UNIVERSITY OF CALIFORNIA

Santa Barbara

Species community distributions: the role of scale-dependent processes and imperfect detection

A dissertation submitted in partial satisfaction of
the requirements for the degree Doctor of Philosophy
in Ecology, Evolution and Marine Biology

by

Marisa Morse

Committee in charge:
Professor Stephen R. Proulx, Chair
Dr. Kevin D. Lafferty, USGS/Adjunct Professor
Dr. John P. McLaughlin, Assistant Researcher
Professor Erika J. Eliason
Professor Deron E. Burkepile

December 2023

The Dissertation of Marisa Morse is approved.

Deron E. Burkepile

Erika J. Eliason

John P. McLaughlin

Kevin D. Lafferty

Stephen R. Proulx, Committee Chair

September 2023

Species community distributions: the role of scale-dependent processes and imperfect detection

by

Marisa Morse

Dedicated to Austin and Piggy

Acknowledgments

I would first like to thank my committee members: To Dr. Stephen Proulx, I am sincerely thankful that you believed in my capabilities; in doing so, you made space for my growth and allowed me to see myself as a scientist. Thank you for embodying true professionalism and putting logic and intention first. To Dr. Kevin Lafferty, thank you for your availability and support in my exploration of new dissertation ventures. I admire your creativity and determined drive to always push the science forward. To Dr. John McLaughlin, thank you for your guidance, your encouragement, and your continued willingness to pick up my frantic phone calls. You have not only been an invaluable resource and gracious mentor, but an exceptional confidant and friend. To Dr. Erika Eliason, thank you for inspiring me in the early days while I sat front row in your 8am lectures. I idolize your intelligence, dedication, and humanity. You are truly a role model. And to Dr. Deron Burkepile, thank you for advocating for me when I was down, I genuinely appreciate it.

To Johanna Fornberg and Jasmine Childress, I am forever grateful that I had such thoughtful and tenacious women standing by my side throughout this roller coaster. Thank you for the teatime Zoom sessions, Office references, soup breaks, and your steadfast support. I would have given up 100 times over if not for the both of you.

For my friends Eduardo Romero, Lisa Mesrop, Taom Sakal, Maxi Navarrette, Zoe Zilz, James Loving-Lichtenstein, Sophia Gimenez, and Shelley Bennett. My most cherished graduate school memories are composed of times spent with you. You are all incredible human beings who enhance this world, and I am changed for the better because our lives crossed paths. Thank you for being your truest selves, as you have allowed me to be mine. For Terra Dressler, Braulio

Castillo, An Bui, Stephanie Ma, Austen Apigo, Terence Leach, and the many others. Thank you for the craft nights, outdoor excursions, and spontaneous lunches that turned into dinners. I have learned strength, compassion, and resilience from each of you. To my non-graduate school friends: Kyrsthy Remigio, Melanie Guardado, James McClure, and Zena Schiff. Thank you for keeping me grounded and giving me perspective throughout the years.

For my mother and father, thank you for your curiosity, endurance, and chicken adobo while I achieved this dream. You set me on this path and have been loyally along for the ride. For my sister, Alex, thank you for listening to my rantings and laughing at all my jokes. The cat pics served as a welcome distraction. And for my Auntie Teeta, Kevin, Lisa, Joseph, Brian, and all the cuzzos. I love you all.

And finally, I would like to thank Austin Prouty for taking this journey with me. Your unabashed goofiness, chocolate chip cookie recipe, and devoted faith in my abilities has pulled me out of many slumps and pushed me when I could not go on. You continue to be my voice of reason, my best friend, and my home. Thank you for your unconditional support and love; I simply would not have made it without you.

Vita of Marisa Morse

EDUCATION

2016 – 2023	**University of California, Santa Barbara** Ph.D. Candidate – Ecology, Evolution and Marine Biology
2008 – 2013	**California State University, Monterey Bay** B.Sc – Marine Science

PROFESSIONAL EXPERIENCE

Oct 2020 – Feb 2021	United States Geological Survey Data Analyst Consultant
Oct 2014 – Oct 2016	California Department of Fish and Wildlife Scientific Aid
Nov 2015 - March 2016	Pacific States Marine Fisheries Commission Fisheries Technician
Sept 2012 – Oct 2014	Hopkins Marine Station, Stanford University Marine Ecology Assistant – Micheli Lab
May 2012 – Aug 2012	Scripps Institution of Oceanography, UC San Diego Summer Undergraduate Research Fellow - Shurin Lab
May 2011 – Aug 2011	Whitney Laboratory of Marine Bioscience, University of Florida Summer Undergraduate Research Fellow - Ptitsyn Lab

PUBLICATIONS

- Miller-ter Kuile, A., Apigo, A., Bui, A., DiFiore, B., Forbes, E. S., Lee, M., Orr, D., Preston, D., Behm, R., Bogar, T., Childress, J., Dirzo, R., Klope. M., Lafferty, K., McLaughlin, J., **Morse, M**., Motta, C., Park, K., Plummer, K., Weber, D., Young, R., & Young, H. (2022). Predator–prey interactions of terrestrial invertebrates are determined by predator body size and species identity. *Ecology*, *103*(5), e3634.

- Lafferty, K. D., Garcia-Vedrenne, A. E., McLaughlin, J. P., Childress, J. N., **Morse, M. F**., & Jerde, C. L. (2021). At Palmyra Atoll, the fish-community environmental DNA signal changes across habitats but not with tides. *Journal of Fish Biology*, *98*(2), 415-425.

PRESENTATIONS AND PROFESSIONAL OUTCOMES

Internal Reports

- M. Morse and D. Stein. 2016. San Nicolas Island Black Abalone Survey 2015-2016. California Department of Fish and Wildlife Internal Report.
- M. Morse and S. Bankston. 2015. Snorkel Survey Report for North Fork Matilija 2015. PSMFC Internal Report.

Research Conferences

- M. Morse, J. McLaughlin, E. Sarnat, H. Young, K. Lafferty. 2018. The influence of ecosystem scale on the distribution and abundance of 11 non-native ant species across Palmyra Atoll. Ecological Society of America. New Orleans, LA. Contributed Talk
- M. Morse, A. Noto, K. Bergesen, J. Shurin. 2013. Influence of tidal elevation on macroinvertebrate distribution in salt marshes of southern California. Association for the Sciences of Limnology and Oceanography. New Orleans, LA. Abstract and Poster Presentation
- M. Morse, A. Noto, K. Bergesen, J. Shurin. 2012. Influence of tidal elevation on macroinvertebrate distribution in salt marshes of southern California. UC San Diego Summer Research Symposium. Contributed Talk
- M. Morse, A. Noto, K. Bergesen, J. Shurin. 2012. Influence of tidal elevation on macroinvertebrate distribution in salt marshes of southern California. Sanctuary Currents Symposium. Monterey Bay, CA. Contributed Talk
- M. Morse, A. Ptitsyn. 2011. Circadian oscillation in sensory systems of *Aedes aegypti*. University of Florida Summer Research Symposium. Gainesville, FL. Contributed Talk

Guest Lectures

- CSU Monterey Bay, MSCI 341 - Conservation Genetics
- UCSB, EEMB 120 – Introduction to Ecology

TEACHING

Fall 2016 – 2019	**Teaching Assistant**, EEMB 112 Invertebrate Zoology, UC Santa Barbara
Winter 2017 – 2020	**Teaching Assistant,** EEMB 111 Parasitology, UC Santa Barbara
Spring 2017 – 2019	**Teaching Assistant,** EEMB 116 Higher Invertebrate Zoology, UCSB
Summer 2018	**Teaching Assistant,** EEMB 120 Introduction to Ecology, UCSB
Spring 2020 – 2023	**Teaching Assistant**, EEMB W22 Biological Controversies, UCSB
Winter 2021 – 2022	**Head Teaching Assistant,** EEMB 138 Behavioral Ecology, UCSB

AWARDS AND GRANTS

2018 Coastal Fund Research Grant
2018 UCSB Research Block Grant
2023 Letters to a Pre-Scientist Pen Pal Compassionate Connection Award

Abstract

Species community distributions: the role of scale-dependent processes and imperfect detection

by

Marisa Morse

Community ecology seeks to understand the assembly processes that structure the abundance, richness, and prevalence of species in the community. However, some major difficulties are that the effect of an assembly process can change between scales and that natural scales can be difficult to identify. The first two chapters of this dissertation aimed to understand the role of assembly processes across biologically defined scales. All empirical data used in this dissertation were previously collected on Palmyra Atoll (McLaughlin *et al.,* 2023; McLaughlin, 2018), a national wildlife refuge located within the Remote Pacific Islands Marine National Monument. Chapter 1 examined the scale-dependent factors that influenced free-living arthropod species abundances across terrestrial spatial scales. We found evidence that of the island- or forest-specific covariates included in the model, only soil cation exchange capacity had population-level effects on arthropod abundances. However, we found species-specific and higher-taxon level ("order") responses to island size, nutrient input, and canopy type. We also explored how species residual associations changed with scale as a possible indicator of biotic interactions with narrowing scale. Chapter 2 focused on the effect of host traits on parasite component and infracommunities in marine sandflat fish species. Parasite species occurrence probabilities varied between and within host species, with parasite occurrences responding to host species generality, density, and host individual weight. Another major difficulty in ecological studies is sampling error when observing species communities. Few species distribution studies account for imperfect detection, but assuming that observations perfectly capture a community can lead to biased results and inferences. In the final chapter, we assessed a method to predict true communities based on estimated false-negative probabilities in 1000 simulated datasets. Based on this analysis, predicted communities were more accurate and shared more mutual information with the true community than the observed community. We then tested these methods in a case study using empirical data from Palmyra Atoll's Hymenoptera community.

Table of Contents

List of Figures and Tables

Chapter 1: Scale-dependent effects on terrestrial arthropod distributions on Palmyra Atoll

1.1 Introduction

1.1.1 Spatial scales on Palmyra Atoll

Community ecology seeks to understand how assembly processes, such as environmental factors, biotic interactions, or neutral forces, impact species distributions. However, the effect of assembly processes will change between global, regional, or local scales (Garzon-Lopez *et al.,* 2014; Viana & Chase, 2019). For example, climatic variables, like mean temperature and precipitation, impact species distributions and abundances at global scales by imposing physiological limitations (Thomas, 2010; Araújo *et al.,* 2005; Diamond *et al.,* 2012; Kearney & Porter, 2009; Helaouët & Beaugrand, 2009; Harsch & HilleRisLambers, 2016). But the importance of temperature or precipitation may decrease at smaller scales. Instead, local community assembly will be driven by local factors such as the effect of windspeed on seed dispersal (Bullock & Clarke 2000; Heydel *et al.,* 2014), soil fertility on plant diversity (Janssens *et al.,* 1998; Tilman *et al.,* 1996; Dybzinski *et al.,* 2008; Holl, 1999), or habitat type (i.e. intertidal versus subtidal, mudflats versus grasslands). The nested structure of these processes means that communities observed at smaller scales have already been narrowed by processes impacting larger scales. Thus, the accumulation of assembly processes will lead to variation in species identities and abundances occurring at each scale (Levin, 1992). Studying the effect of scale can be a challenge because biologically relevant and distinct scales are difficult to identify (Stuber & Gruber, 2020; Pelosi *et al.,* 2010; Bishop *et al.,* 2002; Dormann *et al.,* 2018). In addition, a species community dataset with replications across several scales can be laborious and often infeasible to collect (Gauch, 1982; Marschall & Roche, 1998). As a result, most studies focus on a single scale (McGarigal *et al.,* 2016). To discern the role of assembly processes between natural scales, we will examined terrestrial arthropod communities at three biologically defined spatial scales on Palmyra Atoll: a tree within a forest, a forest within an island, and an island within the atoll (Figure 1).

Palmyra Atoll provides several advantages for understanding how assembly processes impact species distributions. First, terrestrial communities can be assessed at naturally defined spatial scales. At the largest scale, we have Palmyra Atoll itself, which houses all arthropods in the community. Palmyra is a 4km long atoll composed of around 30 islands (depending on tidal height) that vary in size and shape and surround a central deep-water lagoon. Unusually for an atoll, the islands are heavily forested. The bounds of forests within islands are well-defined by natural, but sharp shifts in dominant canopy type (Young *et al.,* 2010). And nested within a forest are individual trees. Thus, our three hierarchical scales of interest are islands, forests, and trees. At each of these scales, we expect variation in arthropod distributions that can be explained by assembly processes at that scale. And because each of these scales are biologically defined, as opposed to quadrats or transects, the observed communities will be a direct reflection of natural processes, allowing more insight into community assembly. Another advantage of Palmyra Atoll is that it holds constant many factors known to influence species distributions. While the islands vary in size over several orders of magnitude, they are all collocated within 4km of each other, and none has an elevation higher than 2 meters. Thus, all sites experience similar disturbance regimes and exposure to large-scale precipitation and climate fluctuations. Palmyra also minimizes recent anthropogenic impacts. Urbanization and human disturbance are known to influence arthropod distributions in other systems, but aside from a few sporadic researchers, Palmyra hasn't maintained human habitation for several decades (Bang & Faeth, 2011; Fenoglio *et al.,* 2020). Historically, Palmyra did not likely support permanent indigenous populations, but there is evidence of occasional Micronesian and Polynesian visits (Dawson, 1959; Wester, 1985). The US military occupied Palmyra during WWII, making several structural changes and likely introducing the many non-native arthropods (and rats) to the terrestrial community (Handler, 2007). But since their departure in the 1960s, Palmyra's species have been left to assemble under limited human activity. Palmyra Atoll was protected as a National Wildlife Refuge in 2001. And researchers that visit now must follow strict protocols to minimize the introduction and dispersal of non-native species (Hathaway & Fisher, 2010). As a result, Palmyra's arthropod communities are not currently influenced by contemporary human disturbance. Another advantage of studying terrestrial species on Palmyra is that the community structure is relatively simple. Following the black rat eradication in 2011, no mammals exist in this system (Wegmann *et al.,* 2012). Besides a few geckos, mollusks, and sea bird species, Palmyra Atoll's terrestrial fauna is dominated by arthropods in qualitative abundance and overall richness (Handler *et al.,* 2007). In some systems, arthropod distributions are influenced by other taxa through competition or predation (Gunnarsson, 1996; vanKlink *et al.,* 2015; Moran & Hurd, 1997; Gunnarsson *et al.,* 2009). However, interactions with other animal phyla likely have a small influence on arthropod distributions on Palmyra Atoll. Further simplifying the system, immigration of new species by natural processes is low because Palmyra Atoll is isolated by over 1000 miles from the nearest inhabited land mass. In sum, due to the distinct spatial scales, restriction of explanatory variables, system simplicity, and geographic isolation, Palmyra Atoll appears to be a suitable system to evaluate how the effect of assembly processes influence arthropod communities with changing scale.

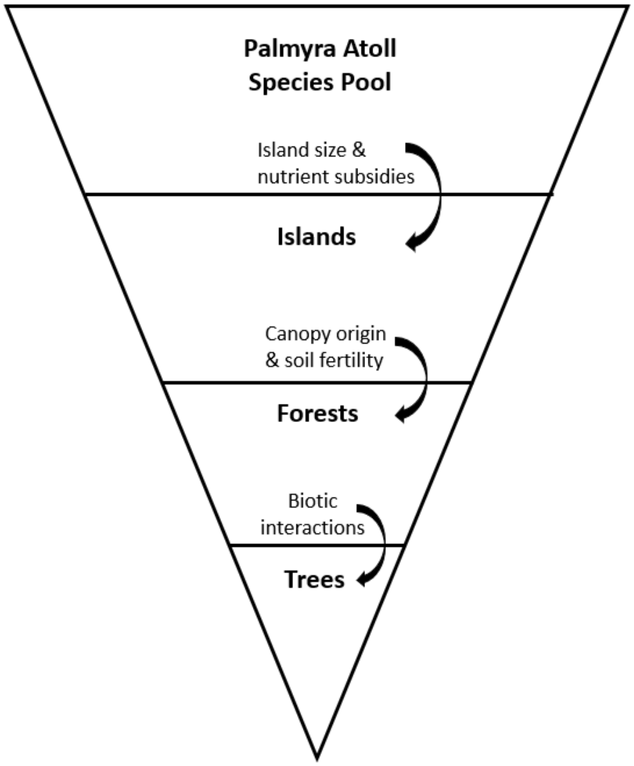

Figure 1. Palmyra Atoll's nested spatial scales and the predicted assembly processes that filter arthropod communities between each.

1.1.2 Atoll Scale Species Pool

At the largest scale, we have Palmyra Atoll which encompasses the entire species pool. The species present have already been filtered by large-scale processes like a tropical climate (low annual temperature variation and high annual precipitation), Palmyra's remoteness, and small footprint. The native species have evolved in conjunction with one another and are adapted to the other native flora and fauna. However, native species on islands are particularly vulnerable to biological invasions because they evolved in geographic isolation, avoiding many forms of predation and competition (Whittaker & Fernández-Palacios, 2007; de'Antonio & Dudley, 1995). The native status for about 1/3 of Palmyra's species is unclear and difficult to determine. Many arthropods present on Palmyra are widely distributed throughout the Pacific, and some are globally invasive tramp species, like *Pheidole megacephala* (the big-headed ant) (Passera, 2021). Previous surveys have estimated that ~86% of Palmyra's arthropod species were accidental introductions and that ~89% of the species overlapped with Hawai'i (Handler *et al.,* 2007; Nishida, 2002). These surveys documented 115 arthropod taxa present on Palmyra (Handler *et al.,* 2007), and although extensive, were not comprehensive. Previous surveys have omitted entire Orders (e.g. Collembola, Thysanoptera) and misidentified members of taxonomically difficult groups (Blattodea). McLaughlin's intensive sampling protocol more than doubled the most recent species count on Palmyra. Additionally, collaboration with taxonomists around the world has revised the native status of many species, with the result that the most abundant members of many Orders (Blattodea, Isopoda, Orthoptera) are not only native, but undescribed endemics (McLaughlin *et al.,* 2023).

Not all terrestrial consumers present on Palmyra are included in this analysis. Currently, Palmyra's entire terrestrial community includes native birds, lizards, crabs, insects, and arachnids, but the data used here only include canopy-dwelling organisms. For example, terrestrial crabs are high biomass arthropods in this system (Howald *et al.,* 2004), but they are not included here. Additionally, the collection methods may disproportionately sample some species over others (weak fliers vs. strong fliers). Despite these biases, this study analyzes relative species abundance so we can still gather insight as to factors that assemble communities at each of these spatial scales.

1.1.3 Environmental Factors of the Island Scale

Many environmental factors are expected to structure species communities at the island scale. The classic theory of island biogeography focuses on island size and island distance from the mainland to predict patterns of species richness (MacArthur & Wilson, 2001). And the unified theory with relative species abundance expects that local abundance of a species is dependent on the number of species in the local community, and that relative abundance reaches an equilibrium depending on the source area (Hubbell, 1997). Because of this equilibrium, we do not expect that island size or distance from the mainland will universally impact relative population abundances on the island. However, the variance in species abundances depends on the rate of immigration, so as an island becomes progressively more isolated, rare species are expected to become rarer, while common species become more common (Hubbell, 1997). Therefore, we *do* expect species-specific responses with some species increasing in abundance in response to island size or distance, while others decrease in abundance. In this system, Palmyra Atoll is isolated with no nearby mainland to acquire new species from. However, Cooper Island contains the only functioning airstrip, deep-water dock, and living quarters, making it the most likely point of entry for non-native species. Cooper Island is also the largest island and contains all habitat types and (almost all) plant species that occur on other islands. Similar to a mainland, Cooper Island may function as a point-of-dispersal for new species to surrounding islands. For our study, island size and distance from Cooper were highly correlated, so we only include island size as a representative of both.

Another environmental covariate expected to influence arthropod communities are nutrient subsidies on islands. On Palmyra Atoll, most sea birds prefer to roost on native plants resulting in higher guano deposits and nutrient inputs on certain islands (McCauley *et al.,* 2012; Young *et al.,* 2010). In other studies, arthropod communities have responded to high nitrogen densities with changes in structure and relative abundance (Wimp *et al.,* 2019; Haddad *et al.,* 2000; Ritchie, 2000), and species-specific responses were partially explained by feeding groups. In the current study, we also expect that nutrient subsidies will alter arthropod communities with species-specific effects. Some species will respond positively to increased nutrient subsidies, while others will respond negatively. However, because of sea bird roosting preferences, nutrient subsidies are highly correlated to the proportion of canopy types on an island. Nitrogen density is positively correlated with the proportion of broadleaf native trees (*Pisonia* and *Scaveola*) on an island, and negatively correlated with the introduced coconut palm (*Cocos nucifera)*. Therefore, the effects of nitrogen density may indirectly explain arthropod variation due to habitat proportions on the island. But, whether due to direct or indirect effects, we expect arthropod abundances to be influenced by increasing nutrient subsidies.

1.1.4 Environmental Factors of the Forest Scale

On Palmyra Atoll, individual forests (delineated by canopy type) are discretely defined creating homogeneity at local scales and heterogeneity at larger scales. As a result, forests on islands are an intermediate scale at which we expect assembly processes to operate. In other systems, vegetation type can be a predictor of arthropod communities with species demonstrating plant preferences (Antunes *et al.,* 2008, Schaffers *et al.,* 2008), and Palmyra's arthropods should have habitat preferences as well. On Palmyra, *Cocos nucifera* is the most common forest type (~43%), followed by *Scaevola sericea* (~29.5%), *Pisonia grandis* (~12%), *Terminalia catappa* (~6.2%), and *Pandanus fischerianus* (Hathaway *et al.,* 2011; Wegmann, 2005) (Figure 2). Three of these forest types were introduced to Palmyra, and we predict that plant origin will affect arthropod distributions. Some arthropod species currently on Palmyra are non-native (Handler, 2007), and for these, we expect positive, negative, and neutral species-specific responses to plant origin. We expect native arthropods will have a positive response to native plants due to a shared evolutionary history. Additionally, as mentioned in the island scale, sea birds preferentially roost and nest in native broad-leaf canopies, which in-turn receive more guano. Therefore, canopy origin may also capture the variation in arthropod abundances as a result of nutrient input at the forest scale. Whether based on habitat preferences or nutrient input, we expect that plant origin will be a defining characteristic that structures abundances and identities of the arthropod communities found at the forest scale.

Independent of nutrient subsidies, there are many soil characteristics expected to correlate with arthropod distributions on Palmyra Atoll. Energy enters an ecosystem through net primary production, and when net primary production is high, more individuals within a taxon are expected to be supported (Kaspari *et al.,* 2000; Wright, 1983). Assuming that net primary production will be higher as a result of increased soil fertility (Malhi *et al.,* 2004; Aragão *et al.,* 2009), Palmyra's arthropod abundances are expected to increase in forests with more fertile soil. Soil fertility is an accumulation of many soil traits such as sediment size, pH, organic matter percentage, and cation exchange capacity. Cation exchange capacity (CEC) is the soil's ability to retain positively charged ions and is estimated based on many soil characteristics. When combined with other soil fertility measurements, it is a good indicator of soil quality and productivity (Ross & Ketterings, 1995). And in this system, CEC is correlated with soil pH, organic matter, and sediment size. Therefore, here we represent soil fertility using cation exchange capacity (CEC) which also serves as a proxy for other correlated soil traits. We expect arthropod abundances to increase with CEC at the forest scale due to higher soil fertility and net primary production.

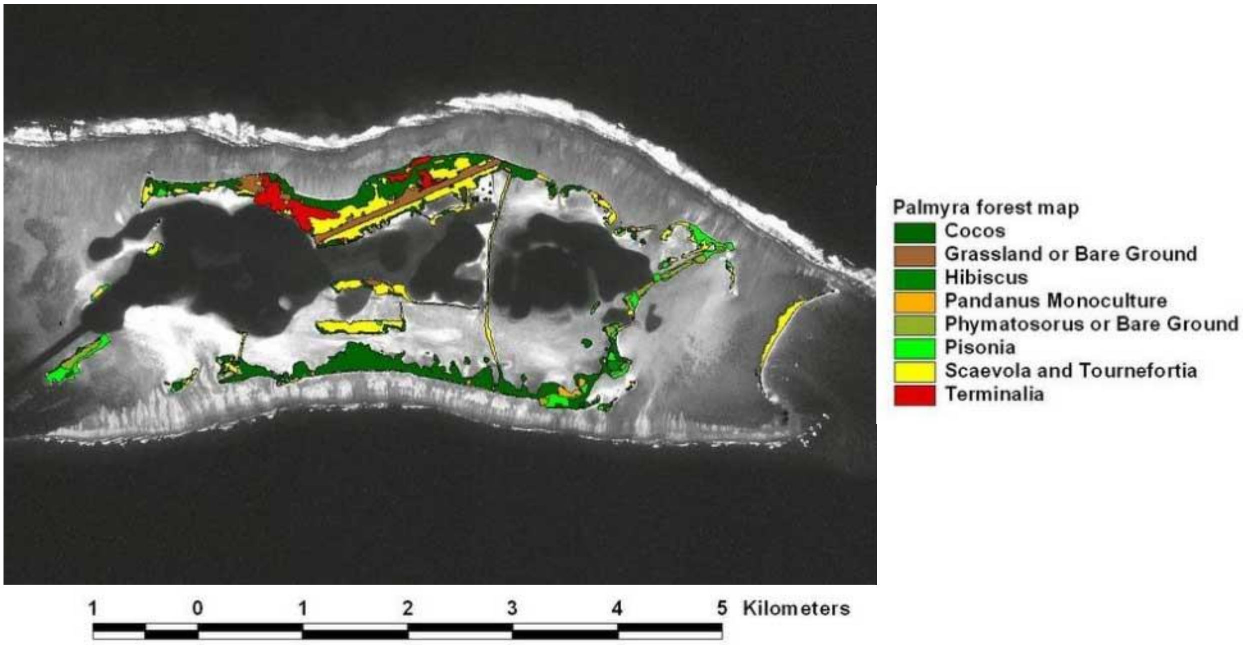

Figure 2. A map of prominent forest types on Palmyra Atoll (Wegmann, 2005)

1.1.5 Biotic Interactions of the Tree Scale

Individual trees were the smallest spatial scale on which we measured community structure. Some analyses have found evidence of a reduced role of environmental forces and an increased role of biotic interactions at smaller scales (Warren *et al.,* 2010; Bell *et al.,* 2010; Nachman & Borregaard, 2010; Gotelli *et al.,* 2010), and we expect this pattern as well. After environmental predictors outline the fundamental niche of a species, biotic interactions are expected to further structure the realized niche. Different types of biotic interactions may be detectable at different scales. Simulations suggest that negative interactions will be more detectable at small scales and positive interactions apparent across scales (Araújo & Rozenfeld, 2014). This is because if species positively interact, we expect co-occurrence patterns at all scales (Araújo & Luoto, 2007). But negatively interacting species may cause local extinctions at small scales that may appear as coexistence at larger scales (Godsoe *et al.,* 2015). But empirical studies have found mixed results with no clear patterns of scale dependence (Bullock *et al.,* 2000; Mod *et al.,* 2020; Whittaker *et al.,*2001; McGill, 2010; Veech, 2006; Russell *et al.,* 2006; Belmaker *et al.,* 2015).

To explore how biotic interactions may influence arthropod communities in this study, we compared species residual associations between and within scales. Residual associations are based on non-random patterns between the unexplained variance of a species pair after controlling for factors in the statistical model. Species residuals that are positively or negatively associated suggest that there is something other than environmental covariates or random effects included in the model that are driving the association patterns between species (Ovaskainen & Abrego, 2020; Ovaskainen *et al.,* 2010). Because measured and scale-dependent unmeasured abiotic factors are accounted for, we suspected that some of these residual associations could be explained by biotic interactions between species that either enhance or inhibit one another's abundances (Ovaskainen & Abrego, 2020; Mod *et al.,* 2020). However, unmeasured assembly processes or spurious patterns, like stochastic events or sampling error, could lead to associated residuals as well (Wisz *et al.,* 2013). To distinguish association patterns from randomness, we could compared observed species matrices against several null models (Gotelli, 2000; Gotelli & McCabe, 2002). Even so, it remains difficult to determine if residual associations are due to biotic interactions, environmental filtering (independent of the model's random effects), or dispersal limitations. But we assume that some proportion of the detected residual associations within scales are a result of biotic interactions. Additionally, we acknowledge that these species comparisons are based on correlations and can only suggest interactions. In this study, to explore the role of biotic interactions with narrowing scale, we examined residual associations between and within each scale. Similar to the predicted patterns for biotic interactions, we expected positive associations to be present across all scales, while negative associations would be most prominent at the tree scale.

As an example, the most abundant arthropod taxa on Palmyra Atoll are ants, all of which are non-native. The 13 ant species found on Palmyra Atoll represented ~30% of the individual arthropods counted in this study. Among ant communities, competition is ubiquitous (Davidson, 1998) and is an important process forming dominance hierarchies (LeBrun, 2005) and ant mosaics (Ribas & Schoereder, 2002; Sanders *et al.,* 2007). Many ant species use aggressive behaviors that seem to hinder or facilitate the abundance, behavior, and spatial distribution of other ants (Hölldobler & Wilson 1990; Andersen *et al.* 1991; Andersen & Patel 1994). Based on the theory of island biogeography, we might expect more abundant ant populations to be found on Palmyra's large and close islands, leading to positive co-occurrence between species on islands. We could also expect that, within those islands, the energy limitation hypothesis drives several ant species to positively co-occur in response to forest soil fertility. However, if two ant species negatively interact, we would expect a negative residual association at the tree scale. This means that, with all else constant, the two ant species would be less abundant when occurring together than when on trees without the other. Because ants (the most abundant arthropod taxon on the atoll), are likely to be affected by biotic interactions, we expect biotic interactions to be a process that influences many species of Palmyra's arthropod community. Biotic interactions could be based on associations between predators and prey, competitors with a shared resource, or mutualists (ants with aphids or scale insects) (Wisz *et al.,* 2013; Tilman, 1994; Stradler & Dixon, 2005; Kulikowski *et al.,* 2020). Some species may demonstrate many positive or negative associations, suggesting a strong impact of biotic interactions, while other species will have mostly neutral associations, providing no evidence for interactions (Benkman, 2013; Ovaskainen *et al.,* 2010; Araújo & Rozenfeld, 2014). Additionally, we expect residual association patterns to shift in magnitude and sign between spatial scales depending on the effect of environmental covariates, random effects, unmeasured environmental traits, and biotic interactions.

1.1.6 Arthropod Taxonomy

Because species traits are acquired from a common ancestor, closely related species are expected to share more morphological and physiological traits. An overlap in inherited traits may lead to similar responses to environmental factors. However, these similar responses cannot simply be attributed to covariates because unmeasured species traits aren't independent of one another. They are instead linked by phylogenetic distance. Thus, to avoid misinterpreting model outputs, we incorporated taxonomic groups to account for unmeasured and phylogenetically associated species traits when examining community organization. We expect that species within the same taxonomic group will have more similar traits, resulting in higher-taxon-level responses to environmental covariates.

1.1.7 Approach

In this study, we examined the interaction between assembly processes and arthropod community structure at three different spatial scales. To test our predictions, we analyzed arthropod abundance data on Palmyra Atoll at the island, forest, and tree scales. This study aimed to answer three questions. First, how do community environmental processes influence species distributions at increasingly narrow spatial scales? To answer this question, we built a hierarchical mixed model with trees nested within forests nested within islands. At the island scale, we expected arthropod abundances to have species-specific responses to island size and increase with nutrient subsidies. At the forest scale, we expected arthropod abundances to increase with CEC and have species-specific effects in response to canopy origin. We also included random intercepts of all three spatial scales to account for unmeasured environmental traits. Second, how did taxonomic relatedness affect arthropod abundances? To answer this question, we nested "order"-specific and species-specific responses within the population effect of each covariate. We then analyzed their mean estimates and 95% credible intervals. Third, how might biotic interactions structure arthropod communities? To answer this, we estimated the number of positive and negative species residual associations at each spatial scale. Species residual association matrices examine unexplained variance after controlling for environmental covariates and random effects in the statistical model. We suspected that some of these residual associations were driven by biotic interactions between species pairs. We then compared residual association matrices across the island, forest, and tree to explore the influence of biotic interactions with narrowing scale, both in frequency and magnitude.

