Accounting for detection bias increases trait-matching in a tropical plant-hummingbird network

Manuscript prepared for “Reports” section of Ecology

Ben G. Weinstein, corresponding author, email: benjamin.weinstein@stonybrook.edu

Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794; USA

Catherine H. Graham, email: Catherine.Graham@stonybrook.edu

Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794; USA

# Abstract

# Introduction

Network ecology is a rapidly developing field that has reinvorgorated research on assemblage stability and complexity. By harnessing graph theory to represent ecological interactions, networks are used to study the evolution of specialization (Armbruster), the importance of complex interactions (Holt) and the vulnerability of mutualisms to climate change (Jordano). Network ecology has matured from a largely theoretical exploration of community stability (May) to an emerging tool to study ecological systems (Waser). However, to validate the mechanisms shaping network structure (Allesina, Vasquez), we must confront new challenges in connecting theory to data (Bluthgen). A central obstacle is our ability to detect species interactions is imperfect in time and space. This is a shared challenge across all types of ecological networks, and can bias results by assuming that non-detection is synonymous with non-interaction. In addition, field data are naturally nested with uneven observations among interacting partners. By embracing hierarchical models, we can account for these challenges, and introduce additional biological realism into the study of ecological networks.

The importance of detection bias is widely appreciated in wildlife sciences (Mackenzie, Royle), but only rarely incoporated into evolutionary and theoretical ecology (xy, but see Bartomeluea). However, the same issues that limit inference in determining population dynamics (), site selection (), and breeding ecology (), also undermine our ability to make inference about the ecology and evolution of ecological systems. The main improvement is to incorporate detection bias directly into our estimates of species interactions using occupancy models. In occupancy modeling, repeated surveys are used to estimate the probability of detection given a species probability of occurrence. The probability of occurrence is a latent variable which cannot be directly infered from the data. If we did not see a species at a site, was it absent or undetected? By analogy, if we failed to find a pollinator interacting with a plant, is it due to detection or is it a 'forbidden link' due mismatch in species timing or functional morphology? The goal of this paper is to use a large dataset on plant-hummingbird interactions from a tropical montane forest to evaluate the effect of imperfect detection on estimating species interactions.

One approach to predicting the frequency of species interactions is to use morphological traits as a proxy for functional niches, and measure the degree of trait-matching among partners (Vasquez). In tropical hummingbirds, species with longer and more curved bills tend to be more specialized and feed on plants with longer and more curved corollas (Temeles, Magliansi). To test the strength of trait-matching between bird bill and flower corolla lengths, studies use the number of visits to a flower as interaction strength among birds and plants (Bugonsi, Petroyama, Magliansi, Sazima). We begin with simulations to demonstrate the importance of accounting for detection when estimating interaction frequency and trait-matching. Before fitting models to an observed dataset, it is crucial to simulate data with known results to show that the models perform as expected. In this analysis, we will use a Bayesian hierarchical approach (O’Hara, Kery). We are ultimately interested in estimating three quantities: the number of interactions between a hummingbird species and plant species, the probability of detection, and the frequency (or 'intensity') of the interaction. To do this, we divide the model into three components: an observation model, a process model, and a hierarcichal set of priors.

## Observation Model

The observation model describes the sampling process that yields our data (Elison). These are the actual number of observations detected by reviewing the cameras and performing transects. We know there is some true, but unknown, intensity of interactions among hummingbird i and plant j, and that we are likely to see some, but not all, interactions. We assume that we are equally likely to see any one interaction, so the number of detections of hummingbird i on plant j in month k is modeled as as a binomial trial where the probability of detecting a species is a species-specific parameter *detect* with *N* trials (Royle).

