Loons as an optimal resource allocation problem. I worked directly with WDFW biologists to determine the objectives and resources available for monitoring. WDFW identified an objective of maximizing their ability to detect trends in occupancy parameters through time, which can be best achieved by minimizing the standard error of occupancy parameter estimates. They also identified a constraint on the survey effort (in hours spent surveying) per season. I used parameter estimates and their associated uncertainty arising from a multi-state occupancy model fitted to available data to simulate outcomes of future survey design alternatives. I then identified alternatives that maximized the survey objective under the effort constraint. Alternatives were characterized by the set of sites that would be visited during a survey season and were constructed based on alternative decision rules for selecting sites using various metrics from the existing model output.

2.3 METHODS

2.3.1 DECISION PROBLEM

The objectives of the optimal survey design problem for Common Loons in Washington are: (1) to minimize the standard error of occupancy model parameter estimates resulting from future surveys, and (2) to minimize the effort spent conducting occupancy surveys each season. Mathematically, this can be written as

$$\min\left(f_1(\theta), f_2(\theta)\right)$$

where $f_1(\theta)$ is the objective function component for the standard error of occupancy and $f_2(\theta)$ is the objective function component for survey effort, for any survey design θ. To simplify this problem, I reframed one of the objectives as a constraint and optimized over the other objective; i.e., using the bounded objective function method (Marler and Arora, 2004). I employ this method here by transforming the effort objective into a constraint, allowing for a fixed amount of survey effort in a season, and optimizing based on the predicted standard error of occupancy model parameter estimates. The optimization problem then becomes

$$\min\left(f_1(\theta)\right)$$

$$subject\ to\colon f_2(\theta) \leq E$$

where E is the available effort for surveying under any survey design θ. Alternative survey designs, θ, are defined by the sites visited during a survey season. Those alternatives that fall outside of the constraint are non-admissible in this context; i.e., WDFW cannot expend more effort than they have available. The task then is to identify the admissible survey design that produces the minimum predicted standard error on occupancy parameter estimates.

The number of feasible survey design alternatives (i.e., within the effort constraint) is large, making it infeasible to directly evaluate each possible alternative due to the scale of the computational task. Instead, promising alternatives can be identified based on candidate decision rules, as follows. First, I used parameter estimates from the existing multi-state occupancy analysis to predict site-specific occupancy for all potential Common Loon sites, both previously sampled and not previously sampled, within Washington (see Chapter 1 for details). Sites were broken down by WDFW management districts, the level at which sampling activities are carried

out. Next, I converted the effort constraint provided by WDFW from the number of hours available to survey in each district to the number of sites that can be surveyed. I then identified alternatives (i.e., collections of sites) based on candidate decision rules that draw on existing site-specific parameter estimates and their associated uncertainty estimates. Finally, I simulated data using the site-specific occupancy parameter estimated under each alternative. The simulated data were then fit to the multi-state occupancy model with informed priors based on the previous analysis, and designs were compared based on the mean of the standard errors of Common Loon occupancy estimates across Washington State.

2.3.2 Parameter Estimates

A multi-state occupancy analysis was conducted using WDFW survey data integrated with citizen science eBird data (see Chapter 1 for details). The results of this analysis supplied parameter estimates that can be applied to all potential Common Loon sites, resulting in site-specific estimates of occupancy. Potential sites in the occupancy analysis were defined as Washington lakes, reservoirs, and rivers greater than 15 acres and below 5000 ft in elevation (n = 2324, see Chapter 1). WDFW is interested in sites where there is a possibility for Common Loon reproduction. Common Loons nest on the shorelines of fresh waterbodies, hence requiring fresh waterbodies that have minimal current to avoid inundation of nest and eggs (McIntyre, 1998; Richardson et al., 2000). Therefore, river sites were removed from the set of sites considered for surveying. Reservoirs were retained, as Common Loons are known to utilize reservoirs for breeding (Richardson et al., 2000). Removing river sites resulted in n = 1268 sites for consideration in survey designs.

Site-specific occupancy estimates, with their associated uncertainty, were generated for lake and reservoir sites using site-specific covariate information and multi-state occupancy

model parameter estimates. Site-specific occupancy can be generated from these parameter estimates as:

$$logit(\Psi_{i,1}) = \alpha_{\Psi,1} + \boldsymbol{X}_{\Psi}\beta_{\Psi,1} + \varepsilon_{\Psi,i,1} \quad (2.1)$$

$$\varepsilon_{\Psi,i,1} = \begin{cases} \varepsilon_{\Psi,i,1} & if\ i \in M1 \\ \varepsilon_{\Psi,i,1} \sim Normal(0, \sigma^{\Psi,1}) & if\ i \notin M1 \end{cases}$$

where $\Psi_{i,1}$ is the occupancy probability estimate for site i, where i indexes sites and $i = 1, \dots, M$, and $M1$ is the subset of the M sites that were included in the initial occupancy analysis. The subscript 1 signifies that parameter estimates are the result of the initial multi-state occupancy analysis. The site-specific covariate information for $\Psi_{i,1}$ is the matrix $\boldsymbol{X}_{\Psi}$. The intercept term, $\alpha_{\Psi,1}$, and the vector of coefficients, $\beta_{\Psi,1}$, are parameter estimates from the multi-state occupancy analysis. The term $\varepsilon_{\Psi,i,1}$ is the random site effect with $Normal(0, \sigma^{\Psi,1})$, the variance term (i.e. $\sigma^{\Psi,1}$) was estimated in the occupancy analysis. For sites included in the initial occupancy analysis (i.e. $i \in M1$), each $\varepsilon_{\Psi,i,1}$ was monitored, resulting in a site-specific random effect distribution; i.e., $\varepsilon_{\Psi,i,1} \sim Normal(\mu_{i,1}, \sigma^{\Psi,1})$, which was used for generating $\Psi_{i,1}$ estimates. For sites that were not included in the occupancy model analysis (i.e. $i \notin M1$), $\varepsilon_{\Psi,i,1}$ was simulated using random samples from a $Normal(0, \sigma^{\Psi,1})$.

