AN INVESTIGATION OF STUDY DESIGN AND PRIOR SELECTION IN BAYESIAN OCCUPANCY MODELS

by

Meaghan Elizabeth Winder

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

May 2025

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor Andrew Hoegh for his commitment, patience, support, and guidance. In many ways Andy is the educator, researcher, and advisor I aspire to become, and I am extremely fortunate to have had the opportunity to work with him during my time at Montana State University. I am grateful for the time, support, and guidance offered by the rest of my committee members: Katharine Banner, Adam Sepulveda, and John Smith.

I would like to express thanks to Christian Stratton for his endless support; I truly could not have done it without him. Thanks to him, the push to the finish line has been fueled by caffeinated beverages and little treats. I thank my sister, Lauren for her true friendship and support; I am lucky so much of our time in Bozeman overlapped. I would like to thank my parents Jason and Kristin, as well as my grandparents, for their love, encouragement, and endless support.

I would also like to thank Mark Greenwood for his guidance and support during my time at SCRS and beyond. A huge thanks to Katie Sutich, Jane Crawford, and the graduate program coordinators for the Department of Mathematical Sciences for their tireless support. I would also like to thank the faculty in the Department of Mathematics and Statistics at Middlebury College for their support, kind advice, and mentorship during this past year.

Without the support of each of these people, none of this would have been possible. I am truly grateful for my time at Montana State University and all of the relationships I have built in that time.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF TABLES – CONTINUED

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

ABSTRACT

Ecological data are inherently complex and often require making inferences about quantities that are not directly observed. To estimate these unobserved quantities, statisticians typically employ hierarchical modeling techniques. Detection/non-detection data, often referred to as occupancy data, are collected by ecologists and biologists to address a wide range of ecological questions; these inquiries may include, but are not limited to, questions about species distributions, changes in species distributions over time, or species detectability. Occupancy data are particularly well suited for analysis using hierarchical occupancy models. In this work, we present novel methods and recommendations for a variety of Bayesian occupancy models, focusing on improving accessibility for practicing scientists; in our work, we emphasize both study design and prior selection as critical components for enhancing analyses to better address ecological questions. Specifically, we identify methodological modifications and sampling recommendations to guide early detection monitoring of invasive dreissenid mussels in the western United States. We provide sampling and modeling recommendations for long-term ecological monitoring projects in the presence of incomplete data. We offer a flexible method for implementing Bayesian regularizing priors in the occupancy modeling framework, and offer an accessible tool for implementing the technique in R.

1

INTRODUCTION

Occupancy data, or detection-nondetection data, are collected by ecologists and biologists to investigate a wide range of ecological questions; including monitoring of invasive species (Sepulveda 2018b; Sepulveda et al. 2019), threatened species (Boone et al. 2023; Campos-Cerqueira and Aide 2016), plants (Middleton and Vining 2022; Penman et al. 2009), and pathogens (Lachish et al. 2012; Mosher et al. 2019). Occupancy data are valuable for learning about species distributions, however they are typically sparse, noisy, or riddled with imperfect detection. Since 2002, the occupancy model developed by MacKenzie et al. (2002) has been a crucial tool for modeling these data, as the model incorporates imperfect detection of the species of interest. Expanding on the framework developed by MacKenzie et al. (2002), models that account for false positives (Chambert et al. 2018; McClintock et al. 2010; Miller et al. 2013, 2011; Royle and Link 2006), multi-scale occupancy models (Emmet et al. 2021; Guillera-Arroita et al. 2011; Hines et al. 2010; Nichols et al. 2008), multi-species models (Dorazio et al. 2011; Guillera-Arroita 2017; MacKenzie et al. 2004; Miller et al. 2012), multi-state models (MacKenzie et al. 2009), and dynamic models (MacKenzie et al. 2003, 2018; Royle and Dorazio 2008; Royle and Kéry 2007) have been developed to answer more complex ecological questions.

Ecological systems are inherently complex and often involve making inference about quantities that are not directly observed; in order to estimate these latent quantities, scientists can utilize a hierarchical sampling design. Poorly designed studies can lead to biased estimates, incorrect ecological inferences, and unreliable predictions (Yoccoz et al. 2001). Alternatively, careful study design helps capture the complexity of ecological systems while minimizing noise (MacKenzie et al. 2002). Similarly, thoughtful Bayesian priors incorporate biological knowledge, while allowing the data to shape inference (McCarthy and Masters 2005). Additionally, when there are many environmental covariates, priors can be used to control model complexity and prevent overfitting. Thoughtful prior specification and study design leads to more reliable results, which are critical for making decisions regarding conservation, management, and policy (Morris et al. 2015).