## Process Model

To estimate the parameter *N*, which is the unknown true rate of interactions, we need a process model. The process model is the mechansism that generates interactions among plants and pollinators in a given time period. In our analysis, we choose a sampling period of one day to match the camera data with the transect data. This means that the units of our interaction is a daily intensity, modeled as a Poisson distribution with . The hypothesis is that should increase with greater matching between plant corolla and hummingbird bill lengths. Since lambda is a intensity of interactions, it must be a positive number. We therefore model as a function of linear predictors with a log link function. In our simulation, we give equal penalty for having a bill too long, or too short for a given corolla. The process model can be written as:

## Hierarchical structure

Each observed interaction () is not independent. Due to foraging behavior (Sandlin), physiological requirements (Powers), or difference in available resources (Weinstein), species may respond individualistically to differences in bill or corolla lengths. However, the aim in this study is not to evaluate any one species, but to capture the overall effect of trait-matching on hummingbirds and plants. We therefore use a hierarchical model to estimate a group-level mean and account for within species variability (Bolker). Similiarly, due to differences in abundance, behavior, and foraging strategy (Colwell), some hummingbird species will be more detectable than others, we therefore allow a fixed species level effect of hummingbird species for detection probability. Additional information on species abundance, observation distance, and vegetation structure could be used to estimate species detection probabilities (Royle, Tingley). By combining an observation model, process model in a hierarchical structure, this approach can accounting for known obstacles in interpreting field data. To illustrate the added information based on this setup, I compare my results to a direct poisson generalized linear model that does not account for detection or non-independence of observations.

# Methods

## Simulation

In the simulated network, there are 8 hummingbird species with a range of bill sizes drawn from a Poisson (lambda=1.5). Each of these species can interact with twenty plant species with a range of corolla sizes drawn from a Poisson (lambda=2). Each hummingbird has a different detection probability, which is fixed for each species. The frequency of interactions between hummingbird and plant species is based on the similarity in corolla lengths. This covariate is hierarchically structured with species level random effects on both slope and intercept. The overall trait -matching (minimizing Bill-Corolla difference) relationship is drawn from a normal distribution among all hummingbird species. There are 24 survey replicates for each plant. For added realism, there is unequal sampling, where each plant species is in flower in 50% of survey periods.

## Observed Data

### Data Collection

Data were collected at the Maquipucuna Research Station and Santa Lucia Ecolodges, Ecuador (0.118 S,-78.612 W) between June 2013 and July 2015 along an elevation gradient from 1300 m to 2500 m. The area contains regenerating secondary and primary cloud-forest, and has a warmer dry season from June to September (Webster and Rhodes 2005). We used time-lapse cameras to monitor hummingbird visitation. Flowers in bloom were noted during floral transects, and cameras were placed 1-3 meters from focal flowers. Cameras turned on automatically at dawn and recorded an image every second for at least one day. We used the computer vision program MotionMeerkat to identify images where hummingbirds were observed (Weinstein 2014). In addition, more traditional hummingbird transects were used to survey flower visitation. Observers walked six pre-established transects at a steady pace and recorded all hummingbird-plant interactions within 5m from the trail. Each of the six transect covered ~200m of elevation gain (1300-2500m). Transects were repeated twice a month for 24 months. Combining these sampling types, we have a large network with over 1500 interactions (Weinstein XXXX). By embracing automated monitoring of plant-animal interactions, we can begin to determine and account for detection bias. The approach outlined here is general for any sampling strategy where observers can return to the same plant within a sampling window.

## Modeling Approach

For both the simulated and observed data, I compared Poisson GLMs and hierarchical occupancy models. The Poisson GLM does not account for detection bias or the non-independence of each observation. In contrast, the hierarchical occupancy model assumes imperfect detection and considers trait-matching covariates with a species-level random effect.