1.2 Methods

1.2.1 Field Sites and Data Collection

This data for this study were collected from the terrestrial habitats of Palmyra Atoll. Located 1680km south of Hawai'i, Palmyra Atoll is a US National Wildlife Refuge within the Pacific Remote Islands Marine National Monument. As part of an NSF DEB (1457371) sponsored project awarded to Hillary Young and Kevin Lafferty, McLaughlin *et al.,* (2023) built a food web describing the terrestrial communities on Palmyra. The methods from this study included many species sampling techniques such as black light surveys, branch clippings, and point sampling. Another one of these sampling techniques was canopy fogging, in which researchers dispersed a pyrethrum-based insecticide (*ExciteR* 6% Pyrethrin) with a bio-diesel fuel carrier on tree canopies. Falling insects were collected in plastic funnels that sampled the column of air directly above them. The contents of these funnels were pooled for each tree for a tree-level estimate of species relative abundance, and the species in the samples were counted and sorted to lowest possible taxonomic level. This study included 87 individual trees that were fogged across 15 islets (McLaughlin *et al.,* 2023). For the current study, we took advantage of this comprehensive fogging data to analyze the factors that influence arthropod distributions across spatial scales. McLaughlin shared this dataset and agreed to collaborate on this analysis. For a more detailed description of field methods and lab sorting, see McLaughlin *et al*. (2023). Measurements of island size/distance and forest soil traits are also described in McLaughlin *et al.,* (2023). Nutrient subsidy measurements were based on data from Young *et al.,* (2010). For forest origin, we classified 5 plant species as native (*Pisonia grandis, Pandanus fischerianus, Tournefortia argentea, Barringtonia asiatica,* and *Scaevola sericea*) and 3 as non-native (*Cocos nucifera, Terminalia catappa,* and *Hibiscus tiliaceus*).

The observed dataset counted 102,870 individuals belonging to 246 species. Almost all were identified to the family level and 148 were identified to species. The most abundant species was the big-headed ant, *Pheidole megacephala,* which composed 21.6% of observed individuals. 23 species observations were singletons. All species were nested within 22 taxonomic groups which consisted of mostly taxonomic orders. As an exception, ants and scale insects were particularly rich and abundant, and they were expected to respond to covariates unlike the other species in their taxonomic order. For this reason, we categorized Formicidae and Coccoidae separately from the rest of the Hymenoptera and Hemiptera and refer to this higher taxonomic group as "order". The most abundant taxonomic group was Formicidae with 31,098 individuals, followed by Coccoidea with 21, 293 individuals. Dipterans were the most diverse taxonomic group with 48 morphospecies represented by 3,068 individuals. The dataset sampled 87 trees in 49 forests on 15 islands. The number of trees fogged in a single forest ranged from 1 to 7 with an average of 1.8. The number of forests sampled within an island ranged from 1 to 8 with an average of 3.2.

1.2.2 Statistical Analysis and Model Specifications

To understand how species distributions were affected with narrowing spatial scale, we built a hierarchical mixed model to analyze arthropod community data on Palmyra Atoll. We used a zero-inflated Poisson distribution with a log link and logit link function to describe arthropod abundances. The model took the form:

$$y_i \sim \text{ZIPoisson}\ (\lambda_i, \pi_i)$$

$$\text{logit}(\pi_i) = \alpha_\pi$$

$$\log(\lambda_i) = \alpha_i + \beta_{1i} * \text{Size} + \beta_{2i} * \text{Nitrogen} + \beta_{3i} * \text{CEC} + \beta_{4i} * \text{Non-Native}$$

$$\alpha_i = \alpha_i + \alpha_{\text{order}_i} + \alpha_{\text{species}_i} + \alpha_{\text{island}_i} + \alpha_{\text{forest}_i} + \alpha_{\text{tree}_i}$$

$$\beta_{1i} = \beta_1 + \beta_{1,\text{order}_i} + \beta_{1,\text{species}_i}$$

$$\beta_{2i} = \beta_2 + \beta_{2,\text{order}_i} + \beta_{2,\text{species}_i}$$

$$\beta_{3i} = \beta_3 + \beta_{3,\text{order}_i} + \beta_{3,\text{species}_i}$$

$$\beta_{4i} = \beta_4 + \beta_{4,\text{order}_i} + \beta_{4,\text{species}_i}$$

$$\begin{bmatrix} \alpha_{\text{order}} \\ \beta_{1,\text{order}} \\ \beta_{2,\text{order}} \\ \beta_{3,\text{order}} \\ \beta_{4,\text{order}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{S}_{\text{order}} \right)$$

$$\begin{bmatrix} \alpha_{\text{species}} \\ \beta_{1,\text{species}} \\ \beta_{2,\text{species}} \\ \beta_{3,\text{species}} \\ \beta_{4,\text{species}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{S}_{\text{species}} \right)$$

$$\mathbf{S}_{\text{order}} = \begin{pmatrix} \sigma_{\alpha_{\text{order}}} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\beta_{1_{\text{order}}}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_{2_{\text{order}}}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{3_{\text{order}}}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\beta_{4_{\text{order}}}} \end{pmatrix} \mathbf{R}_{\text{order}} \begin{pmatrix} \sigma_{\alpha_{\text{order}}} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\beta_{1_{\text{order}}}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_{2_{\text{order}}}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{3_{\text{order}}}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\beta_{4\text{order}}} \end{pmatrix}$$

$$\mathbf{S}_{\text{species}} = \begin{pmatrix} \sigma_{\alpha_{\text{species}}} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\beta_{1_{\text{species}}}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_{2_{\text{species}}}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{3_{\text{species}}}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\beta_{4_{\text{species}}}} \end{pmatrix} \mathbf{R}_{\text{species}} \begin{pmatrix} \sigma_{\alpha_{\text{species}}} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\beta_{1_{\text{species}}}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_{2_{\text{species}}}} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\beta_{3_{\text{species}}}} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\beta_{4\text{species}}} \end{pmatrix}$$

$$\mathbf{R}_{\text{species}} \sim \text{LKJcorr}\ (4)$$

$$\mathbf{R}_{\text{order}} \sim \text{LKJcorr}\ (4)$$

$$(\sigma_{\alpha}\text{species}, \sigma_{\beta_{1\text{species}}}, \sigma_{\beta_{2\text{species}}}, \sigma_{\beta_{3\text{species}}}, \sigma_{\beta_{4\text{species}}}) \sim \text{HalfCauchy}\ (0, 2)$$

$$(\sigma_{\alpha}\text{order}, \sigma_{\beta_{1\text{order}}}, \sigma_{\beta_{2\text{order}}}, \sigma_{\beta_{3\text{order}}}, \sigma_{\beta_{4\text{order}}}) \sim \text{HalfCauchy}\ (0, 2)$$

$$\alpha_{\text{tree}_i} \sim \text{Normal}\ (0, \sigma_{\text{tree}})$$

$$\alpha_{\text{forest}_i} \sim \text{Normal}\ (0, \sigma_{\text{forest}})$$

$$\alpha_{\text{island}_i} \sim \text{Normal}\ (0, \sigma_{\text{island}})$$

$$(\sigma_{\text{island}}, \sigma_{\text{forest}}, \sigma_{\text{tree}}) \sim \text{HalfCauchy}\ (0, 2)$$

$$\alpha \sim \text{Normal}\ (0\ ,\ 3)$$

$$\beta_1 \sim \text{Normal}\ (0\ ,\ 1)$$

$$\beta_2 \sim \text{Normal}\ (0\ ,\ 1)$$

$$\beta_3 \sim \text{Normal}\ (0\ ,\ 1)$$

$$\beta_4 \sim \text{Normal}\ (0\ ,\ 1)$$

where the response variable y_i represented the observed abundance of an invertebrate species on a tree. π_i estimated the probability of extra zeros and was modelled with a logit link function. λ_i was the expected count of a species on a tree and was modelled with a log link function. In the linear part of the model, the predictors included island size, island nitrogen input per day, forest soil cation exchange capacity, and forest plant origin. The estimated covariate effect on the population was described by regression coefficient β_x. These covariates also had random slopes which estimated species-specific effects ($\beta_{x,\text{species}_i}$) nested within “order”-specific effects

($\beta_{x,\mathrm{order}_i}$). Species-specific slopes and intercepts were modelled with a multivariate normal distribution with expected mean effects of zero. “Order”-specific estimates were also modelled with a multivariate normal distribution with expected values of zero. S described the covariance between levels within a random effect and R was the corresponding correlation matrix. To capture the nested spatial structure, we used additive random intercepts for each island, forest, and tree. The expected mean of the tree level intercept (α_{tree}), forest level intercept (α_{forest}), and the island level intercept (α_{island}) was zero.

The prior distributions of all environmental covariates were assumed to follow a normal distribution with a mean of 0 and a standard deviation of 1. The prior distributions of intercepts were described by a normal distribution with a mean of 0 and standard deviation of 3 to allow for increased variance. Variance term priors had a half-cauchy distribution bounded at 0 with a scale parameter of 2.

All numeric predictors were standardized to a mean of 0 and standard deviation of 1 before the analysis to improve chain convergence and model interpretation. This model was fit using the brm() function in the brms package in R (Bürkner, 2017) which uses ‘Stan’ for full Bayesian inference.

To fit this model, we ran 4 MCMC chains with a warmup up period of 500 and total of 3000 samples per chain, resulting in 10,000 post-warm up draws. We assessed chain convergence with potential scale reduction factor and effective sample size statistics. To evaluate and compare the predictive abilities of this model fit, we conducted a leave-one-out cross validation.

To understand how environmental covariates impacted species distributions, we extracted the posterior samples of the population-level regression coefficients and calculated their estimated means and 95% credible intervals. We examined the estimated standard deviations for each random effect (intercepts and slopes) to understand how much variation could be described by "order", species, island, forest, and tree identities. We then extracted the posterior distributions for "order"-specific and species-specific intercepts and regression coefficients. We also extracted the island-specific, forest-specific, and tree-specific random intercepts. For each of these parameters, we calculated their estimated mean and 95% credible interval.

To estimate the residual associations between species pairs, we used the residuals() functions from the brms package to extract 1000 residual estimates for each observation. We then used the CorrelationBF() function in the BayesFactor package in R to draw a posterior distribution of 3000 correlation coefficients between the residual estimates of each species-species combination. Posterior distributions with 95% credible intervals that excluded zero suggested non-random correlations between species pairs. However, these correlations were based on thousands of estimates per species (1000 draws * # of sites), and the 95% credible intervals around the mean correlation coefficients were very narrow, spanning an average distance of 0.01. For this reason, we further required the 95% credible intervals to exclude -0.2 to 0.2 (instead of just 0). We chose this threshold because coefficients within this range provide weak, negligible, or non-existent evidence of a correlation (Dancey & Reidy, 2007; Chan*.,* 2003; Akoglu, 2018). We repeated this process for all three spatial scales to estimate species residual associations on the island, forest, and tree.

We then examined the species residual association matrix for each spatial scale to explore the role of biotic interactions. A residual association matrix illustrated species pairs whose abundances were correlated after controlling for predictors included in the statistical model. Other factors are likely driving these residual associations, and we suspected some biotic interactions. However, residual associations can also arise due to unmeasured environmental traits that are independent of our scale-based random effects. By comparing these matrices, we hoped to reveal some species pairs that directly interact, influencing one another's abundances. We counted the number of positive, negative, and neutral species pairs in each of the species residual association matrices to suggest possible biotic interactions. Then, to explore the role of biotic interactions with narrowing scale, we compared the frequency and magnitude of associations between islands, forests, and trees. For species-specific counts of positive and negative associations, we summed the mean correlation coefficients for an entire species row.

All forest plots were created using the ggplot() function in the ggplot2 package in R. All heatmaps were created using the heatmap.2() function in the gplots package in R. And all histograms were created with the hist() function in base R.

1.3 Results

1.3.1 Computational Results

Before inspecting the outputs of our statistical model, we assessed convergence and efficiency of the MCMC chains. All Gelman diagnostics (r-hats) were below 1.02 confirming chain convergence, and all bulk effective sample sizes (ESS) were at least 200. The smallest ESS was 334, indicating reliable efficiency. We also evaluated pairs plots and correlation coefficients to confirm that no parameters had interacting estimates. Finally, we performed a leave-one-out cross-validation to assess model performance.

1.3.2 Population, Order, and Species-Specific Effects

We evaluated the community dataset using a mixed model to understand the effect of assembly processes on arthropod abundances. The overall population intercept had a mean estimate of -

1.26 (95% CI [-2.16, -0.35]) (Table 2). We found no evidence that island size, island nutrient subsidies, or forest plant origin affected abundances at the population level with mean estimates 0.29 (95% CI [-0.28, 0.87]), -0.01 (95% CI [-0.53, 0.49]), and -0.21 (95% CI [-0.88, 0.49]), respectively. However, we did find evidence that forest CEC had a population-level effect with estimated mean 0.51 (95% CI [0.11, 0.89]).

We further examined taxonomic standard deviations around the population-level effects which could indicate “order”-specific (Figure 3) or species-specific responses. The large estimated standard deviations around the population mean implied that “order” identity (1.51 (95% CI [0.97, 2.24])) and species identity (2.02 (95% CI [1.82, 2.24])) likely explained variation in the population-level intercept. Forest plant origin was the only covariate whose standard deviation (0.77 (95% CI [0.33, 1.32])) suggested “order”-specific responses. Standard deviations of island size (0.19 (95% CI [0.01, 0.50])), island nutrient subsidies (0.37 (95% CI [0.07, 0.73])), and forest CEC (0.27 (95% CI [0.04, 0.54])) did not indicate “order”-specific effects. However, species-based standard deviations suggested that species identity described variation around population means of all covariates. The species-based standard deviations around island size (0.85 (95% CI [0.74, 0.97])), island nutrient subsidies (1.07 (95% CI [0.94, 1.21])), forest CEC (0.90 (95% CI [0.79, 1.02])), and forest plant origin (1.51 (95% CI [1.32, 1.73])) were all high, suggesting species-specific responses.

To further explore the large estimated standard deviations around the population means, we extracted the “order”-specific and species-specific responses for the intercept and corresponding covariates. The “order”-specific intercept estimates indicated that Psocodea, Formicidae, and Coccoidea were more abundant than the average taxonomic group, while Poduromorpha, Isopoda, and Diptera were less abundant than the average taxonomic group. The “order”-specific effect of forest-plant origin implied that Thysanoptera were less abundant on non-native forests than expected by chance, while Isopoda were more abundant in non-native forests than expected by chance. The 95% credible interval ranges of the other 14 "order"-specific responses overlapped zero, meaning that there was not enough evidence to suggest that taxonomic groupings explained the observed variation. We then extracted the species-level posterior draws and found support for many species-specific responses. After summing the population-specific, taxonomic-group-specific, and 246 species-specific effects, island size had a positive effect on the abundance of 83 species and a negative effect on 16 (147 species did not respond to island size). 43 species responded positively to nitrogen deposits, while 37 responded negatively (167 species did not respond to nitrogen deposits). Cation exchange capacity had a positive effect on 74 species, but a negative effect on 13 (159 species did not respond to island size). Finally, we found evidence that 28 species were more abundant in non-native canopies while 46 were less abundant (172 species did not respond to canopy origin).

Table 1. The hierarchical model parameter estimates

	Mean Estimate	Est. Error	Lower 95%	Upper 95%	Rhat	Bulk ESS
Population-Level Effects						
Intercept	-1.26	0.46	-2.16	-0.35	1.00	1874
Island Size	0.29	0.29	-0.28	0.87	1.00	3127
Island Nutrient Subsidy	-0.01	0.26	-0.53	0.49	1.00	2758
Forest CEC	0.51	0.2	0.11	0.89	1.00	2563
Forest Origin	-0.21	0.35	-0.88	0.49	1.00	2499
"Order" Identity Standard Deviations						
Intercept	1.51	0.33	0.97	2.24	1.00	1893
Island Size	0.19	0.13	0.01	0.5	1.01	334
Island Nutrient Subsidy	0.37	0.16	0.07	0.73	1.01	714
Forest CEC	0.27	0.12	0.04	0.54	1.01	581
Forest Origin	0.77	0.25	0.33	1.32	1.00	1405
Species Identity Standard Deviations						
Intercept	2.02	0.11	1.82	2.24	1.00	1365
Island Size	0.85	0.06	0.74	0.97	1.01	1115

Island Nutrient Subsidy	1.07	0.07	0.94	1.21	1.00	1474
Forest CEC	0.9	0.06	0.79	1.02	1.00	1761
Forest Origin	1.51	0.11	1.32	1.73	1.00	1732
Island Identity Standard Deviation						
Intercept	0.55	0.25	0.08	1.1	1.00	1335
Forest Identity Standard Deviation						
Intercept	0.23	0.16	0.01	0.59	1.00	873
Tree Identity Standard Deviation						
Intercept	1.06	0.1	0.87	1.27	1.00	2074
Family Specific Parameter						
zi	0.49	0.01	0.48	0.51	1.00	12813

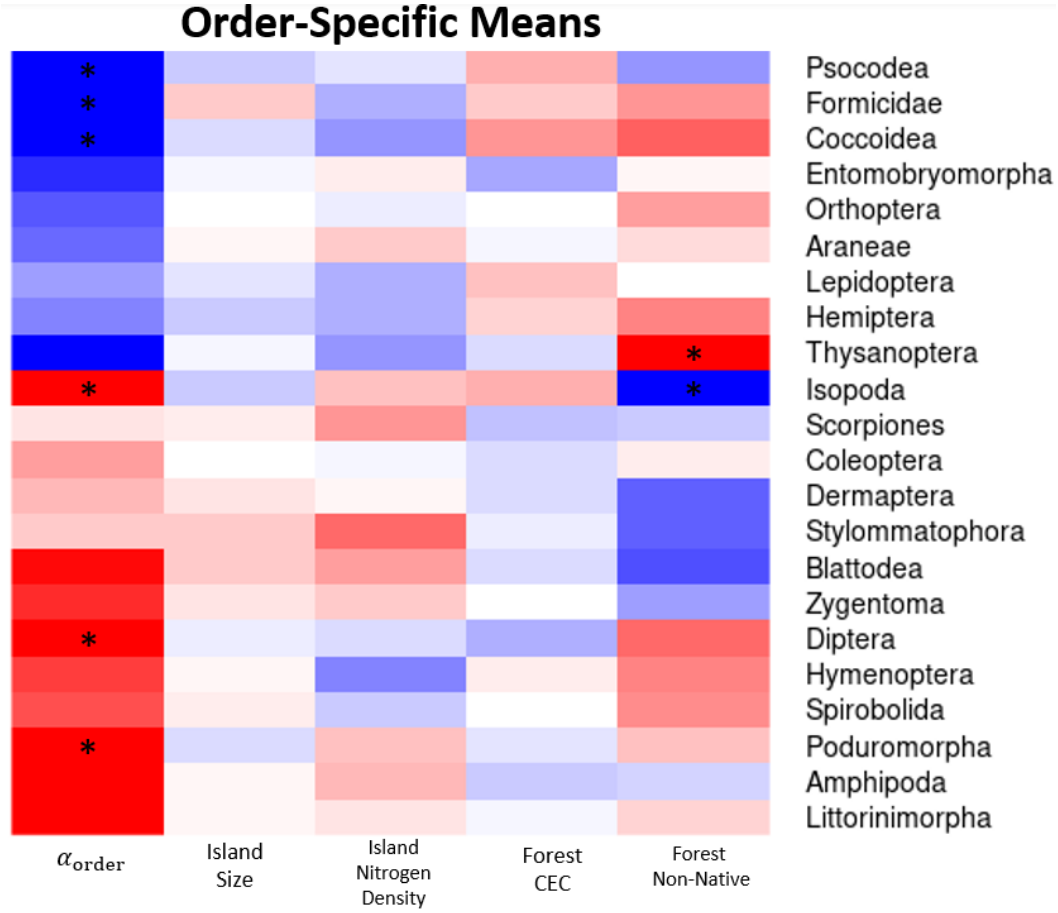

Figure 3. The “order”-specific mean estimates of the intercept and response to covariates. Blue estimates suggest that “order”-specific responses are more positive than the average population, while red estimates suggest that “order”-specific responses are more negative than the average population. Cells with an asterisk denote mean estimates whose 95% credible interval does not include zero.

1.3.3 Scale Random Effects

To account for the hierarchical structure of the spatial scales and unmeasured environmental traits, we included nested random effects for trees within forests within islands. The estimated island-level standard deviation around the population-level intercept suggested some small variation in invertebrate abundances between islands with a 95% credible interval ranging from 0.08 to 1.10. The forest-level standard deviation around the island-specific mean estimated little variation explained by forest identity. Further exploring the island-specific and forest-specific mean random intercepts revealed that all 95% credible intervals overlapped zero, providing no evidence for an effect (Figure 4). In contrast, the tree-level standard deviation around the forest-specific mean effect suggested tree-specific variation with a standard deviation 95% credible interval between 0.87 and 1.27. We examined the tree-specific random effects and found that 12 trees had higher community abundances than expected while 9 trees had less (and 66 trees were not different from the mean).

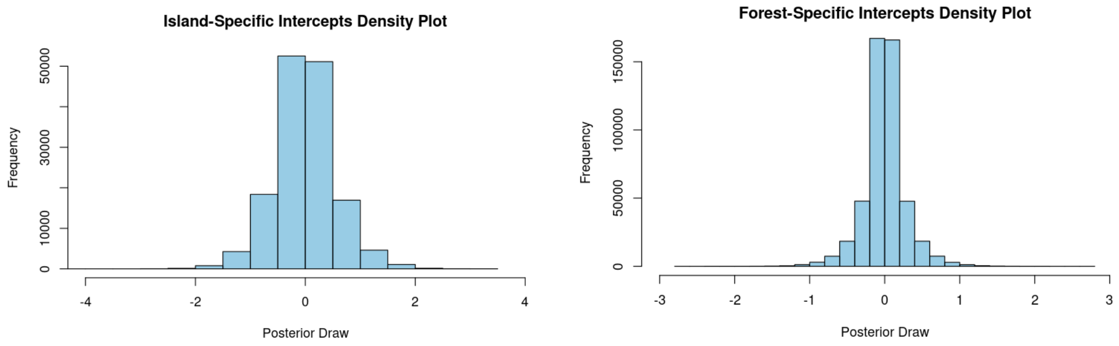

Figure 4. Density plot of the (A) island-specific random effects and (B) forest-specific random effects. The 95% credible intervals of all island-specific and forest-specific random effects overlapped zero.

1.3.4 Species Residual Associations

To explore the influence of biotic interactions, we examined species residual associations both between and within islands, forests, and trees (Figure 5). Our residual association matrices estimated 2403, 2283, and 2181 positively correlated species pairs and 30, 23, and 716 negatively correlated species pairs at the island, forest, and tree scales respectively. As a result, at these same scales, we found no evidence for residual associations between 27702, 27829, and 27238 species pairs (Table 3). Across scales, the number of positively associated species pairs represented between 8.0 to 7.2% of the cells in the association matrix. The number of negatively associated pairs represented 0.0% at the island and forest scale to 2.4% of cells in the tree scale association matrix. Although associations were rare, these results matched the predictions that positively associated pairs would be detectable across scales, while more negative associations would be detected at the smallest scale. In terms of magnitude, the species association estimates centered near zero ranging from -0.30 to 0.95 at the island scale, -0.30 to 0.95 at the forest scale, and -0.66 to 0.95 at the tree scale. And within the tree scale, 198 negative associations were of greater magnitude (less than -.3) than the most negative association in either the island or forest. When comparing the species pair estimates across scales, there was little difference between the island and forest association matrices, which was also supported by the small estimated standard deviation at the forest scale. Their average difference was 0.00 but species pair differences ranged in value from -.61 to .41 (Figure 6). The forest-level association matrix had fewer similarities compared to the tree-level matrix, with an average difference of 0.02 ranging from 0.99 to -.97. From the forest to tree matrix, 0 associations changed from negative to positive and 0 from positive to negative. 20 pairs remained negative and 1721 remained positive. 563 pairs changed from positive to no association, while 3 changed from negative to no association. 462 pairs changed from a neutral to a positive association, while 696 changed from a neutral to a negative association. 26,671 associations remained neutral across scales.

Table 2. The count of species pairs that are estimated as positive, neutral, or negative in the residual association matrix for each spatial scale.

Species Residual Associations

	Positive	Neutral	Negative
Island	2403	27702	30
Forest	2283	27829	23
Tree	2181	27238	716

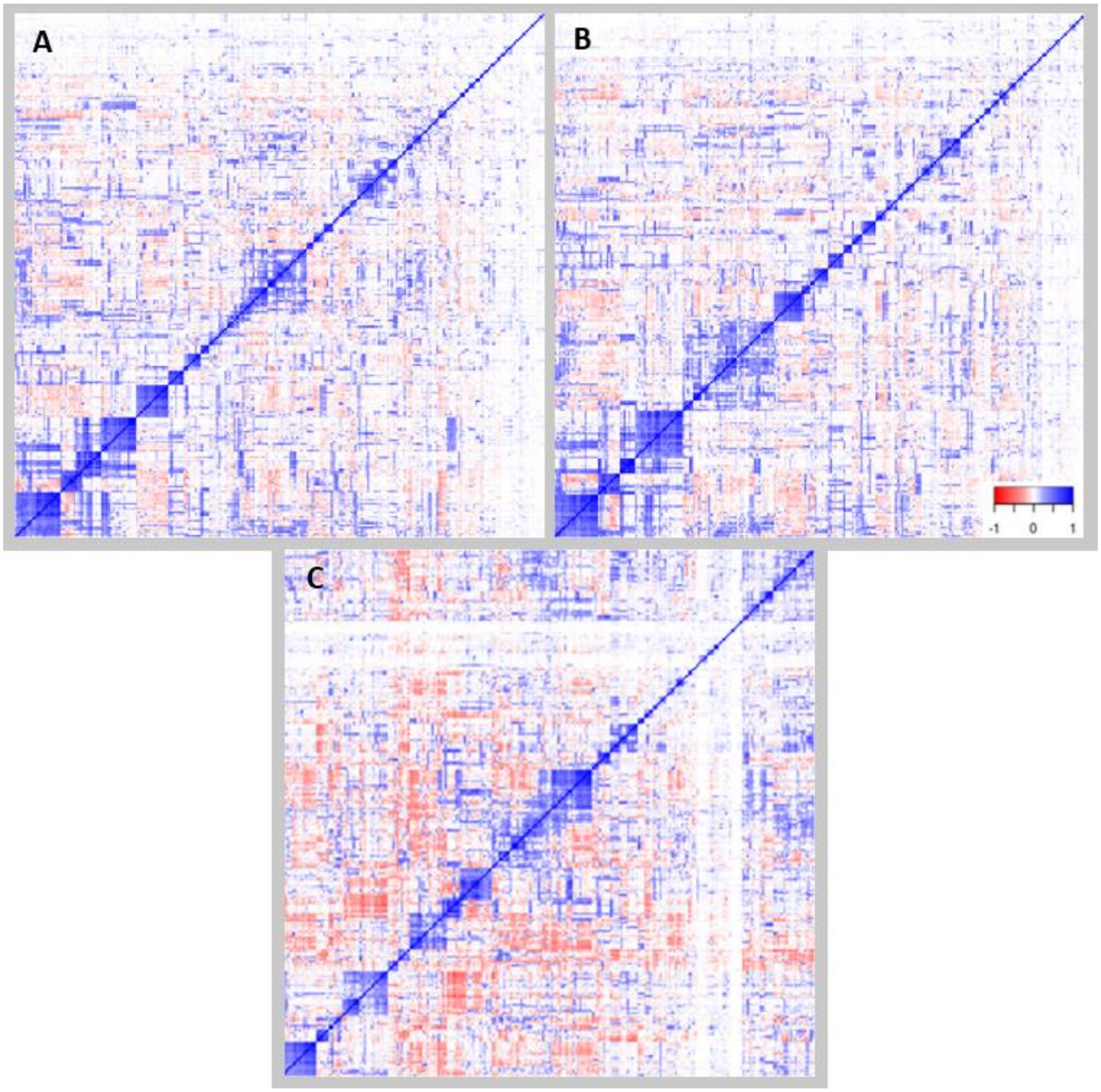

Figure 5. Estimated residual-association matrices for species pairs at the (A) island, (B) forest, and (C) tree spatial scales.

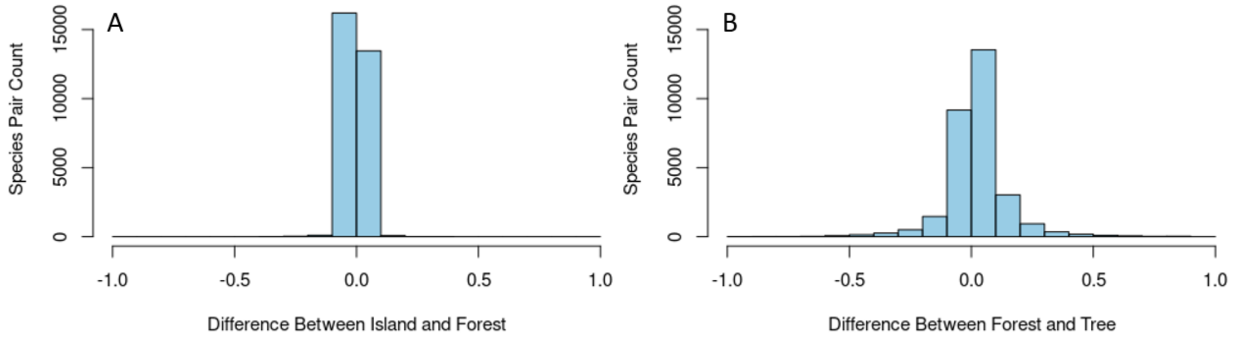

Figure 6. The distribution of species pair differences between the (A) island and forest association matrices and (B) forest and tree association matrices.

In the tree-association matrix, species had an average of 13.2 positive associations, 7.3 negative associations, and 224.5 neutral associations. The species with the highest number of negative associations was the big-headed ant, *Pheidole megacephala*, which had negative associations with 23.4 (9.6%) species (Figure 7). The scale insect, *Aleurodicus disperses,* had the second highest number of negative associations with 23.2 (9.5%), followed by an orb weaver, *Neoscona theisi,* with 19.9 (8.1%) negative associations. A ladybird beetle, *Sticholotis ruficeps,* had the highest number of positive associations with 27.1 (11.1%) pairs. This was followed by the cricket, *Ornebius* sp., with 25.0 (10.2%) positive associations, the ladybird, *Rhyzobius lophanthae*, with 24.4 (10.0%), and the goblin spider, *Opopaea deserticola*, with 24.0 (9.8%). The species pair with the highest estimated positive association was Collembola Morph C and *Scymus* sp. 1 (another ladybird), with an association coefficient of 0.95. This pair was followed by the association between *Trigonotylus brevipes* (a plant bug) and dark tip hemiptera morph with a positive estimate of 0.92. The species pair with the highest estimated negative association was *Coccophagus ceroplastae* (an aphelinid parasitic wasp) and *Neoscona theisi* (an orb weaver spider) that had an association coefficient of -0.66. This was followed by *Coccophagus ceroplastae* and *Karnyothrips melaleucus* (a thrip) which had a negative association of -0.63.