This dissertation is a compilation of four manuscripts that use a variety of hierarchical occupancy models to make inferences about ecological parameters of interest while accounting for imperfect detection. In each of the manuscripts, there is a focus on making these models accessible to practicing scientists, considering study design, and prior selection as methods for improving analyses to better address ecological questions. Additional background information is provided for each manuscript below.

Primer for Chapter 2

In Chapter 2: "An initial assessment of plankton tow detection probabilities for dreissenid mussels in the western United States" we analyze a dreissenid mussel long-term monitoring data set collected by the U.S. Bureau of Reclamation and their partners. Dreissenids are invasive mussels first discovered in the United States in the late 1980s. Since their discovery in the Great Lakes, dreissenids have spread rapidly across the continental United States. Early detection of the invasive mussels plays a crucial role in accelerating action to contain and control the species. If a population of dreissenids becomes established, there is very little that can be done to remove them. Consequently, tens of millions of dollars (USD) are invested annually to prevent, contain, and control the spread of dreissenid mussel in North America; however, the sampling effort required for early detection has received little attention (Department of the Interior 2020).

In Chapter 2, our novel contribution is the guidance about the early detection sampling effort required to ensure a high probability of detecting the mussels if they are present at low densities. In addition to the sampling recommendations, we identify methodological modifications that could improve the accuracy and precision of future detection probability estimates.

Primer for Chapter 3

In Chapter 2, we model the long-term dreissenid mussel monitoring data set with the implicit dynamic occupancy model and provide initial guidance for sampling effort required to detect the species at low densities. Some water bodies were added to the monitoring effort after the project had begun, and others were simply not visited every year (Figure 1); as a result, we saw a need to investigate the impact of incomplete sampling on the modeling of dynamic occupancy data. Here, 'incomplete sampling' refers to a sampling scenario where there are one or more sites that were not visited at least once during each season.

Since sampling all sites repeatedly for multi-year studies can be complicated, due to limited funding, sites being difficult to access, or the addition of sites after monitoring has begun, dynamic occupancy data are often riddled with incomplete sampling. In Chapter 3: "Review of Bayesian dynamic occupancy models and the implications of incomplete sampling," we set a strong foundation for the simulation and data analysis with a review of dynamic occupancy literature. We define the standard dynamic occupancy model (MacKenzie et al. 2002), the seminal dynamic occupancy model (MacKenzie et al. 2003), and the various re-parameterizations of dynamic occupancy models (MacKenzie et al. 2018; Royle and Dorazio 2008; Royle and Kéry 2007). Then, with simulation, we explore the implications of incomplete sampling on the inferences gleaned from the different parameterizations. In addition to simulation, we apply these models to a data set monitoring American Bullfrog breeding populations along the Yellowstone River floodplain in Montana, USA. We compare and synthesize the results from the simulation and the results of the bullfrog data analysis, and make both modeling and sampling recommendations for long-term occupancy monitoring projects.

In Chapter 3, our novel contribution is the exploration of the impact of incomplete sampling on the various parameterizations of dynamic occupancy models. Prior to our work, there has been little to no formal exploration of the ramifications of incomplete sampling on the results gleaned from the various dynamic occupancy modeling frameworks. Using our results, we are able to make both sampling and modeling recommendations for ecological long-term occupancy monitoring projects, and further our understanding about the impact of incomplete sampling on the model results.

Primer for Chapter 4

The Bayesian framework lends itself nicely to regularization; in Chapter 4 we consider regularization of occupancy models through the lens of the regularized horseshoe prior. The horseshoe prior was introduced by Carvalho et al. (2009, 2010) and induces model sparsity with two parameters: a local shrinkage parameter ( ) and a global shrinkage parameter ( ). A few years later, Piironen and Vehtari (2017a,b) adapted the horseshoe prior to create the regularized horseshoe prior, which helps address a few of the key issues with the original horseshoe prior. Along with the regularized horseshoe prior, Piironen and Vehtari (2017a,b) propose an alternative hyper-prior for the global shrinkage parameter that incorporates prior knowledge about the number of relevant coefficients in the model (Eq. 1). In this formulation, p_0 represents the guess for the number of relevant coefficients, D represents the total number of potential coefficients.