## Estimating Occupancy

The number of observations observed between and in is a binomial trial where the probability of detecting a species is a species specific parameter *detect* with N trials. N is the underlying frequency of interactions among each pair of plants and hummingbirds drawn from a Poisson . The detection probability is fixed for each species. Species level priors for the strength of trait-matching is drawn from a normal distribution of all hummingbird species. The full model can be written as,

**Priors**

**Hyperpriors**

**Derived quantities**

### Non-detections

One assumption of an occupancy model is that the underlying state of the interaction does not change throughout the sampling window. Species interactions need to remain consistent to be estimated using repeat visits. However, due to flowering phenology, some species may be present, but unavailable for interaction. Without correcting for this bias, it is possible to inflate the number of non-detections by erroneously filling all non-detections as zeros. To avoid this, We accounted for overestimating non-detection in two ways. First, if no hummingbird was seen on a plant in a sampling period, the data were considered missing, rather than non-detected. This accounts for plant species marked as present, but not in flower. Secondly, if a plant flowered outside the elevation of a species, it was missing rather than a non-detection. These are conservative assumptions that can be relaxed in more complex model formulations (Royle).

## Posterior Checks

To measure the relative fit of each model, We performed posterior goodness of fit checks. This approach allows a standardized way of comparing models of different types (Gelman). We choose chi-squared tests as a fit statistic calculated as (Expected-Observed)/ (Expected^2 + 0.5). For the occupancy model, the expected value for each month is the detection rate \* the estimated intensity of interactions. Given that this is a stochastic process, there will always be underlying variability in the intensity of interactions. To account for this, we created a replicate dataset generated from the posterior predicted intensity. In this setup, better fitting models have smaller discrepancy values and are closer to the 1:1 line. A perfect model would be equal discrepancy for both the observed and replicated dataset at (0,0). However, given the stochasticity of the true process, this is unrealistic expectation.

# Results

* We were able to recover the known values using the 95th credible intervals for all simulation parameters (Figure 1).
* The estimated relationship using the occupancy model closely matches the simulated relationship between trait-matching and interaction frequency (Figure 2)
* Without accounting for detection bias and non-independence, the estimated relationship is much weaker than the true value (Figure 2).
* In the observed dataset, the detection probability, even for the most common species is less than 20% (Table 1). Even if the interaction is known to occur, we need atleast five days of sampling to ensure we would capture it.
* Accounting for detection bias strongly increases the predictive importance of trait matching (Figure 3). Without correcting for detection or non-independence of observations, the predicted relationship would be minimal.
* The occupancy model minimizes the residuals better in both the simulated and observed datasets (Figure 4). This suggest a stronger fit. In the simulated dataset, the residuals were identical for the ‘observed’ data, used to fit the model, as compared to a replicate dataset. This means that we captured the dataset as well as possible given the underlying stochasticity of the process. To visualize the data, we log transformed the values to view on similar axes.