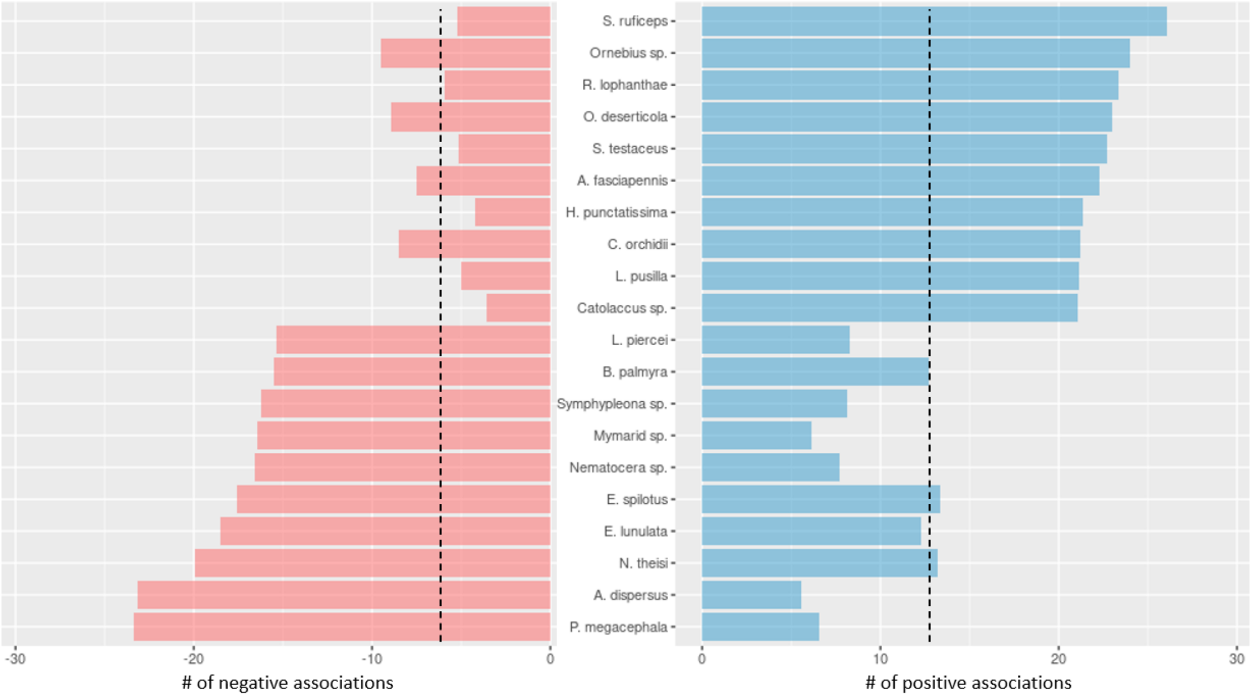

Figure 7. The 20 species with the greatest number of positive and negative associations at the tree level. The dotted lines illustrate the average number of positive/negative associations in the species community.

1.4 Discussion

1.4.1 Population, "order", and species-specific effects

Somewhat unexpectedly, arthropod populations only responded uniformly to one of the four large-scale abiotic community assembly processes we evaluated. Only CEC, an estimate of soil quality, solicited a positive correlation from arthropod populations. This effect occurred regardless of island size, forest type or nutrient deposition rate. This suggests that increasing soil quality has a bottom-up effect on arthropod populations at Palmyra. It also suggests that soil quality and availability may be the factor limiting total arthropod biomass in the system, a situation that makes sense on an atoll.

Native status of the canopy was the only indicator that higher-level (order or similar) taxonomic groups responded to. Native status of groups like scale insects, may also indicate preferences for introduced host plants. Though, this is not the only explanation. For example, Thysanoptera were less abundant in non-native forests, while the opposite was true for Isopoda, and it is not clear why. Another explanation is that our sampling methods may explain the relative abundance of some higher-level groups.

With respect to variation in abundance among groups, it is important to consider that fogging is most effective at sampling canopy species but not all groups are found primarily in the canopy. The three groups that were over-represented in our samples are all strongly associated with trees. At Palmyra, bark lice (Psocodea) and scale insects (Coccoidea) are only found on trees, and due to the shallow soil horizon, most ant nests at Palmyra are in trees or in the rhizome mats of ferns. The three groups that are under-represented in our samples may not be well-sampled by fogging. Springtails (Podumorpha) and isopods (Isopoda) are strongly associated with soil habitats and more abundant in the litter layer on the ground than in the canopy. Diptera are likely abundant in the canopy, but as strong fliers more of them may have been able to escape our sampling prior to succumbing to the fog. Thus, variation among orders in this study is, in part, based on variation in the extent to which fogging captures different groups, more than on their relative abundances on Palmyra per se.

There were strong, species-level responses to all the abiotic assembly processes we evaluated. 40.2%, 32.5%, 35.4%, and 30.1% of the 246 species demonstrated species-specific responses to island size, island nitrogen deposits, forest CEC, and forest canopy type, respectively. Five times as many species responded to island size positively than negatively. Similarly, six times as many species preferred high quality soils (high CEC). Both trends are in line with predictions that larger islands, with higher quality soils should support larger populations of more insects. However, species responses to nitrogen deposition rates were more evenly split. This suggests that about one-third of consumers are partitioning habitats based on nutrient availability and competitive ability. Lastly, as suggested at the "order"-level there appear to be strong preferences within certain groups for native and non-native canopies. For example, while present in all canopies, the native tree crickets strongly prefer native *Pisonia* and *Tournefortia* canopies. On the other hand, non-native cockroaches strongly prefer introduced *Cocos* canopies.

1.4.2 Biotic Interactions

Most observed species correlations were neutral. This suggests that many species in this study do not interact with one another directly or indirectly. Here we will focus on direct interactions (we discuss indirect interactions in a food web context below). To interact directly, two species must encounter each other and be compatible for interaction. Many species at Palmyra likely do not encounter each other because they do not overlap in microhabitat. At Palmyra, while they may be on the same tree, bark lice are generally found on the trunk and scale insects on the underside of leaves. Both were over-represented in our samples but both are unlikely to encounter each other outside of their microhabitat. If they did encounter each other, they would be unlikely to interact directly as neither is a compatible food source or competitor for the other (same tree different parts). The number of neutral correlations was similar across scales, but positive and negative correlations changed with scale. The number of positive correlations was consistently higher than negative correlations across all scales. But the number of positive correlations slightly decreased with narrowing scale, while the number of negative correlations increased as scale decreased. Negative correlations were stronger and 20-30 times more abundant than positive correlations at the smallest scale.

The big-headed ant, *Pheidole megacephala*, was the consumer with the largest number of negative associations. This species also had the largest diet-breadth (generality) in the system, which likely explains the high number of negative correlations if ants reduce prey of competitor density. The three species that had the strongest negative correlations with *P. megacephala* were all detritivores/scavengers that shared complete diet overlap. Additionally, *P. megacephala,* is a predator on each of their larval and juvenile stages. A < 1cm herbivorous snail had the next largest negative correlation, and *P. megacephala* is its main predator. *P. megacephala's* strongest competitor, the stinging ant, *Tetramorium bicarnatum,* had the fifth largest negative correlation. These examples offer mechanisms to support how some negative associations with aggressive ants could be based on biotic interactions.

The spiraling white-fly (*Aluerodicus dispersus*) also had a large number of negatively correlated species, and further exploration suggest that these were a result of interference competition and predator deterrence. Spiraling white-flies are widespread pests in the Pacific (Mani & Krishnamoorthy, 2002; Balikai & Pushpalatha, 2018), and are among the most abundant consumer species at Palmyra. They aggregate in large numbers on the underside of leaves, often covering plants in excessive amounts of white, flocculant, waxy secretions. This has a strong deterrent effect on both competitors and predators. 11 of the 12 strongest negative correlations with *A. dispersus* were generalist predatory arthropods. And six of the top 12 correlations were hunting spiders. This suggests that generalist predators might be deterred by the waxy secretions and that other prey are not found nearby. The major predators of white-fly at Palmyra are dragonflies which can avoid the waxy leaves by hawking them out of the air.

While a few species had many negative correlations, many species had a few strong ones. The two strongest observed negative correlations were both associated with the parasitoid wasp *Coccophagus ceroplastae.* This wasp is a specialist parasitoid on *Pulvinaria urbanicola,* a scale insect found on native *Pisonia* trees. The orb-weaver spider is a direct predator of *C. ceroplaste* and had the strongest negative correlation with it. Surprisingly, the thrip, *Karnyothrips melaleucus* also had a very similar negative correlation with *C. ceroplaste*. This thrip is a specialist predator on armored scale insects like *Pulvinaria urbanicola*. *Karnyothrips melaleucus* is a direct competitor with *C. ceroplaste* and an intraguild predator. *K. melaleucus* likely also attacks scales infected with C. *ceroplaste* larva – thus incidentally predating them. Negative correlations offer a mix of direct and indirect explanations, while positive correlations appear to be primarily indirect.

The ladybird beetle, *Sticholotis ruficeps*, had the highest number of positive correlations of any species on Palmyra and is most abundant in introduced *Cocos* and native *Tournefortia*. *S. ruficeps* was positively correlated with both predators and herbivores. The two species it was most strongly correlated with were, like itself, predators on scale insects. The next two most positively correlated species were herbivores on the plants where it is most common. Next, was a small hunting spider similar in trophic position to *S. ruficeps,* and then another herbivore common to the same host plants. This points to the potential for competitors to be positively associated at some scales due to aggregation to a shared resource.

1.4.3 Assumptions

There were several assumptions that this analysis made about the community and the system. First, because the residual values were based on the effect of environmental predictors and nested scales, the resulting residual associations were dependent on the selected covariates and random effects. Although we evaluated several likely explanatory factors, and believe others (ex. temperature, rainfall, humidity, sunlight, tree height) to vary little, smaller scale effects, and particularly microhabitats, could be sources of unmeasured variation. Including other environmental variables would result in different residuals for each species, which could lead to different association patterns and predicted biotic interactions. In our analysis, the environmental predictors and spatial scales that we chose are ecologically justified, however we acknowledge that other factors might better explain variation in species distributions. Another assumption is based on the additive structure of the random effects. The nested nature means that predicted species-specific effects were expected to have similar effects as their taxonomic "order", and "orders" were expected to follow the population responses. Rare species experienced the largest amount of shrinkage because they provide the model with the least amount of information. As a result, species-specific effects, especially of rare species that matched the "order"-specific or population-specific response, should be interpreted with caution.

1.5 Conclusion

In conclusion, our hierarchical model of empirical observations demonstrated the effect of scale-specific assembly processes on species communities at three naturally-defined spatial scales. Arthropod populations tended to be more abundant on islands with more nutrient subsidies, while arthropod species varied in their response to island nutrient deposits, forest soil fertility, and canopy origin. Residual associations also suggested that biotic interactions impacted the distributions of some species, but not others. Although most negative associations were potentially due to negative interactions between predators and prey, or between competitors, there were also positive associations between predator and prey and between competitors. Such interactions might be discernable by considering even smaller spatial scales where species interact. For future directions, we could explore the effect of functional traits on species responses to assembly processes. For example, functional feeding groups (Wimp, 2019; Lima *et al.,* 2022), origin (native or non-native) (Buckley & Catford, 2016; Sorte *et al.,* 2013), or dispersal abilities (winged or non-winged) (Carlquist, 1974; As, 1984) may help explain species-specific responses. We could also examine the impact of assembly processes at the atoll scale. The factors included in this study may not universally impact distributions on other atolls as there will be variation in community diversity and functionally. Therefore, comparing our results to similar analyses from other atolls could be an insightful next step.

Chapter 2: The influence of host traits on parasite communities in Palmyra Atoll's sand flat fish

Introduction 2.1

2.1.1 Biological Scales and Parasite Communities

A fundamental aim of parasite ecology is to understand the factors that structure parasite communities. However, these factors (like host density, trophic level, or body mass) will have scale-dependent effects with some influencing parasites found in a host species and others determining parasites in a host individual. This nested structure means that parasite communities observed at smaller scales have also been narrowed by processes impacting larger scales (Poulin & Valtonen, 2001; Valtonen *et al.,* 2001). Few studies focus on factors that structure parasite distributions at more than one scale, and the ones that do target a selection of parasites or hosts in an ecosystem (Benavides *et al.,* 2012; Mwita & Nkwengulila, 2008; Vignon & Sasal, 2010; Pence, 1990; Fuentes *et al*., 2004; Linardi & Krasnov, 2013). Additionally, many studies seeking factors that impact parasite distributions use datasets compiled from many different sources across large geographic or temporal ranges (Takemoto *et al.,* 2005) which could add biases and variability to the analyses. These studies do not truly reflect the distribution of the parasite species pool available to infect a sympatric host community in a given location. In this study, we examined how parasites of a single compound community were distributed throughout a fish host community at both the host species and host-individual scales. We used a multi-level model to assess the effect of host traits on parasite occurrences across and within (almost) all fish hosts of a sand flat habitat.

Unlike free-living species, parasite communities are biologically defined units of replication that are intimately housed within hosts (Hechinger, 2013). They can be defined by three hierarchical scales of community organization (Bush *et al.,* 1997): the compound community, the component community, and the infracommunity. The compound community refers to the ensemble of all parasite species infecting a sympatric community of host species. The component community encompasses all parasite species infecting a host population from a single species. And the infracommunity includes all parasite species infecting a single host individual. These biological scales are nested such that a host's infracommunity can exclusively contain a subset of parasites occurring in that host's component community, and a parasites' occurrence probability can differ between these scales. Some studies that examine factors structuring parasite community distributions have been performed on the infracommunity (Guégan & Hugueny, 1994; Poulin, 1996; Fernandez & Esch, 1991, Rohde*,* 1998), whereas other studies have focused on the component community (Poulin, 1995; Fernandez & Esch, 1991; Abu-Madi *et al.,* 2000; Guegan & Kennedy, 1996; Poulin, 1997; Locke *et al.,* 2014). And a few have considered structure at both scales (e.g. Lafferty *et al*., 1994; Goater *et al*., 1987). In this study, a comprehensive dataset allowed us to examine how parasites in a compound community were distributed throughout the component and infracommunities of Palmyra Atoll's sandflat fishes. We explored the host traits that influenced parasite communities across and within these two biological scales.

2.1.2 Host Ecological Characteristics and Taxonomy

Parasite organization at each scale should depend on the encounter and compatibility rates between hosts and parasites (Euzet & Combes, 1980; Lagrue *et al.,* 2011). And these rates are determined by ecological traits and evolutionary history of both parasite and host. Variation in these characteristics should result in variation of parasite occurrences at each scale. In this study, the assembly processes of interest were host species and host individual traits.

Similar to how free-living species are affected by abiotic factors in their surrounding habitat, parasite occurrences should correlate with variation in several host-species characteristics. Host characteristics that represent differences in host behavior, like trophic level, will alter the encounter rates and identities of acquired parasites (Ranta, 1992; Poulin & Fitzgerald, 1989; Timi *et al.,* 2011). Some traits, like vulnerability to predation, diet generality, and habitat density, describe how a host species might be connected with other potential fish host species in the system, alluding to the likelihood of being included in a parasite life cycle (Lafferty *et al.,* 2006; Guegan & Kennedy, 1993; Hudson *et al.,* 1992; Benesh *et al.,* 2021). And individual host body metrics, like length or mass, serve as a proxy for exposure rates and host age, both of which will correspond to parasite accumulation (George-Nacimento *et al.,* 2004; Poulin & George-Nacimento, 2007; Guegan *et al.,* 1992; Poulin & Valtonen, 2001). Several of the above host characteristics will have scale-dependent effects on parasite occurrences. As an example, host species living at high population densities are expected to have higher component community richness (at least for directly transmitted parasites) than low density host species (Morand & Poulin, 1998). This is because high density populations provide more pathways for parasites to invade hosts (Anderson & May, 1978) and can more readily sustain adult parasite populations (Bell & Burt, 1991). However, because individuals of a specific host species will experience the same density, host density cannot explain differences in infracommunity composition. Instead, parasites occurring in the infracommunity may be more influenced by host traits like individual body mass. Therefore, the effect of some host characteristics may be important at one biological scale, but not another. Here, we expected that host species traits (like host density, mean size, or trophic level) should affect parasites within the component community, whereas host individual traits (like individual body size) should affect parasites within an infracommunity.

Host taxonomy could play a dual role in organizing parasite communities. Because hosts that are closely related have more behavioral traits in common (Poulin & Rohde, 1997; Lagrue *et al.,* 2011), closely related hosts were expected to acquire more similar parasite communities simply based on encounter rates. And, since parasites tend to co-evolve with their hosts (Anderson & May, 1982), overlap in parasite compatibility should also elicit parasite community similarity in closely related hosts. Because of host trait overlap and host-parasite co-evolution, not accounting for host taxonomy in a parasite community model could lead to biased inference and predictions (Poulin, 1995). As a result, we accounted for host family in our analysis to capture variation explained by similarities between fish phyla. In our statistical model, we grouped hosts by taxonomic family and allowed variation of a random intercept for each.

2.1.3 Parasite Species-Specific Responses

Host characteristics were not expected to affect all parasite species uniformly. Instead, the effect may be parasite species-specific. As an example, host generality (diet-breadth) was expected to positively correlate with richness of parasites that are acquired through diet because ingesting a wider variety of diet items increases a host's likelihood of encountering a novel parasite species (Guegan & Kennedy, 1993). However, because host generality is a metric of diet, it should not affect directly acquired parasites. Thus, parasite transmission strategy may influence a parasite's response to host diet generality. In this study, we examined species-specific responses to host traits. We also completed a post-hoc analysis to explore if parasite life stages may influence their species response to host traits.

2.1.4 Approach

In this study, we examined the interaction between host traits and parasite occurrences at two biological scales: the component community within a host species and the infracommunity within a host individual. This study analyzed parasite occurrence data from 33 fish host species living in Palmyra Atoll's intertidal sand flats. We aimed to address 4 questions: 1) how do parasite occurrences change across biological scales, 2) how do scale-specific host traits affect parasite occurrences in the component and infracommunities, 3) how does host taxonomic family influence parasite species occupancy, and 4) can parasite species-specific responses to host traits be explained by parasite life stages? For the first three questions, we used a hierarchical occupancy model to estimate occurrence probabilities and the effect of host traits and host family on parasite occupancy. We predicted that component community occurrence probability would increase with host generality (Rasmussen & Randhawa, 2018; Locke *et al.,* 2014), species body mass (Guégan *et al.,* 1992; George-Nacimento *et al.,* 2004; Poulin & George-Nacimento, 2007), habitat density (individuals per hectare) (Morand & Poulin 1998), and trophic level (Lafferty *et al.,* 2006; Chen *et al.,* 2008). We expected that infracommunity occurrence probability would increase with individual weight (g) (Guégan & Hugueny, 1994). We also predicted that host species within the same family would share more parasites such that host family would explain variation between component communities (Poulin & Rohde, 1997; Lagrue *et al.,* 2011; Anderson & May, 1982). For the last question, we conducted a random effects post-hoc analysis to explore if parasite life stages could explain the estimated species-specific responses to host traits.

2.2 Methods

2.2.1 Study system and study design

This study was conducted in the intertidal sand flats of Palmyra Atoll. Located 1680 km south of Hawai'i, Palmyra Atoll became a US National Wildlife Refuge in 2001 and is part of the Pacific Remote Islands Marine National Monument established in 2014. It is a remote and relatively pristine coral atoll that has never supported permanent human habitation or a commercial or subsistence fishery. As a result, Palmyra contains a trophically intact marine community with a high apex-predator biomass (Stevenson *et al.,* 2007). Studies have found that fishes, and most notably sharks, from Palmyra have higher parasite richness, prevalence, and abundance than fishes from a nearby, heavily fished island (Lafferty *et al.,* 2008).

Palmyra Atoll contains 3.14 hectares of intertidal sandflats that provide various ecosystem services and habitat for a rich species assemblage. McLaughlin (2018) examined Palmyra Atoll's intertidal sand flat food web. The resulting data were of unprecedented scope and quality, surveying parasites in 35 fish species, dissecting 642 fish individuals, and quantifying 70 parasite species. This comprehensive dataset provided us with the opportunity to analyze factors that affected parasite occupancy across multiple biological scales. McLaughlin agreed to share these data and collaborate on this analysis. Sites were distributed throughout the atoll's lagoonal system and fishes were sampled by seine and spear. Site selection and study design details are described in McLaughlin (2018).

2.2.2 Dataset and Data Collection

Fish were dissected using a parasitological examination designed to detect most eukaryotic parasites. Details are outlined in McLaughlin (2018). The number of individuals sampled from each fish species ranged from 5 (*Carangoides ferdau*) to 63 (*Valamugil engeli*), with an average of 19. Because parasite life stages are expected to infect different host species, parasite species were further separated by life stage (adult, larva, metacercaria, cystacanth, or plerocercoid). This resulted in 84 unique parasite species-stage identifications. Of the 2,772 possible fish species-parasite stage links (33 x 84), 353 links were observed in dissections (Figure 1). Further, of the possible 53,928 fish individual-parasite stage links (642 x 84), 1,809 were observed in dissections. In total, this dataset counted 81,201 parasite individuals.

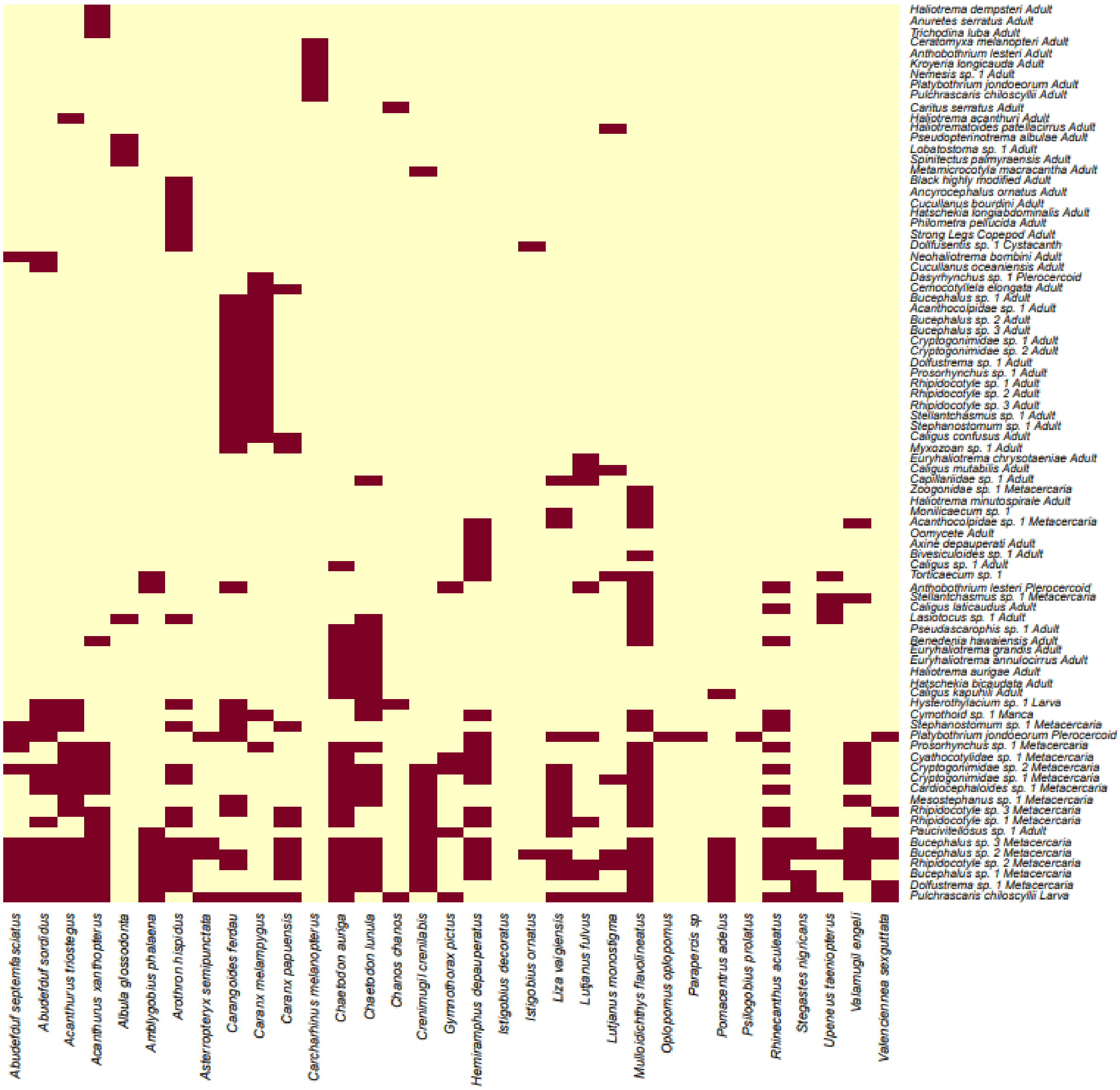

Figure 1. A heatmap of the parasites and host links found in the raw observations

Fish trait measurements (average species weight and individual weight) and estimates (species generality, species vulnerability, species density) were outlined in McLaughlin (2018). In the original dataset, fish identifications and their traits were differentiated by life stage (i.e. adult and juveniles). However, life stages were not evenly sampled across fish species. Therefore, in this analysis, we merged adults and juveniles to represent the same species. This applied to 21 fish species. For fish species traits, we used adult values.

2.2.3 Statistical Analysis and Model Specifications

We used a Bayesian multi-level occupancy model to assess parasite presence/absence data in Palmyra Atoll's sandflat fish species. We used a statistical model (Doser *et al.,* 2022; Dorazio and Royle, 2005) that was originally designed to assess free-living species occurrence based on environmental factors within sites. In the original model framework, sites can be surveyed more than once for replicate samples, which is then used to estimate the detection probability of species at each site. However, the framework of this occupancy model can also be applied to parasites infecting host species. In our analysis, we modeled host-parasite links, with parasite species analogous to free-living species, host species serving as "sites", and an infection representing a site-level occurrence/occupancy. Host individuals within a host species are replicates of the same "site" and correspond to a "visit" in traditional occupancy modelling. Host traits (both at the host species and host individual scale) are analogous to the environmental factors expected to affect occurrence probabilities at the site and visit scales. And finally, analogous to spatial distances among sites, host species that are close relatives are expected to share a more similar parasite community than distantly related hosts. From this model, we examined the differences between occurrence probabilities in the component community and infracommunity. We estimated the effects of host species and host individual traits at both of these scales. We also evaluated parasite species-specific responses.

The nested model was separated into two parts (the component community model and the infracommunity model), which takes the form:

Component Community Model

$$z_{ij} \sim \text{Bernoulli}\ (\psi_{ij})$$

$$\text{logit}(\psi_{ij}) \sim \beta_{0i} + \beta_{1i}\ \text{DietGenerality}_{\text{j}} + \beta_{2i}\ \text{BodyMass}_{\text{j}} + \beta_{3i}\ \text{HabitatDensity}_{\text{j}} + \beta_{4\text{i}}\text{TrophicLevel}_{\text{j}} + u_{if}$$

$$\beta_i \sim \text{Normal}\ (\mu_{\beta i}, \tau_{\beta i}^2)$$

$$\mu_{\beta i} \sim \text{Normal}\ (0, 2.72)$$

$$\tau_{\beta i}^2 \sim \text{InverseGamma}\ (0.1, 0.1)$$

$$u_{\text{if}} \sim \text{Normal}\ (0, \sigma_u^2)$$

$$\sigma_u^2 \sim \text{InverseGamma}\ (0.1, 0.1)$$

Infracommunity Model

$$y_{ijk} \sim \text{Bernoulli}\ (p_{ijk}\, z_{ij})$$

$$\text{logit}(p_{ijk}) \sim \alpha_{0i} + \alpha_{1i} * \text{IndividualWeight}_{ijk}$$

$$\alpha_i \sim \text{Normal}\ (\mu_{\alpha i}, \tau_{\alpha i}^2)$$

$$\mu_{\alpha i} \sim \text{Normal}\ (0, 2.72)$$

$$\tau_{\alpha i}^2 \sim \text{InverseGamma}\ (0.1, 0.1)$$

Where z_{ij} represented the true presence (z_{ij} = 1) or absence (z_{ij} = 0) of a host-parasite link between the ith parasite taxon in the jth fish species, and ψ_{ij} was the expected probability of z_{ij}. β_i was a vector of regression coefficients for each host species-specific covariate predictive of ψ_i. Covariates of ψ_i included fish species generality, species body mass, habitat density (individuals per hectare), and adjusted trophic level. Community-level regression coefficients were modeled using a normal distribution with mean $\mu_{\beta i}$ and variance $\tau_{\beta i}^2$. Parasite species-level regression coefficients were treated as random effects with a normal distribution that assumes community-level mean and variance parameters. And the effect of host family *f* on the occurrence of parasite *i* was described by a random intercept u_{if}.

In the infracommunity formula of the model, response variable y_{ijk} represented the presence/absence of the *i*th parasite taxon in the *k*th fish individual of the *j*th fish species. p_{ijk} was the occurrence probability of parasite *i* in the *j*th fish species in fish individual *k*. Covariates at the individual host scale included weight (g) and length (cm). Similar to the occurrence model, community-level regression coefficients, α_i, were modelled with normal mean $\mu_{\alpha i}$ and variance $\tau_{\alpha i}^2$. Species-level regression coefficients were treated as random effects with a normal distribution assuming community-level mean $\mu_{\alpha i}$ and variance $\tau_{\alpha i}^2$.

All numeric predictors were standardized to have a mean of 0 and standard deviation of 1 before analysis by using the scale() function in base R. We did this to improve model convergence. This model was fit using the multi-species occupancy model function msPGOcc() in the spOccupancy package (Doser *et al.,* 2022) in R. Occurrence regression coefficients were assigned priors with a normal distribution and a mean of 0 and variance of 2.72. The variance parameters (τ^2) had priors with an inverse gamma distribution with shape parameter 0.1 and scale parameter 0.1. This allowed some effects to be much larger or smaller than average. We fit our model using 5 MCMC chains. Each chain had a burn in period of 3000 samples, 43000 samples per chain, and a thinning rate of 5, resulting in 40,000 total posterior samples.

After fitting the model, we confirmed MCMC chain convergence and chain resolution using potential scale reduction factor (PSRF) and effective sample size (ESS) statistics.

To assess the effect of scale on parasite occurrence, we extracted the posterior distributions for the ψ_{ij} and p_{ijk} parameters. For each parameter, we summarized the posterior by calculating the mean, standard deviation, and 95% credible interval. Because ψ_{ij} was parasite-host specific, and p_{ijk} was parasite-host-replicate specific, we calculated the mean and standard deviation of p_{ijk} within each parasite-host combination. To illustrate occurrence probabilities at each scale, we used the heatmap() function in base R. To estimate the expected number of hosts that a parasite species-stage infected, we summed all occurrence probabilities across all hosts. This suggested the broad distribution of the parasite species in the host community.

To evaluate host trait effects on component community occurrence, we summarized the community-level mean, standard deviation, and 95% credible intervals of regression coefficient and variance parameters, β_i and $\tau_{\beta i}^2$. Similarly, for the community-level effect of host traits on the infracommunity, we summarized α_i and $\tau_{\alpha i}^2$. After assessing the effect of host traits across the parasite community, we wanted to understand if parasite species-stages differed in their response. From the above model, we extracted the species-specific posterior distributions and summarized their mean and variance parameters. We illustrated these species-level responses with a forest plot using the ggplot() function from the ggplot2 package in R.

To understand if host family was a predictor of parasite occurrences, we examined the species-specific random intercepts for each host family (u_{if}). The parameter σ_u^2 describes the variance around the mean in parasite occurrences across host species. If the variance was high, there was evidence that species-specific occurrence probabilities were influenced by host family.

Finally, we suspected that the effect of host traits may depend on parasite taxon and life stage groups. To explore this suspicion, we conducted a random effects post-hoc analysis for each host trait with high variance. Our response variables were the species-specific mean intercept and regression coefficients estimated in the above occupancy model. We re-scaled the mean effects for a mean of zero and standard deviation of 1. These post-hoc random effect models took the form:

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{lifestage}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 0.5) \quad \text{for } j = 1...14$$

$$\sigma \sim \text{Exponential}(1)$$

where the y_i represented the parasite species-specific mean estimate of either intercept or host weight. $\alpha_{\text{lifestage}[i]}$ was a random effect and the categorical variable was a concatenated parasite taxonomic group and life stage (ex. Trematode metacercaria, Cestode adult, etc.). There were 14 taxonomic group-life stage identifications in the group. This model was estimated using the quap() function in the Rethinking package in R. All predictors were assumed to be independent.