$$j \qquad \mathrm{C}^+(0; {}^2_0) \text{ where } {}_0 = \frac{p_0}{D \quad p_0} \, p\overline{n} \tag{1}$$

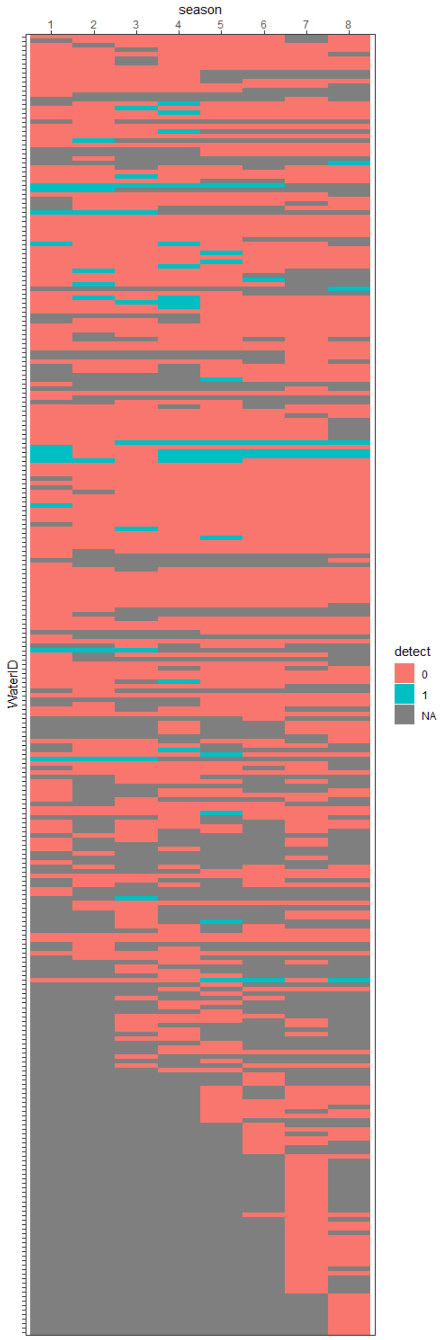

Figure 1: A plot of the naive occupancy estimates for each of the water bodies in the western United States monitored in the Bureau of Reclamation data. Grey indicates that there were no observations collected at the water body during that particular season. A 0 indicates that there were no detections of dreissenid mussel veligers at the water body during that season, whereas 1 indicates that dreissenid mussel veligers were detected at the water body at least once during that particular season.

Though the horseshoe prior has been around for over a decade, there is little agreement for how to carry out inference for the global shrinkage parameter. It is typically agreed upon that a fully Bayesian approach should be used, however there is software available that allows the user to specify a fixed value for using an empirical Bayes estimate (van der Pas et al. 2019). Piironen and Vehtari (2017b) show that the prior recommendation by Carvalho et al. (2009) (Eq. 2) which has become the default choice, tends to favor models that are too large, which is especially apparent in situations where is weakly identified by the data.

$$\mathrm{C}^{+}(0; 1) \tag{2}$$

In Chapter 4 we show the sensitivity of the Piironen and Vehtari (2017b) global shrinkage hyper-prior to each of its components; we show that an incorrect guess for the number of relevant coefficients in the model can have a large impact on the size of the resulting model. We then propose a novel method for the specification of p_0. We show that using a thoughtful prior distribution on the guess for the number of relevant coefficients helps mitigate the risk of an incorrect guess while incorporating uncertainty associated with not knowing p_0. We then demonstrate the mixture prior on both simulated synthetic data as well as koala occupancy data and discuss examination of sensitivity to the chosen mixture prior as an area for future research.

Primer for Chapter 5

Chapter 5 describes the existing R packages for implementing the horseshoe prior or its variants. We discuss the limitations of the existing implementations and show that fitting an occupancy model using the regularized horseshoe prior involves using Stan and `brms` (Bürkner 2021) which requires marginalizing out the discrete latent variables. Alternatively, other probabilistic programming languages must be used to fit occupancy models with the horseshoe prior.

