Running head: BUMBLE BEE WINNERS AND LOSERS

Bee trends in USA

Laura Melissa Guzman[1,2,*], Elizabeth Elle[2], Lora A. Morandin[3],
Neil Cobb, Paige Chesshire, Leithen K. M'Gonigle[2]

1. Marine and Environmental Biology section at the Department of Biological Sciences
   University of Southern California
   Allan Hancock Foundation Building, Los Angeles, CA 90089-0371

2. Department of Biological Sciences
   Simon Fraser University
   8888 University Drive, Burnaby, BC, V5A 1S6, Canada

3. Pollinator Partnership
   600 Montgomery Street, Suite 440 San Francisco, CA 94111

* Corresponding author

Keywords: bumble bees, climate change, land use, species' declines, occupancy models

Acknowledgements: We would like to thank ... and Vaughn Shirey for useful discussions. We also thank all of the data contributors that made this study possible .... We acknowledge funding from Simon Fraser University (to LKM and LMG), the Natural Sciences and Engineering Research Council of Canada (NSERC) (Discovery Grant to LKM), and the Liber Ero Fellowship Program (LMG). This research was enabled in part by support from West Grid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

Running head: Bumble bee winners and losers

Bee trends in USA

Keywords: bumble bees, climate change, land use, species' declines, occupancy models

Abstract

Keywords: bumble bees, climate change, land use, species' declines, occupancy models

1 Introduction

2 Methods

2.1 Data Sources

2.1.1 Bee records

We acquired records of North American occurrence records comprising six bee families (Andrenidae, Apidae, Colletidae, Halictidae, Megachilidae, and Melittidae) from GBIF and SCAN. These records were cleaned to (i) ensure taxonomic names were correct, (ii) erroneous records were removed following [**?**]. Records from *Apis Mellifera* were removed and the data was restricted the contiguous US. The total number of occurrence records obtained form Chesshire et al. **?**] was $1,923,814$ occurrence records for $3,158$ bee species from 1700 to 2021. From these $1,923,814$ occurrence records, many were multiple observations of the same species in the same date, in the same location. The number of unique species x date x location was $634,597$. Because occupancy analyses required the use of month, the data was further cleaned to ensure month was correct. After cleaning the month, the remaining number of unique species x date x location was $631,233$ for $3,080$ bee species. Further, we removed any species that had less than 10 unique observations (unique dates and locations). This filter further reduced the number of species to $2,150$ and $627,793$ observations. Finally, we also removed species that were present in less than 3 years and 10 counties, resulting in $1,579$ bee species with a total of $608,276$ unique records. Finally we filter the data to the year range of this study (1995 to 2014) resulting in XXXX records and XXXX species.

These $1,579$ bee species were distributed such that Andrenidae had 433 species, Apidae had 423 species, Colletidae 92 species, Halctidae 250 species, Megachilidae 372 species, and Melittidae 9 species.

2.1.2 Species Ranges

For each species we constructed plausible species ranges by drawing a convex hull around all observations and finding the counties that fell within the convex hull. This resulted in the plausible set of sites where the occupancy of each species could be modelled.

While this may include counties where a species cannot be found, this greatly reduces the bias of including all possible sites for every species [**? ?**]. This is particularly problematic for models that do not have spatial predictors in occupancy (such as model 1).

2.1.3 Sites

Because the pesticide use data is reported at the level of county (see below), we calculate every other environmental predictor at the level of county and report each county as a ”site”.

2.1.4 Site level environmental predictors

Agriculture:

Land cover data was obtained from the National Land Cover Database (NLCD), which provides national wide data on land cover at a 30*m* resolution for every 2-3 years from 2001 to 2016. The NLDC provides data for 16 land use categories. Of relevance to us is the category of ”Cultivated crops” (hereafter agriculture), which contains areas used for the production of annual crops (such as corn, soybeans, vegetables, etc.), perennial woody crops (such as orchards and vineyards), and land actively tilled [**? ?**]. For each county, we calculate the proportion of the county covered in agriculture. While this database does not cover every year in our study, we interpolate the data within years.

