

EXTREME-VALUE MODELLING OF MIGRATORY BIRD ARRIVAL DATES: INSIGHTS FROM CITIZEN SCIENCE DATA



Jonathan Koh
Joint work with Thomas Opitz

ETH zürich

CMStat 2025
London, 13/12/2025

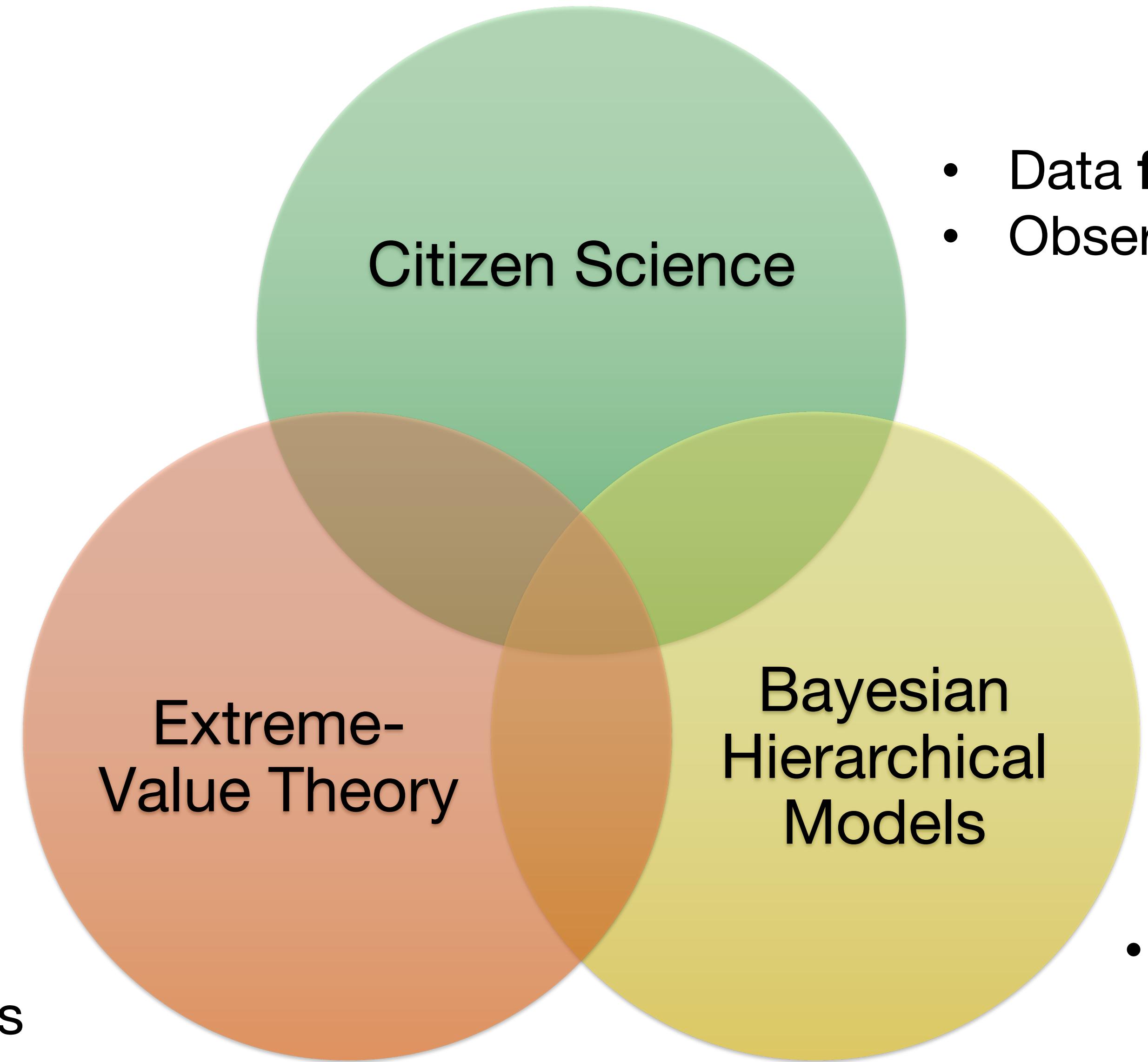
Environmental Stats (includes Ecological Stats)