2.3.3 District Effort Constraint

Washington State is divided into 17 WDFW management districts and surveys are conducted by district biologists, with a single biologist typically responsible for a given district. Therefore, the effort constraint was applied at the district rather than the state level. In consultation with WDFW biologists, the maximum available survey effort for each district was fixed at 20 hours per survey season (the summer months of a given year). Surveys for occupancy

analysis require repeat visits to a site within a season (MacKenzie et al., 2002). The total survey effort for a given district needed to be translated into the number of sites that could be surveyed. I examined effort information from previous surveys by WDFW biologists; only $n = 40$ records included time spent on the survey. The mean time spent at each site was 1 hour and 51 minutes (range = 5 minutes to 10 hours and 30 minutes). Given the low sample size and the high variability in these data, I made the simplifying assumptions that any site can be surveyed twice in 2 hours, and that any site selected would be surveyed twice. Although the mean time spent surveying a site by the WDFW was closer to 2 hours per visit, Common Loons are thought to be relatively rare and also relatively conspicuous to WDFW observers (see Chapter 1), so visiting more sites for less time will be generally preferable (based on guidelines in (MacKenzie and Royle, 2005). Under these assumptions, the maximum number of sites that could be surveyed in any district is 10, with a maximum of 170 sites surveyed across the state. Alternatives can therefore be characterized by the set of up to 170 sites surveyed, up to 10 per district.

2.3.4 Defining Alternatives

I considered 9 alternative decision rules and used each decision rule to identify the sites that would be surveyed under that alternative (*A1*-*A9*). The first decision rule is the status quo, and therefore alternative *A1* is composed of the sites surveyed by WDFW in 2017 and included in the existing occupancy model estimates ($n = 53$ sites). This alternative is included as a baseline. I expected that surveying a relatively small number of sites that have already been surveyed by the WDFW would result in relatively little reduction in overall uncertainty. All other alternatives were designed to select up to 10 sites in each of 17 districts, though the total sites surveyed in these alternatives was $n = 163$, because one district only included 3 potential survey sites.

The next 6 alternatives (*A2-A7*) are based on decision rules that include sites depending on occupancy probability estimates and/or their associated uncertainty, on either the real or logit scale (Figure 2.1a). I hypothesized that surveys of sites with a high degree of uncertainty would have the potential to provide a relatively large reduction in uncertainty at those particular sites, and so would provide a relatively large reduction in the overall uncertainty in occupancy estimates. I also hypothesized that focusing surveys on sites with high occupancy probabilities would lead to more detections of Common Loons, thereby providing more data to reduce overall uncertainty. Finally, I hypothesized that considering uncertainty on the logit scale rather than the real scale would sidestep the inherently greater uncertainty in occupancy estimates near $\Psi_i = 0.5$, and instead would focus survey effort on sites with the greatest uncertainty in the linear predictor of occupancy (Figure 2.1c and Figure 2.1d).

Alternatives *A8* and *A9* were based on a decision rule that selected sites based on a concept similar to the generalized independent variable hull (gIVH) derived in Conn et al., (2015). The gIVH provides information about site-specific predictions that are outside of or highly uncertain with respect to observed covariate relationships. The site-specific predictions for this decision rule are generated similarly to Eq. 2.1, but without the random site effect, i.e.

$$logit\left(\Psi'_{i,1}\right) = \alpha_{\Psi,1} + \boldsymbol{X}_{\Psi}\beta_{\Psi,1} \tag{2.2}$$

These alternatives describe another method for selecting sites with high uncertainty but based strictly on estimated covariate relationships. Collecting additional information at these sites would provide information to update the statistical relationship between occupancy and habitat characteristics. Alternative *A8* selects sites based on the uncertainty of occupancy predictions generated from estimated covariate relationships, without random effects or temporal autocorrelation, on the real scale as in Eq 2.2. Alternative *A9* selects sites based on this same

decision rule as *A8*, but with values on the logit scale. These alternatives describe another method for selecting sites with high uncertainty but based strictly on estimated covariate relationships (Figure 2.1b).

In summary, the alternative decision rules include:

A1 – survey $n = 53$ sites previously surveyed in 2017

A2 – survey up to 10 sites in each district with the highest uncertainty in the occupancy probability estimate, i.e. max(SE($\Psi_{i,1}$))

A3 – survey up to 10 sites in each district with the highest coefficient of variation in the occupancy probability estimate, i.e. max(SE($\Psi_{i,1}$)/ mean($\Psi_{i,1}$))

A4 – survey up to 10 sites in each district with the highest mean occupancy probability estimate, i.e. max(mean($\Psi_{i,1}$))).

A5 – survey up to 10 sites in each district with the highest uncertainty in the occupancy probability estimate on the logit scale, i.e., max(SE(logit($\Psi_{i,1}$))).

A6 – survey up to 10 sites in each district with the highest coefficient of variation in the occupancy probability estimate on the logit scale, i.e., max (SE(logit($\Psi_{i,1}$))/mean(logit($\Psi_{i,1}$))).

A7 – survey up to 10 sites in each district with the highest mean occupancy probability estimate on the logit scale, i.e., max(mean(logit($\Psi_{i,1}$))).

A8 – survey up to 10 sites in each district with the highest uncertainty in occupancy covariate relationships, i.e. max(SE($\Psi'_{i,1}$)).

A9 – survey up to 10 sites in each district with the highest uncertainty in occupancy covariate relationships, i.e. max(SE(logit($\Psi'_{i,1}$)).

2.3.5 Data Simulation and Model Fitting

One year of observation data were simulated from existing site-specific occupancy and detection parameter estimates, under each of the alternatives, for each of $n = 50$ replicate datasets. The sites selected were constant for each of the alternatives. I only considered new detections by WDFW biologists, and did not consider new detections by eBird volunteers (see Chapter 1). The only survey-specific covariate information used to model WDFW detection probability was the area of each site. The posterior distributions from the occupancy model were used as informative priors when fitting the simulated data to the occupancy model, therefore constituting Bayesian updating. Informative priors for coefficients were specified as $Normal(\mu_{j,1}, \sigma_{j,1})$ where $\mu_{j,1}$ is the mean of the posterior distribution for parameter j and $\sigma_{j,1}$ is the standard deviation of parameter j. A normal distribution was selected based on the multi-state occupancy model prior structure and the estimated posterior distribution's shape. I use informed priors to account for the existing information from the full multi-state occupancy model (see Chapter 1). Model parameters included intercepts and coefficients on predictors in the detection, occupancy, and reproduction models, as well as estimated site-specific random effects, and – for use with sites that were not previously surveyed – the standard deviation of the random effect of site in the occupancy and reproduction models. The random effect of year was ignored in the updated analysis.

Models were written using R and NIMBLE, a Bayesian hierarchical interface that compiles and runs models through C++ (de Valpine et al., 2017; R Core Team, 2018). The adapted multi-state occupancy model was fit to each of the simulated data sets using the MCMC

sampler in NIMBLE. Two chains were run for 60,000 iterations, with 30,000 burn-in and a thinning rate of 10 to reduce file size. The adapted model can be found in Appendix 2.