Pesticide use:

Pesticide use data was obtained from USGS Pesticide National Synthesis Project, which provides national data on pesticide use for each county for every year from 1992 to 2021. The pesticide use data is provided as Kg of active ingredient used in each county for 448 types of active ingredients [**?** **?**]. From these active ingredients, we select those that have been shown to be highly toxic or moderately toxic to honey bees based on LD50 ecotoxicity results (EPA Ecotoxocity database). Specifically the active ingredients we included in our analyses were the following neonicotinoids: Acetamiprid, Clothianidin, Dinotefuran, Imidacloprid, and Thiamethoxam, the following pyrethroids: Alpha Cypermethrin, Bifenthrin, Cyfluthrin, Deltamethrin, Esfenvalerate, Gamma Cyhalothrin, Lambda Cyhalothrin, Permethrin, Tefluthrin, and Zeta Cypermethrin. Finally, the following organophospates: Abamectin, Carbaryl, Oxamyl, Pyridaben, Acephate, Chlorethoxyfos, Chlorpyrifos, Diazinon, Dimethoate, Malathion, Fipronil, Sulfoxaflor.

We compile climatic variables using CHELSA high resolution climate data for earth [**?** **?**]. We calculate the mean monthly maximum temperature and the mean precipitation at a given site in a given era. We also quantify floral resources for bees by combining classifications of land use estimates for the Holocene (HYDE) [**?**] with previously established floral resource scores for bees [**?**]. We overlay the HYDE land-use map with the CDL (crop data layer) map to obtain the categories of the CDL that geographically overlap with the HYDE categories for 2008. Then for each HYDE category, we calculate the average floral resources reported by **?**]. **?**] provide expert-opinion derived floral resource scores for many types of crops and other land use categories and we average these for the various land-use types within each site in each era. The temporal variation in floral resources in our data-set, therefore, arises from variation in land-use through time in HYDE and not variation in floral-resource values for each category. We sum floral resources across spring, summer, and fall to provide an overall metric through the season. We denote our floral resources by FR and, while the actual magnitude of these scores are not particularly meaningful, relative values between sites are. In our data-set, our floral resource scores range from 0.80 FR to 1.61 FR. We calculate predictor values for each site in each era.

2.2 Occupancy models

We develop the first multi-species occupancy model for bumble bee occurrence records in North America that directly estimates effects of climate and land-use variables on species' occupancy. In constructing our models, we build on work done by [**?**] that tested the validity of various methods of applying occupancy models to large-scale presence-only data sets. Here, we present two models: one where time is a predictor of occupancy ("Era Model") and the other where climate and land use are predictors of occupancy ("Environmental Model"). Full model details and parameter definitions are provided in the Supplementary Material and we provide a short summary here.

Era model: To test for genus-wide temporal trends in bumble bee occupancy (Q1), we consider a simple model wherein we model the effect of "era" as a direct predictor of each species' occupancy, letting $\mu_{\psi\text{era}}$ denote the mean effect across all species and $\psi_{\text{era}}[i]$ denote the effect for species i.

Environmental model: Next, we replace the effect of era in the above model with environmental predictors that vary across sites and eras (Q2). Specifically, we include linear and quadratic effects of site-averaged maximum temperature (mean linear effect across species denoted by $\mu_{\psi\text{temp}}$, standard deviation by $\sigma_{\psi\text{temp}}$, and species-specific responses by $\psi_{\text{temp}}[i]$; quadratic effect denoted by ψ_{temp2}), a linear effect of site-averaged precipitation (analogously denoted by $\mu_{\psi\text{precip}}$ $\sigma_{\psi\text{precip}}$, $\psi_{\text{precip}}[i]$), and a linear effect of site-averaged floral cover (analogously denoted by $\mu_{\psi\text{floral}}$ $\sigma_{\psi\text{floral}}$, $\psi_{\text{floral}}[i]$). The quadratic effect of temperature allows the model to estimate each species' thermal optima from which deviation in either direction leads to decreases in occupancy. To minimize model complexity, we only estimate a single community-wide quadratic effect of temperature, rather than species-specific quadratic effects. In doing so, we are assuming that all species have approximately the same niche breadth, while still allowing for species-specific responses to temperature. We do not include era in this model because variation in occupancy due to any monotonic increase or decrease in environmental covariates would then be accounted for by this non-environmental temporal variable. However, when we did include all variables in a single model, our conclusions did not change (Supplementary Fig. **??**).

In both of the above models, we model detection probability with a site- and era-specific random effect. This allows detection to vary relatively independently across sites and between eras. We also considered models that included an additional fixed effect of era on detection and, again, our conclusions did not change (Supplementary Fig. **??**).