Environmental Stats

Citizen Science

Extreme-
Value Theory

Bayesian
Hierarchical
Models



- Extremes of **phenological** events

- Data **fusion**
- Observational **bias**

Bayesian
Hierarchical
Models

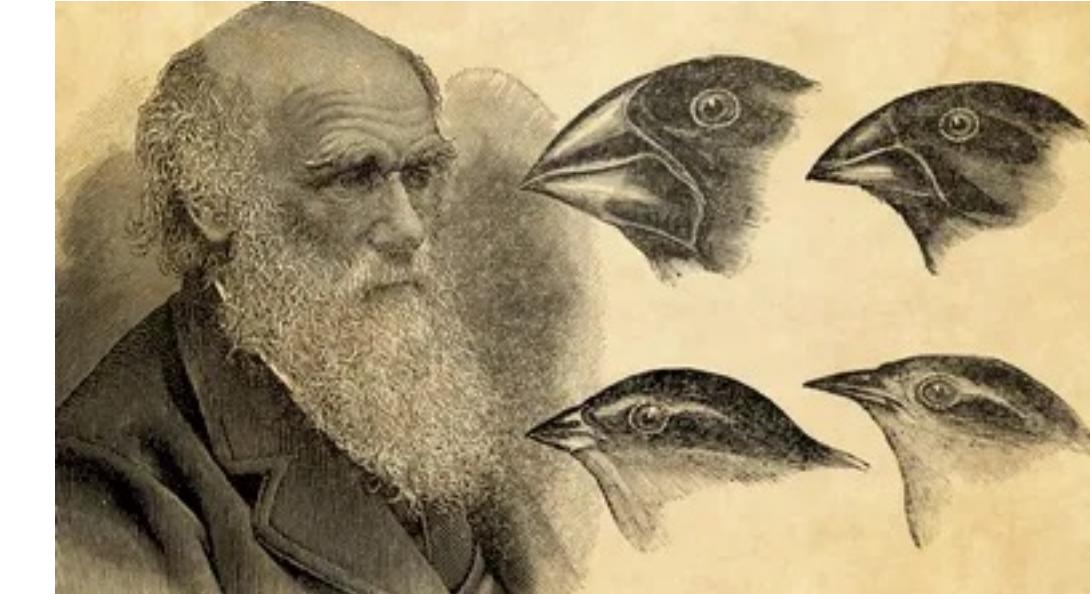
Extreme-
Value Theory

Citizen Science

- Inference on important **latent processes**

What is citizen science?

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new...



What is citizen science?

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new...



What is citizen science?

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new...
3. Form has changed, and significant **traction** over the last two decades

The H word
Science

Geoffrey Belknap

Tue 26 Apr 2016 08.45 CEST

f t m 6

SCIENCE GOSSIP A Zooniverse

HOME CLASSIFY PROFILE PERIODICALS ABOUT BLOG FEEDBACK

Uncover the history of citizen science

In the Victorian period, just like today, scientists and members of the public worked together to further scientific discovery. Before computers and cameras they had to draw what they saw. Their drawings are locked away in the pages of Victorian periodicals, such as *Science Gossip*, *Recreational Science* and *The Intellectual Observer*. Help us draw and map the origins of citizen science.

Get started!

The front page of the [Science Gossip](#) website, a Zooniverse Citizen Humanities project. Citizen science is a digital method, which has been applied to a range of big-data scientific problems. [The Zooniverse](#) is a key player in this; having first sought the help of the crowd in classifying [galaxies](#) almost a decade ago, it now boasts 47 different projects with well over a million users. The projects hosted on their site have been bringing to the forefront concerns over who

About

About Publications Our Team Acknowledgements Resources Contact Us

FAQ Highlights Book Mobile App Donate

What is the Zooniverse?

The Zooniverse is the world's largest and most popular platform for people-powered research. This research is made possible by volunteers – more than a million people around the world who come together to assist professional researchers. Our goal is to enable research that would not be possible, or practical, otherwise. Zooniverse research results in new discoveries, datasets useful to the wider research community, and [many publications](#).

At the Zooniverse, anyone can be a researcher

You don't need any specialised background, training, or expertise to participate in any Zooniverse projects. We make it easy for anyone to contribute to real academic research, on their own computer, at their own convenience.

You'll be able to study authentic objects of interest gathered by researchers, like images of faraway galaxies, historical records and diaries, or videos of animals in their natural habitats. By answering simple questions about them, you'll help contribute to our understanding of our world, our history, our Universe, and more.