Fitting the simulated data to the model resulted in updated parameter estimates, which were then used to predict probability of occupancy at all sites in Washington. Parameter estimates of occupancy were generated similar to Eq. 2.1, but with updated parameter estimates from each alternative. For each alternative,

$$logit(\Psi_{i,2,a}) = \alpha_{\Psi,2,\mathrm{a}} + \boldsymbol{X}_{\Psi}\beta_{\Psi,2,\mathrm{a}} + \varepsilon_{\Psi,i,2,a} \tag{2.3}$$

$$\varepsilon_{\Psi,i,2,a} = \begin{cases} \varepsilon_{\Psi,i,2,a} & if\ i \in A^{*}\ and\ i \in M1 \\ \varepsilon_{\Psi,i,1} & if\ i \notin A^{*}\ and\ i \in M1 \\ \varepsilon_{\Psi,i,2,a} \sim Normal(0, \sigma^{\Psi,2,a}) & if\ i \notin A^{*}\ and\ i \notin M1 \end{cases}$$

where $\Psi_{i,2,a}$ is the updated occupancy estimate for site i under alternative a, where i indexes sites and $i = 1, \ldots, A^{*}$ where A^{*} is the set of sites included in the alternative. The subscript 2 indicates that parameter estimates are the result of the updated analysis. The site-specific covariate information for $\Psi_{i,2}$ is the matrix $\boldsymbol{X}_{\Psi}$. The intercept term, $\alpha_{\Psi,2}$, and the vector of coefficients, $\beta_{\Psi,2}$, are parameter estimates from the updated analysis under alternative a. For any alternative, $\varepsilon_{\Psi,i,2}$ was updated for sites included in that alternative (i.e. $i \in A^{*}\ and\ \ i \in M1$). For sites that were not updated but were included in the initial occupancy analysis (i.e. $if\ i \notin A^{*}\ and\ i \in M1$) random site effect estimates from the initial occupancy analysis were informed by posterior estimates of site effects from the original analysis, used as priors in the updated analysis. Random effects for sites that were included in neither the full model nor a given alternative (i.e. $i \notin A^{*}\ and\ i \notin M1$) were randomly generated from the updated $\sigma^{\Psi,2,a}$ posterior estimates simulated as $Normal(0, \sigma^{\Psi,2,a})$.

2.4 RESULTS

The overall SE and CV from the initial occupancy analysis were 0.486 and 1.312, respectively, with an overall estimated mean occupancy probability of $\Psi_{i,1}$ = 0.366 for *n* = 1268 sites (Table 2.1). The number of sites that were selected for surveying was the same for 8 alternatives (*n* = 163); in these alternatives (*A2*-*A9*) 10 sites were selected from all management districts except district 3, which only contained 3 candidate sites. Alternative *A1* selected sites that were previously surveyed in 2017 (*n* =53), alternatives *A2* - *A9* selected 10 sites from all management district except district 3 (*n* = 163).

The alternative that resulted in the greatest reduction in uncertainty Ψ was *A8*, in which sites were selected based on the uncertainty of occupancy predicted from covariate relationships; i.e. SE($\Psi'_{i,1}$). This resulted in an updated mean $\Psi_{i,2}$ = 0.269, mean SE = 0.283, and mean CV = 0.815 (Table 2.1, Figure 2.2). This alternative reduced the SE of the site-specific occupancy estimate for 87% of sites (Table 2.2). The alternative that performed second best in reducing the uncertainty of the occupancy parameter was *A2*, in which sites were selected based on SE($\Psi_{i,1}$). This resulted in an updated mean $\Psi_{i,2}$ = 0.272, mean SE = 0.297, and mean CV = 0.824. This alternative reduced the SE of the site-specific occupancy estimate for 86% of sites.

The two alternatives that resulted in the least reduction in predicted uncertainty were *A1*, the status quo alternative and *A9*, the alternative that selected sites based on the uncertainty in occupancy predicted from habitat characteristics on the logit scale. Alternative *A1* resulted in an updated mean $\Psi_{i,2}$ = 0.342, mean SE = 0.483, and mean CV = 1.251 (Table 2.1, Figure 2.2). Only 62% of sites saw a reduction in the SE of the site-specific occupancy estimate under alternative *A1* (Table 2.2). Alternative *A9* resulted in an updated mean $\Psi_{i,2}$ = 0.330, mean SE =

0.440, and mean CV = 1.237 (Table 2.1, Figure 2.2), and 72% of sites saw a reduction in the SE of the site-specific occupancy estimate under this alternative (Table 2.2).

2.5 Discussion

Here, I provide an approach for constructing survey designs for occupancy analysis that optimizes management's monitoring objective with consideration for the limited resources available to conduct surveys. Designing monitoring for species data collection is a complex decision, often with competing objectives. The problem of where to survey for Common Loons to decrease overall uncertainty for Washington State was framed as an optimization problem with a constraint on available survey effort. Using this technique allowed me to break the problem down into structural elements, define discrete alternative decision rules, and find a solution. While the use of decision rules means that I cannot guarantee that the selected sites are globally optimal, this method provides a feasible approach and provides a mechanism for easily identifying survey sites in subsequent years, i.e., by updating the analysis and selecting sites based on the same decision rule. The Bayesian methods described here will facilitate updating after future data collection.

I found that surveying sites where the relationship between occupancy and site-specific covariates (i.e., alternative decision rule *A8*) had the greatest uncertainty was the best approach for reducing overall uncertainty in occupancy. Conducting surveys at such sites provides more precise inference to sites not surveyed, increasing the ability to detect trends in occupancy over time. Selecting sites based on the predicted standard error also reduced overall uncertainty, though not as much.

Surveying sites that had previously been surveyed by WDFW, *A1*, resulted in the highest predicted mean occupancy for Washington state, but this design did not provide as great a reduction in the overall uncertainty. The selection of sites previously surveyed by the WDFW likely were influenced by prior knowledge about the state of a site, where sites with a prior detection of a Common Loon were favored. This may result in bias and may occur in any data collection program with a non-random sampling scheme. Sampling at sites preferentially leads to model estimates biased towards the sites surveyed, but not the entire area of interest, and limits the capability to make inference to all sites. Covariate information may not fully account for this bias, although model-based approaches can to a degree improve accuracy in parameter estimates (Conn et al., 2017). This illustrates the risk of non-random sampling and surveying where we expect to find birds, both in terms of uncertainty and bias.

