An Introduction to Occupancy Models with Collections Data

JT Miller
University of Florida

FLORIDA MUSEUM

UNIVERSITY of FLORIDA

iDigBio
Integrated Digitized Biocollections

NSF

Background: Species Distribution Models

Assumes that areas w/no detections are less suitable.

- Presence Only Data: Presence = More habitable

  - Flexibility of these model is limited

  - - Subject to Biased Sampling, Detection Unknown

     - Niche Conservation (long-term modeling)

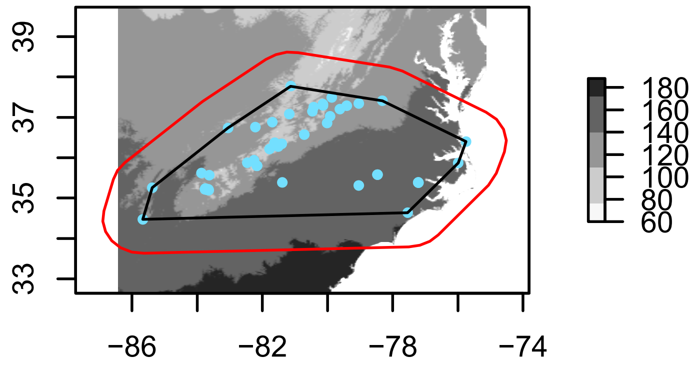


Figure from Gaynor et al. 2018

Not the only way to create a species distribution!

1) Presence only Data
2) ‘Presence/Absence’ Data
3) ‘Presence/Absence’ Data replicated over time
4) Count Data

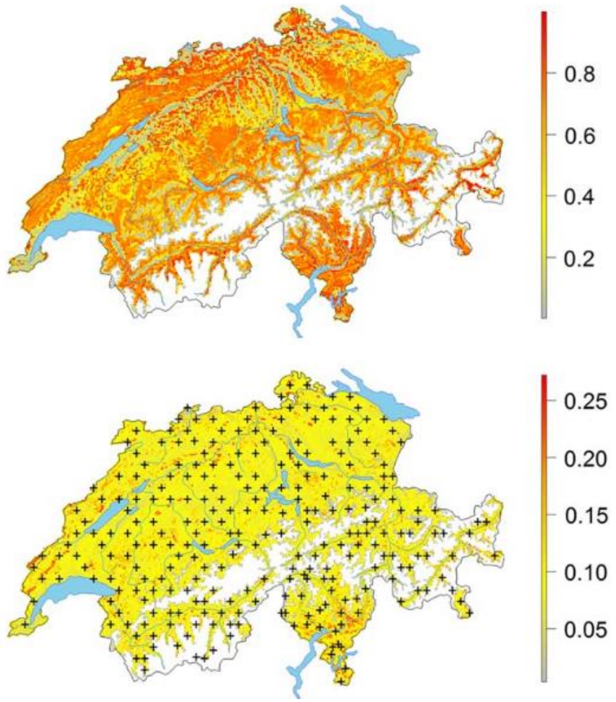
Not the only way to create a species distribution!

ENMs

Occupancy Models (1/0)

1) Presence only Data
2) ‘Presence/Absence’ Data
3) ‘Presence/Absence’ Data replicated over time
4) Count Data

N-Mixture Models



Occupancy Modeling with replicates to model intensity
Kery & Royle 2016

Not the only way to create a species distribution!

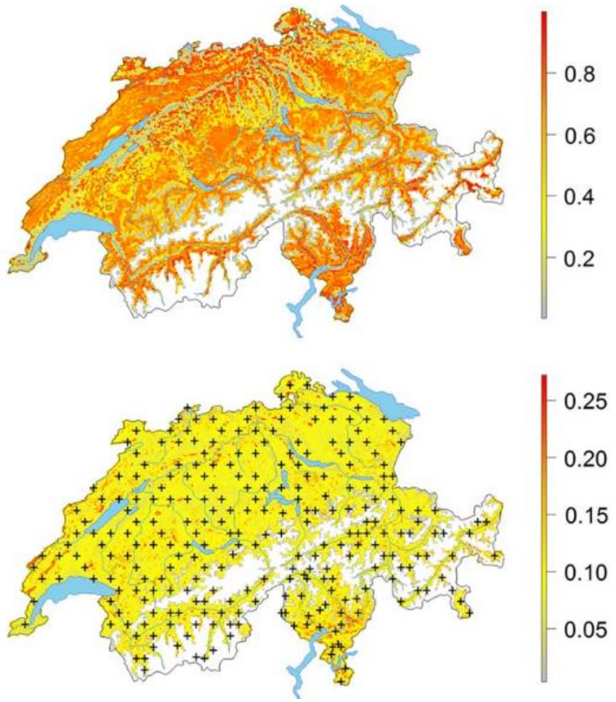
ENMs

Occupancy Models (1/0)

1) Presence only Data
2) 'Presence/Absence' Data
3) 'Presence/Absence' Data replicated over time
4) Count Data

N-Mixture Models

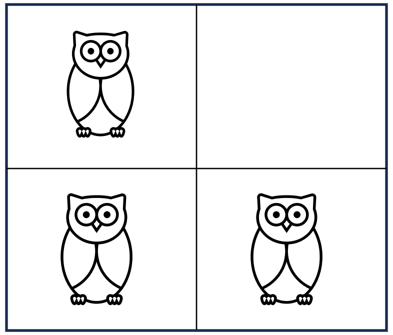
More Data Rich = Access To More Modeling Methods that allow finer inference



Occupancy Modeling with replicates to model intensity
Kery & Royle 2016

Traditional Occupancy Models

- Goal: **Estimate the probability that a species is present in a given site, while accounting for detection error**

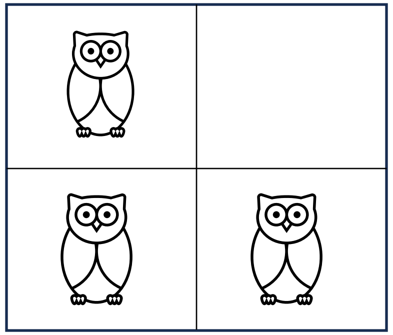


	psi	z	y1	y2	y3
[1,]	0.05	0	0	0	0
[2,]	0.05	0	0	0	0
[3,]	0.07	0	0	0	0
[4,]	0.07	1	0	0	0
[5,]	0.07	0	0	0	0
[6,]	0.08	0	0	0	0

	psi	z	y1	y2	y3
[95,]	0.92	1	0	1	1
[96,]	0.92	1	1	0	0
[97,]	0.93	1	1	0	1
[98,]	0.94	1	1	0	0
[99,]	0.94	1	0	1	0
[100,]	0.95	1	0	0	1