With our wide-ranging and ever-expanding suite of projects, covering many disciplines and topics across the sciences and humanities, there's a place for anyone and everyone to explore, learn and have fun in the Zooniverse. To volunteer with us, just go to the [Projects](#) page, choose one you like the look of, and get started.

We accelerate important research by working together

The major challenge of 21st century research is dealing with the flood of information we can now collect about the world around us. Computers can help, but in many fields the human ability for pattern recognition – and our ability to be surprised – makes us superior. With the help of Zooniverse volunteers, researchers can analyze their information more quickly and accurately than would otherwise be possible, saving time and resources, advancing the ability of computers to do the same tasks, and leading to faster progress and understanding of the world, getting to exciting results more quickly.

What is citizen science?

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new...
3. Form has changed, and significant **traction** over the last two decades
4. **Government** agencies are doing it



6 ways to be a Citizen Scientist

TRACK THE TIDES Report local water levels and flood impacts to help NOAA better understand and communicate about future floods.	WATCH FOR WHALES Share your whale sightings so scientists can track their population trends.	GEOCACHE FOR A GOOD CAUSE Gather field notes, photos, and GPS data at bench marks for location and height data.
MONITOR MARINE DEBRIS Record the type and amount of debris on your beach to help scientists tackle the challenge of marine debris.	FIGHT HARMFUL ALGAL BLOOMS Collect water quality data that helps NOAA respond to harmful algal blooms.	BE A SANCTUARY STEWARD Pitch in at a local marine sanctuary or estuarine research reserve.

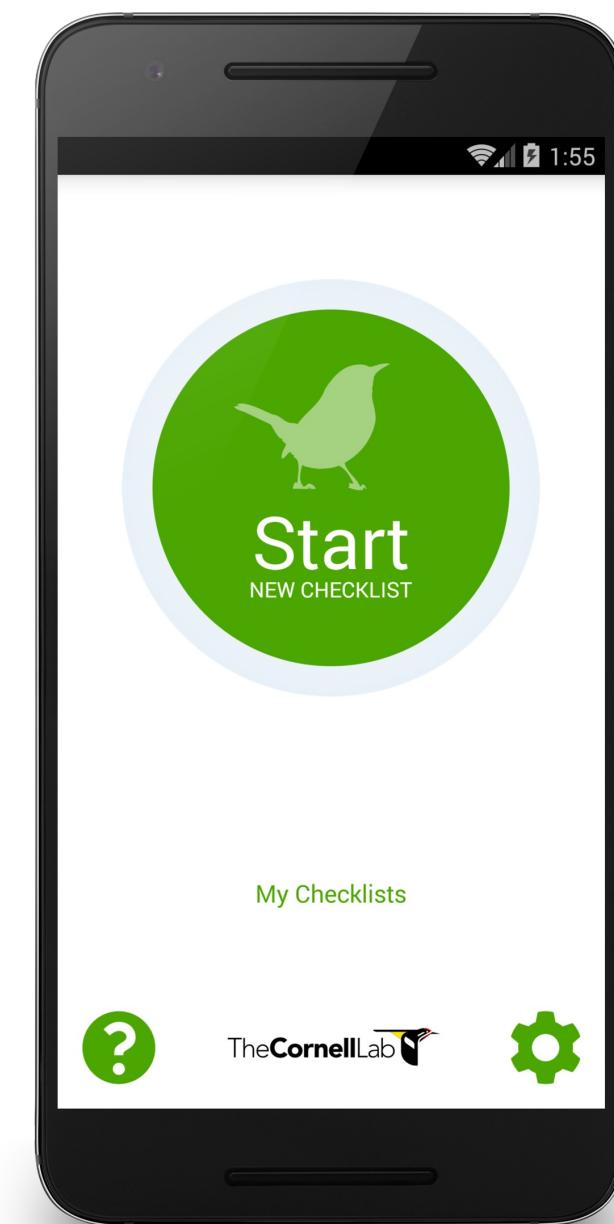
oceanservice.noaa.gov/citizen-science

How does eBird work?