2.3 Results

2.3.1 Computational Results

All potential scale reduction factor statistics (Gelman-Rubin diagnostic values or R-hat) were below 1.02, indicating chain convergence. Additionally, we visually inspected MCMC trace plots for signs of divergence. To assess adequate chain resolution, effective sample sizes (ESS) for each parameter were confirmed to be over >200, with the smallest being 795. Finally, we examined pair plots between the posterior distributions to ensure that no extreme correlations between parameter estimates were present.

2.3.2 Parasite occurrence probabilities between scales

Based on the raw dissection data, the average parasite species infected 4.4 host species in the sandflats. 28 of 84 parasite species-stages were only observed in a single host species. The most commonly observed parasite in the component community was the trematode metacercaria *Bucephalus sp.* 2, infecting 22 of the 33 host species. At the infracommunity scale, parasite species infected an average of 21.5 of 639 host individuals. Of the 84 parasite species-stages, 4 were observed in a single host individual. *Bucephalus sp.* 2 metacercaria was again the most commonly observed parasite, infecting 245 of 639 dissected host individuals. The parasite-host combination with the highest prevalence was the metacercaria *Cyathocotylidae sp.* 1 which infected 20 out of 20 *Hemiramphus depauperatus*, a tropical half-beak fish.

We evaluated the parasite-host dataset with an occupancy model to understand how parasite occurrence probabilities changed across biological scales based on the predicted host trait responses. The mean occurrence probabilities of parasites across all parasite-host species combinations, ψ_{ij}, ranged from 0.002 to 0.96, with an average occurrence probability of 0.16 (95% CI [0.001, 0.76]). Based on occurrence probabilities, host species were estimated to be infected with an average of 13.8 parasite species-stages, ranging from 1.29 parasites species-stages in the host species *Gymnothorax pictus* (peppered moray) to 28.10 in *Chaetodon lunula* (raccoon butterflyfish). The parasite most likely to occur in a randomly selected host species was *Bucephalus sp.* 2 metacercaria with an average mean occurrence probability of 0.68 across all host species. *Bucephalus sp.* 2 metacercaria was estimated to infect 22.46 fish hosts within this system with the broadest host generality. And the least likely parasite was the monogene adults of *Pseudopterinotrema albulae*, expected to occur in only 1.29 host species. We illustrated the parasite species-specific occurrence probabilities in the component community in Figure 2.

The estimated occurrence probability of parasites across host individuals, p_{ijk} * z_{ij}, ranged from an estimated mean of 0.00 to 0.83, with an average occurrence probability of 0.04 (95% CI [0.00, 0.26]) across all hosts. Of expected infracommunities, the highest estimated parasite richness was 8.10 species in an average host individual *C. lunula*, while the lowest was 0.07 parasite species in an average *G. pictus*. At this scale, *Bucephalus sp*. 2 metacercaria was the most likely parasite to occur in a randomly selected host individual with an average probability of 0.33. If one individual of each 33 host species was sampled, *Bucephalus sp*. 2 metacercaria was estimated to occur in 11.12 of them. The parasite *Zoogonidae* sp. 1 metacercaria was the least likely to occur in a randomly selected host individual with a probability of 0.01. We illustrated the parasite species-specific occurrence probabilities in the infracommunity in Figure 3. The sum occurrences across all host species are illustrated in Figure 4.

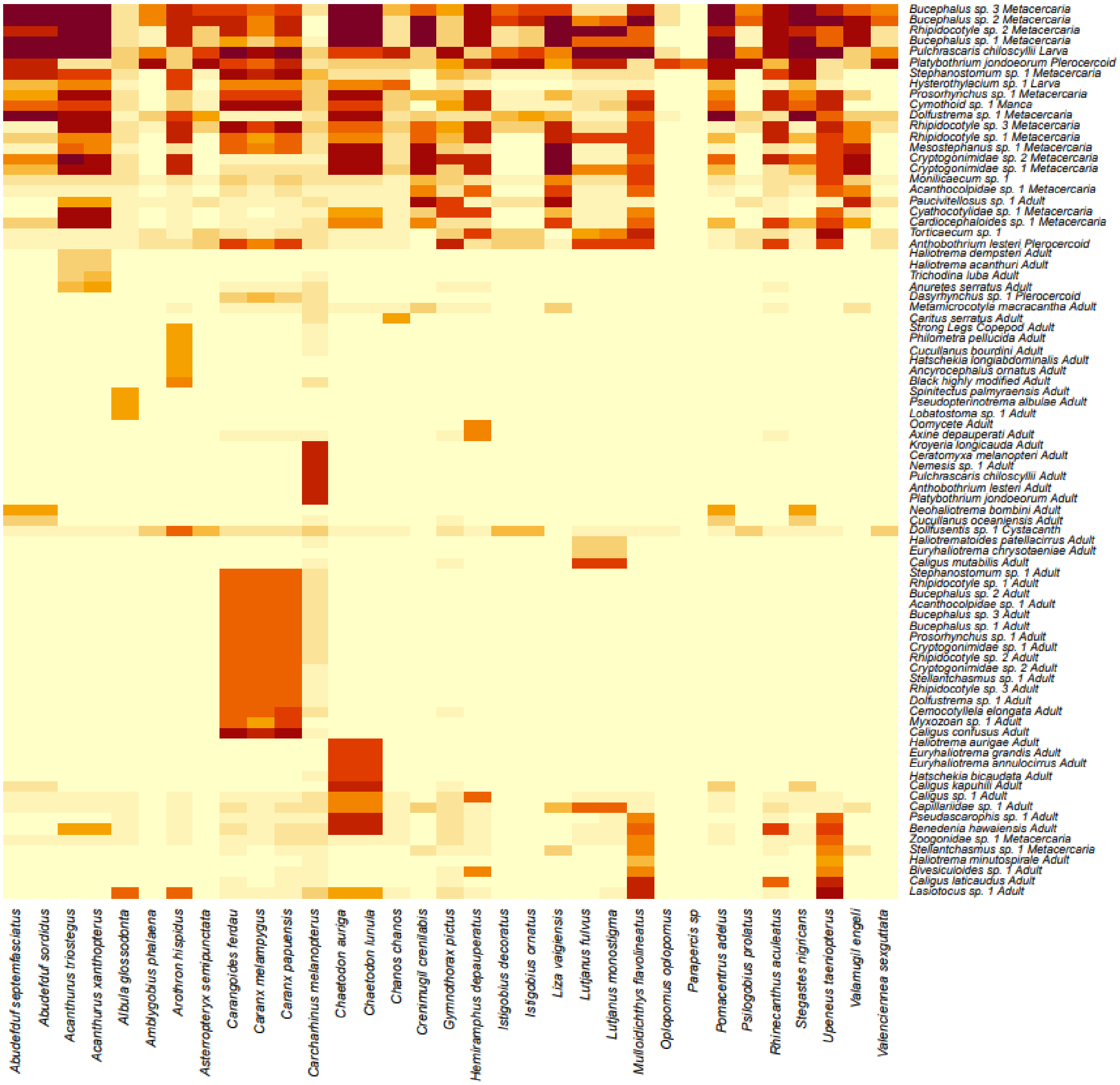

Figure 2. Component community mean occurrence probability heatmap. Cells can range from 0 (lightest) to 1 (darkest)

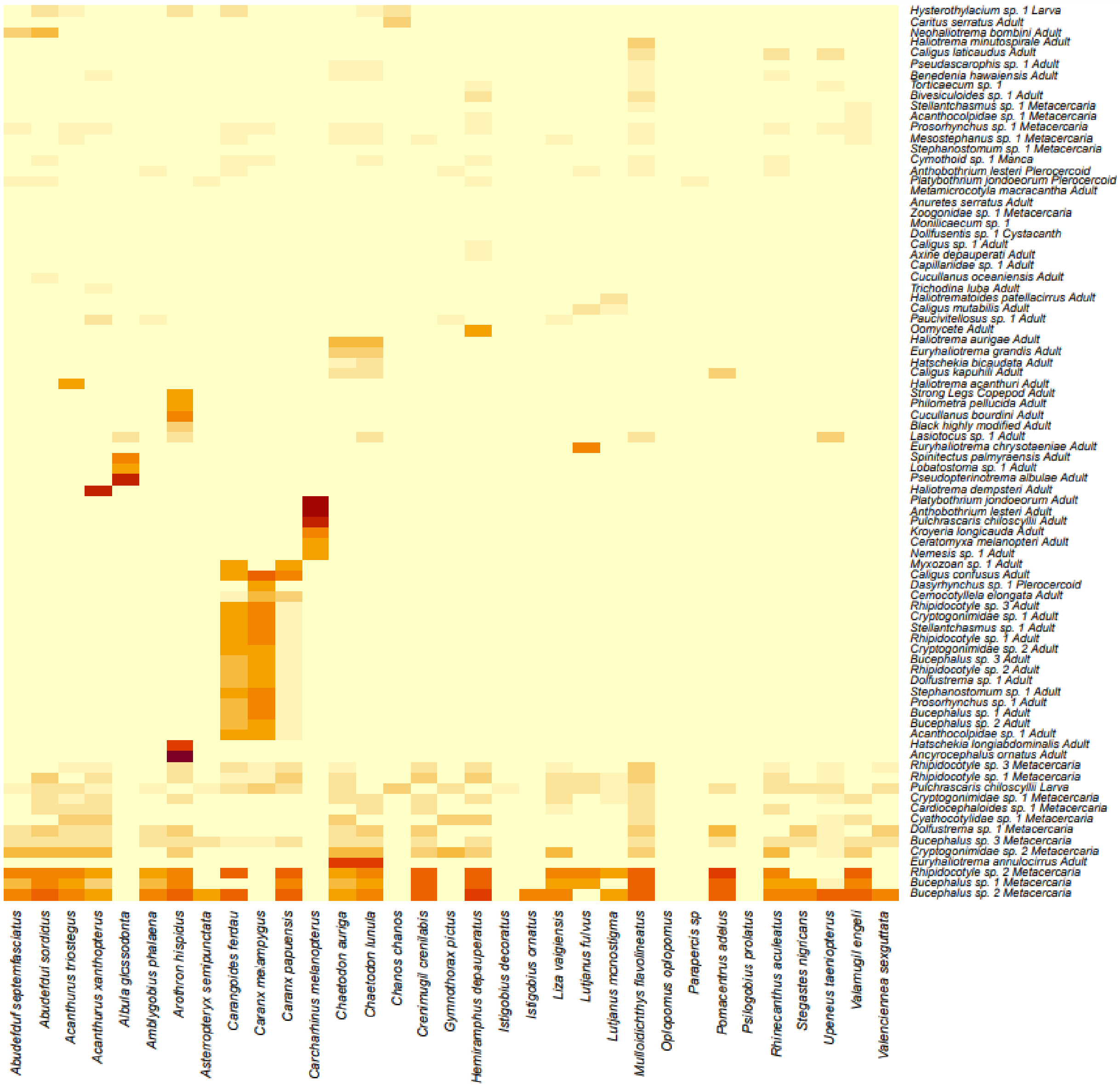

Figure 3. Infracommunity mean occurrence probabilities (prevalence) heatmap. Cells can range from 0 (lightest) to 1 (darkest)

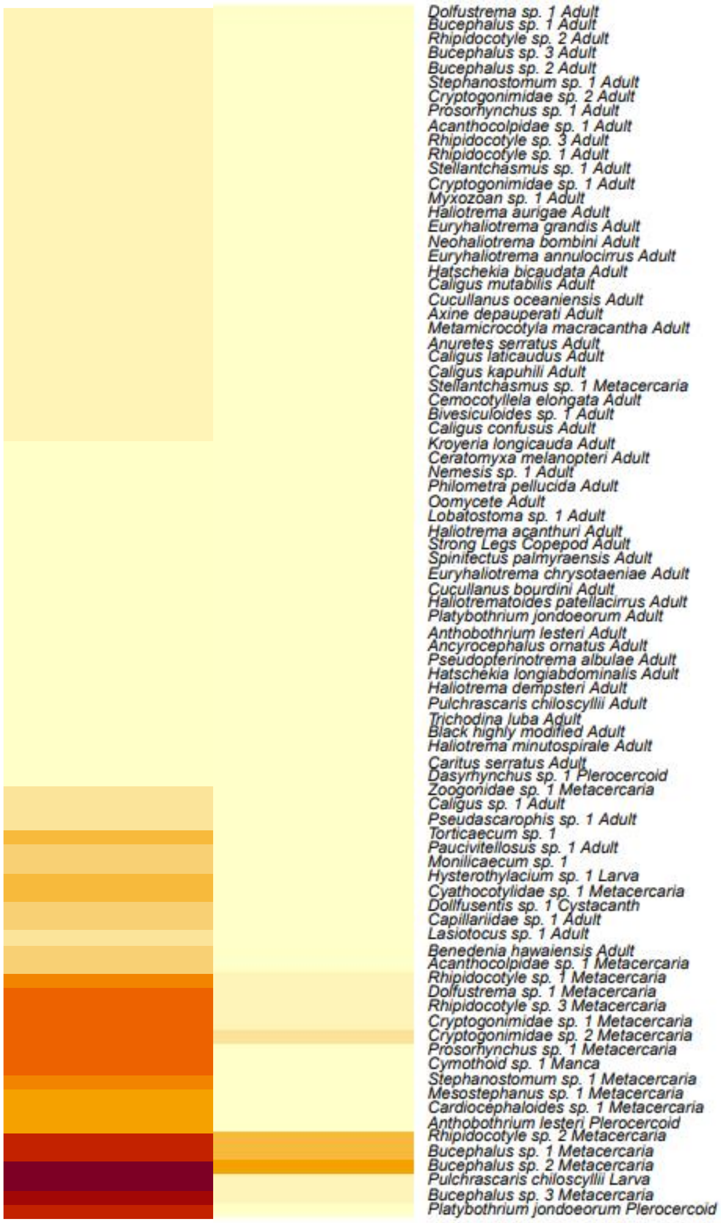

Figure 4. Parasite species-stage occurrence probability sums in the component community (left column) and infracommunity (right column). Dark colors represent species that are expected to occur in more host species in the component community or have high prevalences across many infracommunities.

2.3.3 Host species traits predict parasite community structure among host species

The occupancy model assessed the effect of host traits on the occurrence of parasites in the component community. This analysis indicated that the component community-level mean intercept was -4.51 (95% CI [-5.51, -3.57]) on the logit scale, which translated to an average community occurrence probability of 0.01. This indicated that the average host species had a 0.01 chance of being infected with the average parasite. Our analysis found evidence that several host species traits had population-level effects on component community occurrence probabilities (Table 1). Counter to expectations, host generality and habitat density had a negative effect on component community occurrence with mean effects -0.46 (95% CI [-0.88, -0.05]) and -1.34 (95% CI [-2.04, -0.65]) respectively. We did not find evidence that host species body mass or adjusted trophic level affected component community occurrences with a mean effect of -0.39 (95% CI [-1.02, 0.19]) and -0.02 (95% CI [-0.36, 0.31]). Note that these effects were computed after statistically holding individual host weight constant. Due to the potential for species traits (particularly body mass and trophic level) to correlate with individual mass, these multivariate results do not preclude the possibility that larger species or higher trophic levels have more parasites, just that these effects are not significant after accounting for individual body size.

Table 1. Estimated population-level responses to host traits and the estimated variance around the means

	Mean	SD	2.50%	97.50%	Rhat	ESS
Occurrence Means (logit scale)						
Intercept	-4.51	0.49	-5.56	-3.61	1.00	1261
Diet Generality	-0.46	0.21	-0.89	-0.60	1.00	2340
Body Mass	-0.39	0.31	-1.07	0.15	1.00	4230
Habitat Density	-1.34	0.36	-2.12	-0.71	1.00	1545
Trophic Level	-0.02	0.17	-0.36	0.31	1.00	4575
Occurrence Variances (logit scale)						
Intercept	7.03	2.07	3.91	11.87	1.00	1437
Diet Generality	0.16	0.13	0.03	0.51	1.00	4480
Body Mass	1.76	1.43	0.24	5.37	1.01	1721
Habitat Density	0.35	0.34	0.04	1.25	1.00	2328
Trophic Level	0.23	0.20	0.04	0.75	1.00	3415
Occurrence Random Effect Variances (logit scale)						
Host Family	10.85	2.91	6.30	17.72	1.00	795
Detection Means (logit scale)						
Intercept	-1.34	0.15	-1.63	-1.04	1.00	34575
Individual Weight	0.26	0.06	0.15	0.38	1.00	20269

Detection Variances (logit scale)						
Intercept	1.50	0.33	1.00	2.28	1.00	18471
Individual Weight	0.09	0.04	0.04	0.18	1.00	13092

The community-level variance parameters indicated substantial variability in the component community intercept with a variance of 7.03 (95% CI [3.44, 11.10]). The effect of species body mass on the component community also had a high variance of 1.76 (95% CI [0.04, 4.34]). These high variance parameters suggested that the intercept and effect of host species body mass (after accounting for individual mass) may not have had a universal response across the component community, but instead depended on the individual parasite species-stages. We looked directly at the parasite species-specific effects to further explore this assumption (Figure 5). The parasite species-specific intercept means ranged from -6.94 (95% CI [-10.59, -3.69,]) to 0.59 (95% CI[-1.65, 2.82]) with the greatest distance from the population-level mean being 5.10 suggesting that some parasites were more common than average in the component community. The species-specific effect of host species body mass ranged from -1.77 (95% CI[-4.44, -0.31]) to 1.29 (95% CI[-0.45, 3.14]). There appeared to be less variance in the community level effect of generality (0.16 (95% CI[0.02, 0.42])), habitat density (0.35 (95% CI[0.02, 0.98])), and trophic level (0.23 (95% CI[0.02, 0.61])).

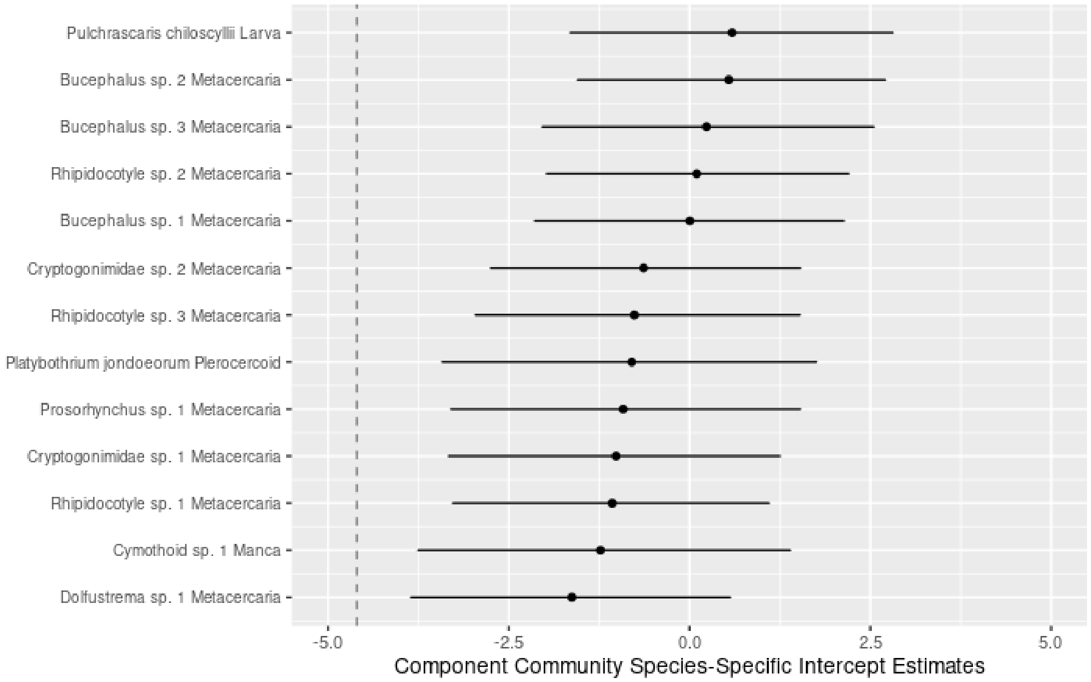

Figure 5. The parasite species-specific mean intercepts with 95% credible intervals excluding the community-level intercept in the component community. The dashed line is the community-level mean intercept.

2.3.4 Individual host traits predict parasite community structure within host individuals

Our study found the infracommunity community-level mean intercept was -1.34 (95% CI [-1.63, -1.03]) on the logit scale, which converts to a mean probability of .23. This indicates that the average parasite had a low probability of occurring in the average infracommunity. We found evidence that individual host weight positively affected infracommunity occurrence probability with a community-level response of 0.26 (95% CI [0.14, 0.37]). The variance parameters in the infracommunity model found high variability for the community-level intercept (1.5 (95% CI[0.95, 2.19])), implying that the intercept estimate depends on the parasite species. This was further supported by directly examining the parasite species-specific mean intercepts, illustrated in Figure 6. The variance parameter for the effect of individual host body weight provided no evidence for parasite species-specific effects with a mean estimate of 0.09 (95% CI [0.03, 0.17]).

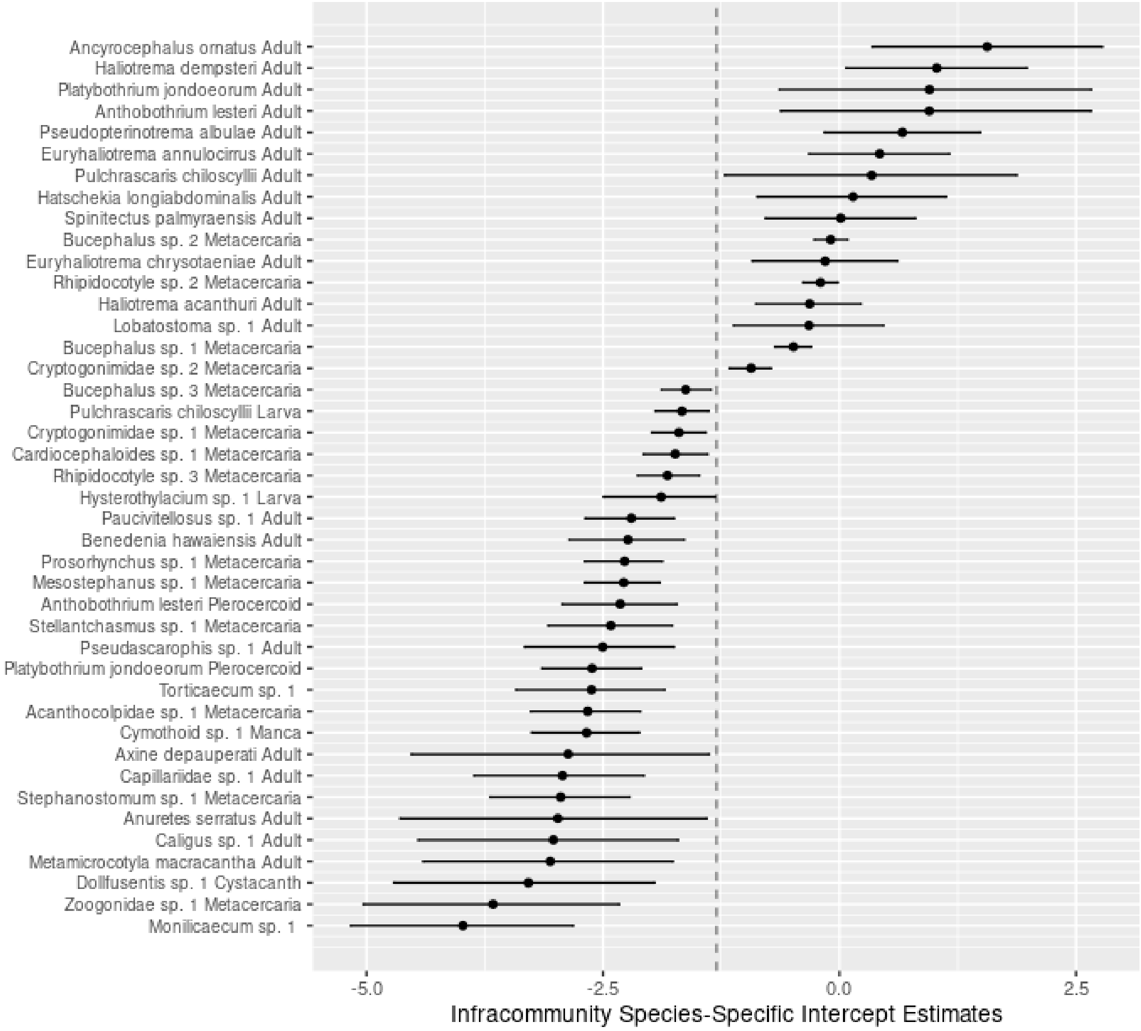

Figure 6. The parasite species-specific mean intercepts with 95% credible intervals in the infracommunity. The dashed line is the community-level mean intercept.

2.3.5 Host Family

In the occupancy model, we included host family as a categorical random effect to account for unmeasured host traits that may influence parasite occurrences. The random effect of host family had an estimated high variance of 10.85 (95% CI [5.92, 16.10]). This suggested that parasite species-specific occurrences varied greatly based on the host family groups, and that some parasite species-stages occurred more often in certain host families than in others. We illustrated the parasite species-specific effects of each host family below (Figure 7).

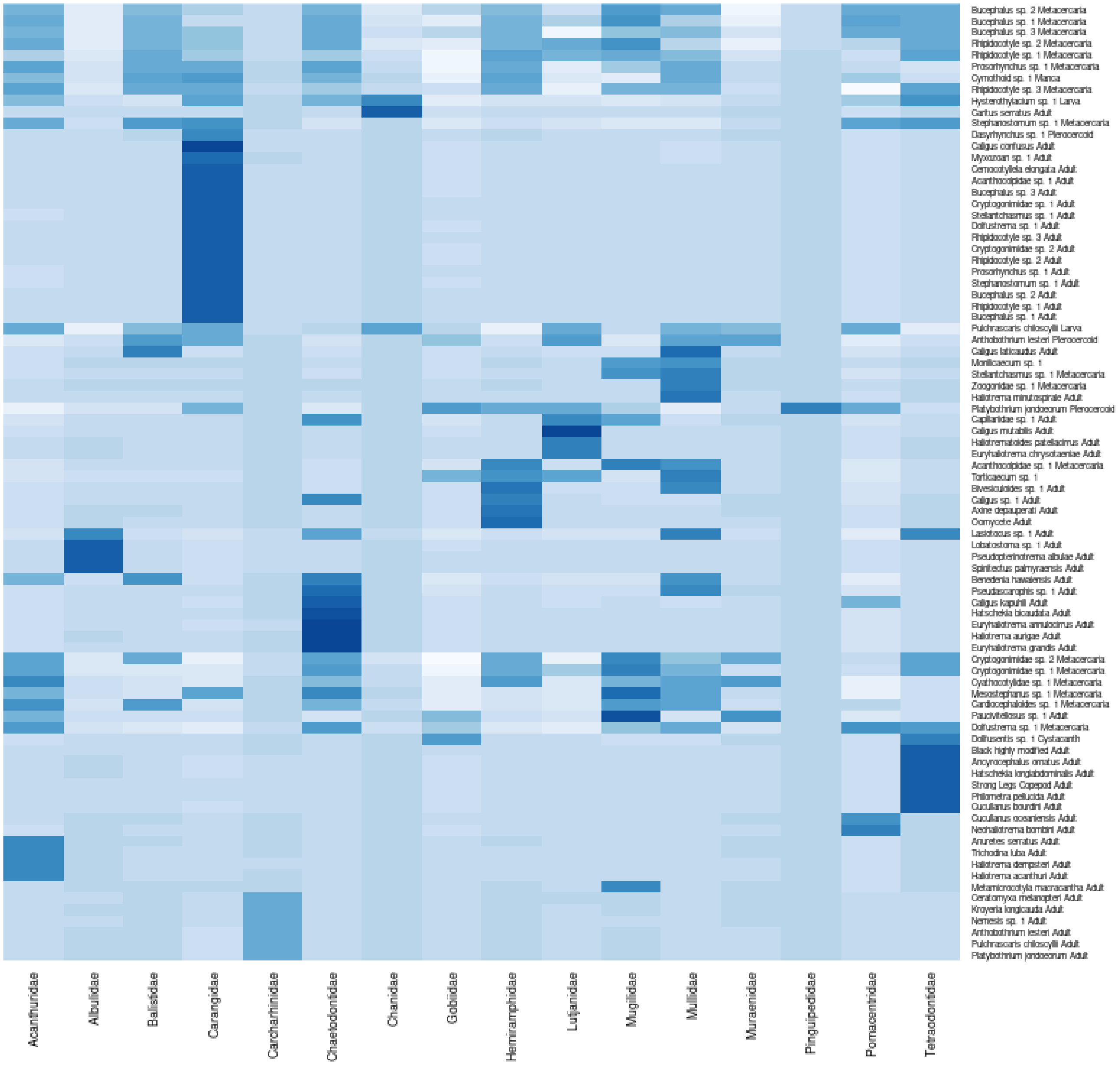

Figure 7. Random effect of family on each parasite species. This heatmap is clustered by column to illustrate the difference between expected parasites in each host family. Dark colors indicate an increase in the occurrence probability of that parasite-host family combination, while light colors suggest a decrease in the occurrence probability.

2.3.6 Post-Hoc Analysis

Finally, the variance parameters in the occupancy model suggested that the component community intercept and the effect of host weight were parasite-taxon specific. We conducted two post-hoc random effect analyses to explore if this variance could be explained by parasite taxonomic group and life stage.

The first post-hoc analysis analyzed if variance in the parasite component community intercept could be explained by parasite taxon and life stage groups (Figure 8). This analysis suggested that cestode adults, copepod adults, monogene adults, nematode adults, and trematode adults had lower estimated intercepts than average, meaning that they were the rarest groups in the component community. In contrast, cestode plerocercoids, isopod adults, nematode larvae, and trematode metacercaria had higher estimated intercepts than expected, suggesting that these groups were the most common across the component community. These post-hoc results supported the estimated variance in the occupancy model and demonstrated that parasite groups vary in their occurrence probabilities in this dataset.

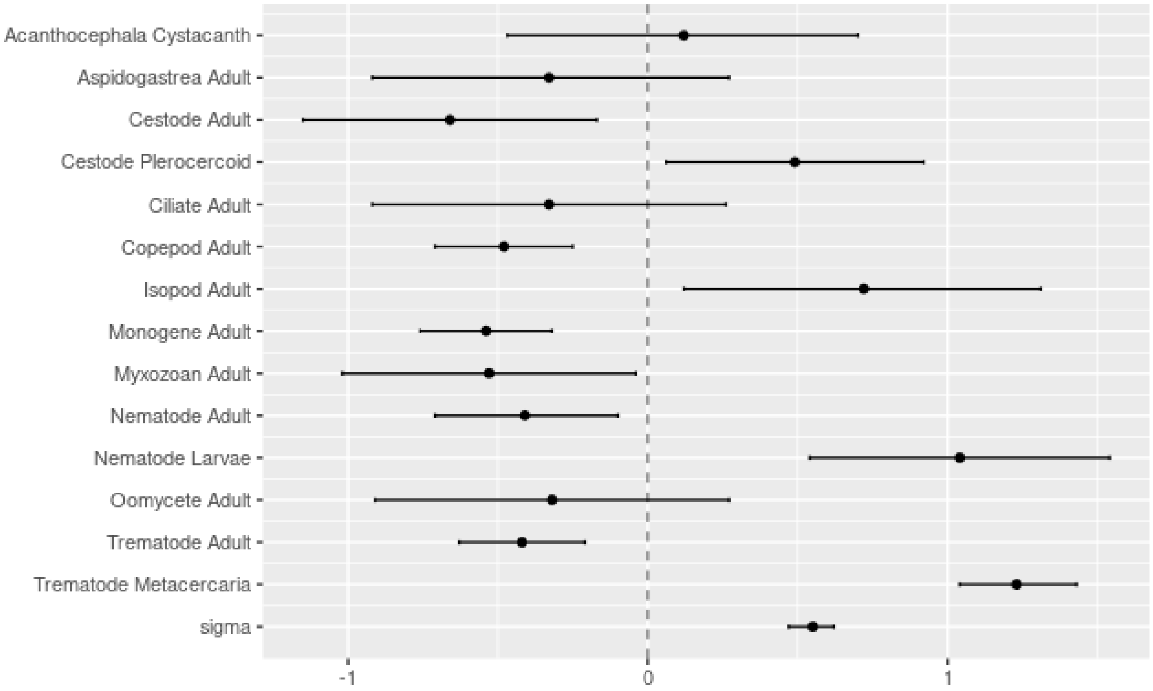

Figure 8. Post-hoc analysis of the effect (and 95% credible interval) of parasite stage on estimated species-specific intercepts. The dotted line represents the estimated population mean intercept.