Traditional Occupancy Models

- Goal: Estimate the probability that a species is present in a given site, while accounting for detection error
- **Requires "Presence/Absence" more accurately "Detection/NonDetection" data.**

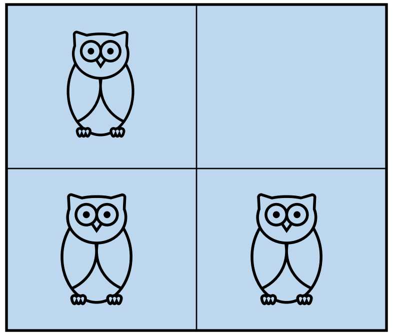


	psi	z	y1	y2	y3
[1,]	0.05	0	0	0	0
[2,]	0.05	0	0	0	0
[3,]	0.07	0	0	0	0
[4,]	0.07	1	0	0	0
[5,]	0.07	0	0	0	0
[6,]	0.08	0	0	0	0

	psi	z	y1	y2	y3
[95,]	0.92	1	0	1	1
[96,]	0.92	1	1	0	0
[97,]	0.93	1	1	0	1
[98,]	0.94	1	1	0	0
[99,]	0.94	1	0	1	0
[100,]	0.95	1	0	0	1

Traditional Occupancy Models

- Goal: Estimate the probability that a species is present in a given site, while accounting for detection error
- Requires “Presence/Absence” or more accurately, “Detection/NonDetection” data.
- **Replicated Across Sites**



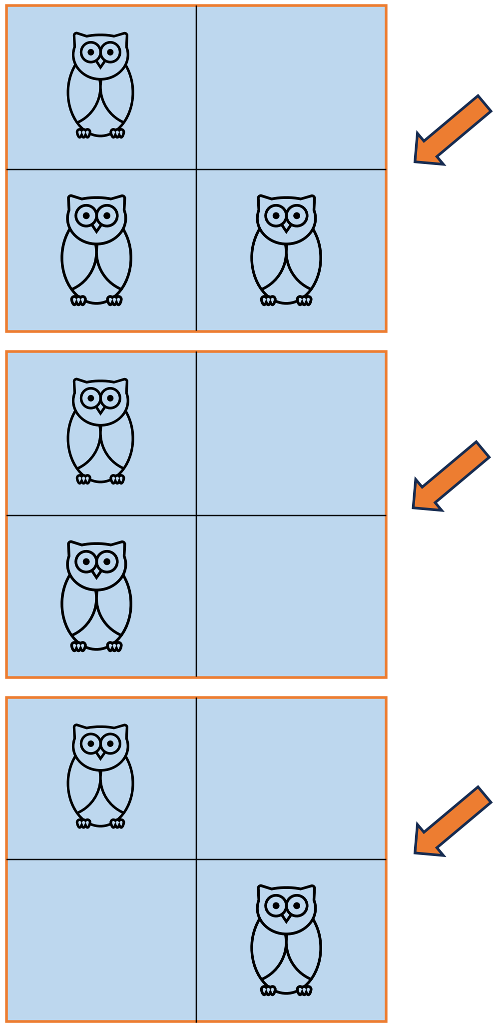
Sites!

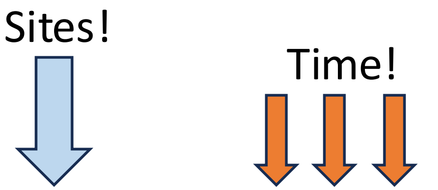
	psi	z	y1	y2	y3
[1,]	0.05	0	0	0	0
[2,]	0.05	0	0	0	0
[3,]	0.07	0	0	0	0
[4,]	0.07	1	0	0	0
[5,]	0.07	0	0	0	0
[6,]	0.08	0	0	0	0

	psi	z	y1	y2	y3
[95,]	0.92	1	0	1	1
[96,]	0.92	1	1	0	0
[97,]	0.93	1	1	0	1
[98,]	0.94	1	1	0	0
[99,]	0.94	1	0	1	0
[100,]	0.95	1	0	0	1

Traditional Occupancy Models

- Goal: Estimate the probability that a species is present in a given site, while accounting for detection error
- Requires "Presence/Absence" or more accurately, "Detection/NonDetection" data.
- Replicated Across Sites
- **Replicated Across Time**

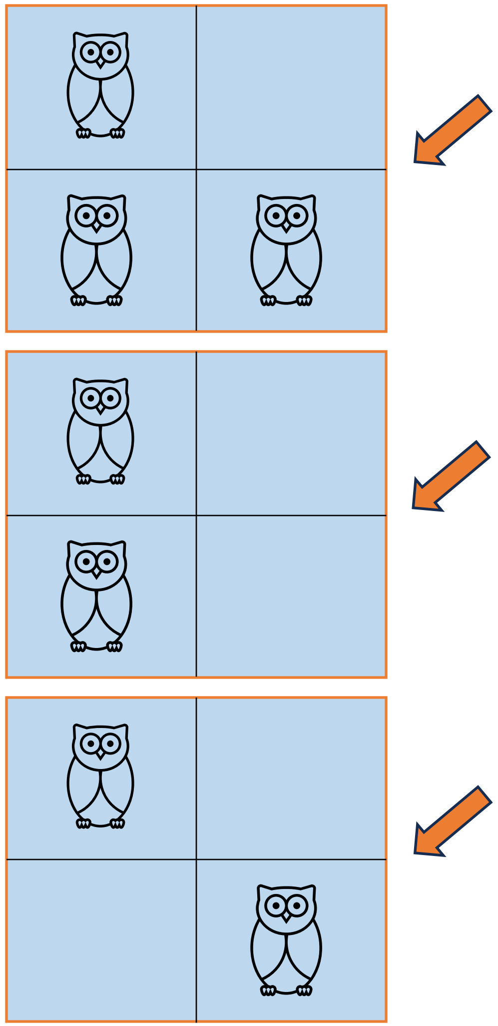





```
     psi z y1 y2 y3
[1,] 0.05 0  0  0  0
[2,] 0.05 0  0  0  0
[3,] 0.07 0  0  0  0
[4,] 0.07 1  0  0  0
[5,] 0.07 0  0  0  0
[6,] 0.08 0  0  0  0
        psi z y1 y2 y3
 [95,] 0.92 1  0  1  1
 [96,] 0.92 1  1  0  0
 [97,] 0.93 1  1  0  1
 [98,] 0.94 1  1  0  0
 [99,] 0.94 1  0  1  0
[100,] 0.95 1  0  0  1
```

Traditional Occupancy Models

- Goal: Estimate the probability that a species is present in a given site, while accounting for detection error
- Requires "Presence/Absence" or more accurately, "Detection/NonDetection" data.
- Replicated Across Sites
- Replicated Across Time
- Helps determine our **uncertainty** in the probability of detection



Sites!