1. Simone is a **birder**
2. While hiking, he opens the eBird app and starts a '**checklist**'. The app notes the **date** and **time** he starts birding, **where** he has travelled during the checklist and **how long** he has been birding

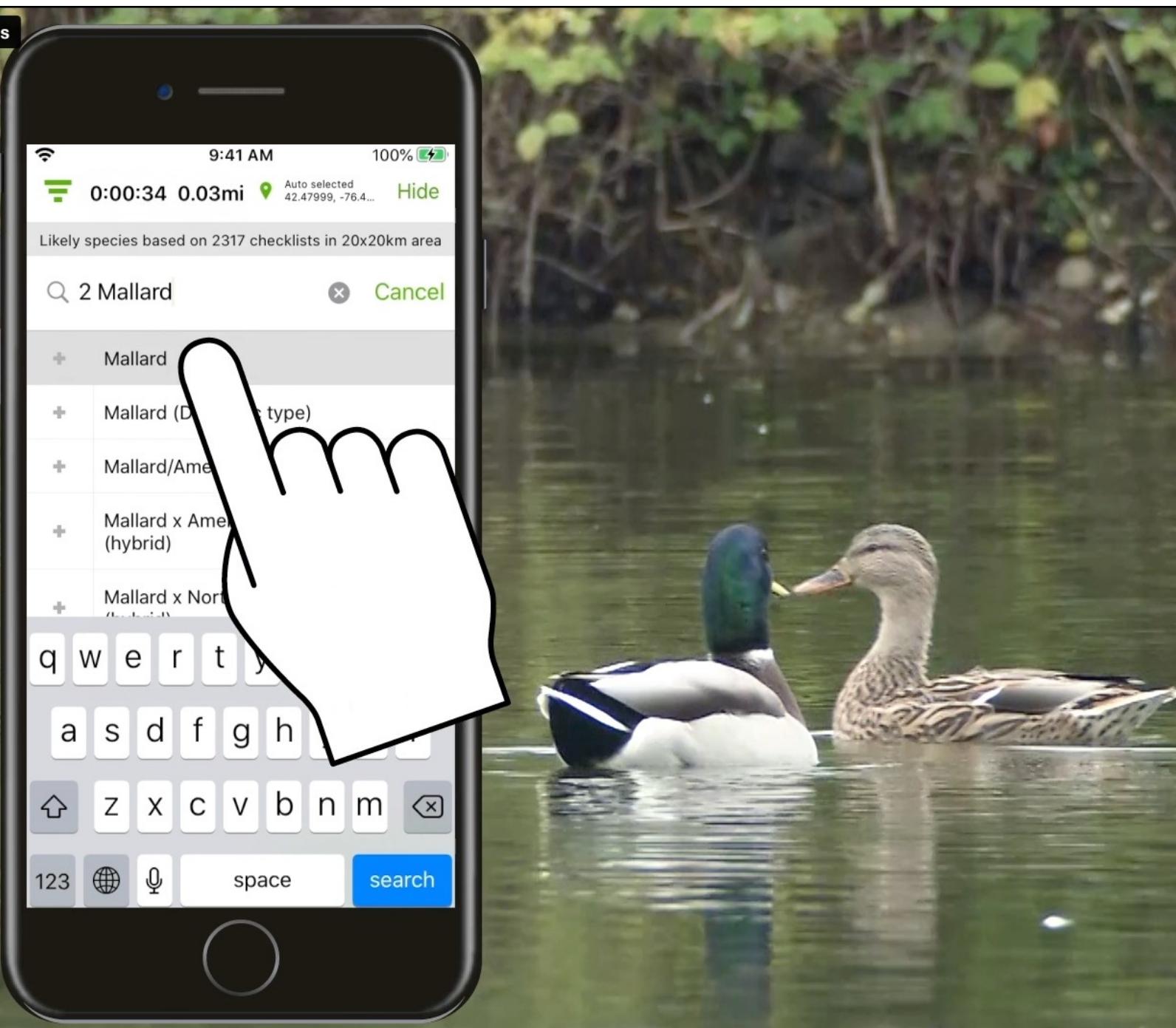
How does eBird work?

1. Simone is a **birder**
2. While hiking, he opens the eBird app and starts a '**checklist**'. The app notes the **date** and **time** he starts birding, **where** he has travelled during the checklist and **how long** he has been birding