We introduce an accessible and flexible framework for regularizing occupancy models using a mixture of truncated Normal distributions as a prior on p_0 with the introduction of the `horseshoeocc` package. The `horseshoeocc` R package offers ecologists and statisticians a user-friendly tool for extracting meaningful insights from high-dimensional data. Chapter 5, contains the manuscript written to showcase the package which provides an important contribution to Bayesian occupancy modeling in R.

AN INITIAL ASSESSMENT OF PLANKTON TOW DETECTION PROBABILITIES FOR DREISSENID MUSSELS IN THE WESTERN UNITED STATES

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Meaghan Winder

Contributions: Responsible for writing of the manuscript and submission, led data analysis.

Co-Author: Adam J Sepulveda

Contributions: Provided motivation for the work, assisted in gathering covariate information, helped with writing the introduction, discussion, and conclusion, provided feedback on manuscript, and assisted with response to reviewer comments.

Co-Author: Andrew Hoegh

Contributions: Provided motivation for the work, offered guidance and feedback on manuscript and submission process, and assisted with response to reviewer comments.

Manuscript Information

Meaghan Winder, Adam J Sepulveda, Andrew Hoegh

Management of Biological Invasions

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
_X_ Published in a peer-reviewed journal; this chapter is presented as the manuscript appears in the journal

REABIC
Submitted 7 February 2022
Published 8 August 2022
Volume 13, Issue 4
DOI: 10.3391/mbi.2022.13.4.05

Abstract

Early detection of dreissenid mussels (*Dreissena polymorpha* and *D. rostriformis bugensis*) is crucial to mitigating the economic and environmental impacts of an infestation. Plankton tow sampling is a common method used for early detection of dreissenid mussels, but little is known about the sampling intensity required for a high probability of early detection using the method. We used implicit dynamic occupancy models to estimate plankton tow detection probabilities of dreissenid mussels from a long-term data set containing plankton tow samples collected across central and western United States. We fit models using a) the entire data set, including water bodies with unknown occupancy status in addition to heavily infested water bodies, b) a data subset that included water bodies with paired water temperature data, and c) a data subset that included water bodies with lower dreissenid densities. For the entire data set, we found that estimated detection probabilities varied by water body size and ranged from approximately 0.10 to 0.86. For the water temperature subset, we observed the same pattern between detection probability and water body size as we did for the full data but additionally found that the estimated detection probabilities were much higher when water temperatures were above 12 C. For the lower dreissenid density subset, we found that the estimated probability of detecting dreissenid mussels with a single aggregated plankton tow sample was near zero. Given these estimates, we conclude that the number of aggregated plankton tow samples taken per water body in the data is far fewer than the number needed to ensure a high probability of detecting dreissenid mussels, especially if they are at low densities. We summarize the analyses with a discussion of plankton tow sampling protocol changes needed to improve estimates of dreissenid detection probabilities.

Introduction

Early detection surveillance is a pillar of invasive species management because early detection can accelerate actions that effectively contain and control the invader (Vander Zanden et al. 2010). However, the effort required to detect an invader shortly after its initial introduction is a critical knowledge gap in many invasive species monitoring programs (Trebitz et al. 2017). Introduction events are characterized by rare individuals at low densities, so sampling efforts that are too low are likely to miss these introductions. Delayed detections could allow an incipient population to become established, prevent any practical control efforts, and increase the magnitude of ecological and socio-economic impacts (Spear et al. 2021). Since resources are typically limited, managers need guidance on cost-effective sampling effort allocation.

The sampling effort required for reliable detection of low-density populations is difficult to determine experimentally since it is intractable to introduce a novel species into most natural systems. Observational studies are also challenged by imperfect detection, where individuals are present at a site, but fail to be detected during a survey because of environmental conditions, behavior, or inadequacies in the sampling method (Bailey et al. 2004). Occupancy models provide a tool for gaining insight about the relationship between sampling effort and detection probability (p) in natural systems. These models incorporate information from independent surveys of the same site to account for imperfect detection and estimate site-level occupancy and survey-level detection probability (MacKenzie et al. 2002). Occupancy models have resulted in considerable variation in low-density detection probabilities as functions of survey methods and site-specific covariates. For example, Schmelzle and Kinziger (2015) found that detection probabilities of endangered tidewater goby (***Eucyclogobius newberryi*** (Girard, 1856)) using environmental DNA (eDNA) sampling-were nearly double that of non-molecular methods, whereas Tank et al. (2021) found a threefold difference in detection probabilities of invasive New Zealand mudsnails (***Potamopyrgus antipodarum*** (Gray, 1843)) across sites using visual surveys. Given this variation, using the incorrect detection probability to guide early detection sampling effort decisions could result in costly oversampling or delayed detections that enable invasive species establishment. Thus, estimates of detection probabilities should be a priority for invasive species monitoring programs.