The second post-hoc analysis explored the high variance found in the effect of host species body mass across the component community (Figure 9). We wanted to know if the species-specific mean effects were correlated to parasite taxon and life stage. The results from this post-hoc analysis suggested that cestode adults and trematode adults had a more positive effect of host size than average, and that trematode metacercaria had a more negative effect of host size than average.

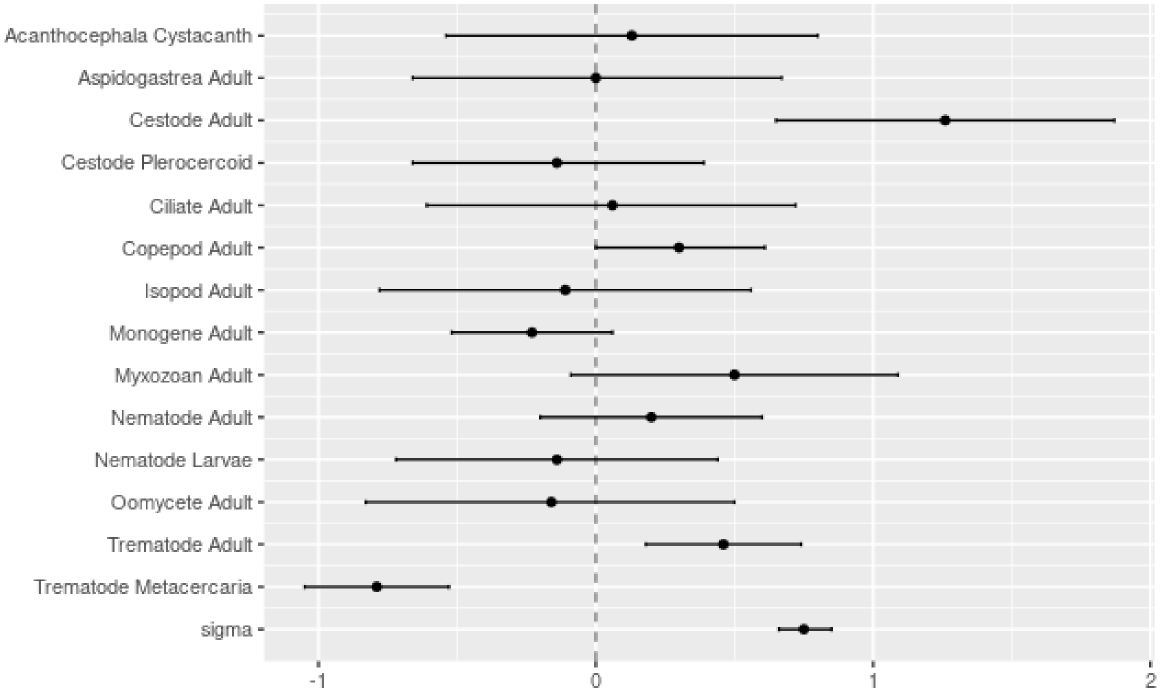

Figure 9. Post-hoc analysis of the effect (and 95% credible interval) of parasite stage on estimated species-specific response to host species body mass. The dotted line represents the estimated population mean effect of host species body mass.

2.4 Discussion

2.4.1 Occurrence Probabilities Between Scales

Our multi-level occupancy model captured the biological scale-dependence of parasite communities. Parasite richness necessarily declines with narrowing host scale because a host's infracommunity can exclusively contain a subset of parasites occurring in that host's component community, but a parasites' occurrence probability can differ between these scales. The probability that an individual fish is infected with a random parasite of the population is equivalent to the probability that a parasite-host combination truly occurs multiplied by the estimated parasite prevalence in the host species. So, unless all parasites in a component community has 100% prevalence, infracommunity variation will occur. Differences between component communities may reflect differences in fish behavior, habitat, or physiology. And at a narrower scale, differences between infracommunities should reflect differences between host individuals.

The majority of parasites were infrequently observed in the average component community, implying many parasite species were specialists. However, larval stages appeared to occur in the greatest number of component communities, and they had relatively low host specificity in comparison to most adult stages. This was further supported by the post-hoc analysis. Cestode plerocercoids, isopod mancae (represented by a single species), nematode larvae, and trematode metacercaria were the most commonly occurring parasite groups in the average component community, which is likely due to the expected low host specificity of each of these parasite stages (Brusca, 1981; Timi & Lanfranchi, 2009). In contrast, cestode adults, copepod adults, monogene adults, nematode adults, and trematode adults occurred less often than average, and similarly these stages should have higher host specificity. However, larval stages are expected to accumulate in hosts over time, while adult parasites may be shorter-lived (Locke et al., 2013). This could lead to higher detection rates of larval parasites in host species, leading to their observed generality in comparison to observed adults, which could also explain why many larval parasites had higher occurrences in the population than expected.

At the infracommunity scale, the majority of parasites that were host specialists had low prevalence within their host species, suggesting they were rare observations. Across infracommunities, there were only three parasite species with high prevalence across many host species. Poulin (1998) suggested that parasite communities are often composed of many rare and a few common species, and this appears to be the case here as well. However, adult trematodes and adult cestodes were specialist parasites that maintained high prevalence in their specific host infracommunities. All adult trematodes whose metacercaria were also present in this fish community (13 parasite species) exclusively occurred in host species of the family Carangidae. Further, all of the adult cestodes exclusively occurred in the elasmobranch *Carcharhinus melanopterus* (the only shark included in this study). Because the larval stages of these adult parasite species are host generalists transmitted through consumption, this suggested the efficiency and efficacy of trophic transmission as an infection strategy. Further, the high infracommunity occurrence probabilities suggested a trade-off between prevalence and specificity in this system. This is supported by the well-documented co-evolution of adult cestodes and elasmobranchs (Palm & Caira, 2008; Caira *et al.,* 2014; Euzet, 1956) and (to a lesser extent) adult trematodes and fish (Salgado-Maldonado et al., 2016; Cameron, 1964). As a result, these parasite stages had high occurrence probabilities in specific infracommunities.

2.4.2 Effect of Host Traits Within Scales

To understand how scale-dependent host traits interacted with parasite occurrences, we analyzed the effect of fish species traits and fish individual traits on occurrence probabilities. Using this occupancy model, we demonstrated that host species traits like diet generality, species body mass, and habitat density, structured parasite distribution in the component community, while host individual traits, like body weight, structured distributions in the infracommunity.

We predicted that fish species with a wide diet generality would be exposed to a wide range of trophically transmitted parasites, leading to higher parasite occurrences in the component community. Other studies have supported these expectations (Rasmussen & Randhawa, 2018) and Locke *et al.* (2014) found that host diet generality had a positive effect on adult and larval parasite richness. But in contrast to these predictions, our mean response and variance estimates suggested that component community occurrences increased with decreasing host generality and that this was a community-level response. In this system, a greater diet breadth may allow hosts to actively switch to parasite-free diet items, resulting in reduced parasite exposure (Becker *et al.,* 2018). Or, host species with high diet generality may simply not encounter specific trophically-transmitted parasite species often enough for competent infective stages to evolve. An alternative explanation is that some larval parasites (the most abundant individuals in our study) are not encountered through feeding (e.g. trematode metacercaria). Larval stages are difficult to identify and often omitted from parasite component community surveys, which may explain how our results differ from other studies.

Because trophic level determines the range of prey a fish consumes, host trophic level should be connected to and constrain the parasites that rely on trophic interactions. Some studies have found that parasite diversity increases with host trophic level (Lafferty *et al.,* 2006; Chen *et al.,* 2008), while others argue that the highest parasite richness will be in mid-range trophic level hosts (Anderson & Sukhdeo, 2011). However, our study found that, after controlling for fish size, host trophic level had no detectable effect on parasite occurrence probabilities in the component community. One reason may be that marine fish species tend to have generally broad diets with high prey switching and diet overlaps (Marcogliese, 2002). As a result, there may not have been enough trophic level variation in this fish community to detect an effect. Further, the majority of larval parasites individuals are not acquired through host diet, but larval parasites had the highest occurrence probabilities across component communities in this system. Their low specificity in intermediate hosts could also reduce the effect of trophic level in fish hosts. In addition, trophic level and body size often covary among fish species (Jennings et al., 2001), with fish body size demonstrated as the better predictor of larval helminth richness (Poulin and Leung, 2011). The raw data for these host traits were slightly correlated, but their parameter posterior distributions were not. Because this study included host generality and body size as predictors, it is possible that host trophic level did not provide new information that improved model fit.

Host body size is often recognized as an important determinant of parasite occurrences (Guegan et al., 1992; Guégan & Hugueny, 1994) because large hosts will be more likely to encounter, eat bigger items (especially in gape-limited predators like fish), and provide more space to house parasites. Some studies have argued that host species body size should not be a determinant of parasite occurrences over evolutionary time, and this has been supported by studies of mammal and bird species (Morand & Poulin, 1998; Poulin, 1995). Our analysis did not find evidence that host species body size affected component community occurrences. In the collected dataset, there was a tendency for the largest host species to have smaller sample sizes, so the parasite species in these component communities were likely underrepresented with many missed parasites. For this reason, we excluded the largest species from this analysis to minimize this effect. Regardless, the large variance parameter indicated that the effect of host species body mass varied across parasite species, with high dispersal around the population mean. The post-hoc random effect analysis suggested that metacercaria trematodes had a more negative effect from host species body mass than the average parasite species. And adult trematodes and adult cestodes had a more positive effect than average, meaning that their occurrences were either not affected or positively affected by host body mass. Although the post-hoc analysis outputs must be viewed with caution, these patterns could be explained from an ecological perspective. For trematode metacercariae to continue their life cycle, they must infect a prey host. Although metacercaria increase in intensity as that prey ages (presumably leading to a positive association with individual size), prey species tend to be smaller than predator species, leading to contrasting associations with size across levels. Similarly, adult cestodes and adult trematodes develop and reproduce in definitive hosts. These definitive hosts should be larger bodied because they must eat small prey items and avoid being eaten themselves. Adult trematodes and cestodes may accumulate in a wider range of medium to large body sized fish. And this may explain their neutral or positive response to host species body size than the average parasite.

Based on previous studies, we predicted that parasite occurrences would increase with host density because dense hosts should have higher encounter rates and can more easily sustain parasite populations (Morand & Poulin, 1998; Anderson & May, 1978; Holmes & Price, 1986). But in this system, fish host species with the highest densities had the lowest estimated parasite occurrences. Because some parasites can reduce host densities (Hudson *et al.,* 1998; Scott, 1987; Anderson & May, 1978), host species with fewer parasites may experience a demographic advantage that allows their populations to reach high densities. However, of the 10 most dense fish host species in this system, 6 were within the Gobiidae family which were infected with consistently few parasites. 2 of these fish species were *Parapercis* sp. and *Albula glossodonta*, which were the only representatives of their family in the host community and also had few parasites. The remaining 2 fish species were within the Mugilidae family which had relatively rich parasite communities. And of the 10 least dense host species, 4 were within the host family Carangidae which had relatively high parasite occurrences. Since there appears to be a connection between host family and species density in this system, this pattern may be a reflection of taxonomy-based parasite competency than of host population dynamics. Another potential driver of negative association between host density and parasite prevalence relates to an encounter dilution effect for parasites with complex life cycles, which can occur when a limited number of infective stages encounters abundant host individuals, resulting in each individual having lower infection risk (Buck *et al*., 2017).

Our analysis found evidence that parasite infracommunity occurrence probabilities increased with fish weight. In other words, large individuals had a richer parasite community than smaller individuals of the same species. Additionally, this effect's low variance meant that this was a broad response from all parasites in the community. Large host individuals within the same species are likely older, resulting in higher parasite accumulation due to time and exposure. And larger hosts also have more area for parasite species to accumulate (Guégan & Hugueny, 1994). Thus, host body size contributed to the nested structure of parasite communities, with parasite species present in the component community further structured by host individual size to determine occurrence probabilities in the infracommunity.

Parasites and fish hosts have well-documented phylogenetic patterns (Poulin & Morand, 1999; Poulin, 2003; Timi *et al.,* 2010; Poulin, 2010; Locke *et al.,* 2013; Chai *et al.,* 2022). Host-switching and co-evolution between parasites and hosts should lead to more closely related hosts sharing more parasites, and more distantly related hosts sharing fewer (Engelstädter & Fortuna, 2019). In the occupancy model, we found that the effect of host family on component community occurrences had a wide variance, indicating that parasite occurrence probabilities varied greatly between host families. As an example, host species in the Carangidae host family had entirely different parasite occurrence probabilities than hosts in the Albulidae family. Some parasites occurred exclusively in a single family, like *Trichodina luba* adults, while others infected hosts more generally like *Bucephalus sp*. 3 metacercaria. As a result, host family was a good predictor of parasite species-stage occurrence probabilities in this system.

2.4.3 Assumptions

In this study, the occupancy model estimated the occurrence probabilities of parasite species across two biological scales. However, this analysis makes several assumptions about the structure of the dataset.

The first assumption from this model framework was based on how the species-level effects were parameterized. Species-specific effects were estimated as random effects that follow a normal distribution with means equivalent to the community-level occurrence parameters. Although there were benefits to this method, this did not necessarily reflect the true species-specific effects. This structure assumed that all parasite species responded similarly to host traits unless further evidence suggested otherwise. Rare parasite species provided the least information regarding factors that influenced their occurrences, so were most affected by this shrinkage and likely followed the community response. The species-level effects of parasites that matched the community response, particularly of rare species, should be interpreted carefully. Species that demonstrated an effect outside of the community response are more reliable because their occurrences provide enough evidence to withstand community shrinkage. Shared patterns across these parasite species are what we attempted to reveal in the post-hoc analysis.

This occupancy model also assumed that only host traits affected parasite occupancy, but in reality parasite traits likely also played a role. As suggested by the variance parameters in the occupancy model and post-hoc analysis outputs, parasite species and groups varied in their response to host traits. Differences in parasite life-histories and other parasite-specific traits may explain these patterns. As an example, trophically acquired parasites are expected to occur more often in hosts with high trophic levels, while directly acquired parasites will not be affected by this covariate. In this case, host trophic level effects could be obscured by failing to differentiate parasites by transmission strategy. There were many parasite traits expected to interact with host traits, and not considering them could muddy results and interpretations. Due to the limitations of the occupancy model used, parasite traits were not included in this study. However, accounting for parasite traits can be an aim for future analyses.

Finally, this model assumed that parasite species had no co-occurrence patterns. Species in communities can have patterns of coexistence or displacement that can be caused directly by biotic interactions or indirectly by correlated responses to environmental factors (Mod *et al.,* 2020). Co-occurrence estimates, like from joint species distribution models, can provide insight into how an entire community will respond with only information on a few species and are particularly advantageous for rare species with little data. Previous research has demonstrated the relevance of co-occurrence in some parasite communities. For example, trematodes in marine snail hosts experienced higher magnitudes of interspecific competition leading to negative interactions (Kuris & Lafferty, 1994). And positive co-occurrences were detected between arthropod ectoparasites on rodent hosts (Krasnov *et al.,* 2010). But other studies have found a lack of evidence to support nonrandom parasite species co-occurrences in specific systems, like metazoan ectoparasites of marine fish (Gotelli & Rohde, 2002). Either way, this occupancy model does not parameterize co-occurrence patterns, so we were unable to do so in this study. In the future, estimating parasite species co-occurrence patterns could lead to further insights.

2.5 Conclusion

In this study, parasite distributions were affected by host covariates at each biological scale, with host species traits structuring component community, and host individual traits structuring infracommunity distribution. Parasite occurrence probabilities decreased with host species density, host species diet generality, and host individual weight. We also determined that host families varied in their parasite assemblages with closely related host species sharing more parasites. Some parasite species had species-specific effects, and there was some support that these differences may be explained by parasite life history. For future work, the role that parasite traits and species co-occurrences play in this system is still unknown, so these factors could be ideal areas for further exploration.

Chapter 3: Estimating and predicting false negatives in simulated species communities

3.1 Introduction

Presence-absence data are collected in biological surveys to understand species occurrences. Results from these studies are often used to predict species ranges, assess the effect of environmental factors on communities, and guide management and conservation decisions. However, assuming that these datasets are collected with perfect detection can lead to biased results and conclusions like underestimating alpha diversity and overestimating beta diversity (Lin, 2018; Ferguson *et al*., 2015; Miller *et al*., 2015; Williams *et al*., 2002; Bayley & Peterson, 2001; Nichols & Karanth, 2002; Ostermiller & Hawkins, 2004). One possible source of error is false negatives, which occur when researchers do not detect a species at a site where it is truly present. False negatives can arise while sampling for several reasons. Factors like local conditions at the time of sampling, methodology, or observer experience can bias measurement error (Kerans *et al*., 1992; Bonneau & Labar, 1997; Dunham *et al.,* 2001; Doser *et al.,* 2022). And species themselves vary in rarity, cryptic-ness, or habitat preference resulting in a range of species-specific detection probabilities (Mao & Colwell, 2005; MacKenzie *et al*., 2005; Williams *et al.,* 2002). Species that are difficult to detect or rare are expected to have higher false-negative probabilities, particularly when sampling effort is low (Mao & Colwell, 2005; Preston, 1948; Colwell & Coddington, 1994). Even low false-error rates can lead to biased estimates of habitat associations and distributions (Tyre *et al*., 2003; Ferguson *et al*., 2015) but few species distribution studies account for imperfect detection (Kellner & Swihart, 2014). The lack of effort towards estimating false negatives is understandable given that modelling imperfect detection is statistically difficult. Thus, new techniques for estimating false negatives would help biologists get more accurate biodiversity estimates.

Some patterns in observed datasets correlate with false negatives. For example, a species that was only observed once could have been easily missed, suggesting that this was a lucky observation. This also suggests that there were probably less lucky species, similarly rare or cryptic, that were not seen at all. Based on this logic, missing species estimators, like the Chao or Jackknife estimator, predict the true number of species (richness) present in a community based on the ratio of observed singletons to doubletons (Chazdon *et al*., 1998; Chao, 1987; Chao, 2016). Another pattern is that false negatives will lead to more 0s than expected. As a result, Tyre *et al*. (2003) estimated the rate of false-negative errors by extending a logistic regression to a zero-inflated binomial model. A different approach is to use MCMC algorithms to match observations to likely combinations of detection and occurrence probabilities. Such occupancy models account for measurement error by modelling the state process and observation process separately (MacKenzie *et al*., 2002; MacKenzie *et al.,* 2003; Dorazio & Royle, 2005; Rota *et al*., 2011; Guillera-Arroita *et al*., 2017; Bled *et al*., 2011). In these two-level systems, a species has a probability of occurring at each site and a conditional probability of detecting that species through sampling if it does occur. Hierarchical occupancy models have demonstrated some success at estimating true parameter values and illustrating biases (Ferguson *et al*., 2015). Although several papers have estimated true richness, occurrence probability, detection probability, and false negative rates (Tyre *et al*., 2003; Ferguson *et al*., 2015; Moilanen, 2002), few have used false-negative probabilities to estimate community composition among sites. Predicting the locations and identities of missed species in an observed community will provide a better understanding of species occurrence dynamics, leading to more accurate community inferences and informed management decisions.

In this study, we validated a method for estimating false-negative occurrences in a metacommunity using a hierarchical occupancy model (Doser *et al.*, 2022). For 1000 simulated datasets, we demonstrated the occupancy model's propensity to correctly and incorrectly predict false negatives based on observed data. We then tabulated these false-negative estimates and observed communities to make predictions about true community composition. We used several metrics to evaluate whether estimated communities were better representations of true communities than observed communities. First, we assessed community accuracy to determine which community (observed or estimated) more closely matched the true community. Then, we compared mutual information to determine which community (observed or estimated) shared more information with the true community. We also compared true, observed, and estimated richness, ignoring species identities and locations. And we then compared our best richness estimate to the commonly used Chao richness estimates. We ended this study by illustrating the method with a case study: false-negative estimates of Hymenoptera species on Palmyra Atoll islets.

3.2 Methods

3.2.1 Data Simulation

To prepare for the analysis, we simulated 1000 datasets representing the true presence/absence and observed detection/non-detection of 10 species at 60 sites. To create simulated data, we simulated a multi-level process using the following equations:

$$\text{Logit(psi)} = \text{B0} + \text{B1} * \text{sitetrait.1} + \text{B2} * \text{sitetrait.2} + \text{B3} * \text{sitetrait.3} \qquad [1]$$

$$\text{True z} \sim \text{Bernoulli (psi)} \tag{2}$$

$$\text{Logit(p)} = \text{a0} + \text{a1} * \text{replicatetrait.1} \tag{3}$$

$$\text{Observed y} \sim \text{Bernoulli (p * true z)} \tag{4}$$

First, to determine the true presence/absence (true z) of a species at a site, we calculated the species' occurrence probability (psi). A species' occurrence probability (psi) at a site was associated with 3 site-level traits and an intercept. We simulated site-level trait values using a uniform distribution with a range from 0 to 100. Each site trait was then rescaled to have a mean of 0 and standard deviation of 1. We then generated a species-specific effect (B) in response to each site-level trait by drawing from a uniform distribution limited between -2 to 2 on the logit scale. The occurrence intercept was generated from the same uniform distribution.

From these simulated site-level traits and species-specific effects, we computed species occurrence probabilities (psi) using a logit link function following equation [1]. We used each psi as the probability of success in a Bernoulli distribution, as demonstrated in equation [2]. A single draw determined the true presence (1) or true absence (0) of each species at each site. From this, we derived the true occurrence matrix (true z) with 2 dimensions matching the number of species times the number of sites.

In addition to variation in occurrence probability at a site, species also vary in detection probability leading to imperfect detection in the observation matrix. In an attempt to parameterize detection probabilities, biological surveys should have replicate observations at each site (Gauch, 1982). To simulate this variation, we modelled replicate-to-replicate variability by re-sampling each site for 3 replicates. To generate the detection/non-detection y data, species detections were associated with a single replicate-level trait. Replicate-level trait values were drawn from a uniform distribution ranging from 0 to 100 and then rescaled using the scale() function in base R. The linear effect of each replicate-level trait on each species at each site was randomly assigned from a uniform distribution with a minimum of -2 and maximum of 2; the detection intercept was generated from a similar uniform distribution. Then, using the simulated traits, intercepts, and effects, we computed detection probabilities with a logit link function following equation [3]. This resulted in p, the probability of detecting the species at a site during each replicate. We multiplied p by the true presence/absence (true z) of that species-site combination so that only truly present species could be detected, and each p * z product was used as the probability of success in a Bernoulli distribution. A single draw resulted in either a 1 (detection) or a 0 (non-detection) representing a detection of the species at a site during a replicate. From this, we derived the detection/non-detection matrix (y) with 3 dimensions equivalent to the number of species times the number of sites times the number of replicates.

3.2.2 Statistical Analysis

To estimate false-negative probabilities from the observed detection/non-detection datasets, we used a Bayesian occupancy analysis for multiple species using the msPGOcc() function in the spOccupancy package in R (Doser *et al*., 2022). The model framework (Dorazio and Royle, 2005) had a hierarchical structure that modeled the state process and observation process separately to account for imperfect detection. The model took the form:

<u>Occurrence Model</u>

$$z_{ij} \sim \text{Bernoulli}\ (\psi_{ij})$$

$$\text{logit}\ (\psi_i) \sim \beta_{0i} + \beta_{1i} * x_1 + \beta_{2i} * x_2 + \beta_{3i} * x_3$$

$$\beta_i \sim \text{Normal}\ (\mu_{\beta i}, \tau_{\beta i}^2)$$

$$\mu_{\beta i} \sim \text{Normal } (0, 2.72)$$

$$\tau_{\beta i}^2 \sim \text{Inverse-Gamma } (0.1, 0.1)$$

Detection Model

$$y_{ijk} \sim \text{Bernoulli } (p_{ijk} z_{ij})$$

$$\text{logit } (p_{ijk}) \sim \alpha_{0i} + \alpha_{1i} * v_1$$

$$\alpha_i \sim \text{Normal } (\mu_{\alpha i}, \tau_{\alpha i}^2)$$

$$\mu_{\alpha i} \sim \text{Normal } (0, 2.72)$$

$$\tau_{\alpha i}^2 \sim \text{Inverse-Gamma } (0.1, 0.1)$$

where the response variable z_{ij} represented the true presence (1) or absence (0) of species i at site j. ψ_{ij} was the probability that species i occurrenced at site j. ψ_{ij} was modelled with a logit link function where β_{ix} were the regression coefficients (including an intercept) that described the species-specific effect of site covariates x. β_{0i} were assumed to have a normal distribution with mean $\mu_{\beta i}$ and variance $\tau_{\beta i}^2$.

In the detection portion of the model, the response variable, y_{ijk}, indicated the sampling observations. Observed detection (1) or non-detection (0) of species during each replicate was modelled with a Bernoulli distribution from detection probability, p_{ijk}, that was conditional on the true state process, z_{ij}. The probability, p_{ijk}, that species i was detected at site j during replicate k was modelled with a logit link function where α_i were the regression coefficients (including an intercept) that described the effect of replicate covariates v. α_i were modelled by a normal distribution with mean $\mu_{\alpha i}$ and variance $\tau_{\alpha i}^2$.

For all fixed effects in the Bayesian model, β_{0i} and α_i, we specified normal prior distributions with mean 0 and variance 2.72. We used an inverse-Gamma prior distribution for the variance parameters with shape 0.1 and scale 0.1.

Using the above occupancy model, we fit each iteration of the simulated datasets. The simulated observed species detection/non-detection matrix (y) was the response variable in the detection model, site traits were the occurrence covariates, and replicate traits were the detection covariates. Each model iteration generated 8000 posterior samples for each parameter with 3 MCMC chains. We used a burn in period of 2000 samples with a thinning rate of 1.

3.2.3 Model Estimates and Defining False Negatives

After the model fit each observed dataset, we extracted the posterior distributions of the z_{ij} samples. A single posterior draw of z_{ij} was the estimated presence (1) or absence (0) of a species i at site j. The mean of z_{ij}'s posterior distribution was a value between 1 and 0, representing the probability of a false negative of species i at site j. Mean z_{ij} differed from ψ_{ij} in that ψ_{ij} was the occurrence probability of a species based on site covariates. z_{ij} predicted true occurrence based both on the site covariates and detection probabilities. This means that if the predicted occurrence probability is low, but many replicates had observed detections, then mean z_{ij} and ψ_{ij} will differ. In contrast, if few replicates observed detections, then mean z_{ij} and ψ_{ij} will be similar if not equivalent. Based on the posterior means of z_{ij}, we defined an estimated false negative as any undetected species-site combination with a false-negative probability greater than 50 percent (mean z_{ij} >.5). We chose a probability threshold of 0.5 simply because these species-site combinations were expected to be truly present more often than not.

By comparing the true presence/absence data to the observed detection/non-detection data and the model estimated mean z_{ij}, we categorized species at each site by 6 outcomes (Table 1). The observed data may indicate that a species was detected (mean y >1) at a particular site. If this observed species was truly present (true z = 1), then this illustrated a **correct presence**. However, if this observed species was truly absent (true z = 0), then this would be an **incorrect presence**, also known as a false positive. False positives were not the focus here, and our simulation methodology did not allow for false positive observations. If the data indicated that a species was unobserved (y = 0), there were four outcomes based on the model's false negative estimates. Studies that do not control for false negatives assume that all undetected species are **correct absences**, meaning the species was not observed (y = 0) and truly absent (true z = 0) with a low false-negative probability (mean z_{ij} < .5). However, an undetected species could simply be truly present (z = 1) but unobserved (y = 0). In this case, the species was considered a **correctly estimated false negative** if the model estimated a high false-negative probability (mean z_{ij} > 0.5). But if the model estimated a low false-negative probability (mean z_{ij} < 0.5), then this undetected, yet present species was an **incorrect absence** that remained a false negative. Finally, an unobserved species may be an **incorrectly estimated false negative** if it was truly absent (z = 0) and not observed (y = 0) with a high estimated false-negative probability (mean z_{ij} > .5). These incorrectly estimated false negatives were the least desirable outcome as this signaled that the model was adding error to the community. As long as the number of correctly estimated false negatives was greater than the number of incorrectly estimated false negatives in a simulated dataset, then the model was decreasing error in the community overall.

Table 1. The six possible categorizations of a species at a site after the occupancy model fits the observed data.

		True Presence	True Absence
Observed		Correct Presence $z = 1$ and $\bar{y} > 0$ Mean $\hat{z} = 1$	Incorrect Presence (aka false positive) $z = 0$ and $\bar{y} > 0$ Mean $\hat{z} = 1$
Unobserved	Estimated False Negative	Correctly Estimated False Negative $z = 1$ and $\bar{y} = 0$ Mean $\hat{z} > 0.5$	Incorrectly Estimated False Negative $z = 0$ and $\bar{y} = 0$ Mean $\hat{z} > 0.5$
	Not Estimated False Negative	Incorrect Absence (aka false negative) $z = 1$ and $\bar{y} = 0$ Mean $\hat{z} < 0.5$	Correct Absence $z = 0$ and $\bar{y} = 0$ Mean $\hat{z} < 0.5$

Each simulated community and model iteration resulted in four species by site community matrices: a true z matrix, an observation matrix, a false-negative probability matrix, and an estimated z matrix (Figure 1). The true z matrix defined the correct state of each species-site combination and contained only true presences and true absences. The true z matrix is usually hidden from researchers and is what the occupancy model is attempting to reveal. The observation matrix was the result of replicate sampling from the true z-matrix. The observation matrix was a truncated version of the detection/non-detection data (y), but it contained a 1 if a site-species combination was detected at least once in any replicate and a 0 if it was never detected in any replicate. Each site-species combination of the observation matrix was classified as either a true presence, a true absence, or an incorrect absence (aka true false negative). The false-negative probability matrix combined the observation matrix with the false-negative probabilities from the occupancy model. Each cell was either a true presence (x = 1) or an estimated false-negative probability (0 < x < 1). And lastly, the estimated z matrix was similar to the false-negative probability matrix but with determined outcomes based on the false-negative probabilities (x = 1 when mean $z_{ij} > 0.5$ or x = 0 when mean $z_{ij} < 0.5$). The estimated z matrix was composed of correct presences, correct absences, correctly estimated false negatives, incorrectly estimated false negatives, and incorrect absences. We compared the true z-matrix to the corresponding observation matrix and estimated z matrix to count and identify the number and locations of the above species-site classifications. We also compared the true z matrix to the false-negative probability matrix to measure the amount of information shared.

We compared the observation matrix, false-negative probability matrix, and estimated z matrix to the corresponding true z matrix to evaluate how well each represented the hidden community.

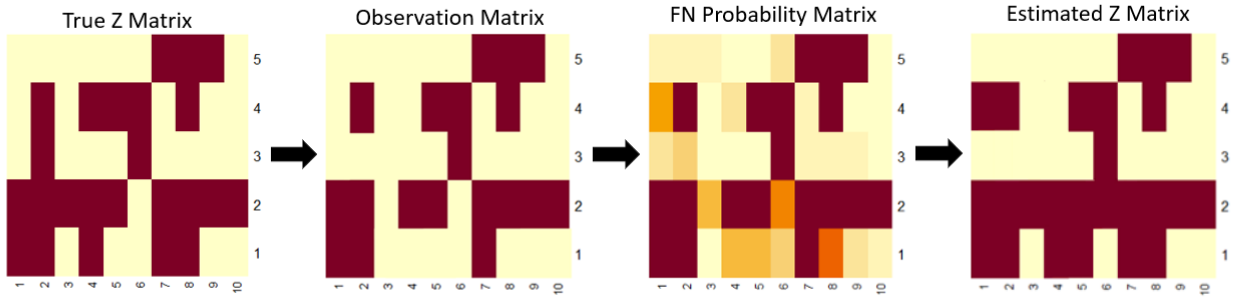

Figure 1. Examples of the true z matrix, observation matrix, false-negative probability matrix, and the estimated z matrix. The true z matrix is usually unknown, and researchers collect imperfect data about the true z matrix during biological surveys resulting in the observation matrix. Using an occupancy model, we can estimate parameter values to get the inferred matrix and estimated z matrix.