Ψ

Time!

	psi	z	y1	y2	y3
[1,]	0.05	0	0	0	0
[2,]	0.05	0	0	0	0
[3,]	0.07	0	0	0	0
[4,]	0.07	1	0	0	0
[5,]	0.07	0	0	0	0
[6,]	0.08	0	0	0	0
[95,]	0.92	1	0	1	1
[96,]	0.92	1	1	0	0
[97,]	0.93	1	1	0	1
[98,]	0.94	1	1	0	0
[99,]	0.94	1	0	1	0
[100,]	0.95	1	0	0	1

Detection Probability Can Be Quantified

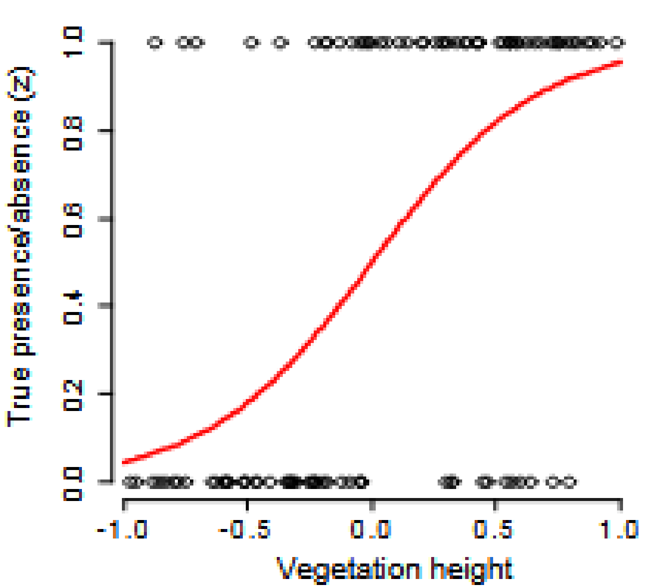
- The aim of these occupancy models are to quantify true occurrence, by accounting for detection complexities
- NonDetections Namely
- Also False Detections if included

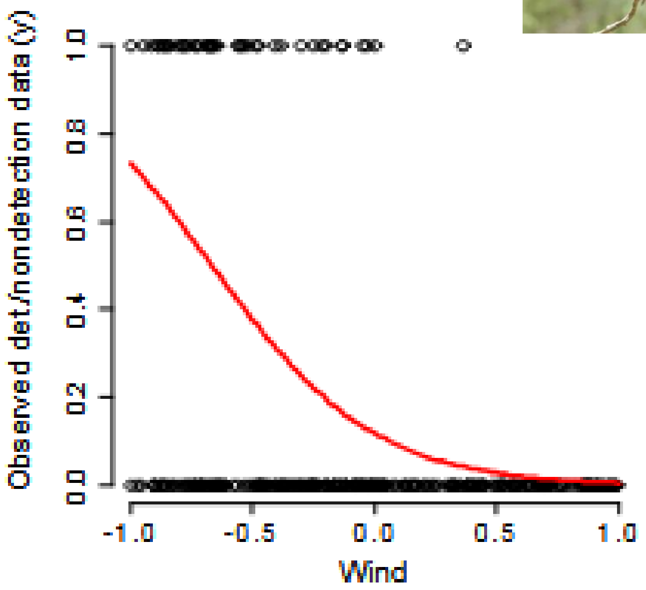



Covariates Can Be Added to the Model

- Covariates (things that we think may affect our detection or the P/A of the organism) can also be added to these models
- Say vegetation height (+) and wind speed (-)





Kery & Royle 2016 Simulation Unmarked

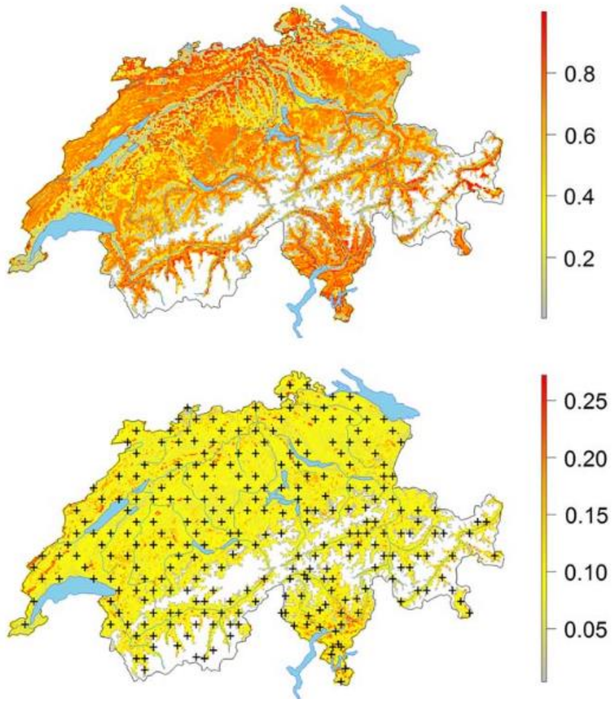
Not the only way to create a species distribution!

ENMs + Occupancy Models?*

Occupancy Models (1/0)

1) Presence only Data
2) 'Presence/Absence' Data
3) 'Presence/Absence' Data replicated over time
4) Count Data

N-Mixture Models



Occupancy Modeling with replicates to model intensity
Kery & Royle 2016

New Methods for Using Museum Specimen Data

- Occupancy can be measured under certain conditions with museum/incidental data
- Provides a new framework for studying Global Change and Biodiversity across unprecedented scales in both time & space


Received: 5 December 2021 | Accepted: 14 April 2022
DOI: 10.1111/2041-210X.13896

Methods in Ecology and Evolution

RESEARCH ARTICLE

Leveraging Natural History Collections to Understand the Impacts of Global Change

Occupancy-detection models with museum specimen data: Promise and pitfalls

Vaughn Shirey[1] | Rassim Khelifa[2] | Leithen K. M'Gonigle[3] | Laura Melissa Guzman[3,4]

[1]Department of Biology, Georgetown University, Washington, DC, USA
[2]Department of Zoology, University of British Columbia, Vancouver, BC, Canada
[3]Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada
[4]Marine and Environmental Biology

Abstract

1. Historical museum records provide potentially useful data for identifying drivers of change in species occupancy. However, because museum records are typically obtained via many collection methods, methodological developments are needed to enable robust inferences. Occupancy-detection models, a relatively new and powerful suite of statistical methods, are a potentially promising avenue because they can account for changes in collection effort through space and time.
2. We use simulated datasets to identify how and when patterns in data and/or modelling decisions can bias inference. We focus primarily on the consequences of contrasting methodological approaches for dealing with species' ranges and inferring species' non-detections in both space and time.
3. We find that not all datasets are suitable for occupancy-detection analysis but, under the right conditions (namely, datasets that are broken into more time periods for occupancy inference and that contain a high fraction of community-wide collections, or collection events that focus on communities of organisms), models can accurately estimate trends. Finally, we present a case study on eastern North American odonates where we calculate long-term trends of occupancy using our most robust workflow.
4. These results indicate that occupancy-detection models are a suitable framework for some research cases and expand the suite of available tools for macroecological analysis available to researchers, especially where structured datasets are unavailable.