Zebra (***Dreissena polymorpha*** (Pallas, 1771)) and quagga mussels (***D. rostriformis bugensis*** (Andrusov, 1897)), hereafter dreissenid mussels, are an example of invasive species that lack ample information about the sampling effort required for effective early detection monitoring. Dreissenid mussels are prolific aquatic invaders that now occur in most major water basins in eastern North America. Once established, dreissenid mussels can cause significant economic and ecological impacts that are predicted to result in annual expenditures of 100s of millions of dollars (USD) for control and mitigation efforts (Nelson 2019; Prescott et al. 2013).

Early detection of dreissenid mussels increases the likelihood that rapid response management actions can prevent population establishment and further spread to additional water bodies. Tens of millions of dollars (USD) are invested annually to prevent, contain, and control the spread of invasive dreissenid mussel in North America, yet the sampling effort required for high probability of detection has been given minimal scrutiny (Department of the Interior 2020). Counihan and Bollens (2017) used species accumulation theory to quantify the sampling effort required for a high probability of detecting rare planktonic taxa. Using these results, they inferred that the sampling effort expended in 2012–2014 was magnitudes too low for reliable early detection of dreissenid mussels in many Columbia River and Snake River reservoirs in the Pacific Northwest. Direct assessments of dreissenid mussel early detection monitoring using occupancy models have been challenged by the absence of appropriate data; many sites lack monitoring data prior to invasion and the sites that do have those data, often lack standardized methods and spatio-temporal replication.

Here, we take advantage of a high frequency, multi-year, standardized monitoring data set from across the western United States that emerged from a multi-jurisdictional, collaborative effort following the discovery of dreissenid mussels in Lake Mead, Nevada in 2007. This collaborative group established a standard protocol for plankton tow sampling of dreissenid mussel veligers (Western Regional Panel on Aquatic Nuisance Species 2020). Veligers are the microscopic planktonic larval form of dreissenids. Unlike adult dreissenids, veligers free swim in the water, which allows for them to be sampled by towing a fine mesh plankton net (64 m) through 1000 L of the water column. Plankton tows are collected at multiple locations per water body per season, typically sampling locations where mussel introduction is likely (e.g., boat docks and moorings). At each location, the debris from all plankton tows is aggregated into a single sample and examined using cross-polarized light microscopy for dreissenid mussel veligers. Following the microscopic evaluation, species identification is confirmed using a molecular technique such as polymerase chain reaction (Western Regional Panel on Aquatic Nuisance Species 2020).

We used implicit dynamic occupancy models to evaluate these data; our objectives were to provide initial estimates of the dreissenid mussel detection probabilities of plankton tow sampling and identify environmental covariates that may influence detection probabilities. Using model results, we provide initial guidance about early detection sampling effort allocation in the western United States and identify methodological improvements that could bolster the accuracy and precision of future detection probability estimates.

Materials and Methods

Data

Plankton tow data consisting of 11,442 samples from 325 water bodies across central and western United States were collected by the Bureau of Reclamation (hereafter referred to as Reclamation) and their state, federal, and tribal partners from fall 2011 to fall 2020. The plankton tow samples were analyzed for veliger presence by Reclamation's Technical Service Center in Denver, Colorado (USA). Multiple plankton tows collected at a given location, from the same water body, on a given day were aggregated to a single sample, therefore each sample refers to the aggregate of multiple tows. Consequently, our detection probability estimates are in terms of aggregated samples and not individual tows. Most samples lacked a description of the number of tows aggregated per sample, so we were unable to account for differences in sampling effort.

We removed water bodies that did not fit our scope of interest, such as hatcheries, water storage and treatment facilities, and one water body in Canada. Additionally, we removed observations prior to October 1, 2012 and after September 30, 2020 because the seasonal data were not complete for those time periods.