3 Results

4 Discussion

Figure 1: Species-specific occupancy trends are variable but, on average, increase through time (a), peak at intermediate temperature (b), and are highly variable as a function of precipitation (c), and floral resources (d). In all cases, species-specific trends (grey curves; only shown over the range of values experienced by that species) are variable and not well characterized by the genus-level trajectories (solid lines). Shaded regions denote 95% Bayesian credible intervals. Output in (a) is from the Era model and (b)-(d) the Environmental model. To highlight that these are two separate models we have plotted the mean line(s) for the Era model in red and the Environmental model in black.

Supplementary Information

S1 Methods

S1.0.1 Bee records

S1.0.2 Spatial and temporal classification of bee occurrence

We model each species only over the sites that we infer to be plausibly within that species' range. To construct a species' range, we trace a convex hull around all sites containing observations of that species, regardless of when those observations occurred, and consider all sites within the resultant polygon to be within that species' range. By only modeling each species over the sites at which it could plausibly occur, we generate meaningful estimates of occurrence, while also ensuring that effects of climate and floral resources are only based on the relevant sets of sites and values of environmental variables.

S1.0.3 Climate data

We compile climatic variables using CHELSA high resolution climate data for earth [**? ?**], which contains monthly global temperature and precipitation values at a spatial resolution of 1×1km. To calculate the maximum temperature at a given site in a given era, we calculate the average maximum temperature (using only data for July and August, as these months will often record the highest temperature in the year) across all of the 1×1km cells within that site and across all the years within that era. To calculate the mean precipitation, we similarly average monthly mean precipitation (for all 12 months) across the same cells and years. Because the climate data records are only available until 2016, climate values in our final era are based on 15 years of data (2001-2016) rather than the full 20 years.

S1.0.4 Floral resource data

To quantify floral resources for bees, we combine classifications of land use estimates for the Holocene (HYDE) [?] with floral resource scores for bees [?]. Land use estimates for the Holocene spans 10000 BCE - 2015 CE worldwide. From 1900s to 2000, HYDE provides land-use categories on a decade basis and from 2000 to 2015, yearly. HYDE's spatial resolution is 5 arc minutes which is approximately 9.26Km at the equator and 4.6Km at latitude 60. HYDE land use categories contain, for example, cropland, urban, rangeland, wild-remote woodlands [?]. While these categories are useful for understanding the form that land conversion has taken over the past century, it is unclear how transitions between these categories might impact bees. ?] quantified expert knowledge to estimate bee abundance based on land uses, including a variety of crops and other land types such as pasture and forest, using the Cropland Data Layer (CDL) from 2008. Using these values of floral resources does not allow for temporal variation in floral resources per se (i.e., the value of 'corn' does not change through time). Variation in floral resources through time stems from changes in land values from HYDE. The Cropland Data Layer, produced by the National Agricultural Statistics Service (NASS), provides geo-referenced crop land cover data for the continental United States at a 30m resolution.

We overlay the HYDE land-use map with the CDL map to obtain the categories of the CDL that geographically overlap with the HYDE categories for 2008. Then for each HYDE category, we calculate the average floral resources reported by ?]. ?] leveraged expert opinion to create a range of floral resources availability for 45 land-use cover types from the CDL. We add floral resources for spring, summer, and fall to provide an overall metric of floral resources through the season, as these are more relevant for bumble bees, which have long flight periods. While ?] also produced expert estimates for nesting resources, we only used floral resources here as these are more likely to apply to all bumble bee species. By overlaying the HYDE land-use map and the CDL map, we are implicitly assuming that the floral resources provided by a given crop are consistent across the continent and have been through the last century; an assumption that is probably not true. However, we believe that this metric still likely captures a course estimate of available floral resources.

S1.1 Occupancy models

We assume that the probability that species i is detected at site j in era k, x_{ijk}, is drawn from a Bernoulli distribution (0 or 1) with probability (y_{ijk}),

$$x_{ijk} \sim \text{Bernoulli}(y_{ijk}) \tag{S1}$$

where y_{ijk} is the product of detection probability (p_{ijk}) and the unknown, but true occupancy state, z_{ijk},

$$y_{ijk} = p_{ijk} * z_{ijk} \tag{S2}$$

The true but unknown site occupancy for species i at site j, z_{ijk} is equal to 1 if that site is occupied and 0 if it is not. We assume that this true site occupancy is drawn from a Bernoulli distribution with mean equal to the species' occupancy probability at that site,

$$z_{ijk} \sim \text{Bernoulli}(\psi_{ijk}) \tag{S3}$$

Both occupancy probability, ψ, and detection probability, p, can be formulated as functions of covariates, and we do this in two different ways for the former.