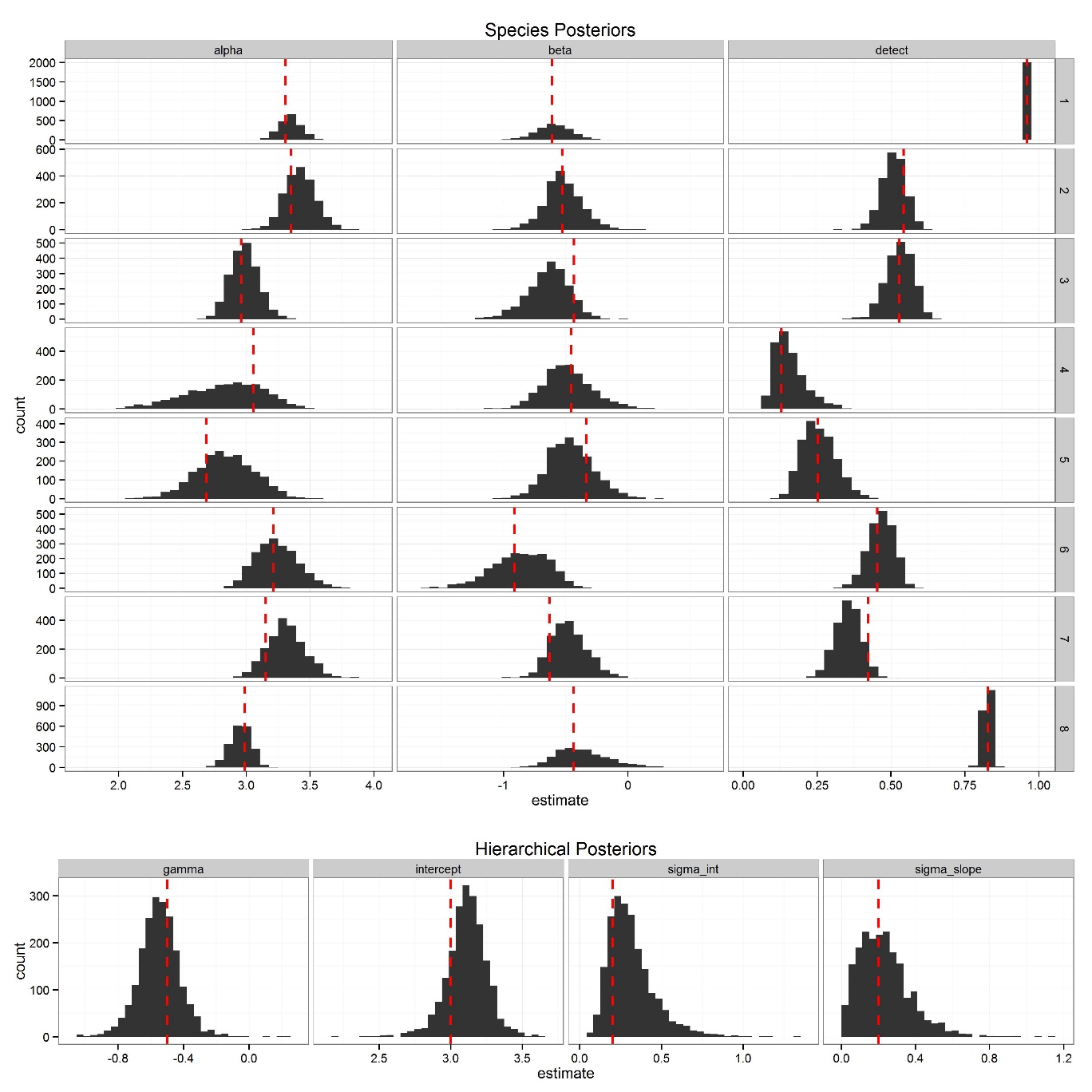