To examine if water body specific factors are associated with plankton tow sample detection probabilities, we obtained potential covariate information from multiple data sources. Few covariates of interest were collected with the Reclamation data. We focused on five covariates: water temperature, dissolved calcium, pH, water body size, and water year. Rationale for each covariate is provided below. Importantly, the associations between dreissenid mussel physiology and water temperature, dissolved calcium, and pH are approximate, as the local conditions experienced by dreissenids are difficult to accurately characterize and vary across geographic regions, environmental conditions, and seasons (Claudi et al. 2012; Claxton and Mackie 1998; Jones and Ricciardi 2005; White et al. 2015; Whittier et al. 2008). Water temperature, dissolved calcium, and pH were extracted from the Water Quality Portal (WQP, https://www.waterqualitydata.us/), as described in Supplementary material Appendix A (United States Geological Survey (USGS) 2012). For each covariate of interest, the WQP data was joined with the Reclamation data as detailed in the Appendix A. The WQP data were collected for other purposes and likely do not reflect the exact values at the time of the plankton tow samples. Consequently, we split each of the covariates into distinct categories to mitigate uncertainty in the measurements.

For each of the covariates, the categories and their associated covariate values are displayed in Table 1; justification for the category choices and a brief description of the biological relevance of each of the covariates follows.

	Category	Value
3*Water Temperature (C)	Low	12
	Mid	$>$ 12 to 28
	High	$>$ 28
3*pH	Minimal	$<$ 7 or $>$ 9.6
	Moderate	7.0 to 7.3 or 9.4 to 9.6
	High	$>$ 7.3 to $<$ 9.4
4*Dissolved Calcium (mg/L)	Minimal	$<$ 8
	Moderate	8 to 20
	High	$>$ 20
	Unknown	Missing
5*Water Body Size (km^2)	Stream	Stream
	Small	5.105
	Medium	$>$ 5.105 to 12.533
	Large	$>$ 12.533 to 40.465
	Larger	$>$ 40.465

Table 1: Water body temperature, pH, dissolved Calcium, and water body size categories used for modeling. The pH and dissolved calcium category names correspond to the presumed risk for dreissenid mussel presence, whereas the category names for water temperature and water body size describe the variables themselves. To create the water body size variable, surface area measurements were divided using quantiles so that the number of samples per category was approximately the same.

Dreissenid survival and reproduction can be limited by water temperature. Mortality increases as water temperatures exceed 28 C (Karatayev et al. 2010) and reaches acute thermal limits between 30 and 32 C (Garton et al. 2013). Temperatures that are too cool ($<$ 12 C) can inhibit gametogenesis; primary spawning occurs when temperatures are 16–18 C (Garton et al. 2013). We treated water temperature as a sample-specific covariate since it varies seasonally within a water body. However, on a given day, water temperatures were treated as constant across all sampling locations within a water body.

Dreissenids are also sensitive to pH, however the physiological mechanisms are unclear. Changes in acidification can lead to differing calcification rates, shell strength, and mortality rates (Claudi et al. 2012). To ensure there were pH data for all water bodies, pH was treated as constant for a given water body across the entire study period. pH category breaks are based on data presented in Ramcharan et al. (1992) and Claudi et al. (2012).

In both native and invaded areas, dreissenids are sensitive to dissolved calcium values because calcium is required for shell building and for metabolic functions (Jones and Ricciardi 2005; Whittier et al. 2008). For a given water body, we considered dissolved calcium as constant across the entire study period.

Water body size is another factor that may influence plankton tow detection probabilities. If dreissenids occur at similar abundance in small and large water bodies, then all else being equal, more samples are needed to detect dreissenids in a large water body. Larger water bodies are also likely to attract a larger number of motorized watercraft recreation, and so may be at increased risk of dreissenid mussel introductions (Robertson et al. 2020). Surface area (km^2) estimates were taken from the USGS National Hydrography Dataset's (NHD) Waterbody feature class, which is comprised of polygons representing water bodies in the NHD. The surface areas are approximations since water bodies have fluctuating water levels and source data were compiled over a period of several decades, therefore we split water body size into categories to mitigate the uncertainty in the measurements. With the exception of the stream category, we divided the surface area measurements using quantiles so that the number of samples per category was approximately the same.

Finally, water year was added as a seasonal variable describing the 12-month period that spans from October 1 of a given year to September 30 of the following year. This variable was added to allow for potential changes in occupancy status of a water body over time, assuming within a season the occupancy status of water bodies does not change. Additionally, the use