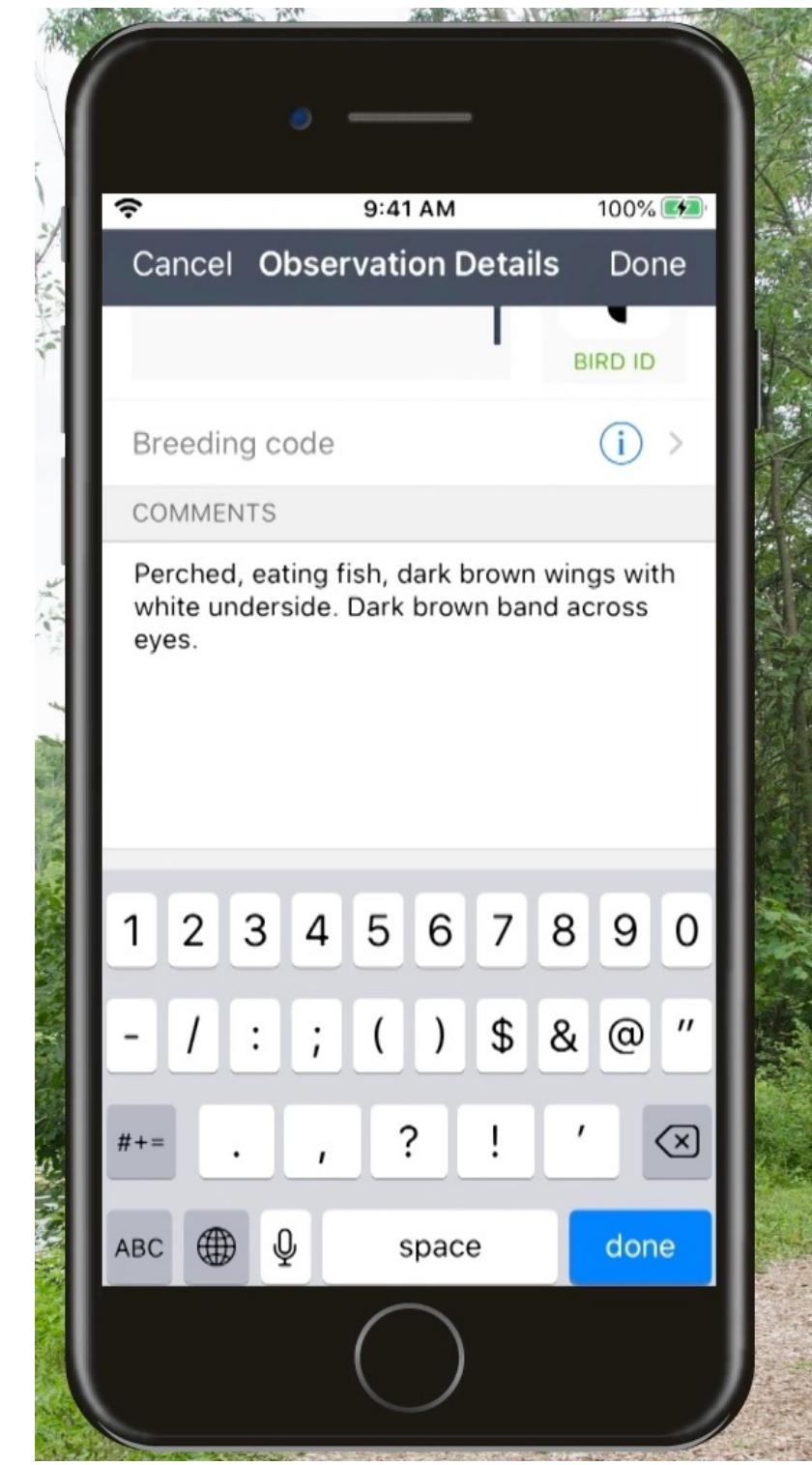
How does eBird work?

- He spots a Mallard, and can easily **record** it in the app (based on a pack of recommended bird species)