My approach considered discrete alternatives that fully utilized the available survey resources. Any survey design that did not fully use resources should be dominated by at least one survey design that did fully use resources (alternatives are said to be dominated if there is at least one other alternative that is better with respect to at least one characteristic and equal on all others). In other words, it can never be optimal to underuse available survey resources. I identified 9 decision rules used for selecting alternatives and can only identify the survey design that is best across the 9 resulting alternatives; i.e., my approach leads to a quasi-optimal strategy. Determination of the globally optimal survey design for all sites in the state of Washington would require every feasible non-dominated combination of sites to be evaluated through model fitting. In this application, the number of all possible combinations, selecting 10 sites from each district, would require evaluation through model fitting of approximately $3.657 * 10^{28}$ possible

unique survey design alternatives, which is clearly infeasible. I therefore used information on existing parameter estimates to limit the number of designs to test.

I used a bounded objective function approach and considered only a single constraint on effort, which was provided by WDFW management. A simple extension to this method is the epsilon constraint method (Haimes et al., 1971), which considers a suite of effort constraints and assesses the gains in precision for the effort spent during a survey season. Examining the relationship between increasing survey effort and occupancy parameter precision could provide managers with additional information for making refined decisions about monitoring Common Loons in Washington. Understanding this relationship would also serve as a sensitivity analysis of the optimal solution to changes in constraint values (Williams and Kendall, 2017).

The initial multi-state occupancy analysis was formulated using a Bayesian framework, and the results provided informative priors for parameters when fitting simulated data generated under the alternative designs. Fitting the simulated survey data for each alternative allowed me to update the posterior parameter estimates that were then used to predict occupancy at all sites in Washington state. Although the updated parameter estimates were from simulated data in this application, Bayesian updating could also be used after surveys are conducted. Iteratively updating the occupancy model with data and examining the resulting uncertainty will allow decision makers to dynamically choose survey sites (Wade, 2000). Each season, the survey sites that optimize management's monitoring objective given the new data could be selected based on the decision rule I identified as optimal (i.e., alternative decision rule *A8*). In this scheme, each year the Bayesian analysis would be updated, and the sites with the greatest uncertainty in covariate relationships would be identified for survey the following year.

The optimal survey design provides each district biologist with a survey strategy specifying which sites should be visited twice in a season. The overall strategy aims to produce high precision in parameter estimates, while remaining within the effort constraint defined by management. The approach of optimally allocating survey resources provides a practical approach to surveying Common Loons in Washington. Future Common Loon survey designs may be adjusted based on modifications to the available budget, updating of information gained through sampling, or in response to management needs or actions. Acquiring finer details about the effort spent surveying for Common Loons or travel time to sites, would increase the achievability of the survey design.

This work shows that designing surveys for use in data analysis and monitoring can be tailored in a way that accounts for species-specific details, management objectives, and constraints. In general, the decision rule discovered here would result in decreased uncertainty in other applications, i.e. collecting more information at sites where the relationship between the parameter of interest and relevant habitat covariates is uncertain will lead to decreased uncertainty in parameter estimates. A Bayesian approach was essential for capturing prior understanding and integrating it with future information. An interesting extension to this work would be to consider other objectives and compare the decision rule that is found to be optimal. Likewise, it would be possible to compare the optimal decision rule with the true optimal survey design alternative, where the number of unique alternatives is reasonable to evaluate. In comparing the truly optimal alternative and the optimal decision rule, further refinement to the approach here would be realized.

2.6 References

Ascough, J.C., Maier, H.R., Ravalico, J.K., Strudley, M.W., 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecol. Modell. 219, 383–399. https://doi.org/10.1016/j.ecolmodel.2008.07.015

Bailey, L.L., Hines, J.E., Nichols, J.D., MacKenzie, D.I., 2007. Sampling design trade-offs in occupancy studies with imperfect detection: Examples and software. Ecol. Appl. 17, 281–290. https://doi.org/10.1890/1051-0761(2007)017[0281:SDTIOS]2.0.CO;2

Barata, I.M., Griffiths, R.A., Ridout, M.S., 2017. The power of monitoring: Optimizing survey designs to detect occupancy changes in a rare amphibian population. Sci. Rep. 7, 1–9. https://doi.org/10.1038/s41598-017-16534-8

Buckland, S.T., Johnston, A., 2017. Monitoring the biodiversity of regions: Key principles and possible pitfalls. Biol. Conserv. 214, 23–34. https://doi.org/10.1016/j.biocon.2017.07.034

Clemen, R.T., 1996. Making Hard Decisions: An Introduction to Decision Analysis, 2nd ed. Duxbury Press.

Conn, P.B., Johnson, D.S., Boveng, P.L., 2015. On extrapolating past the range of observed data when making statistical predictions in ecology. PLoS One 10. https://doi.org/10.1371/journal.pone.0141416

Conn, P.B., Thorson, J.T., Johnson, D.S., 2017. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. Methods Ecol. Evol. 8, 1535–1546. https://doi.org/10.1111/2041-210X.12803

Converse, S.J., Shelley, K.J., Morey, S., Chan, J., LaTier, A., Scafidi, C., Crouse, D.T., Runge, M.C., 2011. A decision-analytic approach to the optimal allocation of resources for endangered species consultation. Biol. Conserv. 144, 319–329. https://doi.org/10.1016/j.biocon.2010.09.009

de Valpine, P., Turek, D., Paciorek, C.J., Lang, D.T., Bodik, R., 2017. Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. J. Comput. Graph. Stat. 26(2), 403–413.

Di Stefano, J., 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. Funct. Ecol. 17, 707–709. https://doi.org/10.1046/j.1365-2435.2003.00782.x

Ehrgott, M., 2005. Multiobjective programming., in: Figueria, J., Greco, S., Ehrgott, M. (Eds.), Multiple Criteria Decision Analysis: State of the Art Surveys. Springer, Boston, USA, pp. 667–722.

Ellis, M.M., Ivan, J.S., Schwartz, M.K., 2014. Spatially Explicit Power Analyses for Occupancy-Based Monitoring of Wolverine in the U.S. Rocky Mountains. Conserv. Biol. 28, 52–62. https://doi.org/10.1111/cobi.12139

Ellison, A.M., 2004. Bayesian inference in ecology. Ecol. Lett. 7, 509–520. https://doi.org/10.1111/j.1461-0248.2004.00603.x

Field, S.A., O'Connor, P.J., Tyre, A.J., Possingham, H.P., 2007. Making monitoring meaningful.