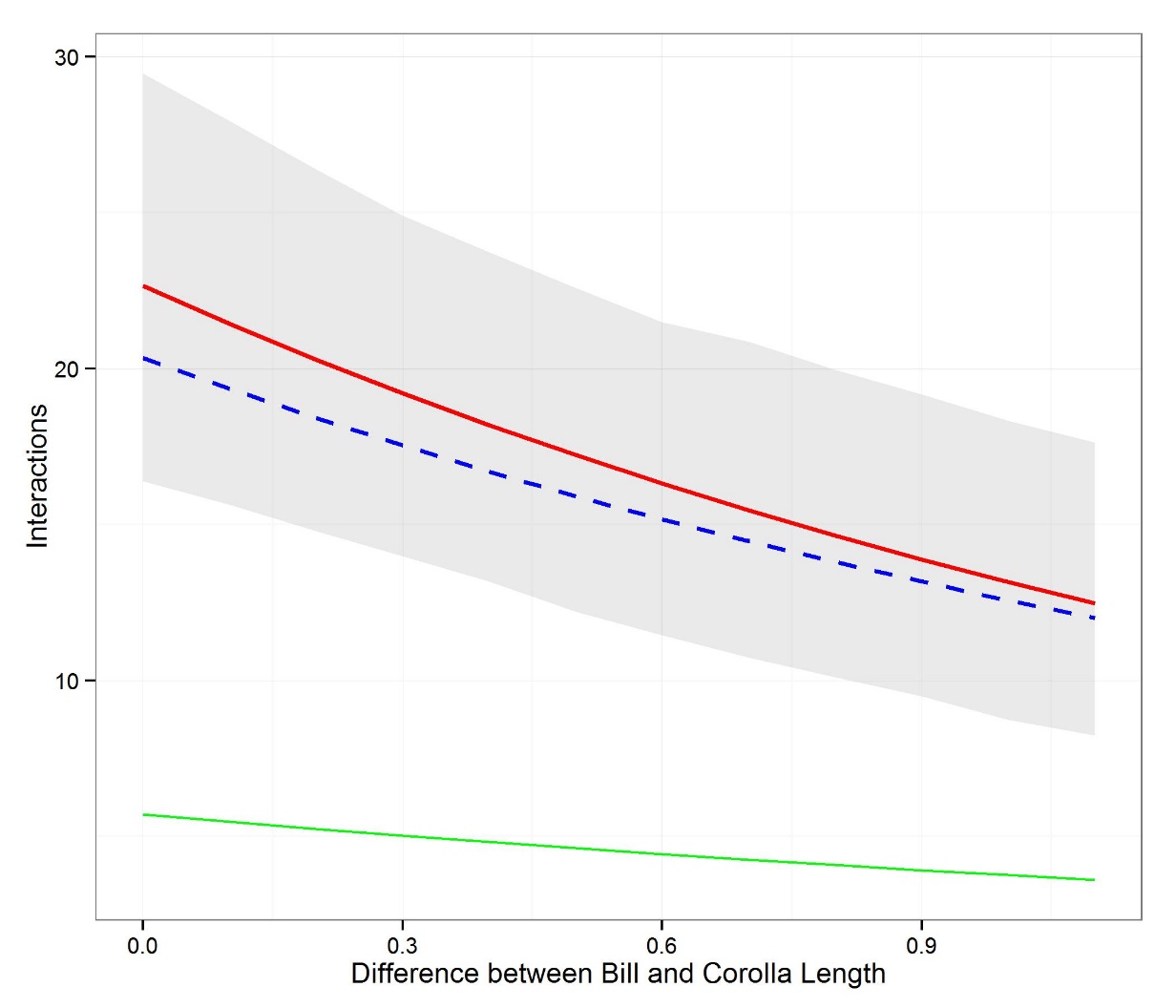