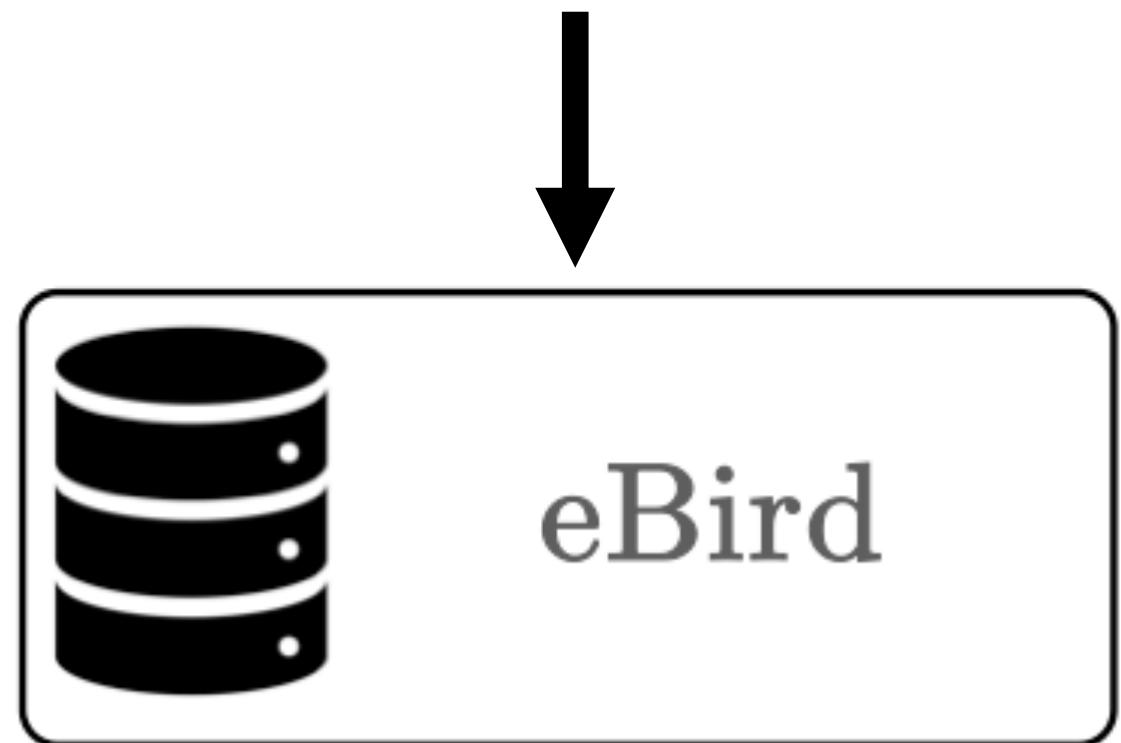
How does eBird work?

- He spots an Osprey, and records it in the app. He can also supply more information (including media files)