3.2.4 False Negative Estimates

To examine how well the model predicted the identities and locations of false negatives, we compared the number of true false negatives in the observation matrix to the number of estimated false negatives in the estimated z matrix. In the estimated z matrix, we quantified the number of correctly estimated false negatives, the number of incorrectly estimated false negatives, and the total number of estimated false negatives (sum of correctly and incorrectly estimated false negatives). To assess how the model adjusted error present in the observation matrix, we calculated the difference between the number of correctly and incorrectly estimated false negatives in each estimated z matrix. We summarized these values with several histograms using the hist() function in base R and the ggplot() function in the ggplot2 package of R.

3.2.5 Overall Community Estimates

To evaluate how well the occupancy model predicted the true z matrix, we examined community accuracy before and after controlling for false negative estimates. To calculate accuracy in the observation matrix, we divided the sum of correct presences and correct absences by the total number of site-species combinations. To calculate accuracy in the estimated z matrix, we divided the sum of correct presences, correct absences, and correctly estimated false negatives by the total number of site-species combinations. Note that the number of correct absences decreases between the observation matrix and estimated z matrix because of incorrectly estimated false negatives. Then we compared community accuracy of the observation matrix to community accuracy in the estimated z matrix.

We also measured the amount of mutual information that the observed matrix and false-negative probability matrix shared with the true z matrix. In general, mutual information is the amount of information that one random variable reveals about another random variable. It is measured in natural units of information (nat) which is a unit of information entropy. In this study, measuring mutual information is beneficial because it can be used to directly compare true z to the estimated false-negative probabilities in the false-negative probability matrix. Therefore, we did not need to define a false negative by a probability threshold (> 0.5), but instead used the estimated probabilities themselves without transformation. By comparing mutual information, we quantified the amount of information that the occupancy model revealed about true z.

Finally, we were interested in how well the occupancy model predicted overall species richness. For every model iteration, we calculated species richness in each of the 4 community matrices. True species richness was the count of true site-species combinations present in each true z matrix, while observed matrix richness was the count of site-species combinations observed in the detection/non-detection data. To estimate richness from the false-negative probability matrix, we summed the estimated probability of every species-site combination in the community. As an example, if a matrix was composed of 2 possible species at 2 sites with respective false-negative probabilities 0.45, 0.76, 1, and 0.14, then the estimated richness would be 2.35 site-species combinations. For the estimated z matrix, richness was the count of observed and estimated false negative site-species combinations. Note that this does not distinguish between correctly and incorrectly estimated false negatives; all were included in the richness estimates. To determine the best predictor of true species richness, we compared richness estimates from the observed matrix, probability matrix, and the estimated z matrix. And to understand how well the best predictor performed in comparison to other benchmarks in the literature, we compared our model estimates to the improved iChao2 richness estimator (Chiu *et al*., 2014). To calculate the Chao richness estimates, we used the ChaoSpecies() function in the SpadeR package in R.

All scatterplots were created using the ggplot() function in the ggplot2 package in R. All histograms were created using the hist() function in base R. All R code used in this study are located at: https://github.com/mmorse8976/Dissertation-Code.git

3.2.6 Arthropod Dataset Case Study

Finally, we demonstrated estimating false-negative probabilities in a species community with a case study. We obtained a dataset from a study of arthropod communities on Palmyra Atoll National Wildlife Refuge (McLaughlin *et al.*, 2023). Researchers fogged 89 trees on 16 islets of Palmyra Atoll and documented the abundance of 240 arthropod species. Field collection and laboratory sorting methods were outlined in McLaughlin *et al.* (2023). This dataset is nested such that fogged trees were grouped within an island identification. Trees that were on the same island were considered replicates of the same site. Across islands, there was an average of 5.56 trees sampled per island. 2 of 16 islands only had one tree fogged, so in these cases there was no distinction between island and tree level sampling. Because sampling effort was low on some islands, but high on others (maximum of 23), we felt this was a prime dataset to consider false-negative probabilities. Multilevel models are particularly well suited for uneven and low sampling. From the 240 total arthropod species, we narrowed our analysis to members of the Order Hymenoptera, representing 44 species total.

For this case study, we estimated false-negative probabilities and predicted the true community of Hymenoptera species on 15 islets of Palmyra Atoll. In the occupancy model, we included six island-level predictors of Hymenoptera occurrence: distance to the mainland (Cooper Island), island area, native canopy proportion, average cation exchange capacity, average soil pH, and average nitrogen release per acre (Figure 2). In the detection model, we included one predictor of Hymenoptera detection: area of the fogged canopy. We also incorporated a random effect of canopy type. All numeric predictors were standardized to have a mean of 0 and standard deviation of 1 before the analysis.

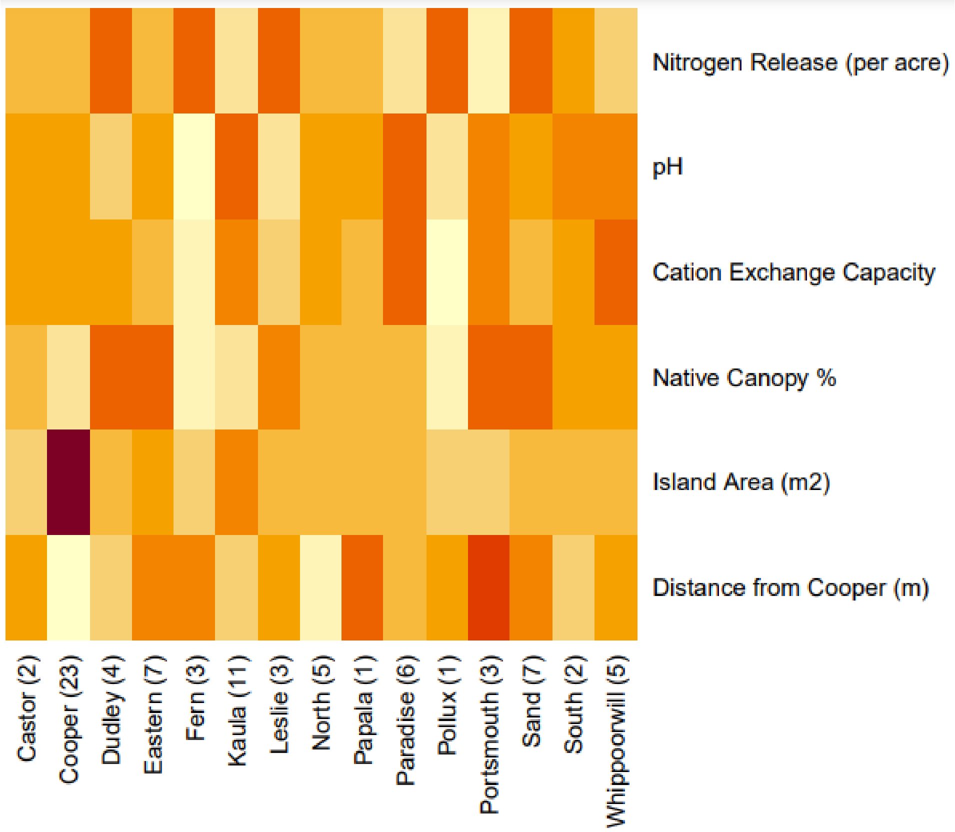

Figure 2. The variation in island traits scaled by row. Sample size for each island are displayed in parentheses

To analyze this dataset, we used the msPGOcc() function from the Spoccupancy package in R. The MCMC algorithm sampled from the posterior distribution with 3 chains. Each chain ran for 50,000 samples with a burn-in period of 2,000 samples and a thinning rate of 3. This resulted in 48,000 total posterior samples. We evaluated MCMC chain convergence and resolution using Gelman-Rubin diagnostics and the effective sample size. We then extracted the posterior distributions for each species-site Z parameter. We calculated the mean of each distribution, then explored the FN probability matrix and estimated Z matrix.

3.3 Results

3.3.1 Simulated True Z and Observation Matrices

Each simulated true z matrix represented the presence/absence of 600 possible site-species combinations. In our 1000 simulated datasets, the average true z matrix was composed of 298.8 present site-species with a standard deviation of 32.80. Across all true z matrices, there were 298,845 total true presences and 301,155 total true absences.

The average observation matrix detected 80% of species present in the true z matrix, with a minimum of 53% (181 out of 341) and a maximum of 97% (254 out of 262). Across all datasets, the observation matrices detected 241,754 out of 298,845 total present species-site combinations. These observation matrices did not detect 57,091 present species, resulting in an average of 57.1 true false negatives per observation matrix with a minimum of 8 and maximum of 160.

3.3.2 False Negative Estimates

After the model fit each simulated observation matrix, we incorporated the false-negative probabilities, and counted the number of estimated false negatives in each estimated z matrix. The estimated z matrices averaged a total of 44.3 estimated false negatives per simulated dataset with a standard deviation of 25.2. On average, 66.3% were correctly estimated false negatives and 33.7% were incorrectly estimated false negatives (Figure 4). Of the 57,091 total true false negatives in the observation matrices, the estimated z matrices correctly predicted 28,576 of them, leaving 28,515 total incorrect absences (Figure 3). The occupancy model estimated a total of 15,672 incorrect false negatives across all simulated datasets. On average, the occupancy model predicted 15.7 incorrect false negatives per simulated dataset with a standard deviation of 11.8 (Figure 5). Additionally, the number of both correctly and incorrectly estimated false negatives increased with the number of true false negatives in the observation matrix.

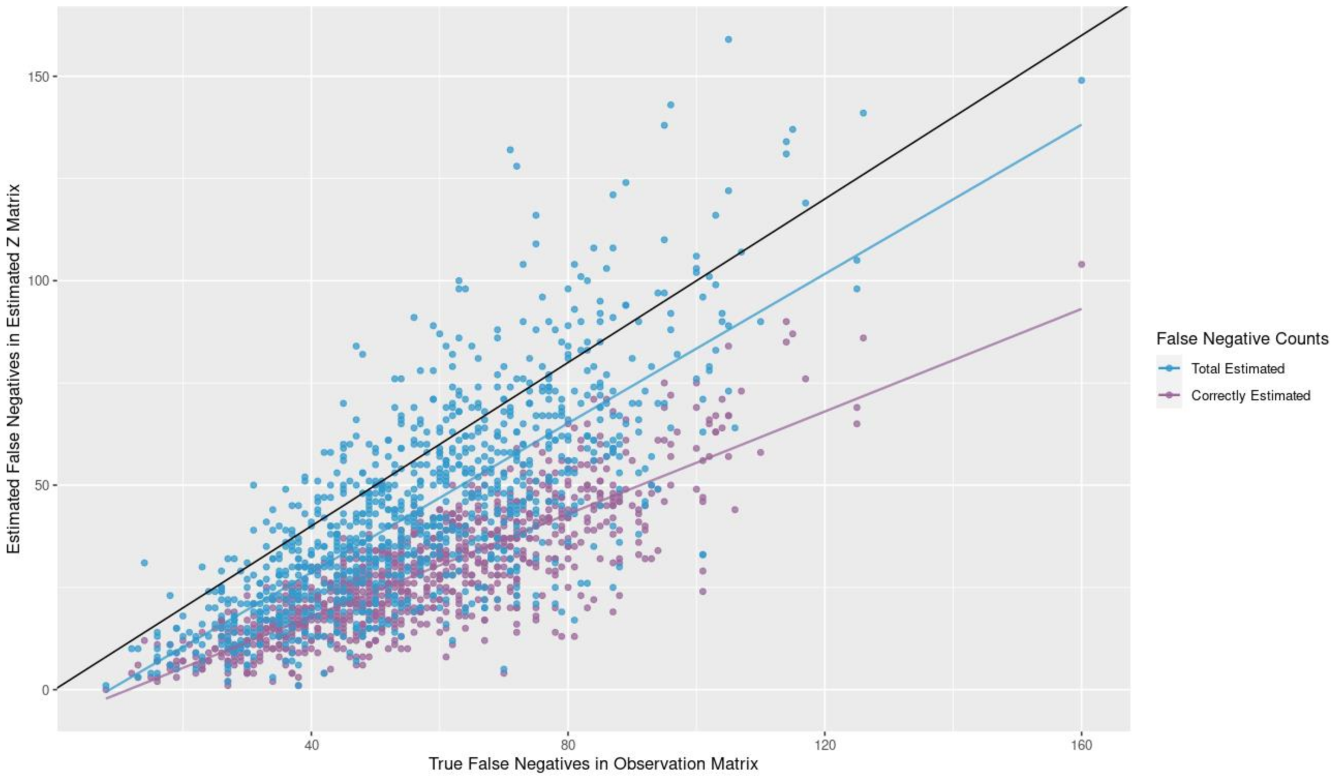

Figure 3. The total number of true false negatives on the x a-axis are compared to the number of estimated false negatives on the y-axis. The purple points represent the number of correctly estimated false negatives in each model. This number must be less than or equal to the true number of false negatives in each observation matrix, so these points will always be under the one-to-one line. Blue points represent the total number of false negatives estimated in the occupancy model. Above the one-to-one line means that there were more estimated false negatives than were true false negatives. Below the line means there were fewer estimated false negatives than were true false negatives.

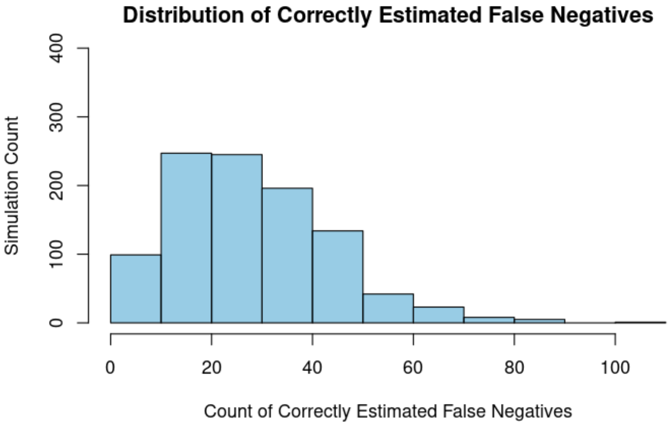

Figure 4. The number of correctly estimated false negatives added to each observation matrix. The occupancy model predicted an average of 29.3 correct false negatives in the estimated z matrix.

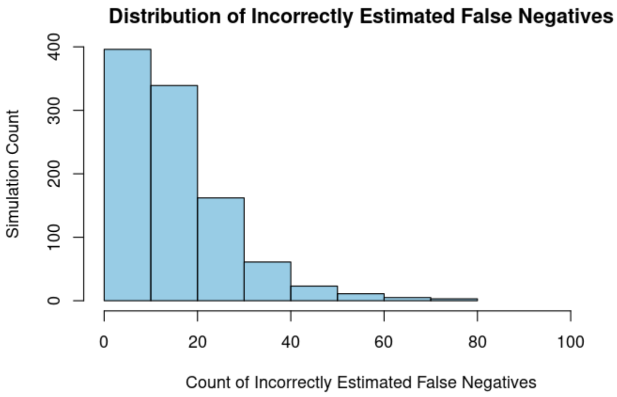

Figure 5. The number of incorrectly estimated false negatives added to each observation matrix. The occupancy model predicted an average of 15.7 incorrect false negatives in the estimated z matrix.

When examining how error was adjusted after controlling for false negative estimates, the occupancy model estimated more correct false negatives than incorrect false negatives in 901 out of 1000 model fits (Figure 6). In these cases, the model estimates decreased error between the observation matrix and estimated z matrix. 15 model fits had an equal number of correctly and incorrectly estimated false negatives. And 84 models predicted more incorrectly than correctly estimated false negatives, meaning that 8.4% of simulated datasets resulted in an increase in error.

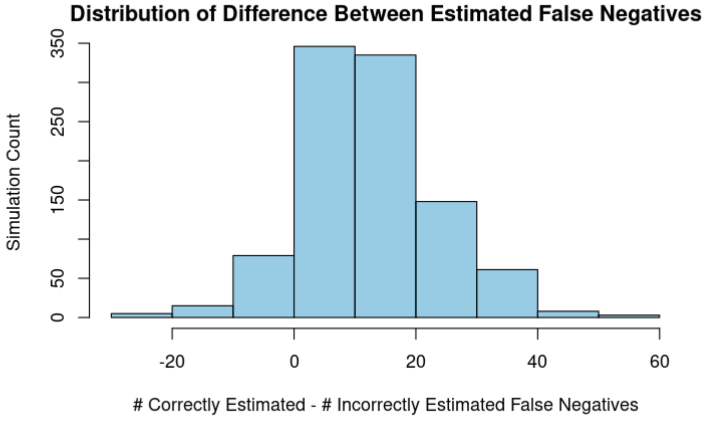

Figure 6. The difference between correctly and incorrectly estimated false negatives in each model iteration. Positive values indicate that the occupancy model estimated more correct false negatives than

incorrect false negatives, decreasing error overall. Negative values indicate that error was increased by the occupancy model estimates.

3.3.3 Overall Community Estimates

Next, we analyzed how well the true community was predicted by the observation and estimated matrices by comparing community accuracy before and after controlling for false negative estimates (Figure 7). When compared to the corresponding true z matrices, the observation matrices averaged an accuracy of 90.4% with a standard deviation of 3.4%. After the occupancy model estimates were incorporated, the estimated z matrices had an average accuracy of 92.6% with a standard deviation of 2.7%. Overall, controlling for false negative estimates increased community accuracy by 2.2% from the observed matrices to the estimated z matrices.

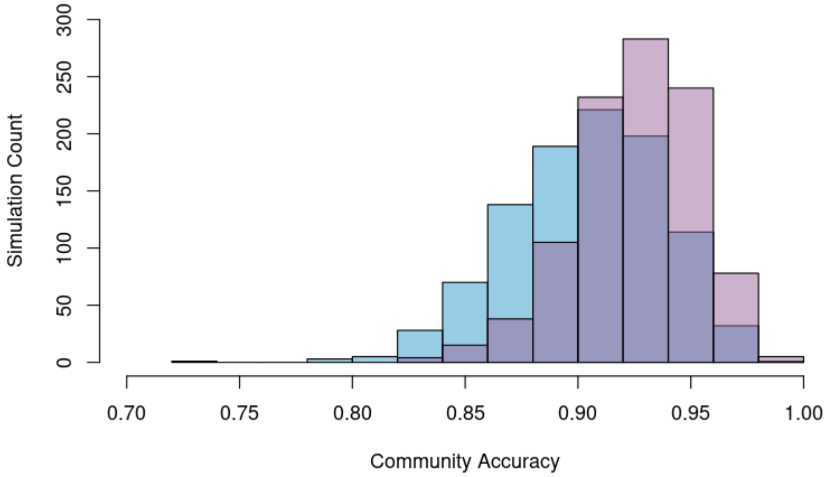

Figure 7. Community accurracy in comparison to the true z matrix. Blue bars illustrate the distribution of community accuracy in the observation matrices, while the purple bars illustrate accuracy of the estimated z matrices.

To understand how much information the occupancy model revealed about the true community, we compared the amount of mutual information shared between the true z matrix and the corresponding observation and false-negative probability matrix (Figure 8). The observation matrix alone shared an average of 0.43 nats of mutual information with the true z matrix with a standard deviation of 0.06. There was a minimum of 0.22 nats and maximum of 0.62 nats. After including model estimates, the false-negative probability matrix shared an average of 0.54 nats of mutual information with the true z matrix with a standard deviation of 0.05. There was a minimum of 0.35 and maximum of 0.68. Our results indicated that both the observation matrix and false-negative probability matrix obtained information from the true z matrix. However, the mutual information shared between every false-negative probability matrix was higher than the information shared with the comparable observation matrix (Figure 9). On average, the false-negative probability matrix revealed 0.11 more nats of information about the true z matrix than the observation matrix alone. The minimum difference was 0.03, while the maximum was 0.20.

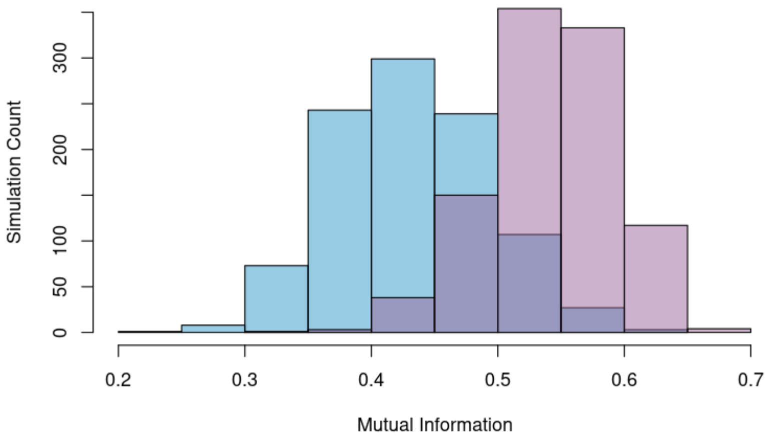

Figure 8. The mutual information shared with the true z matrix of 1000 simulated datasets. The blue bars are mutual information shared with the observation matrix. Purple bars are the mutual information shared with the false-negative probability matrix.

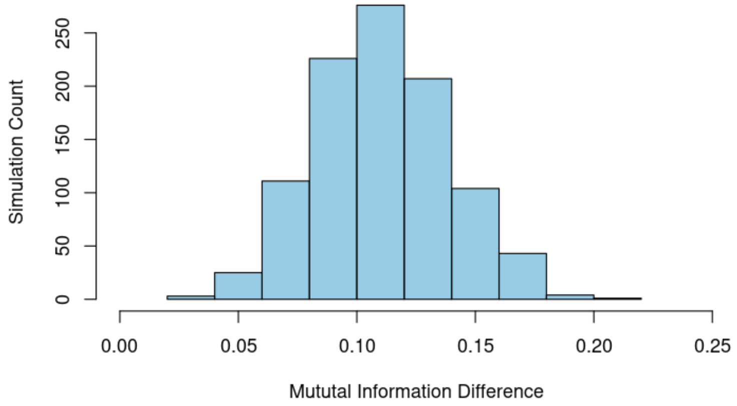

Figure 9. The difference in mutual information that the true z matrix shares with the false-negative probability matrix and the comparable observation matrix. Positive values indicate that the true z matrix shared more mutual information with the false-negative probability matrix than the observation matrix.

Finally, we investigated if the occupancy model improved species-site richness estimates. Although all species identities and locations may not be correct in the false-negative probability matrix and estimated z matrix, we totaled the number of estimated site-species combinations (Figure 10). When comparing true z richness to observed richness, the observed data averaged 57.1 fewer species-site combinations than true richness, with a standard deviation of 20.6. Richness estimates based on the false-negative probabilities averaged 3.7 more site-species combinations than true richness with a standard deviation of 13.0. And richness predicted from the estimated z matrix averaged 12.8 fewer present site-species combinations than true richness with a standard deviation of 16.9. Our results indicated that both the false-negative probability matrix and the estimated z matrix were better predictors of metacommunity richness than the observation matrix alone (Figure 11). And the false-negative probability matrix was the best predictor of species-site richness, even though species identities were inaccurate.

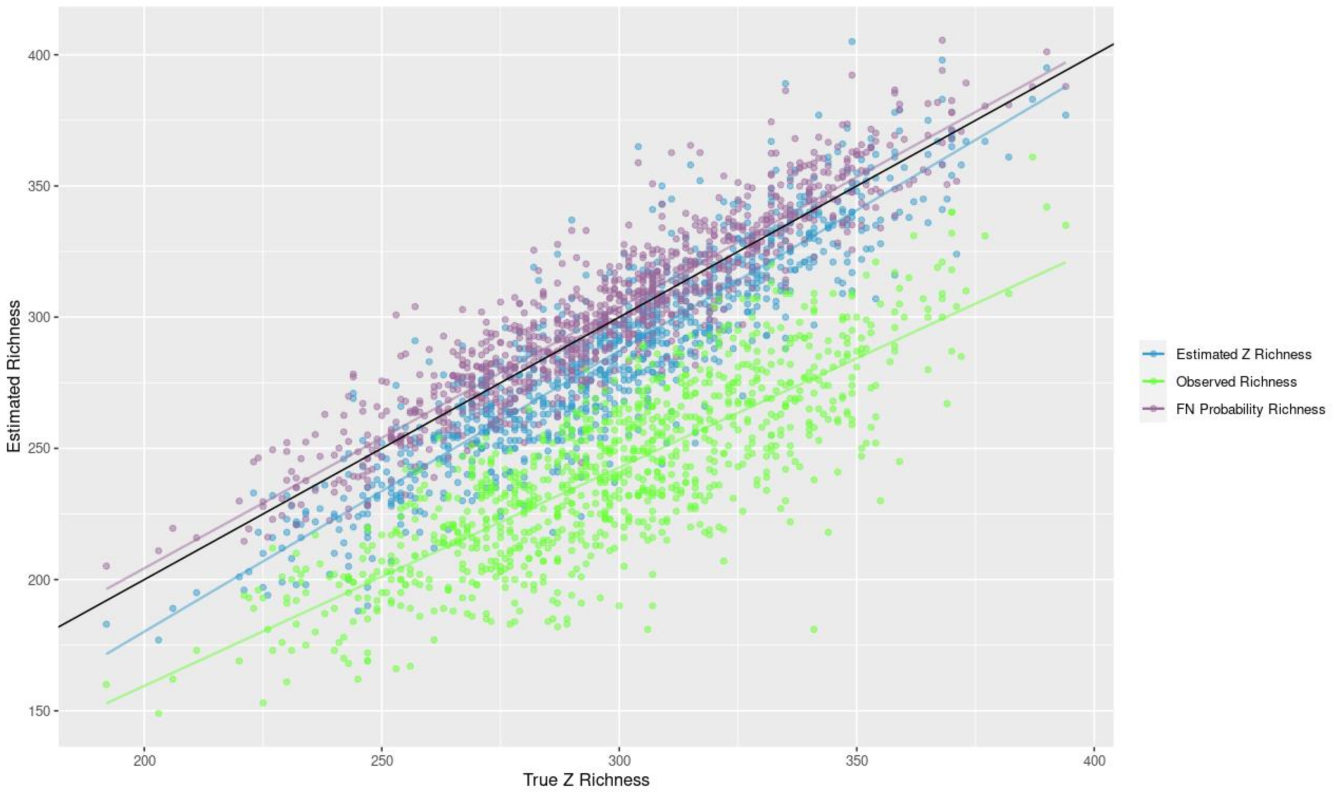

Figure 10. Species richness in the true z matrix and the estimated z matrix. Data points on the one-to-one line indicate that the estimated species richness was equivalent to the true species richness in that model iteration. Data points below the line mean that species richness was underestimated in that model, while points above the line were overestimated.

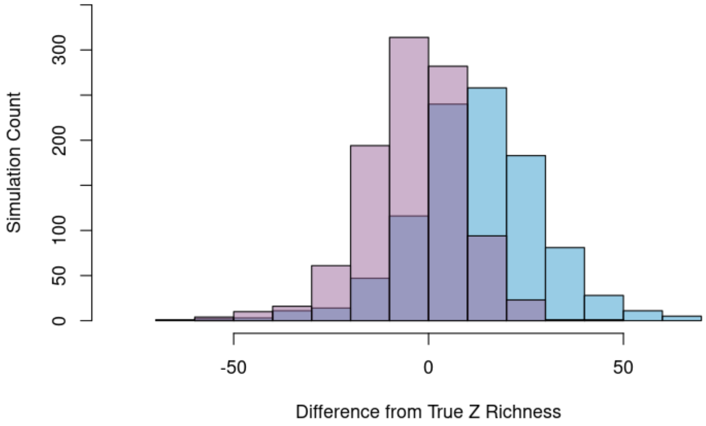

Figure 11. The difference in species-site richness from the true z matrix richness. Blue bars represent the richness difference between the true z matrix and the estimated z matrix. Purple bars represent the richness difference between the true z matrix and the false-negative probability matrix. Positive values mean the true z matrix has higher richness than the estimated matrices, while negative values indicate estimated matrices had higher richness than the true z matrices.

To evaluate the occupancy model's richness estimates compared to published estimators, we calculated richness of each simulated observed dataset using the iChao2 species estimator (Chiu *et al*., 2014). On average, Chao overestimated richness by 7.9 site-species combinations with a standard deviation of 18.4. Therefore, the false-negative probability matrix was better at predicting true richness than the iChao2 richness estimator (though it required considerably more information) (Figure 12).

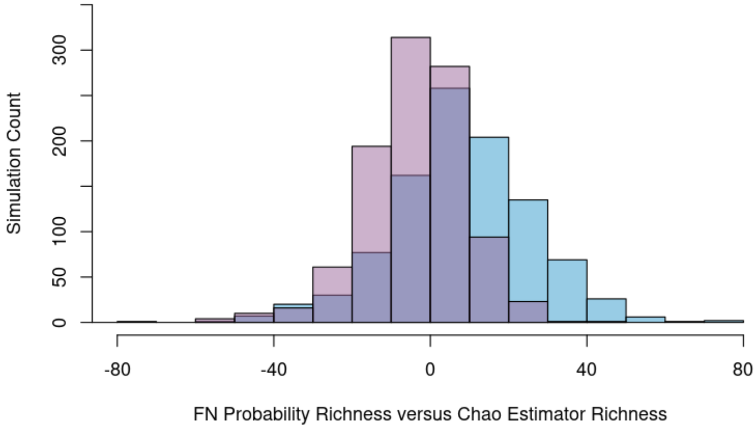

Figure 12. Blue bars are the difference between true and Chao richness. Purple bars are the difference between true and FN probability matrix richness. A difference of zero indicates that true richness and estimated richness are the same.

3.3.4 Case Study – Hymenoptera Data

In the dataset used for the case study, there were 87 trees fogged across 15 islets, and researchers documented the presence-absence of 44 Hymenoptera species on each tree. Of the 3828 possible detections (44 species x 87 trees), there were 765 observed species-tree combinations (Figure 13). The ant *Pheidole megacephala* was the most common species, detected in 77 of 87 fogging samples. And the parasitoid wasps, *Ceraphron sp*. 1, *Chelonus blackburni*, and *Ganaspsi sp*. 2 were the least common, with a single detection each. Of the 660 possible observations at the islet scale (44 species x 15 islets), there were 266 Hymenoptera-island combinations observed.

For the case study, we estimated false-negative probabilities of Hymenoptera species on Palmyra Atoll and predicted true occurrences on each sampled islet. After the occupancy model fit the data, we first confirmed chain convergence of all parameters. All Gelman-Rubin diagnostics were below 1.02 and the lowest ESS was 656. The posterior draws of some regression coefficients were correlated, most notably the effect of distance to Cooper Island and island area, distance to Cooper Island and native canopy proportion, and average soil pH and average nitrogen release. However, we were not interested in the individual effects of each of these covariates, but instead the combined effect of all predictors on species occupancy. As a result, correlated regression coefficients did not bias model predictions and false-negative probabilities.

In this metacommunity, the community-level occurrence intercept was 2.49 with a standard deviation of .88 indicating that the average occurrence in the community was high. The community-level intercept variance was 8.23 with a standard deviation of 6.00 suggesting that average occurrence did depend on the Hymenoptera species. This was further supported by the estimated species-level occurrence probabilities, which ranged from 0.99 (*N. bourbonica* on Dudley) to 0.08 (*Probaryconus sp*. on Pollux). The non-native ant species, *Pheidole megacephala,* had the highest average estimated occurrence probability across islands with a mean of 0.92 and a standard deviation of 0.12. The parasitoid wasp species, *Blepyrus insularis*, had the lowest average estimated occurrence probability across islands with a mean of 0.41 and standard deviation of 0.23.

