Globi Database:

**Running title (< 45 characters):** XXX

Authors: Graziella V. DiRenzo1, Michelle Lee2, XXX….., & Katja Seltmann

1 U.S. Geological Survey, Massachusetts Cooperative Fish and Wildlife Research Unit, University of Massachusetts Amherst, USA

2 ….

\* Corresponding author; email xxxx

**Target journal:** Methods in Ecology & Evolution? Ecology Letters (depends on the results)?

Type of article: Standard article

Number of words (6000 - 7000 words):

Number of words abstract (max = 350 words):

Number of figures & tables: X figs. & Y table

**Competing interest statement**

We have no competing interests.

*This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.*

*Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.*

**Abstract**

1. Species

**Keywords:** imperfect

**Introduction**

* Occupancy models are powerful tools that take into account imperfect detection
  + Require standardized data collection and meeting specific assumptions
  + If violated- grave consequences on parameter estimation and ecological inference
  + However, researchers have come up with ways of using occupancy models for non-standardized methods typically used for opportunistic collections
  + For example, to use historical data, authors directly re-survey sites that were surveyed in historical periods or constrained analyses to locations where 2 or more sampling events occurred within a calendar year
  + Replication ensures that occupancy and detection can be estimated
* Opportunistic community science sampling, atlas data, historical museum records contain presence-only data, species non-detections must be inferred
  + One way to do this infer a non-detection for a particular species if a different species was observed at that same site on the same date
  + Issues that may arise in using community science with occupancy models:
    - 1. Sampling (taxonomic) bias
      * Particular species may be sampled more frequently than others because more is known about them or inference is desired on that species
      * Apidae is over-represented with many more records for this family than others
    - 2. Detection bias
      * Species detectability changes over time and space as a result of observers or number of surveys
      * Number of observers, quality of observers, length of survey, survey conditions
      * Solution: add covariates
      * Solution: filter the data by removing vistis with only a single observation
      * Solution: constrain analyses to species that meet some minimum number of observations
    - 3. Spatial bias
      * Particular locations are more easily visited
      * Some areas are never visited
    - 4. How to define a site?
      * Spatial resolution?
    - 5. A museum record is only a subset of a field collection
      * When and where was the specimen collected?
      * Sometimes common species are not curated
  + Scientists have been trying to use these data sets to learn about species distributions across space and time.
    - Describe population declines, etc.
    - Range shifts + contractions
    - This requires a lot of sampling
    - But these datasets can be used to learn other biologically relevant information.
  + Here, we are interested in exploring species interactions (bee-plant) and what kind of bias there is in documenting these relationships.
    - * What are the next steps for these aggregated databases?
      * What holes are there left in our biological understanding?
      * Which bees do we not have enough data for?
        + Only have data for large and common bees?
        + Taxonomic biases?
      * Looking at the number of collectors in the data- looking to see what they predominately collected
* Objectives
  + Objective: We explored hypotheses related to bee and flower characteristics that relate to the ecological process (i.e., what characteristics relate to a species degree of generalism?) and the detection process (i.e., what characteristics relate to species detectability?). Specifically, we explored the following hypotheses:
    - Ecological hypotheses
      * Probability a bee-plant interact depends on bee size
      * Probability a bee-plant interact depends on bee sociality
      * Probability a bee-plant interact depends on flower color
      * Probability a bee-plant interact depends on flower shape
    - Detection hypotheses
      * Probability a bee-plant interaction is detected depends on bee strippiness
      * Probability a bee-plant interaction is detected depends on bee size
      * Probability a bee-plant interaction is detected depends on month of observation (quadratic term)
      * Probability a bee-plant interaction is detected depends on citation type (observation vs collection)
      * Probability a bee-plant interaction is detected depends on flower color
      * Probability a bee-plant interaction is detected depends on flower shape
      * Probability a bee-plant interaction is detected depends on plant family (asteridae vs. non-asteridae)
  + We used a hierarchical community occupancy model and the GloBI database to estimate the following:
    - * psi = The probability a bee species interacts with a plant species
      * p = The probability that a study documented the bee-plant interaction

**Methods**

*Data background*

We downloaded all unique interactions with bee species from the Globi database on XX 2021, which consisted of 300,465 unique observations. The Globi database is populated with self-reported observations that are obtained from museum collections, community-science projects with singleton observations (e.g., iNaturalist), and research studies (both meta-analyses and ecological studies) worldwide. Note that research studies vary in terms of objectives, study design, etc, and the Globi database consists of presence-only data (no absence data are reported). Global Biotic Interactions is a .... We were only interested in bee interactions with plant species, which could be coded as a number of different interactions depending on which species is listed as the “source” and which is the “target” of the interaction.

*Data gathering*

To explore the hypotheses related to bee and flower characteristics that relate to ecological (i.e., what characteristics relate to a species degree of generalism?) and detection (i.e., what characteristics relate to species detectability?) processes, we compiled several pieces of data before starting the data cleaning process.