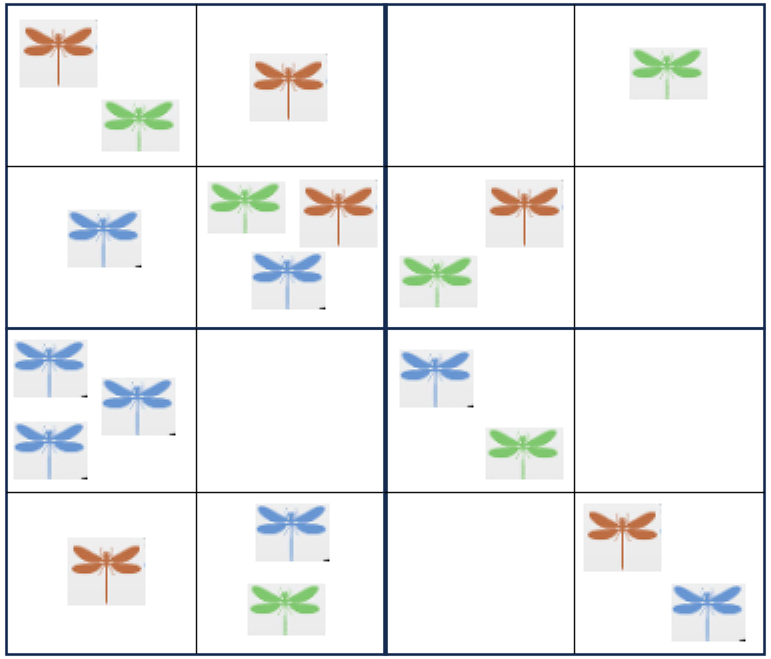
KEYWORDS
global change, hierarchical model, museum specimens, occupancy-detection models

New Methods for Using Museum Specimen Data

- Simulations with specimen/incidental based data suggests that we can reach the gold standard of correctly estimating the probability of detection by using community level observations, if they meet a certain standard

Simulations Suggest:

- Estimate Community Visits vs Target Sampling events. 50% safest, 25% can work with many temporal bins
- Restrict Analyses to only those places and times where collecting is known or likely to have occurred. Don't model all species at all sites/time intervals
- Infer non-detections from community (multi-species) visit data only, do not infer absence everywhere
- Split time across 'Occupancy Intervals', the more intervals the better (provided that you check how visit and detection probability changes in these intervals)

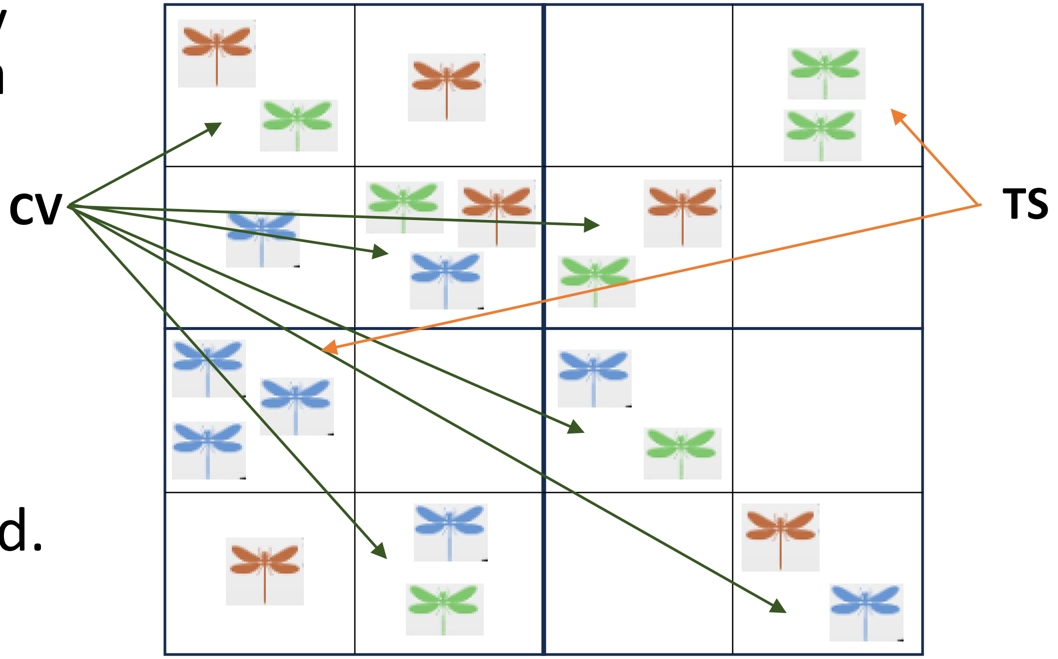


New Methods for Using Museum Specimen Data

- Simulations with specimen/incidental based data suggests that we can reach the gold standard of correctly estimating the probability of detection by using community level observations, if they meet a certain standard

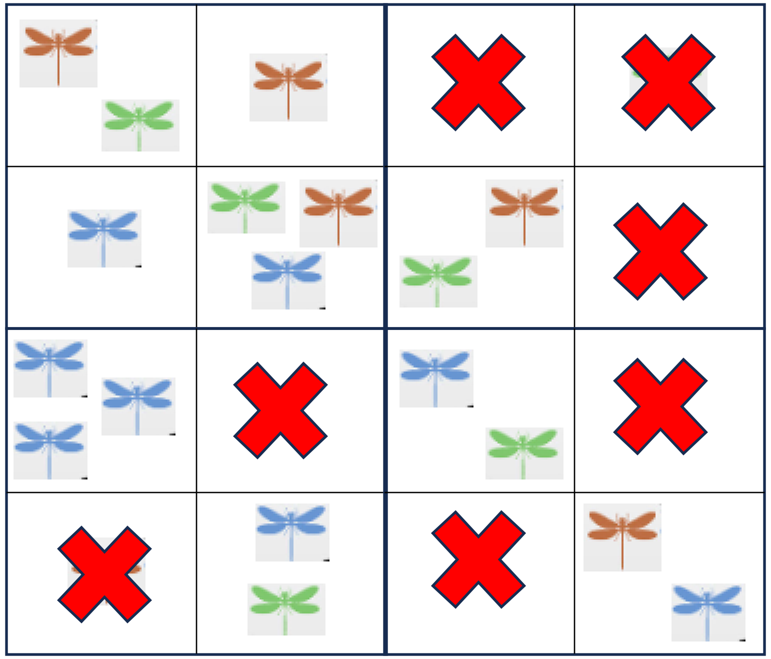
Simulations Suggest:

- Estimate **Community Visits** vs **Target Sampling** events. 50% safest, 25% can work with many temporal bins
- Restrict Analyses to only those places and times where collecting is known or likely to have occurred. Don't model all species at all sites/time intervals
- Infer non-detections from community (multi-species) visit data only, do not infer absence everywhere
- Split time across 'Occupancy Intervals', the more intervals the better (provided that you check how visit and detection probability changes in these intervals)



New Methods for Using Museum Specimen Data

- Simulations with specimen/incidental based data suggests that we can reach the gold standard of correctly estimating the probability of detection by using community level observations, if they meet a certain standard



Simulations Suggest:

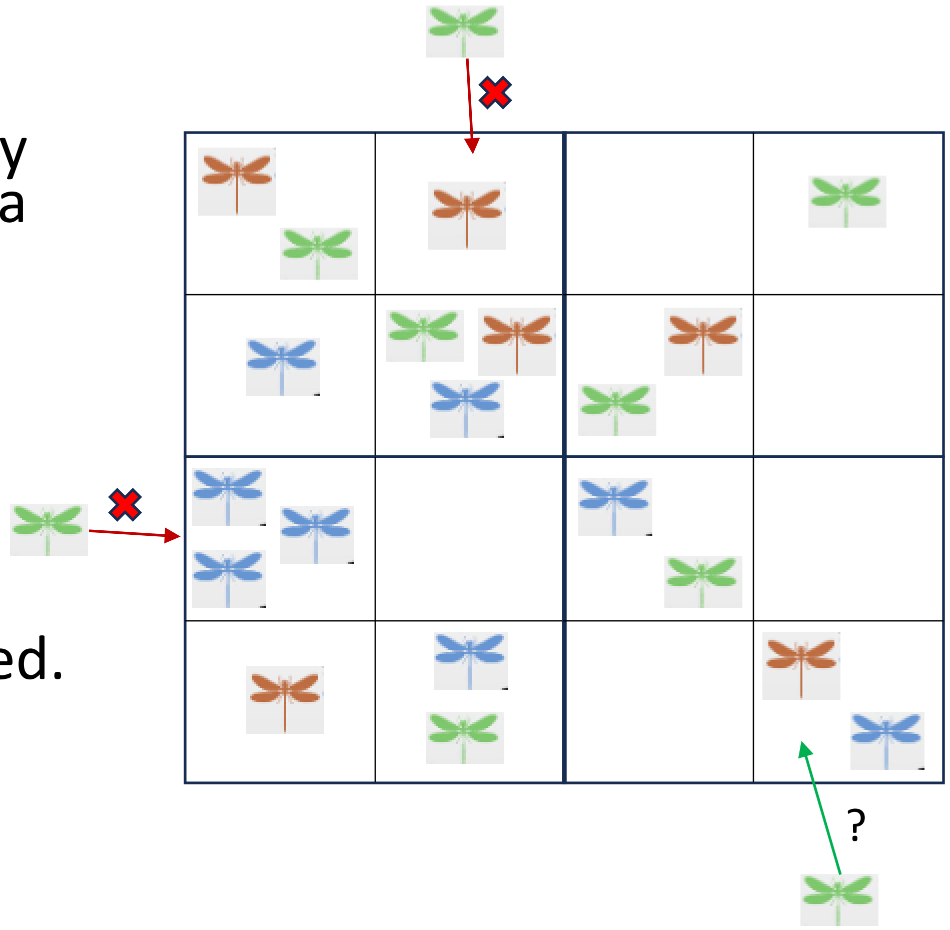
- Estimate Community Visits vs Target Sampling events. 50% safest, 25% can work with many temporal bins
- **Restrict Analyses** to only those **places** and **times** where collecting is known or likely to have occurred. **Don't model all species at all sites/time intervals**
- Infer non-detections from community (multi-species) visit data only, do not infer absence everywhere
- Split time across 'Occupancy Intervals', the more intervals the better (provided that you check how visit and detection probability changes in these intervals)

New Methods for Using Museum Specimen Data

- Simulations with specimen/incidental based data suggests that we can reach the gold standard of correctly estimating the probability of detection by using community level observations, if they meet a certain standard

Simulations Suggest:

- Estimate Community Visits vs Target Sampling events. 50% safest, 25% can work with many temporal bins
- Restrict Analyses to only those places and times where collecting is known or likely to have occurred. Don't model all species at all sites/time intervals
- Infer non-detections from **community** (multi-species) visit data only, **do not infer absence everywhere**
- Split time across 'Occupancy Intervals', the more intervals the better (provided that you check how visit and detection probability changes in these intervals)