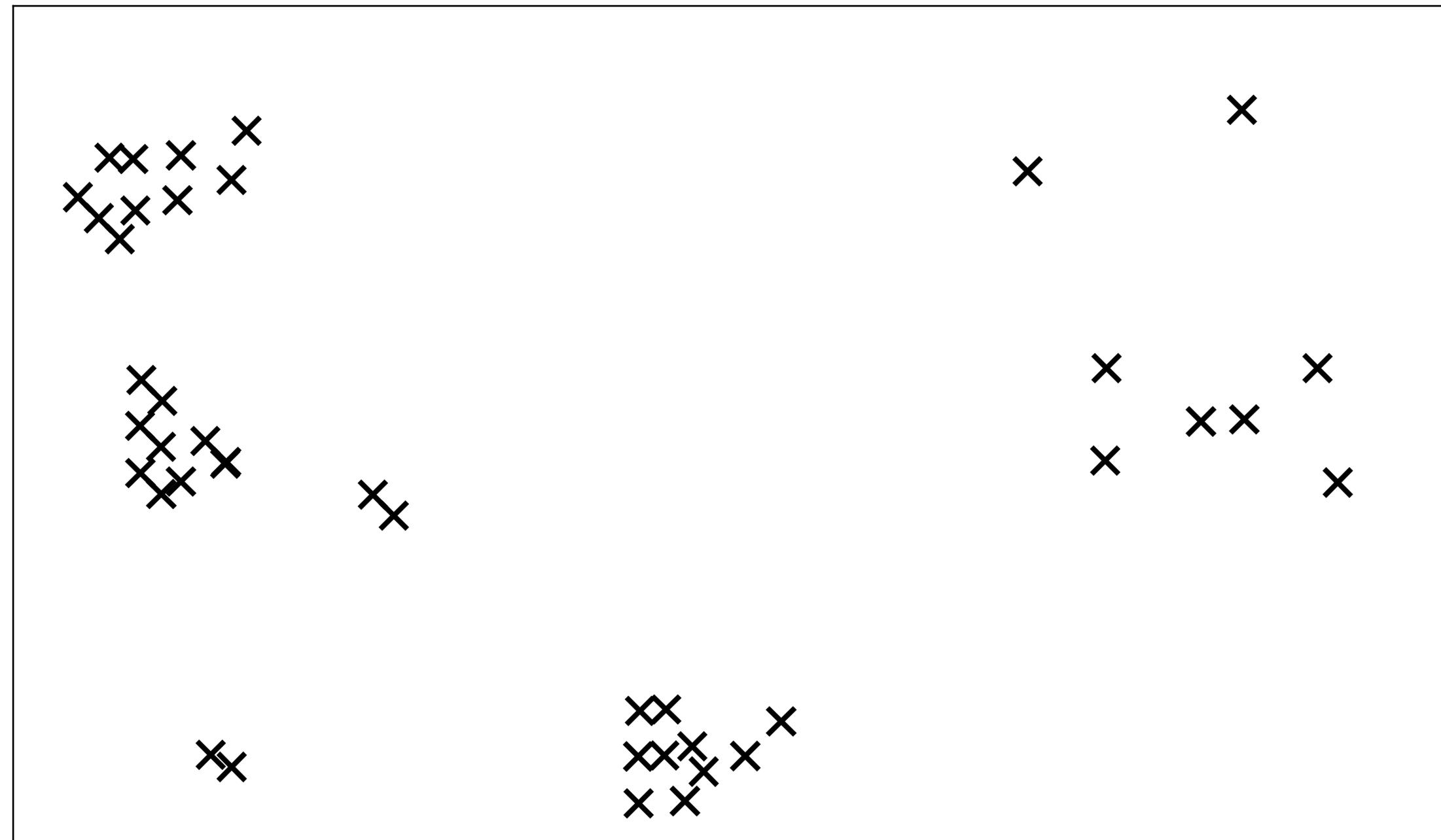
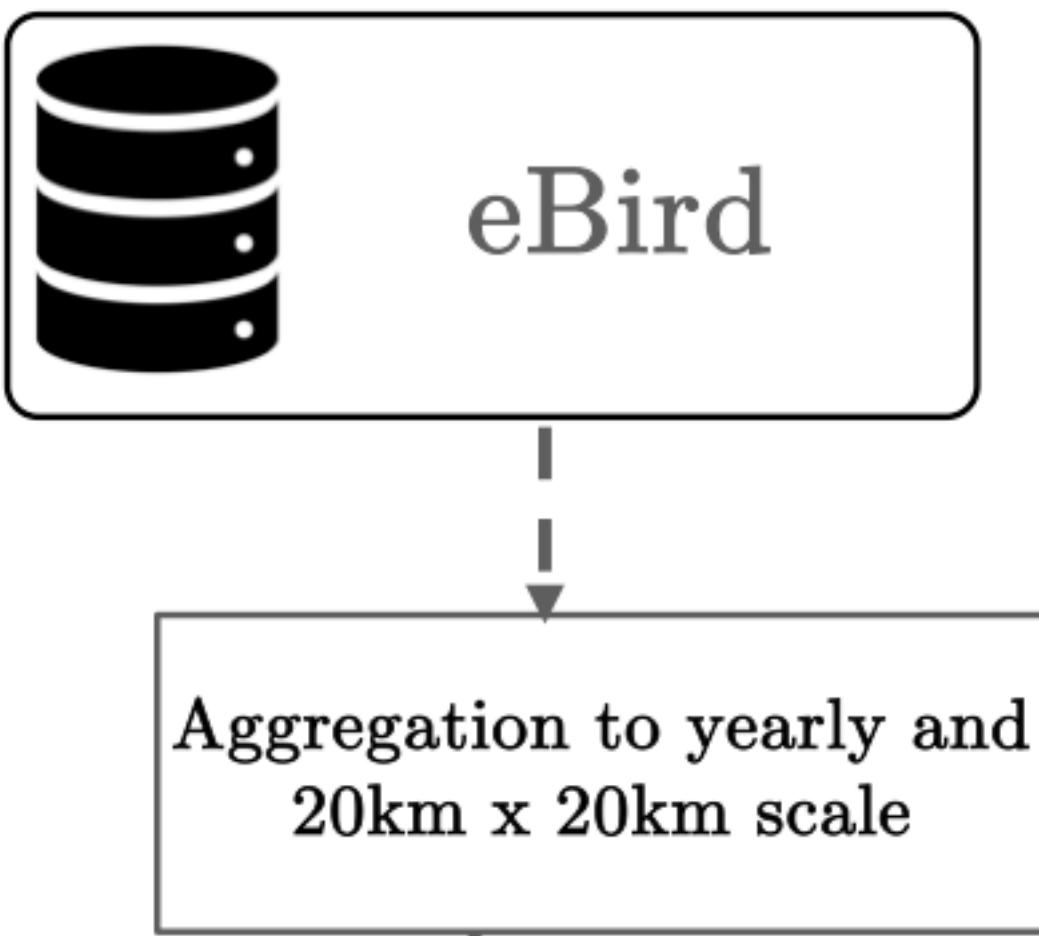
How does eBird work?