First, we defined a spatial range to explore the hypotheses. We decided to define the area to a rectangular region around Santa Cruz Island, CA. Specifically, we defined the rectangle from San Luis Obispo to San Diego, CA (latitude from 30 – 36; longitude from -150 to -116), which is an area that KS has worked extensively in and has expertise in the bee fauna. Our questions relate to determining possible bee-flower interactions (but not specific to a site).

Next, we compiled a bee species and plant species list for this area. We excluded any species in our checklists not identified to species level – i.e., genus is identified but listed as “sp.”. We also excluded *Apis mellifera* from the checklist because… Our original plant list included a total of 562 but we trimmed the list to 150 by …. Our final list of bee species included 139 species, and our final plant list included 150 species. We will use the checklists to retroactively assign species non-detections (similar to the Kery et al. 2009 paper where they ad hoc assigned bird non-detections using site checklists).

Additionally, we compiled and identified a bee synonym list for the bee species in this area. For the plant species, we used the package taxize in R.

Then, we compiled details on bee and plant phenology.

*Data cleaning*

To start the data cleaning process, we standardized the columns where bee and plant species names appeared, given that they could be in either the target or source columns. After the species names were standardized, we ran the taxize package on the plant column to obtain the most current accepted name per plant species. We removed any rows that did not include both a bee and plant species ID. This brought the total number of observations down to 157,893.

Next, we used the bee synonym list from zenodo and a list that KS put together for species that did not appear in the zenodo list to update the Globi database with the most current accepted name per bee species.

Then, we determined which rows in the GloBI file matched both a bee AND plant species in our checklist, which filtered down the number of observations to 7,884. Next, we filtered the Globi data to our pre-defined spatial extent (latitude from 30 – 36; longitude from -150 to -116), which filtered our list down to 566 observations.

We were interested in exploring the detection process for different source citations listed in the Globi database, and we realized that one source citation (SCAN) was a compilation of several institutions. In an effort to obtain higher resolution on the institutions compiled in the SCAN citation, we used the source catalog number column. At this point, we have a total of seven unique source citations (with the SCAN citation resolved to its constituent institutions).

Resolved duplicates in the dataset....

Next, we formatted the event date unix epoch column in Globi to year, month, and day the observation occurred; and using the bee species, plant species, and phenology data, we created a list of bee-plant species interactions that are possible. Likewise, we could have created a list of the “forbidden-links”, meaning that either the bee or plant species are not active during the same time so this interaction is impossible. We removed all Globi observations without month information, which reduced our dataset to 515 observations.

To format the data for the analysis, we created a 4-D array with bee species as the first dimension (139 species), plant species as the second (562 species), month as the third dimension (12 months), and source citation as the fourth dimension (7 citations). This created an array with 6,631,968 possible combinations, but we found that there are only 216,385 total possible bee-plant interactions. Given that not every source citation goes out to the field every month, there was only a total of 611,237 possible bee-plant observations within the months that each source citation was in the field.

Lastly, we populated the 4-D array with the observed bee-plant interactions using the Globi observations. We populated the 4-D array with a value of 1 for each bee-plant interaction during the month that each source citation was in the field. We filled in the rest of the possible bee-plant interactions with a non-detection (= 0). Note that in some cases there the same source citation documented the same bee-plant interaction during the same month. In total, there were 278 unique detections of bee-plant-month-citation. Also note that we collapsed all opportunistic records by source citation, such as those from iNaturalist - such that if a bee-plant interaction was ever documented, then it received a 1 (detection) in the final 4-D array, and all others were marked 0.

*Model covariates*

<ADD TEXT HERE>

*Model assumptions*

Given that we are using the single-level occupancy model in a different way, we explicit outline the assumptions we are making with the data and model formulation:

1. We assume that all bee and plant species occur across our entire geographic window. Given this, an implicit assumption we are making is that bee-plant interactions are static and do not vary across space or time (i.e., no extinction or colonization events- a bee species always interacts with a plant species regardless of location and time).
2. We assume that not all bee and plant species interact with one another. We are using bee and plant phenology data to determine when each bee and plant species is active or flowering, respectively. If a bee is active and a plant is flowering, then we assume that there could be a bee-plant interaction.
3. We assume that each source citation had the opportunity to document all bee-plant interactions. We inferred non-detections for bee-plant interactions even when there is no evidence that each plant species was visited. This assumption allows us to do the following: use the different source citations as the 'repeated survey' for each bee-plant interaction.
4. We assume that all bee-plant interactions remain constant across all observations (i.e., a generalist bee remains generalist). Bees do not start interacting with new/ different plant species.
5. We used source citation as a replicate ‘survey’ of each bee-plant interaction. Traditionally, replicate surveys in occupancy models are either temporal or spatial; but here, we use source citation in GloBI as independent replicate surveys, as is the case during double observer surveys.

*Model formulation*

We used a hierarchical community occupancy model to estimate the probability of bee-plant interactions and the total number of bee species interacting with each plant species (Dorazio et al. 2006). Community occupancy models are powerful tools because they preserve the identity of individual species while accounting for variable and imperfect species detection, which is not always the case for traditional methods for quantifying biodiversity (Iknayan et al. 2014; Kery & Royle 2016). Community occupancy models also lead to increased precision on inferences of rare and elusive species by assuming that species-specific parameters are random effects drawn from community-level distributions (Zipkin et al. 2009).