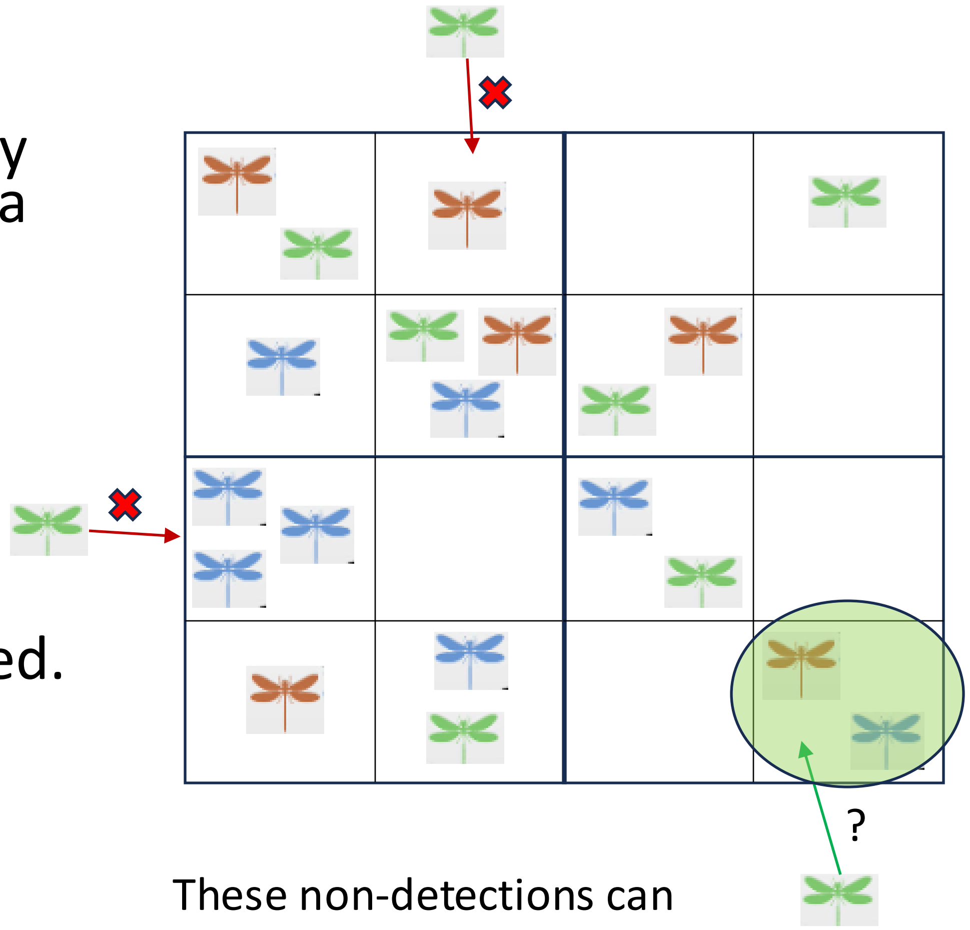


New Methods for Using Museum Specimen Data

- Simulations with specimen/incidental based data suggests that we can reach the gold standard of correctly estimating the probability of detection by using community level observations, if they meet a certain standard

Simulations Suggest:

- Estimate Community Visits vs Target Sampling events. 50% safest, 25% can work with many temporal bins
- Restrict Analyses to only those places and times where collecting is known or likely to have occurred. Don't model all species at all sites/time intervals
- Infer non-detections from **community** (multi-species) visit data only, **do not infer absence everywhere**
- Split time across 'Occupancy Intervals', the more intervals the better (provided that you check how visit and detection probability changes in these intervals)



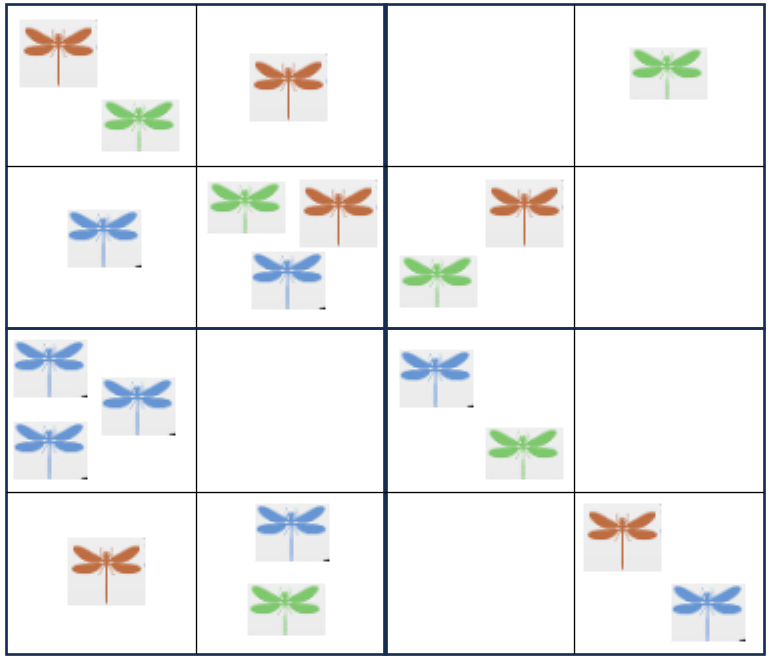
These non-detections can be modeled more accurately if we know something about the range of the species prior

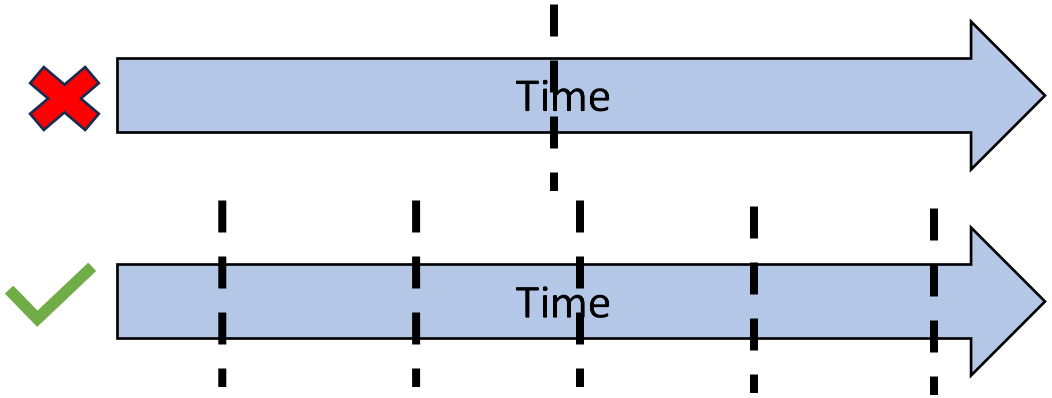
New Methods for Using Museum Specimen Data

- Simulations with specimen/incidental based data suggests that we can reach the gold standard of correctly estimating the probability of detection by using community level observations, if they meet a certain standard