Austral Ecol. 32, 485–491. https://doi.org/10.1111/j.1442-9993.2007.01715.x

Field, S.A., Tyre, A.J., Jonzén, N., Rhodes, J.R., Possingham, H.P., 2004. Minimizing the cost of environmental management decisions by optimizing statistical thresholds. Ecol. Lett. 7, 669–675. https://doi.org/10.1111/j.1461-0248.2004.00625.x

Field, S.A., Tyre, A.J., Possingham, H.P., 2005. Optimizing Allocation of Monitoring Effort Under Economic and Observational Constraints. J. Wildl. Manage. 69, 473–482. https://doi.org/10.2193/0022-541x(2005)069[0473:oaomeu]2.0.co;2

Gerber, L.R., Beger, M., McCarthy, M.A., Possingham, H.P., 2005. A theory for optimal monitoring of marine reserves. Ecol. Lett. 8, 829–837. https://doi.org/10.1111/j.1461-0248.2005.00784.x

Gerber, L.R., Runge, M.C., Maloney, R.F., Iacona, G.D., Drew, C.A., Avery-Gomm, S., Brazill-Boast, J., Crouse, D., Epanchin-Niell, R.S., Hall, S.B., Maguire, L.A., Male, T., Morgan, D., Newman, J., Possingham, H.P., Rumpff, L., Weiss, K.C.B., Wilson, R.S., Zablan, M.A., 2018. Endangered species recovery: A resource allocation problem. Science (80-.). 362, 284–286. https://doi.org/10.1126/science.aat8434

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

Guillera-Arroita, G., Ridout, M.S., Morgan, B.J.T., 2010. Design of occupancy studies with imperfect detection. Methods Ecol. Evol. 1, 131–139. https://doi.org/10.1111/j.2041-210x.2010.00017.x

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A., Zimmermann, N.E., 2006. Using niche-based models to improve the sampling of rare species. Conserv. Biol. 20, 501–511. https://doi.org/10.1111/j.1523-1739.2006.00354.x

Haimes, Y., Lasdon, L., Wismer, D., 1971. On a bicriterion formulation of the problems of integrated system identification and system optimization. IEEE Trans. Syst. Man, Cybern. 1 296–297.

Legg, C.J., Nagy, L., 2006. Why most conservation monitoring is, but need not be, a waste of time. J. Environ. Manage. 78, 194–199. https://doi.org/10.1016/j.jenvman.2005.04.016

Lindenmayer, D.B., Likens, G.E., 2010. The science and application of ecological monitoring. Biol. Conserv. 143, 1317–1328. https://doi.org/10.1016/j.biocon.2010.02.013

Lyons, J.E., Runge, M.C., Laskowski, H.P., Kendall, W.L., 2008. Monitoring in the Context of Structured Decision-Making and Adaptive Management. J. Wildl. Manage. 72, 1683–1692. https://doi.org/10.2193/2008-141

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, A.B., 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84, 2200–2207. https://doi.org/10.1890/02-3090

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.A., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83, 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., Hines, J.E., 2018. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence, 2nd ed. Academic Press.

MacKenzie, D.I., Royle, J.A., 2005. Designing occupancy studies: General advice and allocating survey effort. J. Appl. Ecol. 42, 1105–1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x

Marler, R.T., Arora, J.S., 2004. Survey of multi-objective optimization methods for engineering. Struct. Multidiscip. Optim. 26, 369–395. https://doi.org/10.1007/s00158-003-0368-6

Martin, T.G., Chadès, I., Arcese, P., Marra, P.P., Possingham, H.P., Norris, D.R., 2007. Optimal Conservation of Migratory Species 3–7. https://doi.org/10.1371/journal.pone.0000751

McCarthy, M.A., 2014. Contending with uncertainty in conservation management decisions. Ann. N. Y. Acad. Sci. 1322, 77–91. https://doi.org/10.1111/nyas.12507

McCarthy, M.A., Thompson, C.J., Hauser, C., Burgman, M.A., Possingham, H.P., Moir, M.L., Tiensin, T., Gilbert, M., 2010. Resource allocation for efficient environmental management. Ecol. Lett. 13, 1280–1289. https://doi.org/10.1111/j.1461-0248.2010.01522.x

McIntyre, J.W., 1998. The Common Loon: Spirit of Northern Lakes. Minneapolis: University of Minnesota Press.

Mendoza, G.A., Martins, H., 2006. Multi-criteria decision analysis in natural resource management: A critical review of methods and new modelling paradigms. For. Ecol. Manage. 230, 1–22. https://doi.org/10.1016/j.foreco.2006.03.023

Moilanen, A., Runge, M.C., Elith, J., Tyre, A., Carmel, Y., Fegraus, E., Wintle, B.A., Burgman, M., Ben-Haim, Y., 2006. Planning for robust reserve networks using uncertainty analysis. Ecol. Modell. 199, 115–124. https://doi.org/10.1016/j.ecolmodel.2006.07.004

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L., Sauer, J.R., 2002. Large scale wildlife monitoring studies: Statistical methods for design and analysis. Environmetrics 13, 105–119. https://doi.org/10.1002/env.514

Possingham, H.P., Andelman, S.J., Noon, B.R., Trombulak, S., Pulliam, H.R., 2001. Making smart conservation decisions. Research Priorities for Conservation Biology. Island Press, Washington, D.C.

R Core Team, 2018. R: A language and environment for statistical computing.

Reynolds, J.H., Thompson, W.L., Russell, B., 2011. Planning for success: Identifying effective and efficient survey designs for monitoring. Biol. Conserv. 144, 1278–1284. https://doi.org/10.1016/j.biocon.2010.12.002

Richardson, S., Hays, D., Spencer, R., Stofel, J., 2000. Washington state status report for the common loon. Washingt. Dep. Fish Wildlife, Olympia. 53.

Sewell, D., Guillera-Arroita, G., Griffiths, R.A., Beebee, T.J.C., 2012. When is a species declining? optimizing survey effort to detect population changes in reptiles. PLoS One 7. https://doi.org/10.1371/journal.pone.0043387

Wade, P.R., 2000. Bayesian Methods in Conservation Biology. Conserv. Biol. 14, 1308–1316.

Williams, B.K., 2011. Adaptive management of natural resources-framework and issues. J.