The mean community-level detection intercept was -1.63 with a standard deviation of 0.24 suggesting low average detection in the community. The variance parameter describing the detection intercept had a mean of 2.02 with a standard deviation of 0.55, so detection probabilities varied from species to species. This was further supported by the species-level detection probabilities, ranging from .04 to .90. The species with the highest average detection probability across sampling replicates was *P. megacephala* with a mean of .86 and standard deviation of 0.02. The parasitoid wasp, *Ceraphron sp*., had the lowest average detection probability across sites with a mean of 0.04 and standard deviation of 0.002. This meant that even when present, *Ceraphron sp*. was unlikely to be detected.

We extracted the species-islet specific posterior distributions of z and examined the FN probability matrix (Figure 14) and resulting estimated Z matrix (Figure 15). In the FN probability matrix, estimated false negatives ranged from 0.02 (*Aprostocetus hagenowii* on Fern) to .99 (*Hypoponera punctatissima* on South Fighter). Including observed species, the mean value in the FN matrix was 0.71 with a standard deviation of .32. Of the estimated false negatives in the FN probability matrix, the average false-negative probability was .54 with a standard deviation of .29. The estimated Z matrix predicted 250 total false negatives in this metacommunity. In terms of overall richness, the FN probability matrix estimated the true community had 495.35 species-islet combinations, while the estimated Z matrix expected 506 species-island combinations. The iChao2 estimated richness as 411.60 species-island combinations with a standard error of 16.64.

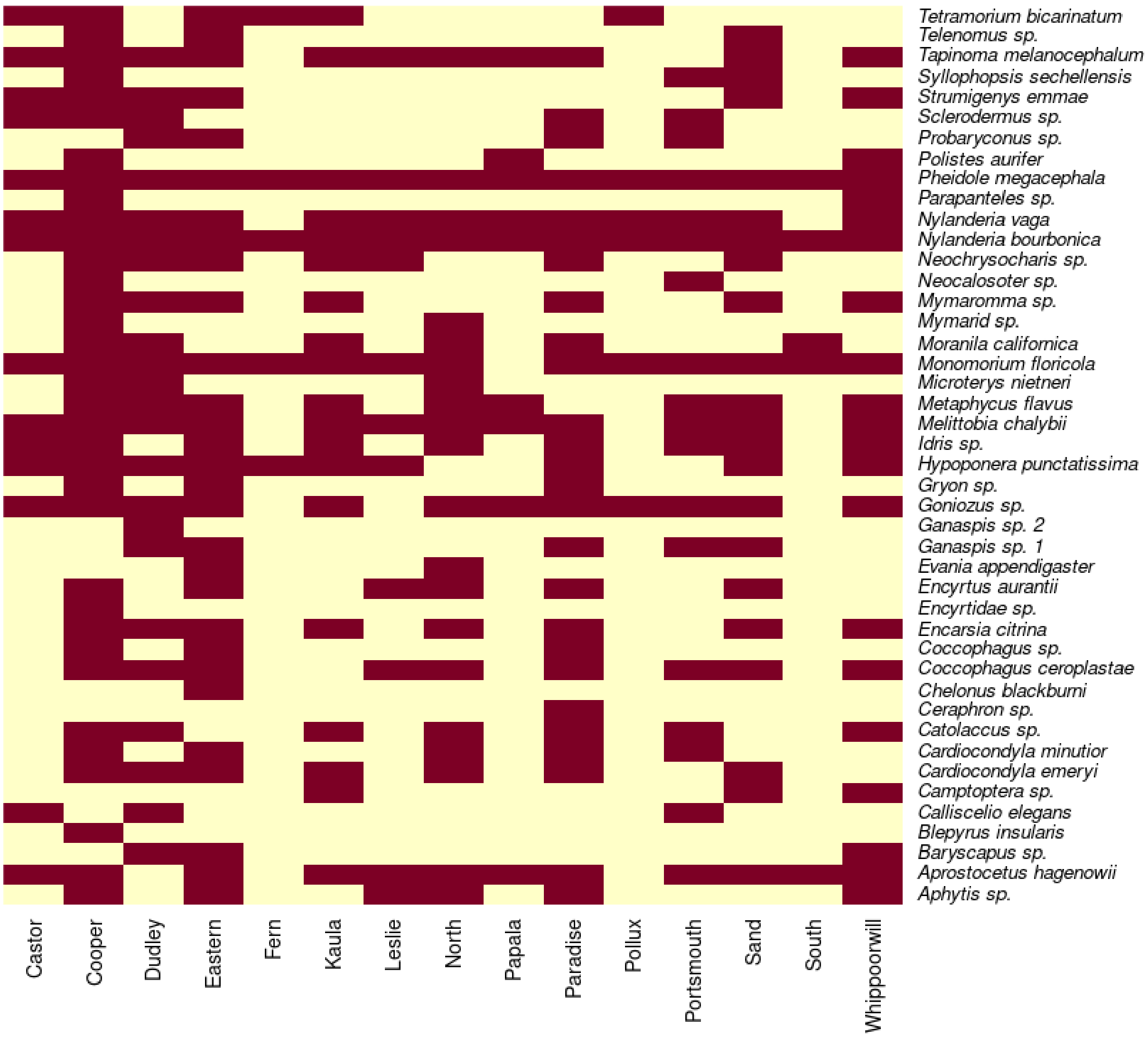

Figure 13. The fogging samples resulted in the above observation matrix. Light cells indicate that a species was not observed on that islet based on the fogging sample detections. Dark cells mean that a species was observed.

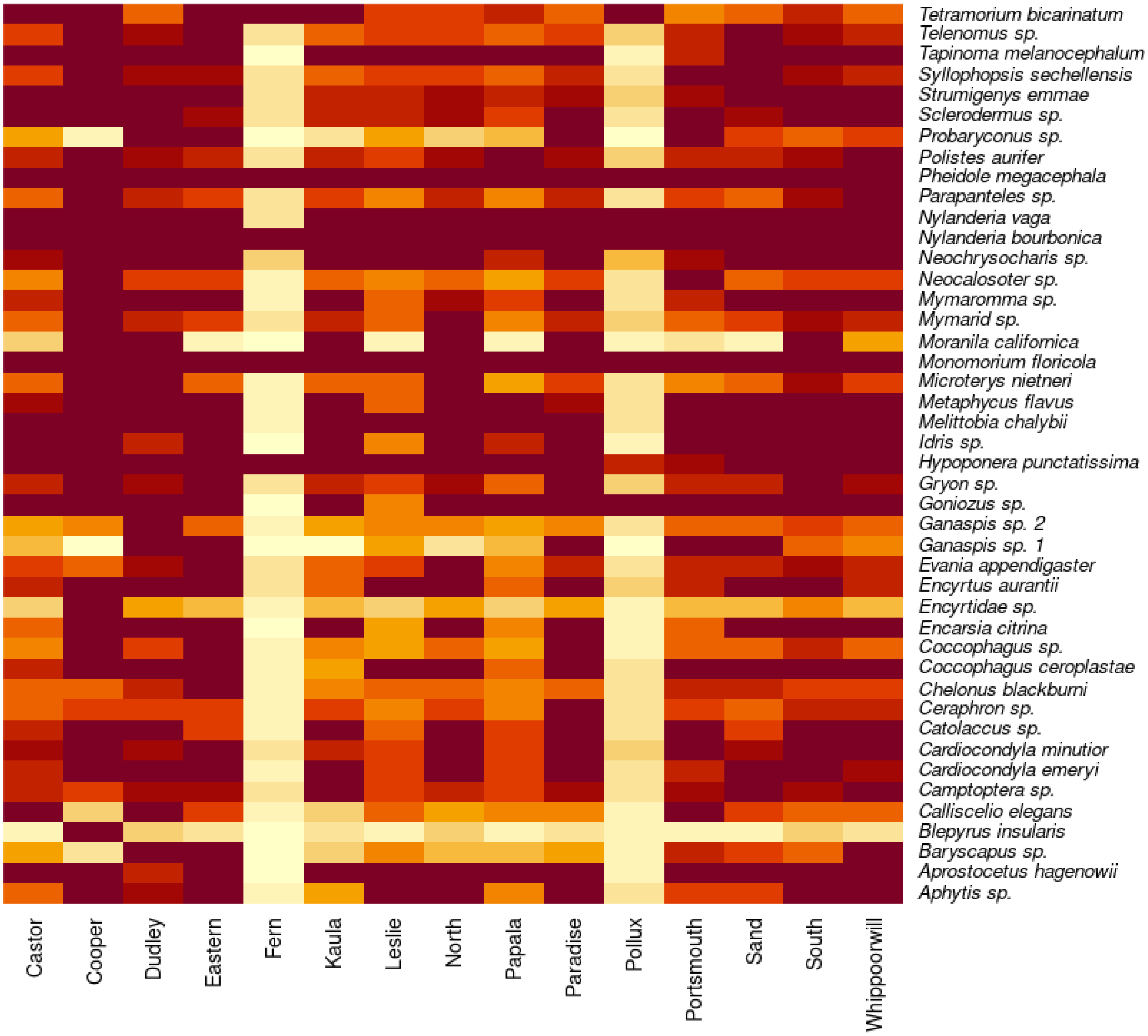

Figure 14. The occupancy model estimated false-negative probabilities for each species-island combination that was not observed. This resulted in the above false-negative probability matrix. Observed species are the darkest color and have the highest probability of occurring (1). The color of unobserved species cells illustrate the estimated false-negative probability from low (lightest) to high (darkest)

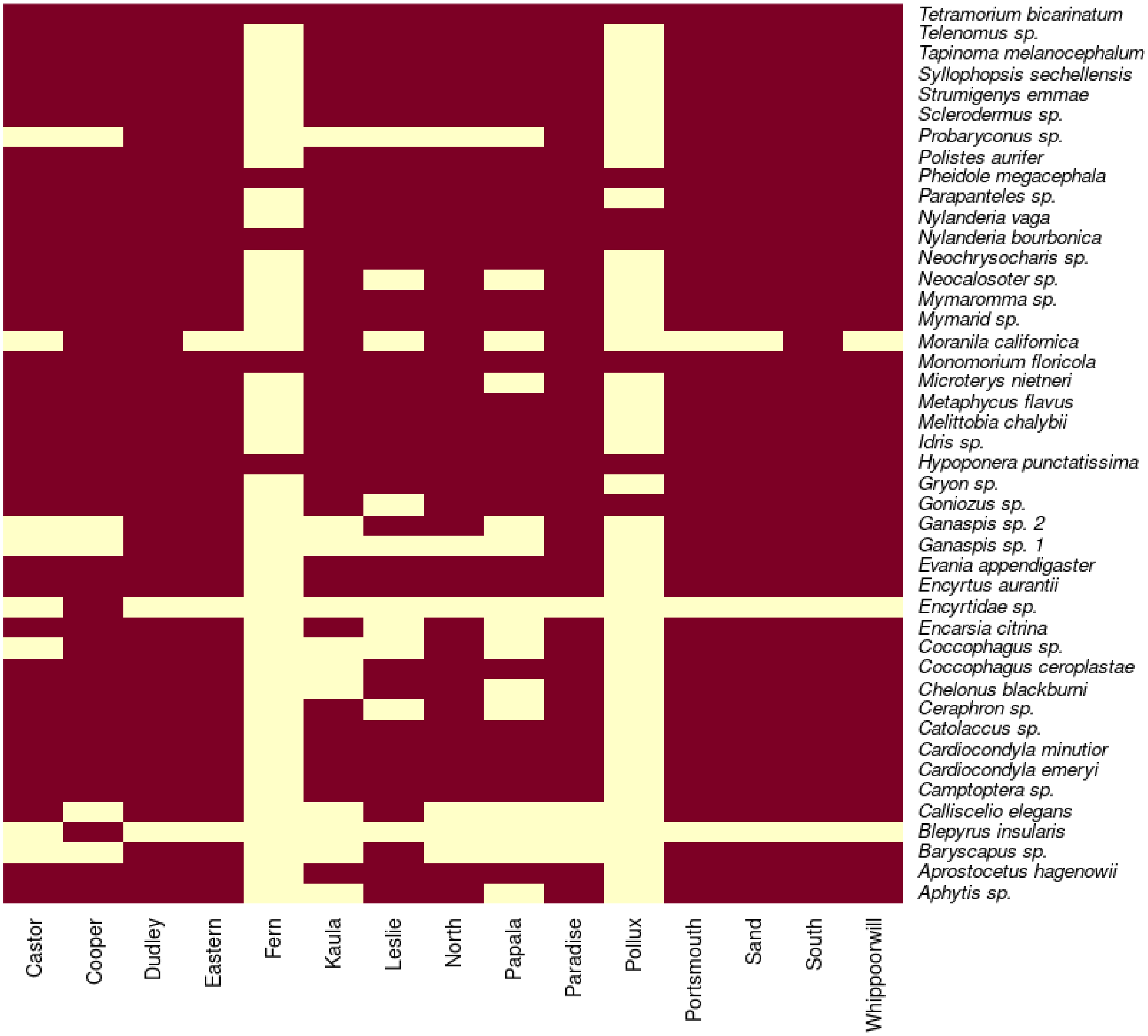

Figure 15. Applying a probability threshold of 0.5 to the FN probability matrix resulted in the above estimated Z matrix.

3.4 Discussion

3.4.1 Simulation Discussion

In our simulation study, we asked whether an occupancy model could predict the location and identities of false negative species. Based on our 1000 simulated datasets, we presented evidence to suggest that it partially could. Across the datasets, the occupancy model estimated more correct false negatives than incorrect false negatives, reducing error in the majority of community estimates. This occupancy model did not predict all true false negatives, leaving about 49.94% as incorrect absences, but 66.3% of the estimated false negatives were correct. After controlling for false negatives, 8% of estimated z matrices had more error when compared to the observation matrix. Increasing error was the least ideal outcome, and more work needs to be done to understand the datasets where this occurs and how to make better predictions in these scenarios. After taking false-negative estimates into account, community accuracy improved overall. The estimated z matrices were closer representations of true z's community composition than the observed matrices alone. And even though many incorrect false negatives were included in the estimated z matrix, community accuracy increased because, on average, the occupancy model predicted more false negatives correctly. In addition to community accuracy, we quantified mutual information in order to directly measure the amount of information that the occupancy model obtained. In every model iteration, the true z matrix shared more information with the false-negative probability matrix than the observation matrix alone. In real biological surveys, the true z matrix is unknown, and researchers must rely on the observed data to make inferences about the true state of the system. Although this occupancy model did not predict any simulated community with 100% accuracy, it did reveal more information about each "hidden" z matrix. These results suggest that occupancy models are a promising tool that could potentially improve our estimates of present community compositions.

In addition to predicting the identities and locations of species, we also examined the occupancy model's ability to predict species richness. True richness was compared to the observed richness, false-negative probability richness, and estimated richness. Both richness estimates from the occupancy model performed better than observed richness. And of the three, richness calculated from the false-negative probability matrix was the closest predictor of true richness. These false-negative probability richness estimates even surpassed a commonly used richness estimator, the iChao2 (Chiu *et al*., 2014).

Our simulated study results found that as the number of false negatives increased in the observation matrix, so too did the number of correctly and incorrectly estimated false negatives in the estimated z matrix. Because each of our simulated datasets had 10 species and 60 sites, there were a maximum of 600 possible false negatives (given all species were present but none were observed), and this number decreased with every observation. As the number of true false negatives increased in the observation matrix, the occupancy model was provided an increasingly less accurate y matrix to estimate parameter values. When true false negatives were high, the y matrix did not contain enough information about true species occurrences for the model to make accurate predictions, leading to more incorrect false-negative estimates. Although estimated false-negative identities were less accurate, species richness was still closely predicted in these cases. Thus, estimated richness remained a reliable predictor of true richness even when true false-negative counts were high.

In our estimated z matrices, we categorized an unobserved species at a site as an estimated false negative when mean z was greater than 0.5. We chose this threshold simply because these species are predicted to be truly present more often than not. However, there are costs associated with this probability threshold. When the probability of a false negative is 0.5, the alternative event (an absence) is equally probable, resulting in maximum uncertainty of the outcome. We cannot confidently make a guess one way or the other. Having a mean probability of 0.51 is not statistically different from 0.50, yet our estimated z matrix confidently determined these species were false negatives. With this threshold, we risked including many incorrectly estimated false negatives in our estimated z matrix. We could increase the probability threshold for more conservative false negative predictions, but we jeopardize defining more species as absent when truly present. The decision comes down to if it's more detrimental to miss present species or include absent species. Using conservation ecology as an example, having an incorrect absence at a site could result in vulnerable populations that are not being managed. But a species that was absent but assumed present could lead to wasted financial resources and efforts. In every scenario, the probability threshold to estimate a false negative should be considered with care. Regardless of the threshold used, basing false negatives on a threshold has the drawback of increasing error in the estimate of the total false-negative rate to the extent that the average false-negative probability departs from 0.5. Namely if the average probability of a false negative is <0.5, the total number of false negatives will be underestimated whereas if the average probability of a false negative if >0.5, the total number of false negatives will be overestimated. A logical alternative to thresholds is to use false-negative probabilities instead. Our simulation study illustrated that the false-negative probability matrix shared more mutual information with true z and was better at predicting true richness than the estimated z matrix. Basing management decisions on the false-negative probability matrix may be less straightforward but could provide more insight into the true community. And if resource managers were interested in conserving species richness, as opposed to a specific species, then richness estimates from the false-negative probability matrix is likely the better option.

Although not specifically highlighted in this study, the probability of a false negative is dependent on the interaction of two parameters: the probability of occurrence (psi) and the probability of detection (p). The value of psi and p imply different types of species and observations. Therefore, combinations of psi and p will result in different false-negative probability distributions. As an example, a high psi and low p describes a species that has favorable local conditions but is cryptic or rare. This will result in a high false-negative probability because the species is likely present, but difficult to find. In contrast, a low psi and high p indicates a species that is incompatible with local conditions, yet easy to observe if present. Therefore, the probability of a false negative is low because the species is unlikely present and detectable if it was. Less intuitively, a species with a high psi and high p is likely present and easy to observe, making a non-detection unexpected, and therefore meaningful. Because the model predicts an observation where there wasn't one, the model cannot confidently estimate the false-negative probability, resulting in a wide distribution centered around 0.5. Particularly in these cases, replicate samples can help distinguish between an estimated true absence or false negative. Finally, species with a low psi and low p have a low mean false-negative probability and a distribution with a long tail. In this occupancy model, correctly estimating the psi and p parameters will dictate the success of the occupancy model's false negative predictions. In the future, we could explore which combinations of psi and p interact to predict correct or incorrect false negatives.

3.4.2 Case Study Discussion

Our case study of Hymenoptera surveys illustrated how an occupancy model can be used to predict false-negative probabilities and metacommunity occurrence. In comparison to our simulated dataset, the case study estimates resulted in a higher ratio of estimated to observed species (from 250 observed species-island combinations to 495 estimated species-island combinations). One reason for this could be that in the simulated datasets, occurrence and detection probabilities were calculated based on uniform distribution draws ranging from -2 to 2 on the logit scale. As a result, the species in the simulated communities represented the range of possible occurrence and detection probability combinations, with mean occurrence and detection probabilities centered around 0.5. However, in our case study dataset, the community-level average occurrence and detection was 2.4935 and -1.634 on the logit scale, or .9237 and .1633 on the probability scale respectively. This meant that most Hymenoptera species in the case study community had a high occurrence probability and low detection probability. These types of species are the most likely to be missed in a community survey resulting in high false-negative probabilities. And as a result, the occupancy model estimated a much higher number of false negative species on islands in the case study community than any simulated community. The most common species, *Pheidole megacephala*, was observed on every island during sampling, so false-negative probability estimates were not necessary here. We gained more insight when species had fewer observations. For example, *Tetramorium bicarinatum* was only observed on 6 islands during sampling. But based on the false-negative probabilities, this species was predicted to truly occur on all islands.

3.4.3 Applications

There are many applications for models that parameterize the probability and predict the identities of false negative species. Estimating false-negative probabilities can be a useful tool for sampling design. By estimating false-negative probabilities, researchers can make more informed decisions about where to allocate additional sampling efforts. For conservation ecology, considering missed species could inform management decisions when preserving habitat, especially of rare or cryptic species (Gaston, 1994; MacKenzie *et al*., 2005; Williams *et al*., 2002). False negative estimates could also help predict the spread of non-native species that threaten biodiversity (Kikillus *et al*., 2009; Pimentel *et al*., 2002; Sikder, 2006; Bled *et al*, 2011).

In disease ecology, false negatives can arise when distinguishing between infected or uninfected individuals or populations. Controlling for imperfect detection could help more accurately assess disease risk and evaluate population dynamics when transmitting to humans, agriculture, or livestock (McClintock, 2010; Dobson & Foufopoulos, 2001; Webster *et al.,* 2006). And in community ecology in general, correct inferences depend on the observed data accurately representing the true communities. Previous studies have demonstrated that not accounting for imperfect detection can lead to spurious patterns (Jennelle *et al.,* 2007; Conn & Cooch 2009). So, assessing detection probabilities and false-negative probabilities should be standard procedure before interpreting results of a community analysis. If false negatives are estimated to be high, researchers may choose to rely more on density estimates rather than presence-absence estimates.

3.4.4 Assumptions and Future Directions

In our simulation study, the occupancy model predicted the locations and identities of false negative species. However, as with any simulated dataset and statistical model, we made several assumptions.

The first set of assumptions arise based on how we generated our simulated datasets. We limited the effect of all site traits and replicate traits to between -2 to 2 on the logit scale which transforms to .12 to .89 on the probability scale. Because the logit scale truncates near both ends of the probability scale, sampling the logit scale uniformly to encapsulate probabilities from .00 to .99 would result in a bimodal distribution. To ensure that we sampled a wide range of probabilities without overrepresentation on either end, we limited our logit scale to sample uniformly between -2 to 2. This range of effect sizes is also ecologically realistic, allowing either positive, negative, or no effects. In reality, covariates could have more extreme effects on species occurrence or detection, however this possibility is excluded here. Another assumption in our simulated study is that the included covariates exclusively determined species occurrence and detection probabilities. In biological surveys, species observations are likely dependent on several covariates, and researchers are often unable to collect data on all of them. Therefore, there may be influential covariates not informing the occupancy model's predictions, which may lead to less accurate false negative estimates. A next step to explore this assumption would be to simulate traits that influence occurrence and detection probabilities but omit them from the occupancy model to see how unexplained variance affects false negative predictions.

The second set of assumptions is based on the bounds of this occupancy model framework. For example, this occupancy model is indexed only for site traits, not species traits or species co-occurrences. As a result, we did not structure our simulated datasets with either of these factors. However, previous empirical work has demonstrated that species traits and co-occurrences can often influence species distributions and observations (Bhowmik *et al*., 2015; Tremlová & Münzbergová, 2007; Syphard & Franklin, 2009; Hanspach *et al*., 2010; Wisz *et al*., 2013; Araújo & Luoto, 2007). So, including species-based effects may improve false negative estimates. Additionally, this occupancy model framework cannot account for false positives in the observation matrix, so we structured our simulated data such that all observed species were truly present. False positives are expected to be less common than false negatives because they arise when an observed species is misidentified. However, these misidentifications should be limited by observer experience and good protocols (Tyre *et al*., 2003). Nonetheless, false positives are likely prevalent sources of error in biological datasets and previous studies have demonstrated their bias (Tyre *et al*., 2003; Ferguson *et al*., 2015; Miller *et al*., 2015; Williams *et al*., 2002). As a result, false positives should not be ignored. To remedy both assumptions, we need to use a different occupancy model framework that can parameterize species trait effects and false positive estimates.

For our last set of assumptions, we assumed that the simulated and case study datasets were structured in the same way, but this is unlikely to be the case. There are several differences that could influence the accuracy of false negative estimates. As previously discussed, large effect sizes, unmeasured covariates, species traits, species co-occurrences, and false positives are potential factors affecting the case study observations and model predictions. Each of these factors must be considered. There are additional differences between our simulated and case study datasets. For example, the simulated datasets were restricted to 3 replicate observations per site in order to generate true false negatives even when detection probabilities were high. But in our case study dataset, the number of replicates varies with site (between 1 to 23), so some sites had greater representation than others. The more replicate samples from a site without an observation, the lower the probability of a false negative at those sites. This may lead to biased estimates depending on the site.

Before we can predict species communities with high accuracy, this study suggests many next steps and future directions. Each of the above assumptions must be fleshed out to understand their influence on model predictions.

3.5 Conclusion

Overall, this occupancy model increased simulated community accuracy by an average of only 2%. So we ask the question, is estimating false negatives important enough to be worth the effort? We argue that it is. Although this occupancy model may not be the solution to predicting all false negatives, it did improve our understanding of the true species community. In our simulation study, true community composition was better matched by the estimated community than observations alone. Mutual information also improved after adding the estimated false negative probabilities to the observed datasets. And richness estimates from the occupancy model resulted in closer approximations of true richness. There are a myriad of future directions to explore before we can confidently predict false negative estimates, but this study is a baseline for next steps. More research and future advancements will only enhance predictions. Occupancy models are promising tools that can help reveal occurrence dynamics and control for imperfect detection. These insights can inform conservation decisions, improve disease detections (McClintock, 2010), and predict invasive species spread (Kikillus *et al*., 2009, Bled *et al.,* 2011). This study focused on an important problem: false negatives are ubiquitous sources of error in biological datasets that can perpetuate misleading inferences if ignored (McClintock, 2010; Tyre *et al*., 2003). Adding model extensions to address estimator assumptions will be an important next step.

References

1. Abu-Madi, M. A., Behnke, J. M., Lewis, J. W., & Gilbert, F. S. (2000). Seasonal and site specific variation in the component community structure of intestinal helminths in Apodemus sylvaticus from three contrasting habitats in south-east England. *Journal of Helminthology*, *74*(1), 7-15.
2. Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91-93.
3. Andersen, A. N., Blum, M. S., & Jones, T. H. (1991). Venom alkaloids in Monomorium "rothsteini" Forel repel other ants: is this the secret to success by Monomorium in Australian ant communities?. *Oecologia*, *88*, 157-160.
4. Andersen, A. N., & Patel, A. D. (1994). Meat ants as dominant members of Australian ant communities: an experimental test of their influence on the foraging success and forager abundance of other species. *Oecologia*, *98*, 15-24.
5. Anderson, R. M., & May, R. M. (1978). Regulation and stability of host-parasite population interactions. *Journal of Animal Ecology*, *47*(1), 219-247.
6. Anderson, R. M., & May, R. M. (1982). Coevolution of hosts and parasites. *Parasitology*, *85*(Pt 2), 411-426.
7. Anderson, T. K., & Sukhdeo, M. V. (2011). Host centrality in food web networks determines parasite diversity. *PLoS One*, *6*(10), e26798.
8. Antunes, S. C., Pereira, R., Sousa, J. P., Santos, M. C., & Gonçalves, F. (2008). Spatial and temporal distribution of litter arthropods in different vegetation covers of Porto Santo Island (Madeira Archipelago, Portugal). *European Journal of Soil Biology*, *44*(1), 45-56.
9. Aragão, L. E. O. C., Malhi, Y., Metcalfe, D. B., Silva-Espejo, J. E., Jiménez, E., Navarrete, D., ... & Vásquez, R. (2009). Above-and below-ground net primary productivity across ten Amazonian forests on contrasting soils. *Biogeosciences*, *6*(12), 2759-2778.
10. Araújo, M. B., & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, *16*(6), 743-753.
11. Araújo, M. B., Pearson, R. G., & Rahbek, C. (2005). Equilibrium of species' distributions with climate. *Ecography*, *28*(5), 693-695.
12. Araújo, M. B., & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. *Ecography*, *37*(5), 406-415.
13. As, S. (1984). To fly or not to fly? Colonization of Baltic islands by winged and wingless carabid beetles. *Journal of Biogeography*, 413-426.
14. Balikai, R. A., & Pushpalatha, D. (2018). Bio-ecology and management of spiraling whitefly, Aleurodicus dispersus Russell through insecticides: A review. *Farming and Management*, *3*(1), 56-65.
15. Bang, C., & Faeth, S. H. (2011). Variation in arthropod communities in response to urbanization: seven years of arthropod monitoring in a desert city. *Landscape and Urban Planning*, *103*(3-4), 383-399.
16. Bayley, P. B., and J. T. Peterson. (2001). An approach to estimate probability of presence and richness of fish species. *Transactions of the American Fisheries Society* 130:620–633.

17. Becker, D. J., Streicker, D. G., & Altizer, S. (2018). Using host species traits to understand the consequences of resource provisioning for host–parasite interactions. *Journal of Animal Ecology*, *87*(2), 511-525.
18. Bell, J. R., Andrew King, R., Bohan, D. A., & Symondson, W. O. (2010). Spatial co-occurrence networks predict the feeding histories of polyphagous arthropod predators at field scales. *Ecography*, *33*(1), 64-72.
19. Bell, G., & Burt, A. (1991). The comparative biology of parasite species diversity: internal helminths of freshwater fish. *The Journal of Animal Ecology*, 1047-1064.
20. Belmaker, J., Zarnetske, P., Tuanmu, M. N., Zonneveld, S., Record, S., Strecker, A., & Beaudrot, L. (2015). Empirical evidence for the scale dependence of biotic interactions. *Global Ecology and Biogeography*, *24*(7), 750-761.
21. Benavides, J. A., Huchard, E., Pettorelli, N., King, A. J., Brown, M. E., Archer, C. E., ... & Cowlishaw, G. (2012). From parasite encounter to infection: Multiple-scale drivers of parasite richness in a wild social primate population. *American Journal of Physical Anthropology*, *147*(1), 52-63.
22. Benesh, D. P., Parker, G. A., Chubb, J. C., & Lafferty, K. D. (2021). Trade-offs with growth limit host range in complex life-cycle helminths. *The American Naturalist*, *197*(2), E40-E54.
23. Benkman, C. W. (2013). Biotic interaction strength and the intensity of selection. *Ecology Letters*, *16*(8), 1054-1060.
24. Bhowmik, A. K., & Schaefer, R. B. (2015). Large scale relationship between aquatic insect traits and climate. *PLOs one*, *10*(6), e0130025.
25. Bishop, M. J., Underwood, A. J., & Archambault, P. (2002). Sewage and environmental impacts on rocky shores: necessity of identifying relevant spatial scales. *Marine Ecology Progress Series*, *236*, 121-128.
26. Bled, F., Royle, J. A., & Cam, E. (2011). Hierarchical modeling of an invasive spread: the Eurasian Collared-Dove Streptopelia decaocto in the United States. *Ecological Applications*, *21*(1), 290-302.
27. Bonneau, J. L., and G. LaBar. (1997). Interobserver and temporal bull trout redd count variability in tributaries of Lake Pend Oreille, Idaho: Completion Report. Department of Fisheries and Wildlife, University of Idaho, Moscow.
28. Brusca, R. C. (1981). A monograph on the Isopoda Cymothoidae (Crustacea) of the eastern Pacific. *Zoological Journal of the Linnean Society*, *73*(2), 117-199.
29. Buck, J. C., Hechinger, R. F., Wood, A. C., Stewart, T. E., Kuris, A. M., & Lafferty, K. D. (2017). Host density increases parasite recruitment but decreases host risk in a snail–trematode system. *Ecology*, *98*(8), 2029-2038.
30. Buckley, Y. M., & Catford, J. (2016). Does the biogeographic origin of species matter? Ecological effects of native and non-native species and the use of origin to guide management. *Journal of Ecology*, *104*(1), 4-17.
31. Bullock, J. M., & Clarke, R. T. (2000). Long distance seed dispersal by wind: measuring and modelling the tail of the curve. *Oecologia*, *124*, 506-521.
32. Bullock, J. M., Edwards, R. J., Carey, P. D., & Rose, R. J. (2000). Geographical separation of two Ulex species at three spatial scales: does competition limit species' ranges?. *Ecography*, *23*(2), 257-271.
33. Bürkner P (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1), 1–28.