We start by defining bee species *i* interaction with plant species *j* during month *t* (Zi,j,t) as a binary variable in which Zi,j,t = 1 if bee species *i* interacts with plant species *j* during month *t*, and zero otherwise. The interaction state is assumed to be the outcome of a Bernoulli random process where:

Zi,j,t ~ Bernoulli(Ѱi,j)

Here, Ѱi,j is the probability that bee species *i* interacts with plant species *j*. We can included bee size, bee sociality, flower color, and flower shape as covariates in the bee-plant interaction model using a logit-link function, where:

logit(Ѱi,j) = ui + β1 \* *Bee.sizei* + β2 \* *Bee.socialityi* + β3 \* *Flower.colorj* + β4 \* *Flower.shapej*

We standardized bee size by subtracting the mean and dividing by the standard deviation. The other three covariates (bee sociality, flower color, and flower shape) were categorical with two levels each as follows: bee sociality = social vs non-social bee, flower color = yellow or non-yellow, flower shape = bowl vs non-bowl. We assume that each of the intercept values (ui, bee species interaction rateson the logit scale) were drawn from a community-level normal distribution with mean, μѰ, and variance, 𝜎Ѱ2, such that:

ui ~ Normal(μѰ, 𝜎Ѱ2)

The data, yi,j,t,k, consist of detection/non-detection observations. If bee species *i* is observed interacting with plant species *j* during month *t* by source citation *k*, then yi,j,t,k = 1 and a zero otherwise. True bee species interactions with plant species, Zi,j,t, are observed imperfectly, and thus, repeated sampling over a period of time when the community remains closed (i.e., no changes in bee-plant interactions) allows for the distinction between true species non-interaction and non-detection of the observation.

We define the detection model as:

yi,j,t,k ~ Bern(pi,j,t,k· Zi,j,t)

where pi,j,t,k is the probability of detecting the interaction of bee species *i* with plant species *j* during month *t* by source citation *k* given that the bee-plant interaction does occur (Zi,j,t = 1). Thus, detection is a fixed zero when a species is not present because Zi,j,t = 0 in that case. We can model detection probabilities using a logit-link function where:

logit(pi,j,t,k) = vi + δ1 \* *Bee.strippinessi* + δ2 \* *Bee.sizei* + δ3 \* *Survey.montht* + δ4 \* *Survey.month2t* + δ5 \* *Citation.typek* + δ6 \* *Flower.colorj* + δ7 \* *Flower.shapej* + δ8 \* *Plant.familyj*

Again, bee size was standardized. We also included a quadratic term for survey month, assuming that detection probability of bee-plant interactions peak during the year and decline. The other five covariates (bee strippiness, citation type, flower color, flower shape, and plant family) were categorical with two levels each as follows: bee strippiness = stripped vs non- stripped, citation type: observation vs collection, flower color = yellow or non-yellow, flower shape = bowl vs non-bowl, and plant family = asteridae vs non-asteridae. We assume that each of the intercept values (vi, bee species detection probability on the logit scale) were drawn from a community-level normal distribution with mean, μp, and variance, 𝜎p2, such that:

vi ~ Normal(μp, 𝜎p2)

To estimate the total number of unique bee species interacting with each plant species (Nj), including the number of species not observed (but present) during sampling, we sum across Zi,j,t.

*Stochastic search variable selection*

To determine which covariates significantly influences the ecological and detection parts of the model above, we used stochastic search variable selection (Hooten & Hobbs, 2018; George & McCulloch 1993). In stochastic search variable selection, we specify alternative priors such that for covariate *n*:

β*n* | γ*n* ~ γ*n* N(0, cτ2) + (1 - γ*n*)N(0, τ2)

Both c and τ2 are tuned such that τ2 is quite small, providing an effective spike at zero while cτ2 is larger, creating a slab around zero. The slab then provides the prior for β*n* when the variable β*n* is in the model (i.e., when γ*n* = 1). We used a similar approach for the δn covariates.

We retained covariates where γ*n* was > 0.50. Then, using that subset of covariates, we re-ran the model to obtain parameter estimates.

*Model run*

We analyzed the data set using a Bayesian approach in programs R (R Core Team, 2019) and NIMBLE (package *nimble*; <citation>). We ran each model for 75000 iterations with a burn-in of 25000 iterations and thinning by 10. We ran a total of three chains. To aid in model convergence, we used the following priors. We assessed convergence using the Ȓ statistic (Brooks & Gelman 1998) and visually inspected traceplots.

**Results**

*Bee-plant interactions*

*Detection of bee-plant interactions*

**Discussion**

…..

**Acknowledgments**

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

**Author contributions**

GVD contributed to project development, wrote the model, analyzed data, and wrote the first draft of the paper.

XXX contributed to project and model development.

XXX contributed to project and model development.

All co-authors edited the manuscript.

**Data Accessibility**

All data and code for analyses can be reproduced and accessed at either the github repository: XXXX or a USGS server following publication.

**References**

…..

**Figures & Tables**