Environ. Manage. 92, 1346–1353. https://doi.org/10.1016/j.jenvman.2010.10.041

Williams, P.J., Kendall, W.L., 2017. A guide to multi-objective optimization for ecological problems with an application to cackling goose management. Ecol. Modell. 343, 54–67. https://doi.org/10.1016/j.ecolmodel.2016.10.010

Wilson, K.A., McBride, M.F., Bode, M., Possingham, H.P., 2006. Prioritizing global conservation efforts. Nature 440, 337–340. https://doi.org/10.1038/nature04366

Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time. Trends Ecol. Evol. 16, 446–453. https://doi.org/10.1016/S0169-5347(01)02205-4

2.7 FIGURES AND TABLES

Table 2.1 Mean coefficient of variation (CV), mean standard error (SE), and mean occupancy probability estimates across all sites for the initial occupancy analysis (*M1*) and each alternative (*A1-A9*). For each alternative the updated parameter estimates after fitting simulated data were used to generate occupancy probability estimates for all sites and their associated standard errors.

Alternative	CV	SE	Mean Occupancy
M1	1.3124	0.4863	0.3662
A1[1]	1.2508	0.4828	0.3416
A2[2]	0.8238	0.2968	0.2724
A3[3]	0.9845	0.3688	0.2908
A4[4]	0.9096	0.3204	0.2753
A5[5]	0.8661	0.3116	0.2777
A6[6]	0.9275	0.3406	0.2841
A7[7]	0.8700	0.3044	0.2699
A8[8]	0.8150	0.2833	0.2688
A9[9]	1.2372	0.4400	0.3296

1. Sites previously surveyed by the WDFW biologists in the year 2017, n = 53 sites
2. Sites with highest standard errors, max(SE($\Psi_{i,1}$))
3. Sites with highest coefficients of variation, max(SE($\Psi_{i,1}$)/ mean($\Psi_{i,1}$))
4. Sites with highest estimated mean occupancy probability, max(mean($\Psi_{i,1}$)))
5. Sites with highest standard errors on the logit scale, max(SE(logit($\Psi_{i,1}$)))
6. Sites with highest coefficients of variation on the logit scale, max (SE(logit($\Psi_{i,1}$))/mean(logit($\Psi_{i,1}$)))
7. Sites with highest estimated mean occupancy probability on the logit scale, max(mean(logit($\Psi_{i,1}$)))
8. Sites with highest variance in occupancy probability estimated from covariate relationships, max(SE($\Psi'_{i,1}$))
9. Sites with highest variance in occupancy probability estimated from covariate relationships on the logit scale, max(SE(logit($\Psi'_{i,1}$))

Table 2.2 The number and percentage of sites with decreased mean standard error compared with the mean initial occupancy analysis standard error, with the number of sites sampled and the percentage of the total sites sampled (total sites = 1268). The number of sites sampled is $n = 53$ in *A1*, and $n = 163$ in *A2 – A9*.

Alternative	Number of sites where $SE(\Psi_{i,2,a}) - SE(\Psi_{i,1}) > 0$	Proportion of sites where $SE(\Psi_{i,2,a}) - SE(\Psi_{i,1}) > 0$
A1[1]	788	0.62
A2[2]	1096	0.86
A3[3]	1024	0.81
A4[4]	1046	0.82
A5[5]	1082	0.85
A6[6]	1066	0.84
A7[7]	1069	0.84
A8[8]	1099	0.87
A9[9]	914	0.72

1. Sites previously surveyed by the WDFW biologists in the year 2017, $n = 53$ sites
2. Sites with highest standard errors, $\max(SE(\Psi_{i,1}))$
3. Sites with highest coefficients of variation, $\max(SE(\Psi_{i,1})/ \text{mean}(\Psi_{i,1}))$
4. Sites with highest estimated mean occupancy probability, $\max(\text{mean}(\Psi_{i,1})))$
5. Sites with highest standard errors on the logit scale, $\max(SE(\text{logit}(\Psi_{i,1})))$
6. Sites with highest coefficients of variation on the logit scale, $\max (SE(\text{logit}(\Psi_{i,1}))/\text{mean}(\text{logit}(\Psi_{i,1})))$
7. Sites with highest estimated mean occupancy probability on the logit scale, $\max(\text{mean}(\text{logit}(\Psi_{i,1})))$
8. Sites with highest variance in occupancy probability estimated from covariate relationships, $\max(SE(\Psi'_{i,1}))$
9. Sites with highest variance in occupancy probability estimated from covariate relationships on the logit scale, $\max(SE(\text{logit}(\Psi'_{i,1}))$