Figure 1. Species level (top) and hierarchical (bottom) posterior distributions for the simulated dataset. The red dashed lines are the true value used to create the parameter in the simulation. Alpha is the species level intercept, Beta is the species level trait-matching coefficient and detect is the detection probability for each species. For each species alpha (top) was drawn from the hierarchical distribution Intercept (bottom) with known variance sigma\_int. For each species Beta (top) was drawn from the hierarchical distribution gamma (bottom) with known variance sigma\_slope.

 Figure 2. Predicted relationship between trait-matching and interaction strength between simulated plants and hummingbirds (top). The red line is the posterior mean with 95th credible intervals in shaded grey for the occupancy model. The green line is the Poisson GLM that does not account for detectability or non-independence among observations.

# 

Figure 3. Predicted relationship for the field data between trait-matching (in centimeters) and interaction frequency using the hierarchical occupancy model (mean value in red, 95th credible interval in gray) and the Poisson generalized linear model (blue).

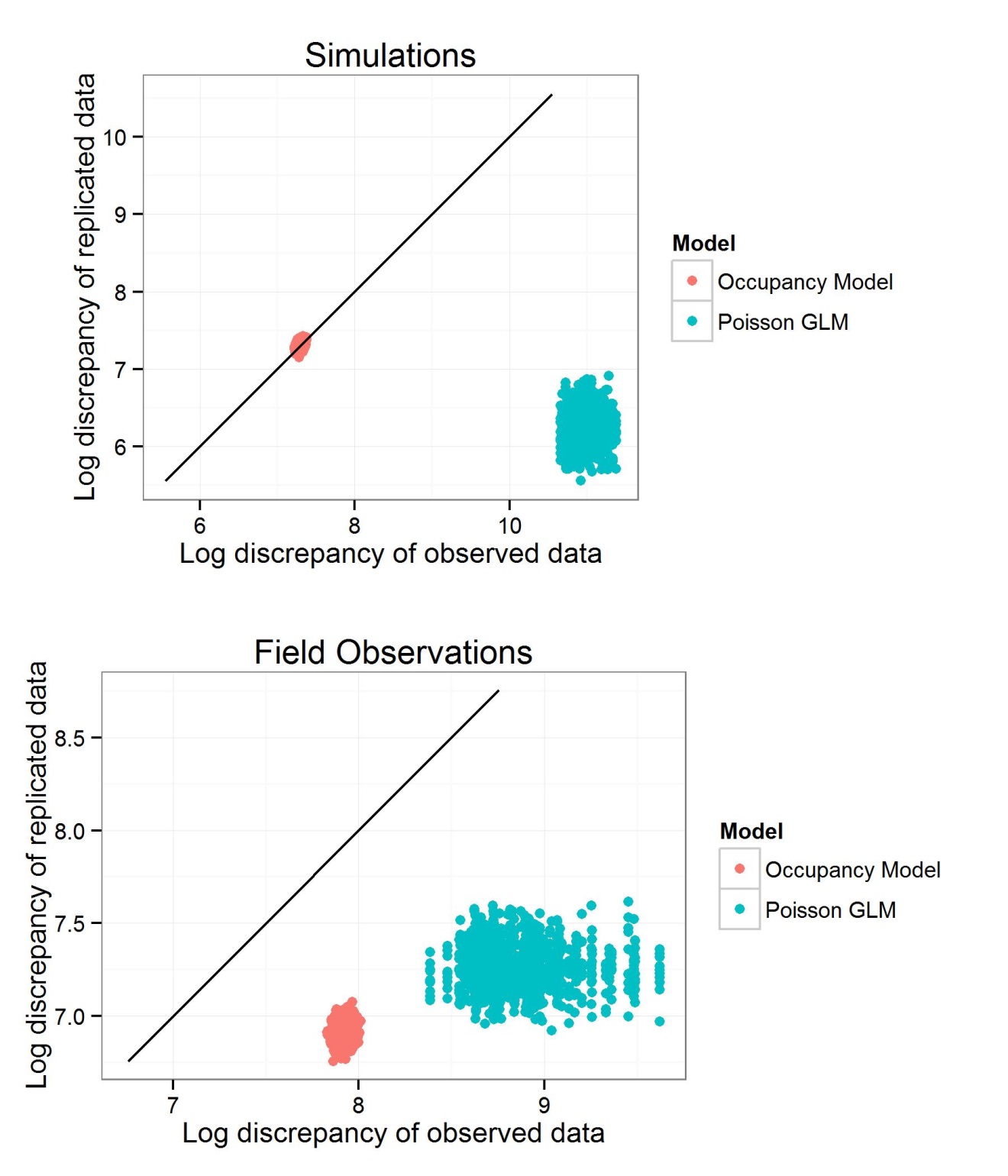


Figure 4. Posterior predicted distributions for the simulated (top) and field data (bottom) using log chi-squared residuals as a measure of discrepancy. The test statistic for the observed data versus a replicated dataset is shown with a 1:1 line. Stronger fitting models will fall closer to the 1:1 line.