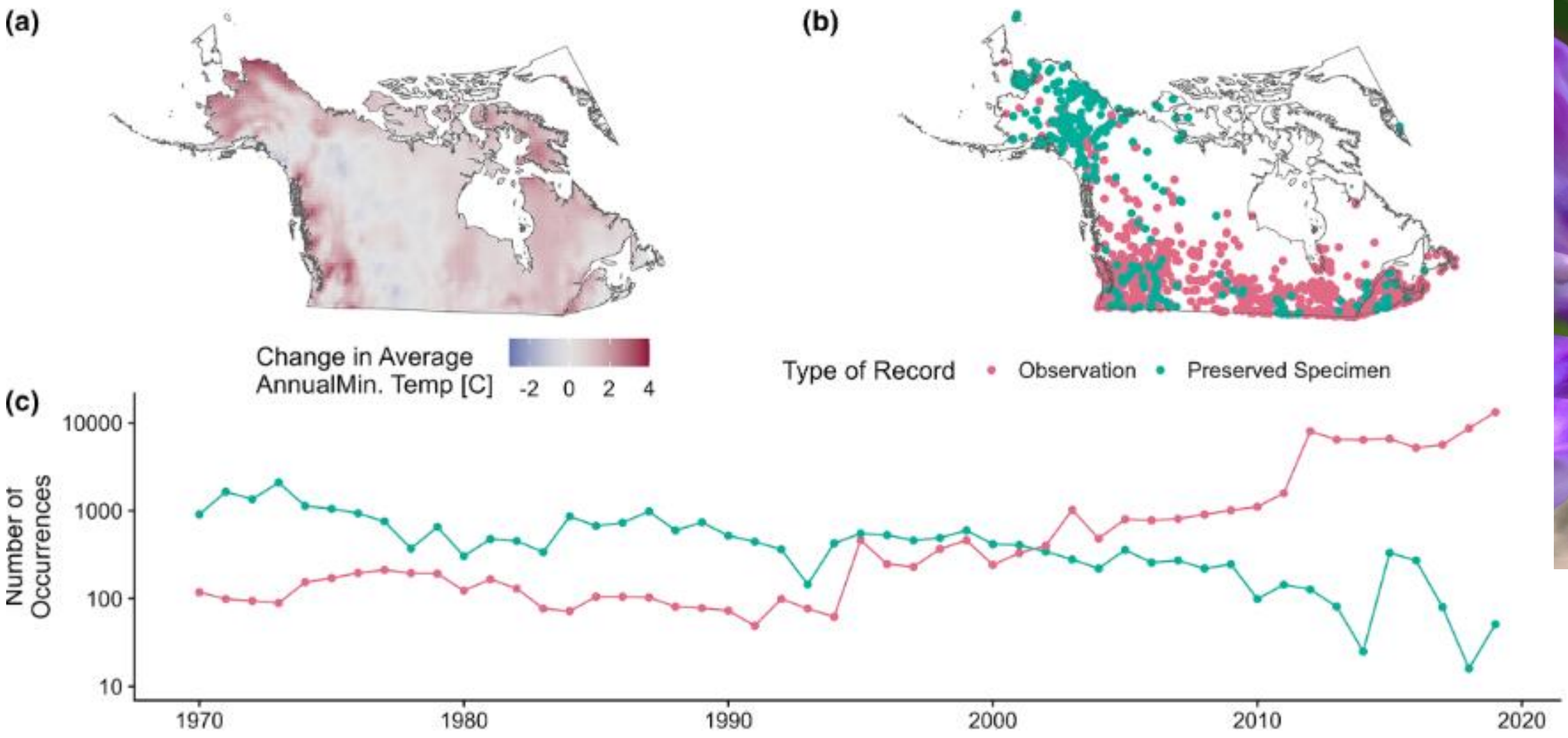
Simulations Suggest:

- Estimate Community Visits vs Target Sampling events. 50% safest, 25% can work with many temporal bins
- Restrict Analyses to only those places and times where collecting is known or likely to have occurred. Don't model all species at all sites/time intervals
- Infer non-detections from community (multi-species) visit data only, do not infer absence everywhere
- Split time across 'Occupancy Intervals', the **more intervals the better** (provided that you check how visit and detection probability changes in these intervals)





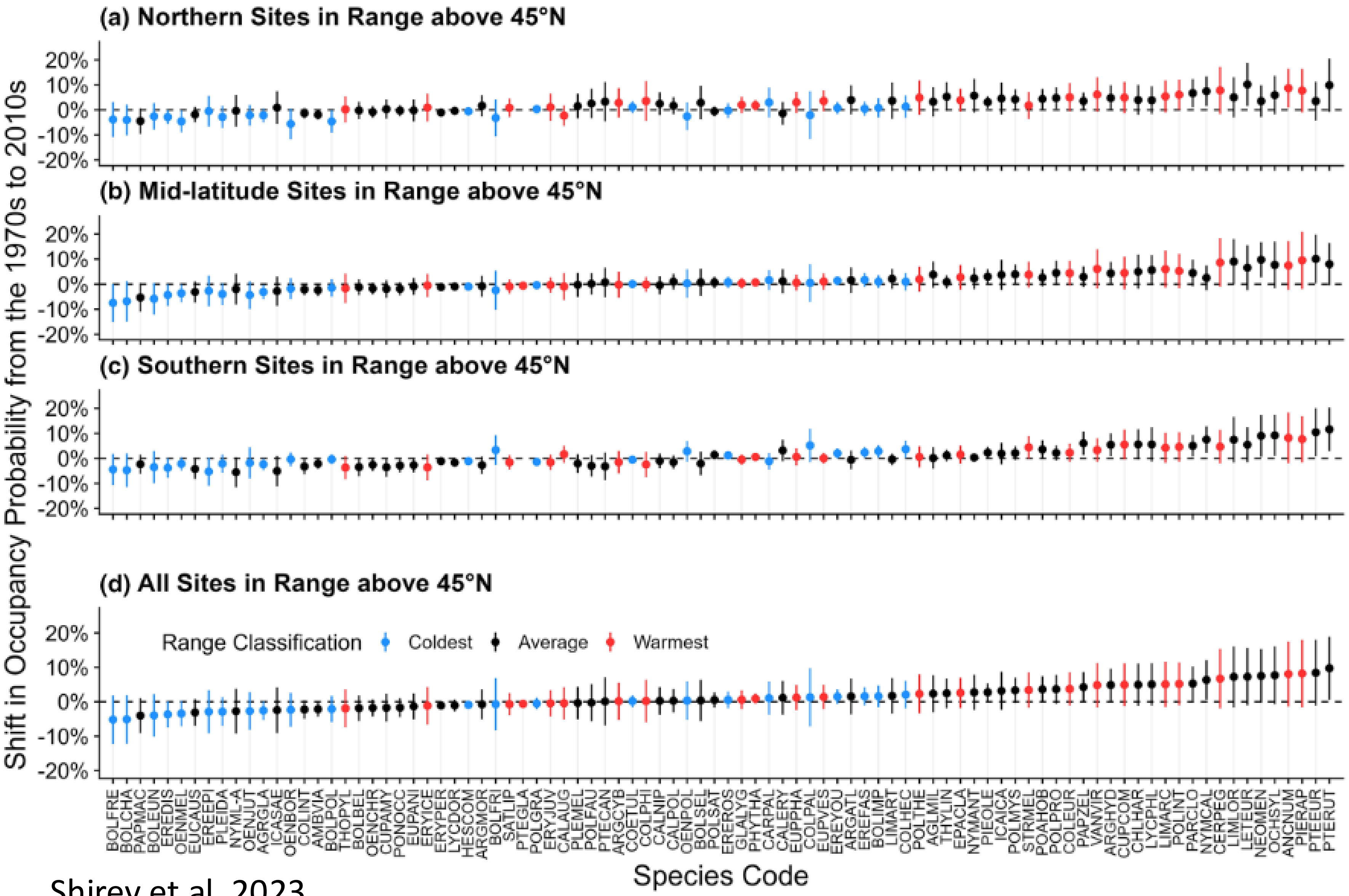
Multi-Species Occupancy Models: Case-study with Arctic or Boreal Butterflies