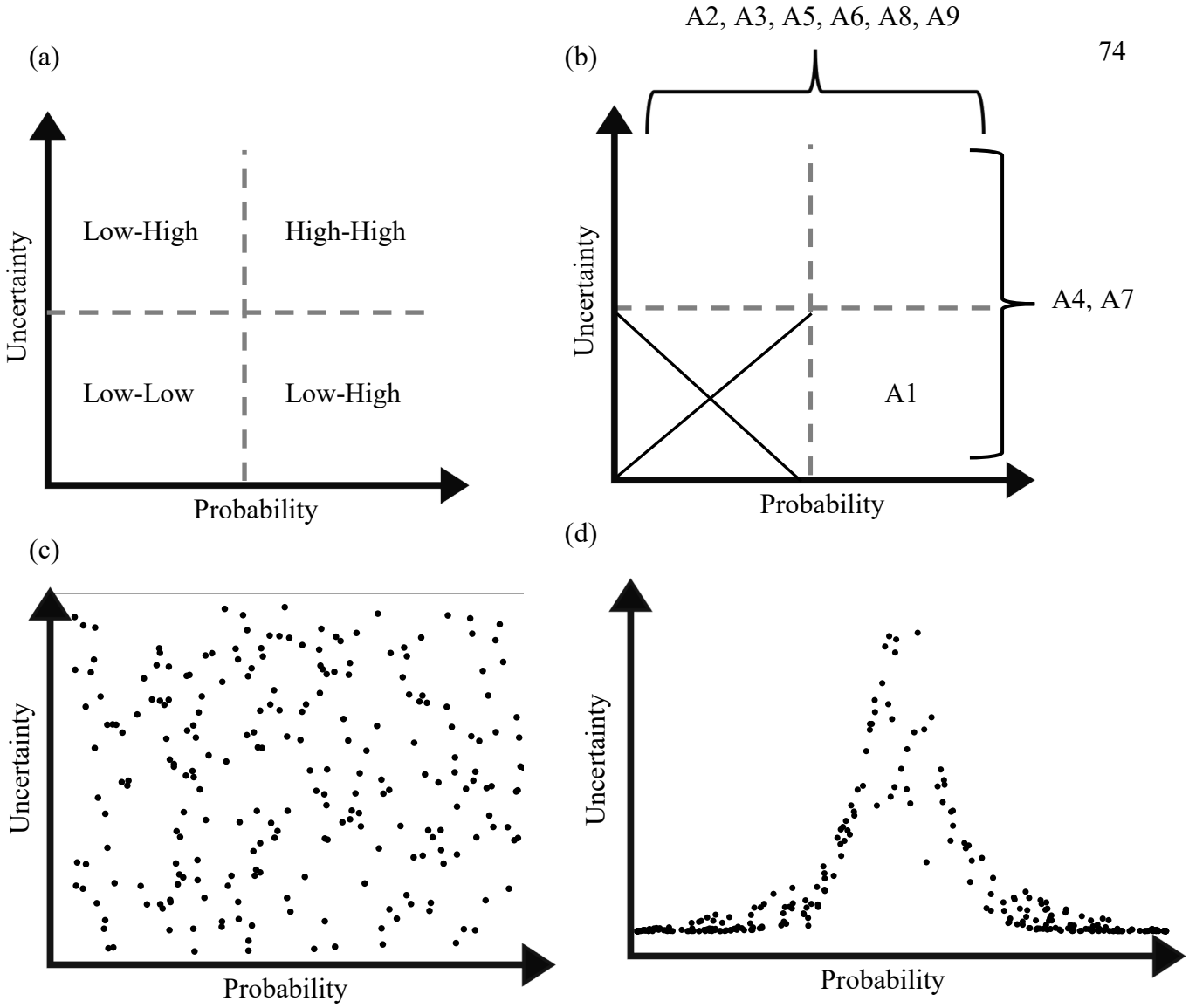

Figure 2.1 Categorization of site-specific occupancy parameter estimates and their associated uncertainty. (a) Sites with low probability and low uncertainty are not considered a priority for survey design, as they will not serve to reduce uncertainty or be likely to have Common Loons present. Surveying sites with high uncertainty (with low or high probability) potentially represent the best opportunity to reduce uncertainty in occupancy parameter estimates for sites surveyed and in overall model estimates. Sites with high probability and high uncertainty may be valuable to decrease uncertainty, and may also lead to discoveries of occupied sites that have not previously been surveyed by trained biologists. Sites with high probability and low uncertainty are sites that are likely to be occupied, if they have not previously been surveyed by the WDFW they may represent an opportunity for validating model predictions. (b) Where the alternatives fall within the categorization of site-specific probability and uncertainty, notice no alternatives represent the 'low-low' category. Alternatives *A2, A3, A5, A6, A8,* and *A9* represent sites with high uncertainty, alternatives *A4* and *A7* represent sites with high probability, and *A1* is the alternative representing the WDFW surveying the same sites as they have previously. (c) An example of occupancy probability estimates and their uncertainty on the logit scale and (d) those same estimates transformed to the real scale.

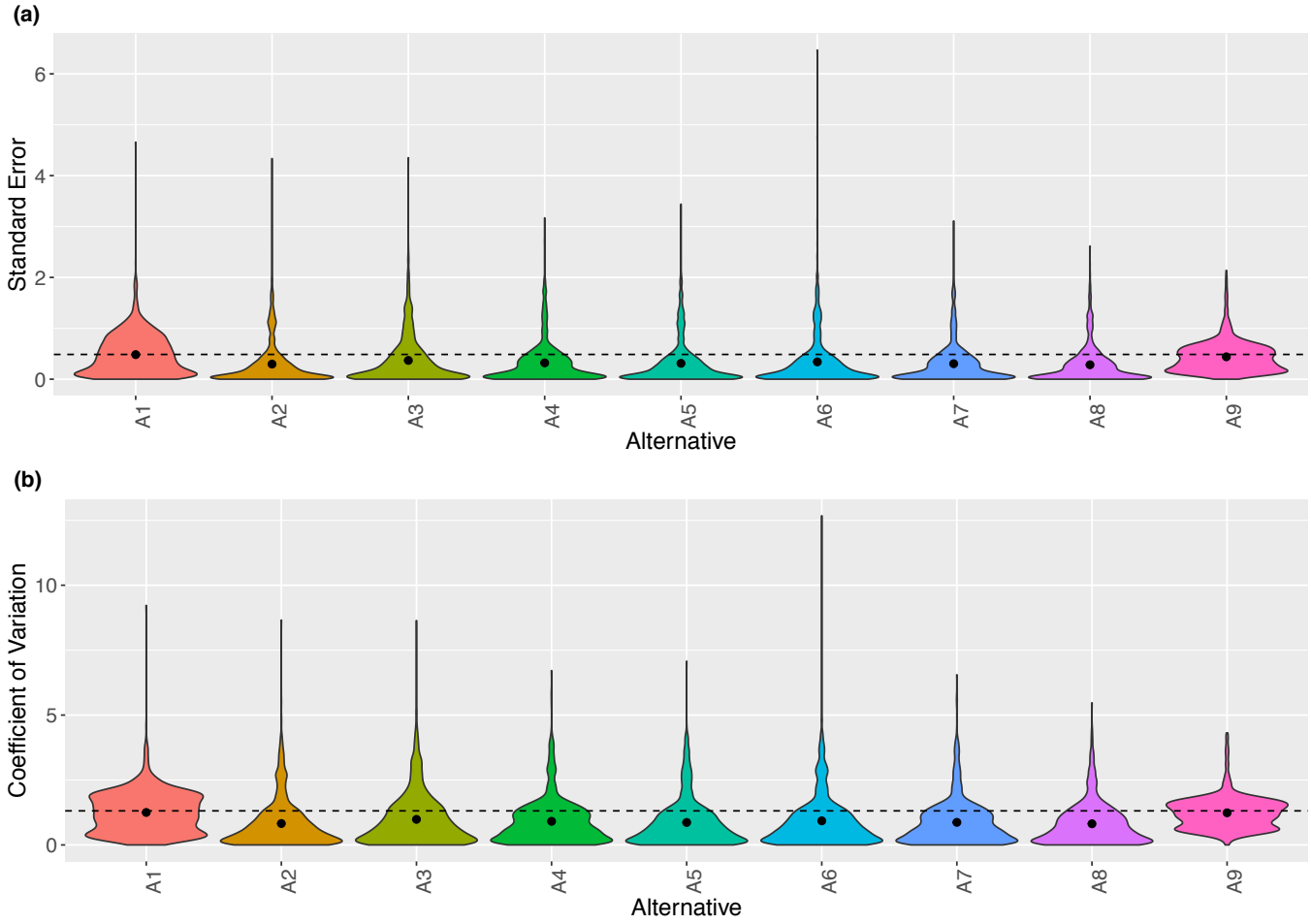

Figure 2.2 Violin plots of (a) the estimated mean standard error of sites by alternative and (b) the estimated mean coefficient of variation of sites by alternative. The dotted line in each represents the standard error and coefficient of variation from the initial occupancy analysis (SE = 0.486 and CV = 1.312).