First, to test for genus-wide temporal trends in bumble bee occupancy (Q1), we consider a simple model wherein we model "era" directly. Specifically, we model occupancy as

$$\begin{aligned}\text{logit}(\psi_{ijk}) =&\psi_0+\\&\psi_{\text{species}}[i]+\\&\psi_{\text{area}} \times \text{area}[j]+\\&\psi_{\text{era}}[i] \times k\end{aligned} \tag{S4}$$

Here, ψ_0 denotes mean occupancy, $\psi_{\text{species}}[i]$ denotes a species-specific random effect, ψ_{area} denotes a fixed effect of site area to account for the fact that some sites are truncated by water and smaller (area[j] denotes the area of site j), and $\psi_{\text{era}}[i]$ denotes a species-specific effect of era. We call this the **Era model**.

Second, we consider a model wherein we replace the effect of era in the above model with era-level environmental predictors (Q2). Specifically, we include site-averaged maximum temperature, site-averaged precipitation, and site-averaged floral cover, such that our model for occupancy becomes

$$\begin{aligned}\text{logit}(\psi_{ijk}) =&\psi_0+\\&\psi_{\text{species}}[i]+\\&\psi_{\text{area}} \times \text{area}[j]+\\&\psi_{\text{temp}}[i] \times \text{temp}[j,k]+\\&\psi_{\text{temp2}} \times \text{temp}[j,k]^2+\\&\psi_{\text{precip}}[i] \times \text{precip}[j,k]+\\&\psi_{\text{floral}}[i] \times \text{floral}[j,k]\end{aligned} \tag{S5}$$

Here, ψ_0, $\psi_{\text{species}}[i]$, and ψ_{area} are as defined above in Eq. **??** and $\psi_{\text{temp}}[i]$, $\psi_{\text{precip}}[i]$, and $\psi_{\text{floral}}[i]$ denote species-specific linear effects of temperature, precipitation, and floral resources, respectively and ψ_{temp2}, denotes a quadratic effect of temperature (not species-specific). We call this the **Environmental model**.

We assume that species-specific slopes in both of the above models are normally distributed about some mean. Specifically,

$$\begin{aligned} \psi_{\text{era}}[i] &\sim \mathcal{N}(\mu_{\psi\text{era}}, \sigma_{\psi\text{era}}) \\ \psi_{\text{temp}}[i] &\sim \mathcal{N}(\mu_{\psi\text{temp}}, \sigma_{\psi\text{temp}}) \\ \psi_{\text{precip}}[i] &\sim \mathcal{N}(\mu_{\psi\text{precip}}, \sigma_{\psi\text{precip}}) \\ \psi_{\text{floral}}[i] &\sim \mathcal{N}(\mu_{\psi\text{floral}}, \sigma_{\psi\text{floral}}), \end{aligned} \tag{S6}$$

where $\mu_{\psi\text{era}}$, $\mu_{\psi\text{temp}}$, $\mu_{\psi\text{precip}}$, $\mu_{\psi\text{floral}}$ denote the mean effect of each corresponding predictor, across species, and σ terms denote the variances about these means.

In both of the above models, we model detection probability as

$$\text{logit}(p_{ijk}) = p_0 + p_{\text{site.era}}[j,k] \tag{S7}$$

where p_0 denotes the mean detection probability and $p_{\text{site}}[j,k]$ denotes a site-specific random effect that is era-specific. This latter term allows detection to vary relatively independently across sites and between eras. Specifically, we assume

$$p_{\text{site.era}}[j,k] \sim \mathcal{N}(\mu_{p\text{site.era}}, \sigma_{p\text{site.era}}). \tag{S8}$$

In addition, we ran our “era” model without splitting *B. occidentali*s into *B. occidentalis* and *B. mckayi* to assess the effect this species split has on their trend through time.

We fit models in JAGS [**?**] and assess model convergence both by visually inspecting chains and checking The Gelman-Rubin statistic (we ensured that Rhat was < 1.1 for all parameters). We use flat, uninformative priors for all parameters and ran models for 20,000 iterations, discarding the first 10,000 iterations and thinning by 10 across 3 chains. For all analysis we used R V4.0.4 [?]. For spatial manipulations we used the packages `raster` [?], `rgeos` [?], `maptools` [?], `rgdal` [?], `sp` [?], `spatstat` [?]; for data manipulation we used `stringr` [?] and `data.table` [?]; for running models, we used `rjags` [?], `R2jags` [?], and `runjags` [?].

S2 Supplementary Results

Figure S1: Change in mean temperature, mean precipitation, and mean floral resources, through time at a spatial resolution of 250×250km.