# Discussion

* Species are difficult to observe in the natural world.
  + By allowing for observation error, we can better estimate the underlying mechanisms shaping ecological communities. In this paper, I compared Poisson generalized linear models with a hierarchical occupancy model that accounted for detection bias and non-independence among observations. Both the estimate of trait-matching, as well as the model fit, increased using the occupancy approach. Estimating interaction frequency without accounting for detection bias or non-independence may underestimate the important of species traits in determining network interactions.
* By using the raw number of interactions among plants and pollinators, we assume that each species sampled evenly, and that each interaction was equally detectable.
  + The automated cameras used to moniter the majority of these interactions turn on and off at dawn and dusk and provide a much longer duration of observation than is possible with a human observer (). Nevertheless, hummingbird-plant interactions were detected less than 20% of the sampling periods. Even if the interaction is on average to occur we need atleast 5 days of sampling to ensure we would capture it. This suggest that the problem of detection bias would be even more severe when the sampling window is even shorter.
* Knowing about imperfect detection helps us design network studies
  + Moving network ecology forward to provide greater realism
* The simulations demonstrate the flexibility of occupancy models to capture complex processes. I was able to recover the known values using the 95th credible intervals for all simulation parameters. The occupancy models presented here can be extended to include time-varying components (Royle), and covariates. For example, detection may decrease as plant height increases, since observations may be harder to confirm at greater distances.
* By adopting an occupancy modeling approach, there are a myriad of new avenues to increase the realism and complexity of estimating species interactions. There is increasing interest in the temporal processes shaping network properties.
  + By using dynamic occupancy models, it is possible to explicitly model interactions as a function of time and space. **More here.**
* Detractors of occupancy models
  + Cannot account for singletons. Excluding single observations, an occupancy approach bolsters rare data because it makes a more informed use of the non-detections and differentiates species that are genuinely rare from species that are difficult to detect.
  + Estimating sampling periods
* Hierarchical models not only help create more robust statistical inference, but give us new tools in teasing out the multiple effects structuring networks.
  + The role of abundance in creating observed network patterns remains a central area of research. Abundance can be seen both as a null process shaping interaction strength and species vulnerability (Winfree) or the underlying mechanisms. Occupancy models provide a third alternative – more abundant species are more likely to be sampled and detected.
  + By accounting for this bias, abundance can be evaluated both as a predictor of detection, as well as the process model shaping interactions. Occupancy models are closely related to other approaches dealing with zero-inflated data (Zipkin), where the occurrence probability and interaction frequency are modeled separately. For networks, this approach was proposed by O’Hara et al (2013,) and although they do no explicitly include detection bias – it could be included if individuals are being consistently tracked. For example, In pollination ecology, the emergence the use of RFID tags (Hadley, ) could allow a wealth of information on the temporal and spatial pattern of resource selection that could tease apart the relative factors such as abundance and morphology in shaping network structure.

**Tables**

Table 1. Detection probability for each hummingbird species based on the hierarchical occupancy model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | Observations | Mean Detection  Probability | 2.5% Quantile Detection  Probability | 97.5% Quantile  Probability |
| Andean Emerald | 22 | 4 | 1.4 | 7.3 |
| Speckled Hummingbird | 68 | 1.9 | 0.6 | 3.9 |
| Stripe-throated Hermit | 88 | 9.2 | 5.9 | 12.7 |
| Tawny-bellied Hermit | 221 | 12.4 | 8.7 | 16.1 |
| Violet-tailed Sylph | 186 | 11 | 7.7 | 14.3 |
| White-whiskered Hermit | 118 | 9.3 | 5.9 | 12.9 |
| Booted Racket-tail | 109 | 7.8 | 5 | 10.7 |
| Brown Inca | 167 | 7.4 | 4.7 | 10.3 |
| Buff-tailed Coronet | 38 | 8 | 2.6 | 15 |
| Collared Inca | 47 | 9.2 | 4.4 | 14.7 |
| Gorgeted Sunangel | 89 | 7.3 | 3.8 | 11.7 |
| Green-crowned Woodnymph | 32 | 5.7 | 1.9 | 11.2 |
| Green-fronted Lancebill | 23 | 11.6 | 5.4 | 18.6 |
| Rufous -tailed Hummingbird | 71 | 3.6 | 1.2 | 7.5 |

**Works Cited**