Shirey et al. 2023

Multi-Species Occupancy Models: Case-study

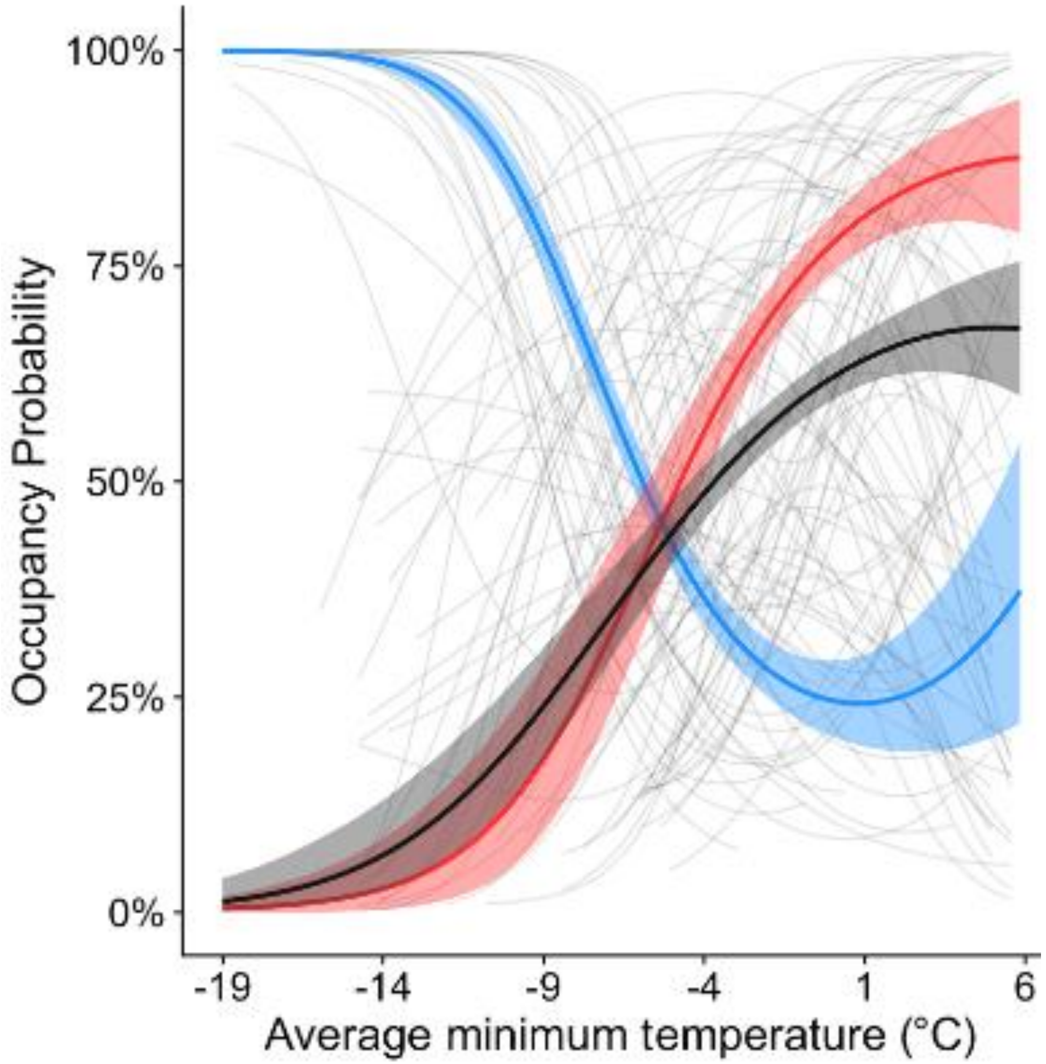
- Used Species Traits & Historical Temperature data as covariates in the model

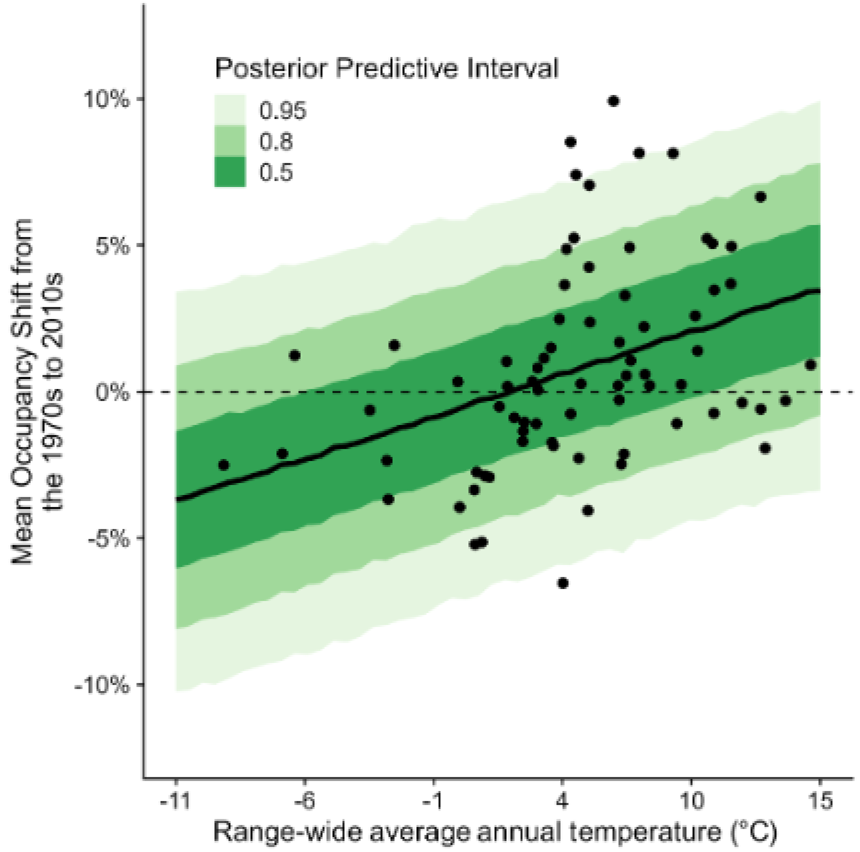


Shirey et al. 2023

Multi-Species Occupancy Models: Case-study

- Used Species Traits & Historical Temperature data as covariates in the model
- Range wide average for species is the best predictor for occupancy increase or decrease
- “winners & losers” of climate change





Shirey et al. 2023

So much to do!

- Occupancy modeling is generally rarer in plants than animals, and has yet to be published with this presence-only approach
- Endless amount of covariates & hypotheses to test in both what determines detection, and what influences occupancy of species
- Species Ranges (SDMs) are highly recommended.