2.8 APPENDIX 2

The updated multi-state occupancy model NIMBLE code, covariates are expressed as in Appendix 1. The prior information from the initial multi-state occupancy analysis and model was adapted to only use observations from one type of data.

```
mod<-nimbleCode({
  delta ~ dnorm(0.7659094,0.06103023)
  l.psi~ dnorm(-0.4460710, 0.2870509)
  b.area~dnorm( 1.981322, 1.583209)
  b.peri~dnorm(-0.4083596, 0.9207903)
  b.elev~dnorm(-1.1713474, 0.5762513)
  b.hii~dnorm(-1.0022639, 0.3477032)
  b.pc~dnorm(0.7811388, 0.5270083)
  b.canopy~dnorm(-0.01366882, 0.36577280)
  b.focal~dnorm(-0.9161131, 0.7452319)
  b.11~dnorm(-0.9671331, 1.1371946)
  b.2~dnorm(-0.4628712, 1.0767284)
  b.31~dnorm(-0.9368070, 0.8520912)
  l.R ~ dnorm(-0.2063482,0.3120782)
  r.area~dnorm(0.2097241, 1.0433757)
  r.hii~dnorm(-1.3088967, 0.8658458)
  r.elev~dnorm(0.8242658, 0.8278252)
  r.peri~dnorm(-1.197198,1.657806)
  r.pc~dnorm(0.0964534, 0.6775097)
  r.focal~dnorm(25.53898, 19.93839)
  r.canopy~dnorm(0.5553999, 0.6366451)
  r.11~dnorm(-1.669633, 2.425677)
  r.2~dnorm( 0.2226298, 1.3009147)
  r.31~dnorm(-0.5002281, 1.0731928)
  l.p1 ~ dnorm(-2.5523798, 0.1301073)
  l.p2 ~ dnorm(-0.2362560, 0.2159415)
  p1.area~dnorm(-0.03066881, 1.97512209)
  p2.area~dnorm(-1.01529, 3.37934)
  b.surv1~dnorm(3.111267, 2.006689)
  b.surv2~dnorm(3.166029, 1.121168)
  sd_alpha_psi~dunif(0,10)
  tau_alpha_psi<-pow(sd_alpha_psi, -2)
  sd_alpha_R~dunif(0,10)
  tau_alpha_R<-pow(sd_alpha_psi, -2)
  alpha_psi~dnorm(0,tau_alpha_psi)
  alpha_R~dnorm(0,tau_alpha_R)
  sd_alpha_psiN~dnorm(3.464766,0.7154881)
  tau_alpha_psiN<-pow(sd_alpha_psiN, -2)
  sd_alpha_RN~dnorm(1.098694, 0.749066)
  tau_alpha_RN<-pow(sd_alpha_RN, -2)
  for(i in 1:N){
    mu_e[i]<-mu_alphaN(a1=Amu[i])
    sd_e[i]<-sd_alphaN(a2=Asd[i], tau=tau_alpha_psi)
```

```
    alpha_psiN[i]~dnorm(mu_e[i],sd_e[i])
    mu_eR[i]<-mu_alphaN(a1=ARmu[i])
    sd_eR[i]<-sd_alphaN(a2=ARsd[i], tau=tau_alpha_RN)
    alpha_RN[i]~dnorm(mu_eR[i],sd_eR[i])
   }
   for(i in 1:N){
    phi1[i,1]<-(1-psipred[i])
    phi1[i,2]<-psipred[i]*(1-Rpred[i])
    phi1[i,3]<-psipred[i]*Rpred[i]
    zt1[i]~ dcat(phi1[i,1:3])
    f[i]<-getFocalOcc(zt1[i])
    fr[i]<-getFocalR(zt1[i])
    logit(psi[i])<-l.psi +alpha_psi+alpha_psiN[i]+
      b.focal*f[i]+
      b.hii*hii[i]+b.elev*elev[i]+b.pc*pc[i]+
      b.area*area[i]+b.peri*peri[i]+b.canopy*canopy[i]+
      b.11*x11[i]+b.2*x2[i]+b.31*x31[i]
    logit(R[i]) <- l.R +alpha_RN[i]+alpha_R+r.area*area[i]+r.hii*hii[i]+r.elev*elev[i]+
      r.focal*fr[i]+
      r.peri*peri[i]+r.pc*pc[i]+ r.canopy*canopy[i]+
      r.11*x11[i]+r.2*x2[i]+r.31*x31[i]
   }
  for(i in 1:N){
    z[i] ~ dcat(phi[i,1:3])
    phi[i,1]<-1-psi[i]
    phi[i,2]<-psi[i]*(1-R[i])
    phi[i,3]<-psi[i]*R[i]
   }
   for(i in 1:N){
    for(j in 1:J1){
      logit(p1[i,j]) <- l.p1+b.surv1+p1.area*area[i]
      logit(p2[i,j]) <- l.p2+b.surv2+p2.area*area[i]

      pvec[i,j,1:3]<-obsFun(p11=p1[i,j], p22=p2[i,j],z=z[i],delta=delta)
      y[i,j] ~ dcat(pvec[i,j,1:3])
    }
   }
})
obsFun<-nimbleFunction(
  run=function( p11=double(), p22=double(), z=integer(), delta=double()){
    p<-matrix(nrow=3, ncol=3)
    p[1,1:3]<-c(1,0,0)
    p[2,1:3]<-c((1-p11), p11,0)
    p[3,1:3]<-c((1-p22), p22*(1-delta), p22*delta)
    pvec<-p[z,1:3]
    return(pvec[1:3])
    returnType(double(1))
  })
getFocalOcc<-nimbleFunction(
  run=function(z=double()){
    if(z>1) return(1)
```

```
    else return(0)
    returnType(integer())
  })
getFocalR<-nimbleFunction(
  run=function(z=double()){
    if(z>2) return(1)
    else return(0)
    returnType(integer())
  })
mu_alphaN<-nimbleFunction(
  run=function(a1=double()){
    if(a1< -999) return(0)
    else return(a1)
    returnType(double())
  })
sd_alphaN<-nimbleFunction(
  run=function(a2=double(), tau=double()){
    if(a2< -999) return(tau)
    else return(a2)
    returnType(double())
  })
```