34. Bush, A. O., Lafferty, K. D., Lotz, J. M., & Shostak, A. W. (1997). Parasitology meets ecology on its own terms: Margolis et al. revisited. *The Journal of Parasitology*, 575-583.
35. Caira, J. N., Jensen, K., & Healy, C. J. (2014). Interrelationships annong tetraphyllidean and lecanicephalidean cestodes. *Interrelationships of the Platyhelminthes*, 135.
36. Cameron, T. W. (1964). Host specificity and the evolution of helminthic parasites. *Advances in parasitology*, *2*, 1-34.
37. Carlquist, S. (1974). Island biology. Columbia University Press, New York.
38. Chai, X., Bennett, J., & Poulin, R. (2022). Decay of parasite community similarity with host phylogenetic and geographic distances among deep-sea fish (grenadiers). *Parasitology*, *149*(13), 1737-1748.
39. Chan, Y. H. (2003). Biostatistics 104: correlational analysis. *Singapore Med J*, *44*(12), 614-619.
40. Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783-791.
41. Chao, A., & Chiu, C. H. (2016). Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, *1*, 26.
42. Chazdon, R.L., Colwell, R.K., Denslow, J.S. & Guariguata, M.R. (1998). Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In: *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies*, Dallmeier F. & Comiskey J.A. (ed.), pp. 285-309. Parthenon Publishing, Paris, France.
43. Chen, H. W., Liu, W. C., Davis, A. J., Jordán, F., Hwang, M. J., & Shao, K. T. (2008). Network position of hosts in food webs and their parasite diversity. *Oikos*, *117*(12), 1847-1855.
44. Chiu, C. H., Wang, Y. T., Walther, B. A., & Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, *70*(3), 671-682.
45. Colwell, R. K., and J. A. Coddington. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B,* 345:101–118.
46. Conn, P. B., & Cooch, E. G. (2009). Multistate capture–recapture analysis under imperfect state observation: an application to disease models. *Journal of Applied Ecology*, *46*(2), 486-492.
47. D'Antonio, C. M., and T. L. Dudley. (1995). Biological invasions as agents of change on islands versus mainlands. Pages 103– 121 in P. M. Vitousek, H. Andersen, and L. L. Loope, eds. Islands: *Biodiversity and ecosystem function*. Springer-Verlag, Berlin.
48. Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.
49. Davidson, D. W. (1998). Resource discovery versus resource domination in ants: a functional mechanism for breaking the trade-off. *Ecological Entomology*, *23*(4), 484-490.
50. Dawson, Y. E. (1959). Changes in Palmyra Atoll and its vegetation through the agency of man, 1913–1958. *Pac. Nat*. 1:1–51.
51. Diamond, S. E., Nichols, L. M., McCoy, N., Hirsch, C., Pelini, S. L., Sanders, N. J., ... & Dunn, R. R. (2012). A physiological trait-based approach to predicting the responses of species to experimental climate warming. *Ecology*, *93*(11), 2305-2312.

52. Dobson, A.P. & Foufopoulos, J. (2001). Emerging infectious pathogens of wildlife. *Philos. T. Roy. Soc*. B, 356, 1001–1012.
53. Dorazio, R. M., & Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, *100*(470), 389-398.
54. Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., ... & Kraan, C. (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography*, *27*(9), 1004-1016.
55. Doser, J. W., Finley, A. O., Kéry, M., & Zipkin, E. F. (2022). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, *13*(8), 1670-1678.
56. Dunham, J., Rieman, B., & Davis, K. (2001). Sources and magnitude of sampling error in redd counts for bull trout. *North American Journal of Fisheries Management*, *21*(2), 343-352.
57. Dybzinski, R., Fargione, J. E., Zak, D. R., Fornara, D., & Tilman, D. (2008). Soil fertility increases with plant species diversity in a long-term biodiversity experiment. *Oecologia*, *158*, 85-93.
58. Engelstädter, J., & Fortuna, N. Z. (2019). The dynamics of preferential host switching: Host phylogeny as a key predictor of parasite distribution. *Evolution*, *73*(7), 1330-1340.
59. Euzet, L. (1956). Recherches sur les Cestodes Tétraphyllides des Sélaciens des côtes de France. *Nat. Monspel., Ser. Zool*. 3:1-263.
60. Euzet, L., & Combes, C. (1980). Les problèmes de l'espèce dans le règne animal. *Memoires Societe Zoologique de France*, *40*, 238-285.
61. Fenoglio, M. S., Rossetti, M. R., & Videla, M. (2020). Negative effects of urbanization on terrestrial arthropod communities: A meta-analysis. *Global Ecology and Biogeography*, *29*(8), 1412-1429.
62. Ferguson, P. F., Conroy, M. J., & Hepinstall-Cymerman, J. (2015). Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. *Methods in Ecology and Evolution*, *6*(12), 1395-1406.
63. Fernandez, J., & Esch, G. W. (1991). Guild structure of larval trematodes in the snail Helisoma anceps: patterns and processes at the individual host level. *The Journal of Parasitology*, 528-539.
64. Fernandez, J., & Esch, G. W. (1991). The component community structure of larval trematodes in the pulmonate snail Helisoma anceps. *The Journal of Parasitology*, 540-550.
65. Fuentes, M. V., Sáez, S., Trelis, M., Galán-Puchades, M. T., & Esteban, J. G. (2004). The helminth community of the wood mouse, Apodemus sylvaticus, in the Sierra Espuña, Murcia, Spain. *Journal of Helminthology*, *78*(3), 219.
66. Garzon-Lopez, C. X., Jansen, P. A., Bohlman, S. A., Ordonez, A., & Olff, H. (2014). Effects of sampling scale on patterns of habitat association in tropical trees. *Journal of Vegetation Science*, *25*(2), 349-362.
67. Gaston, K. J. (1994). What is rarity? *Springer Netherlands*, 1-21.
68. Gauch, H. G. (1982). *Multivariate analysis in community ecology* (No. 1). Cambridge University Press.

69. George-Nascimento, M., Muñoz, G., Marquet, P. A., & Poulin, R. (2004). Testing the energetic equivalence rule with helminth endoparasites of vertebrates. *Ecology Letters*, *7*(7), 527-531.
70. Goater, T. M., Esch, G. W., & Bush, A. O. (1987). Helminth parasites of sympatric salamanders: ecological concepts at infracommunity, component and compound community levels. *American Midland Naturalist*, 289-300.
71. Godsoe, W., Murray, R., & Plank, M. J. (2015). The effect of competition on species' distributions depends on coexistence, rather than scale alone. *Ecography*, *38*(11), 1071-1079.
72. Gotelli, N. J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, *81*(9), 2606-2621.
73. Gotelli, N. J., & McCabe, D. J. (2002). Species co-occurrence: a meta-analysis of JM Diamond's assembly rules model. *Ecology*, *83*(8), 2091-2096.
74. Gotelli, N. J., & Rohde, K. (2002). Co-occurrence of ectoparasites of marine fishes: a null model analysis. *Ecology Letters*, *5*(1), 86-94.
75. Gotelli, N. J., Graves, G. R., & Rahbek, C. (2010). Macroecological signals of species interactions in the Danish avifauna. *Proceedings of the National Academy of Sciences*, *107*(11), 5030-5035.
76. Guégan, J. F., & Hugueny, B. (1994). A nested parasite species subset pattern in tropical fish: host as major determinant of parasite infracommunity structure. *Oecologia*, *100*(1-2), 184-189.
77. Guégan, J. F., & Kennedy, C. R. (1993). Maximum local helminth parasite community richness in British freshwater fish: a test of the colonization time hypothesis. *Parasitology*, *106*(1), 91-100.
78. Guégan, J. F., & Kennedy, C. R. (1996). Parasite richness/sampling effort/host range: the fancy three-piece jigsaw puzzle. *Parasitology Today*, *12*(9), 367-369.
79. Guégan, J. F., Lambert, A., Lévêque, C., Combes, C., & Euzet, L. (1992). Can host body size explain the parasite species richness in tropical freshwater fishes?. *Oecologia*, *90*(2), 197-204.
80. Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, *8*(9), 1081-1091.
81. Gunnarsson, B. (1996). Bird predation and vegetation structure affecting spruce-living arthropods in a temperate forest. *Journal of Animal Ecology*, 389-397.
82. Gunnarsson, B., Heyman, E., & Vowles, T. (2009). Bird predation effects on bush canopy arthropods in suburban forests. *Forest Ecology and Management*, *257*(2), 619-627.
83. Haddad, N. M., Haarstad, J., & Tilman, D. (2000). The effects of long-term nitrogen loading on grassland insect communities. *Oecologia*, *124*, 73-84.
84. Handler, A. T., Gruner, D. S., Haines, W. P., Lange, M. W., & Kaneshiro, K. Y. (2007). Arthropod surveys on Palmyra Atoll, Line Islands, and insights into the decline of the native tree Pisonia grandis (Nyctaginaceae) 1. *Pacific Science*, *61*(4), 485-502.
85. Hanspach, J., Kühn, I., Pompe, S., & Klotz, S. (2010). Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, *12*(3), 219-225.

86. Harsch, M. A., & HilleRisLambers, J. (2016). Climate warming and seasonal precipitation change interact to limit species distribution shifts across Western North America. *PloS one*, *11*(7), e0159184.
87. Hathaway, S. A., & Fisher, R. N. (2010). *Biosecurity plan for Palmyra Atoll*. US Department of the Interior, Geological Survey.
88. Hathaway, S. A., McEachern, A. K., & Fisher, R. N. (2011). *Terrestrial forest management plan for Palmyra Atoll*. Reston (VA): US Department of the Interior, US Geological Survey.
89. Hechinger, R. F. (2013). A metabolic and body-size scaling framework for parasite within-host abundance, biomass, and energy flux. *The American Naturalist*, *182*(2), 234-248.
90. Helaouët, P., & Beaugrand, G. (2009). Physiology, ecological niches and species distribution. *Ecosystems*, *12*, 1235-1245.
91. Heydel, F., Cunze, S., Bernhardt-Römermann, M., & Tackenberg, O. (2014). Long-distance seed dispersal by wind: disentangling the effects of species traits, vegetation types, vertical turbulence and wind speed. *Ecological Research*, *29*, 641-651.
92. Holl, K. D. (1999). Factors limiting tropical rain forest regeneration in abandoned pasture: Seed rain, seed germination, microclimate, and soil 1. *Biotropica*, *31*(2), 229-242.
93. Hölldobler, B., & Wilson, E. O. (1990). *The ants*. Harvard University Press.
94. Holmes, J. C. and Price, P. W. (1986). Communities of parasites. In Community Ecology: Pattern and Process (ed. Anderson, D. J. and Kikkawa, J.), pp. 187–213. Blackwell Scientific Publications, Oxford.
95. Howald, G., Samaniego, A., Buckelew, S., McClelland, P., Keitt, B., Wegmann, A., ... & Barclay, S. (2004). Palmyra Atoll rat eradication assessment trip report, August 2004. *Report to USFWS. Island Conservation, Santa Cruz, CA*.
96. Hubbell, S. P. (1997). A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, *16*, S9-S21.
97. Hudson, P. J., Dobson, A. P., & Newborn, D. (1992). Do parasites make prey vulnerable to predation? Red grouse and parasites. *Journal of animal ecology*, 681-692.
98. Hudson, P. J., Dobson, A. P., & Newborn, D. (1998). Prevention of population cycles by parasite removal. S*cience*, *282*(5397), 2256-2258.
99. Janssens, F., Peeters, A., Tallowin, J. R. B., Bakker, J. P., Bekker, R. M., Fillat, F., & Oomes, M. J. M. (1998). Relationship between soil chemical factors and grassland diversity. *Plant and Soil*, *202*, 69-78.
100. Jennelle, C. S., Cooch, E. G., Conroy, M. J., & Senar, J. C. (2007). State-specific detection probabilities and disease prevalence. *Ecological Applications*, *17*(1), 154-167.
101. Jennings, S., Pinnegar, J. K., Polunin, N. V., & Boon, T. W. (2001). Weak cross-species relationships between body size and trophic level belie powerful size-based trophic structuring in fish communities. *Journal of Animal Ecology*, 934-944.
102. Kaspari, M., Alonso, L., & O'Donnell, S. (2000). Three energy variables predict ant abundance at a geographical scale. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1442), 485-489.
103. Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*(4), 334-350.

104. Kellner, K. F., & Swihart, R. K. (2014). Accounting for imperfect detection in ecology: a quantitative review. *PloS one*, *9*(10), e111436.
105. Kerans, B.L., Karr, J.R. & Ahlstedt, S.A. (1992). Aquatic invertebrate assemblages: spatial and temporal differences among sampling protocols. *Journal of the North American Benthological Society*, 11, 377–390.
106. Kikillus K.H., Hare K.M., Hartley S. (2009). Minimizing false-negatives when predicting the potential distribution of an invasive species: a bioclimatic envelope for the red-eared slider at global and regional scales. *Anim Conserv,* 12(suppl 1):1–11
107. Krasnov, B. R., Matthee, S., Lareschi, M., Korallo-Vinarskaya, N. P., & Vinarski, M. V. (2010). Co-occurrence of ectoparasites on rodent hosts: Null model analyses of data from three continents. *Oikos*, *119*(1), 120-128.
108. Kulikowski, A. J. (2020). Ant–scale mutualism increases scale infestation, decreases folivory, and disrupts biological control in restored tropical forests. *Biotropica*, *52*(4), 709-716.
109. Kuris, A. M., & Lafferty, K. D. (1994). Community structure: larval trematodes in snail hosts. *Annual Review of Ecology and Systematics*, *25*(1), 189-217.
110. Lafferty, K. D., Dobson, A. P., & Kuris, A. M. (2006). Parasites dominate food web links. *Proceedings of the National Academy of Sciences*, *103*(30), 11211-11216.
111. Lafferty, K. D., Sammond, D. T., & Kuris, A. M. (1994). Analysis of larval trematode communities. *Ecology*, *75*(8), 2275-2285.
112. Lafferty, K. D., Shaw, J. C., & Kuris, A. M. (2008). Reef fishes have higher parasite richness at unfished Palmyra Atoll compared to fished Kiritimati Island. *EcoHealth*, *5*, 338-345.
113. Lagrue, C., Kelly, D. W., Hicks, A., & Poulin, R. (2011). Factors influencing infection patterns of trophically transmitted parasites among a fish community: host diet, host–parasite compatibility or both?. *Journal of Fish Biology*, *79*(2), 466-485.
114. LeBrun, E. G. (2005). Who is the top dog in ant communities? Resources, parasitoids, and multiple competitive hierarchies. *Oecologia*, *142*, 643-652.
115. Levin, S. A. (1992). The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology*, *73*(6), 1943-1967.
116. Lima, M., Firmino, V. C., de Paiva, C. K. S., Juen, L., & Brasil, L. S. (2022). Land use changes disrupt streams and affect the functional feeding groups of aquatic insects in the Amazon. *Journal of Insect Conservation*, *26*(2), 137-148.
117. Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PloS one*, *13*(9), e0204056.
118. Linardi, P. M., & Krasnov, B. R. (2013). Patterns of diversity and abundance of fleas and mites in the Neotropics: host-related, parasite-related and environment-related factors. *Medical and Veterinary Entomology*, *27*(1), 49-58.
119. Locke, S. A., Marcogliese, D. J., & Tellervo Valtonen, E. (2014). Vulnerability and diet breadth predict larval and adult parasite diversity in fish of the Bothnian Bay. *Oecologia*, *174*, 253-262.
120. Locke, S. A., McLaughlin, J. D., & Marcogliese, D. J. (2013). Predicting the similarity of parasite communities in freshwater fishes using the phylogeny, ecology and proximity of hosts. *Oikos*, *122*(1), 73-83.

121. MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, *84*(8), 2200-2207.

122. MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248-2255.

123. MacKenzie, D. I., Nichols, J. D., Sutton, N., Kawanishi, K., & Bailey, L. L. (2005). Improving inferences in population studies of rare species that are detected imperfectly. *Ecology*, *86*(5), 1101-1113.

124. MacArthur, R. H., & Wilson, E. O. (2001). *The theory of island biogeography* (Vol. 1). Princeton university press.

125. Malhi, Y., Baker, T. R., Phillips, O. L., Almeida, S., Alvarez, E., Arroyo, L., ... & Lloyd, J. (2004). The above‐ground coarse wood productivity of 104 Neotropical forest plots. *Global Change Biology*, *10*(5), 563-591.

126. Mao, C. X., & Colwell, R. K. (2005). Estimation of species richness: mixture models, the role of rare species, and inferential challenges. *Ecology*, *86*(5), 1143-1153.

127. Mani, M., & Krishnamoorthy, A. (2002). Classical Biological Control of the Spiralling whitefly, Aleurodicus dispersus Russell—An Appraisal. *International Journal of Tropical Insect Science*, *22*(4), 263-273.

128. Marcogliese, D. J. (2002). Food webs and the transmission of parasites to marine fish. *Parasitology*, *124*(7), 83-99.

129. Marschall, E. A., & Roche, B. M. (1998). Using models to enhance the value of information from observations and experiments. *Experimental ecology: issues and perspectives. Oxford University Press, New York*, 281-297.

130. McCauley, D. J., DeSalles, P. A., Young, H. S., Dunbar, R. B., Dirzo, R., Mills, M. M., & Micheli, F. (2012). From wing to wing: the persistence of long ecological interaction chains in less-disturbed ecosystems. *Scientific Reports*, *2*(1), 409.

131. McClintock, B. T., Nichols, J. D., Bailey, L. L., MacKenzie, D. I., Kendall, W. L., & Franklin, A. B. (2010). Seeking a second opinion: uncertainty in disease ecology. *Ecology Letters*, *13*(6), 659-674.

132. McGarigal, K., Wan, H. Y., Zeller, K. A., Timm, B. C., & Cushman, S. A. (2016). Multi-scale habitat selection modeling: a review and outlook. *Landscape Ecology*, *31*, 1161-1175.

133. McGill, B. J. (2010). Matters of scale. *Science*, *328*(5978), 575-576.

134. McLaughlin, J.P. (2018). The food web for the sand flats at Palmyra Atoll. [Doctoral dissertation, University of California, Santa Barbara]. ProQuest Dissertations and Theses database.

135. McLaughlin, J. P., Miller-ter Kuile, A., Bui, A., Klope, M., Lee, M., Bogar, T., … & Young, H.. Submitted. The food web for the terrestrial habitats of Palmyra Atoll. *Nature Scientific Data*.

136. Miller, D. A., Bailey, L. L., Grant, E. H. C., McClintock, B. T., Weir, L. A., & Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, *6*(5), 557-565.

137. Mod, H. K., Chevalier, M., Luoto, M., & Guisan, A. (2020). Scale dependence of ecological assembly rules: Insights from empirical datasets and joint species distribution modelling. *Journal of Ecology*, *108*(5), 1967-1977.

138. Moilanen, A. (2002). Implications of empirical data quality to metapopulation model parameter estimation and application. *Oikos*, *96*(3), 516-530.

139. Moran, M. D., & Hurd, L. E. (1997). A trophic cascade in a diverse arthropod community caused by a generalist arthropod predator. *Oecologia*, *113*, 126-132.

140. Morand, S., & Poulin, R. (1998). Density, body mass and parasite species richness of terrestrial mammals. *Evolutionary Ecology*, *12*(6), 717-727.

141. Mwita, C., & Nkwengulila, G. (2008). Determinants of the parasite community of clariid fishes from Lake Victoria, Tanzania. *Journal of Helminthology*, *82*(1), 7-16.

142. Nachman, G., & Borregaard, M. K. (2010). From complex spatial dynamics to simple Markov chain models: do predators and prey leave footprints?. *Ecography*, *33*(1), 137-147.

143. Nichols, J. D., & Karanth, K. U. (2002). Statistical concepts; assessing spatial distribution. Pages 29–38 in K. U. Karanth and J. D. Nichols, editors. Monitoring tigers and their prey. Centre for Wildlife Studies, Bangalore, India.

144. Nishida, G. M. (2002). *Hawaiian terrestrial arthropod checklist 4th edition*. Bishop Museum Technical Report. Hawaii Biological Survey Bishop Museum.

145. Ostermiller, J.D. & Hawkins, C.P. (2004) Effects of sampling error on bioassessment of stream ecosystems: application to RIVPACS-type models. *Journal of the North American Benthological Society*, 23, 363–382.

146. Ovaskainen, O., & Abrego, N. (2020). *Joint species distribution modelling: With applications in R*. Cambridge University Press.

147. Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*(9), 2514-2521.

148. Palm, H. W., & Caira, J. N. (2008). Host specificity of adult versus larval cestodes of the elasmobranch tapeworm order Trypanorhyncha. *International Journal for Parasitology*, *38*(3-4), 381-388.

149. Passera, L. (2021). Characteristics of tramp species. *Exotic Ants*, 23-43.

150. Pelosi, C., Goulard, M., & Balent, G. (2010). The spatial scale mismatch between ecological processes and agricultural management: Do difficulties come from underlying theoretical frameworks?. *Agriculture, Ecosystems & Environment*, *139*(4), 455-462.

151. Pence, D. B. (1990). Helminth community of mammalian hosts: concepts at the infracommunity, component and compound community levels. In *Parasite Communities: patterns and processes* (pp. 233-260). Dordrecht: Springer Netherlands.

152. Pimentel, D., McNair, S., Janecka, J., Wightman, J., Simmonds, C., O'Connel, C., Wong, E., Russel, L., Zean, J., Aquino, T. & Tsomondo, T. (2002). Economic and environmental threats of alien plant, animal, and microbe invasions. In Biological invasions: 307–329. Pimentel, D. (Ed.). Boca Raton, FL, USA: CRC Press.

153. Poulin, R. (1995). Phylogeny, Ecology, and the Richness of Parasite Communities in Vertebrates: Ecological Archives M065-001. *Ecological Monographs*, *65*(3), 283-302.

154. Poulin, R. (1996). Richness, nestedness, and randomness in parasite infracommunity structure. *Oecologia*, *105*(4), 545-551.

155. Poulin, R. (1997). Species richness of parasite assemblages: evolution and patterns. *Annual Review of Ecology and Systematics*, *28*(1), 341-358.
156. Poulin, R. (1998). Comparison of three estimators of species richness in parasite component communities. *The Journal of Parasitology*, 485-490.
157. Poulin, R. (2003). The decay of similarity with geographical distance in parasite communities of vertebrate hosts. *Journal of biogeography*, *30*(10), 1609-1615.
158. Poulin, R. (2010). Decay of similarity with host phylogenetic distance in parasite faunas. *Parasitology*, *137*(4), 733-741.
159. Poulin, R., & FitzGerald, G. J. (1989). Shoaling as an anti-ectoparasite mechanism in juvenile sticklebacks (Gasterosteus spp.). *Behavioral Ecology and Sociobiology*, *24*(4), 251-255.
160. Poulin, R., & George-Nascimento, M. (2007). The scaling of total parasite biomass with host body mass. *International Journal for Parasitology*, *37*(3-4), 359-364.
161. Poulin, R., & Leung, T. L. F. (2011). Body size, trophic level, and the use of fish as transmission routes by parasites. *Oecologia*, *166*, 731-738.
162. Poulin, R., & Morand, S. (1999). Geographical distances and the similarity among parasite communities of conspecific host populations. *Parasitology*, *119*(4), 369-374.
163. Poulin, R., & Rohde, K. (1997). Comparing the richness of metazoan ectoparasite communities of marine fishes: controlling for host phylogeny. *Oecologia*, *110*(2), 278-283.
164. Poulin, R., & Valtonen, E. T. (2001). Nested assemblages resulting from host size variation: the case of endoparasite communities in fish hosts. *International Journal for Parasitology*, *31*(11), 1194-1204.
165. Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology,* 29:254–283.
166. Ranta, E. (1992). Gregariousness versus solitude: another look at parasite faunal richness in Canadian freshwater fishes. *Oecologia*, *89*(1), 150-152.
167. Rasmussen, T. K., & Randhawa, H. S. (2018). Host diet influences parasite diversity: a case study looking at tapeworm diversity among sharks. *Marine Ecology Progress Series*, *605*, 1-16.
168. Ribas, C. R., & Schoereder, J. H. (2002). Are all ant mosaics caused by competition?. *Oecologia*, *131*, 606-611.
169. Ritchie, M. E. (2000). Nitrogen limitation and trophic vs. abiotic influences on insect herbivores in a temperate grassland. *Ecology*, *81*(6), 1601-1612.
170. Rohde, K. (1998). Is there a fixed number of niches for endoparasites of fish?. *International Journal for Parasitology*, *28*(12), 1861-1865.
171. Ross, D. S., & Ketterings, Q. (1995). Recommended methods for determining soil cation exchange capacity. *Recommended soil testing procedures for the northeastern United States*, *493*(101), 62.
172. Rota, C. T., Fletcher Jr, R. J., Evans, J. M., & Hutto, R. L. (2011). Does accounting for imperfect detection improve species distribution models?. *Ecography*, *34*(4), 659-670.
173. Russell, R., Wood, S. A., Allison, G., & Menge, B. A. (2006). Scale, environment, and trophic status: the context dependency of community saturation in rocky intertidal communities. *The American Naturalist*, *167*(6), E158-E170.
174. Salgado-Maldonado, G., Novelo-Turcotte, M. T., Caspeta-Mandujano, J. M., Vazquez-Hurtado, G., Quiroz-Martínez, B., Mercado-Silva, N., & Favila, M. (2016). Host

specificity and the structure of helminth parasite communities of fishes in a Neotropical river in Mexico. *Parasite*, *23*.

175. Sanders, N. J., Crutsinger, G. M., Dunn, R. R., Majer, J. D., & Delabie, J. H. (2007). An ant mosaic revisited: Dominant ant species disassemble arboreal ant communities but co-occur randomly. *Biotropica*, *39*(3), 422-427.
176. Schaffers, A. P., Raemakers, I. P., Sýkora, K. V., & Ter Braak, C. J. (2008). Arthropod assemblages are best predicted by plant species composition. *Ecology*, *89*(3), 782-794.
177. Scott, M. E. (1987). Regulation of mouse colony abundance by Heligmosomoides polygyrus. *Parasitology*, *95*(1), 111-124.
178. Sikder, I., Mal-Sarkar, S. & Mal, T. (2006). Knowledge-based risk assessment under uncertainty for species invasion. *Risk Anal*. 26, 239–252.
179. Sorte, C. J., Ibáñez, I., Blumenthal, D. M., Molinari, N. A., Miller, L. P., Grosholz, E. D., ... & Dukes, J. S. (2013). Poised to prosper? A cross-system comparison of climate change effects on native and non-native species performance. *Ecology Letters*, *16*(2), 261-270.
180. Stadler, B., & Dixon, A. F. (2005). Ecology and evolution of aphid-ant interactions. *Annu. Rev. Ecol. Evol. Syst.*, *36*, 345-372.
181. Stevenson, C., Katz, L. S., Micheli, F., Block, B., Heiman, K. W., Perle, C., Weng, K., Dunbar, D., & Witting, J. (2007). High apex predator biomass on remote Pacific islands. *Coral Reefs*, *26*, 47-51.
182. Stuber, E. F., & Gruber, L. F. (2020). Recent methodological solutions to identifying scales of effect in multi-scale modeling. *Current Landscape Ecology Reports*, *5*, 127-139.
183. Syphard, A. D., & Franklin, J. (2009). Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, *32*(6), 907-918.
184. Takemoto, R. M., Pavanelli, G. C., Lizama, M. A. P., Luque, J. L., & Poulin, R. (2005). Host population density as the major determinant of endoparasite species richness in floodplain fishes of the upper Parana River, Brazil. *Journal of Helminthology*, *79*(1), 75-84.
185. Thomas, C. D. (2010). Climate, climate change and range boundaries. *Diversity and Distributions*, *16*(3), 488-495.
186. Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology*, *75*(1), 2-16.
187. Tilman, D., Wedin, D., & Knops, J. (1996). Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature*, *379*(6567), 718-720.
188. Timi, J. T., & Lanfranchi, A. L. (2009). The importance of the compound community on the parasite infracommunity structure in a small benthic fish. *Parasitology Research*, *104*(2), 295-302.
189. Timi, J. T., Luque, J. L., & Poulin, R. (2010). Host ontogeny and the temporal decay of similarity in parasite communities of marine fish. *International Journal for Parasitology*, *40*(8), 963-968.
190. Timi, J. T., Rossin, M. A., Alarcos, A. J., Braicovich, P. E., Cantatore, D. M. P., & Lanfranchi, A. L. (2011). Fish trophic level and the similarity of non-specific larval parasite assemblages. *International Journal for Parasitology*, *41*(3-4), 309-316.
191. Tremlová, K., & Münzbergová, Z. (2007). Importance of species traits for species distribution in fragmented landscapes. *Ecology*, *88*(4), 965-977.

192. Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, *13*(6), 1790-1801.
193. Valtonen, E. T., Pulkkinen, K., Poulin, R., & Julkunen, M. (2001). The structure of parasite component communities in brackish water fishes of the northeastern Baltic Sea. *Parasitology*, *122*(4), 471-481.
194. van Klink, R., van der Plas, F., Van Noordwijk, C. G. E., WallisDeVries, M. F., & Olff, H. (2015). Effects of large herbivores on grassland arthropod diversity. *Biological Reviews*, *90*(2), 347-366.
195. Veech, J. A. (2006). A probability-based analysis of temporal and spatial co-occurrence in grassland birds. *Journal of Biogeography*, *33*(12), 2145-2153.
196. Viana, D. S., & Chase, J. M. (2019). Spatial scale modulates the inference of metacommunity assembly processes. *Ecology*, *100*(2), e02576.
197. Vignon, M., & Sasal, P. (2010). Multiscale determinants of parasite abundance: a quantitative hierarchical approach for coral reef fishes. *International Journal for Parasitology*, *40*(4), 443-451.
198. Warren, M., Robertson, M. P., & Greeff, J. M. (2010). A comparative approach to understanding factors limiting abundance patterns and distributions in a fig tree–fig wasp mutualism. *Ecography*, *33*(1), 148-158.
199. Webster, R.G., Peiris, M., Chen, H. & Guan, Y. (2006). H5N1 outbreaks and enzootic influenza. *Emerg. Infect. Dis*., 12, 3–8.
200. Wegmann, A. (2005). Palmyra Atoll National Wildlife Refuge forest type map: Honolulu, Hawai`i, U.S. Fish and Wildlife Service.
201. Wegmann, A., Flint, E., White, S., Fox, M., Howald, G., McClelland, P., ... & Griffiths, R. (2012). Pushing the envelope in paradise: a novel approach to rat eradication at Palmyra Atoll. In *Proceedings of the Vertebrate Pest Conference* (Vol. 25, No. 25).
202. Wester, L. (1985). Checklist of the vascular plants of the northern Line Islands. Atoll Res. Bull. 287:1–38.
203. Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography: ecology, evolution, and conservation*. Oxford University Press.
204. Whittaker, R. J., Willis, K. J., & Field, R. (2001). Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography*, *28*(4), 453-470.
205. Williams, B. K., J. D. Nichols, and M. J. Conroy. (2002). *Analysis and management of animal populations.* Academic Press, San Diego, California, USA.
206. Williams, S. E., Marsh, H., & Winter, J. (2002). Spatial scale, species diversity, and habitat structure: small mammals in Australian tropical rain forest. *Ecology*, *83*(5), 1317-1329.
207. Wimp, G.M., Lewis, D. & Murphy, S.M. (2019). Impacts of Nutrient Subsidies on Salt Marsh Arthropod Food Webs: A Latitudinal Survey. *Front. Ecol. Evol.* 7:350.
208. Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., ... & Svenning, J. C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, *88*(1), 15-30.
209. Wright, D. H. (1983). Species-energy theory: an extension of species-area theory. *Oikos*, 496-506.

210. Young, H. S., McCauley, D. J., Dunbar, R. B., & Dirzo, R. (2010). Plants cause ecosystem nutrient depletion via the interruption of bird-derived spatial subsidies. *Proceedings of the National Academy of Sciences*, *107*(5), 2072-2077.

211. Young, H. S., Raab, T. K., McCauley, D. J., Briggs, A. A., & Dirzo, R. (2010). The coconut palm, Cocos nucifera, impacts forest composition and soil characteristics at Palmyra Atoll, Central Pacific. *Journal of Vegetation Science*, *21*(6), 1058-1068.