5. He finishes his checklist and submits his data to eBird
6. eBird internally verifies it, and it goes into their database



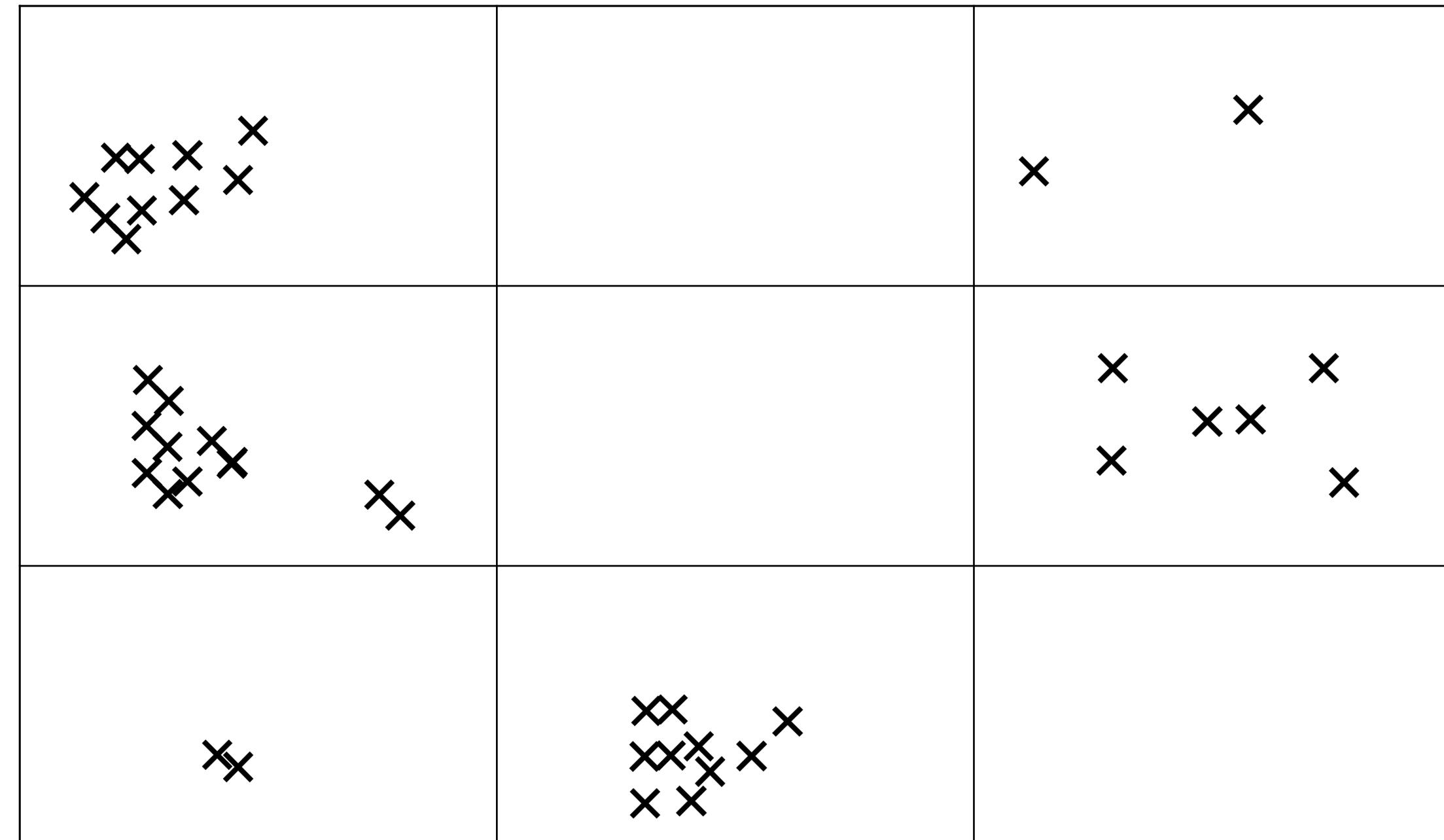
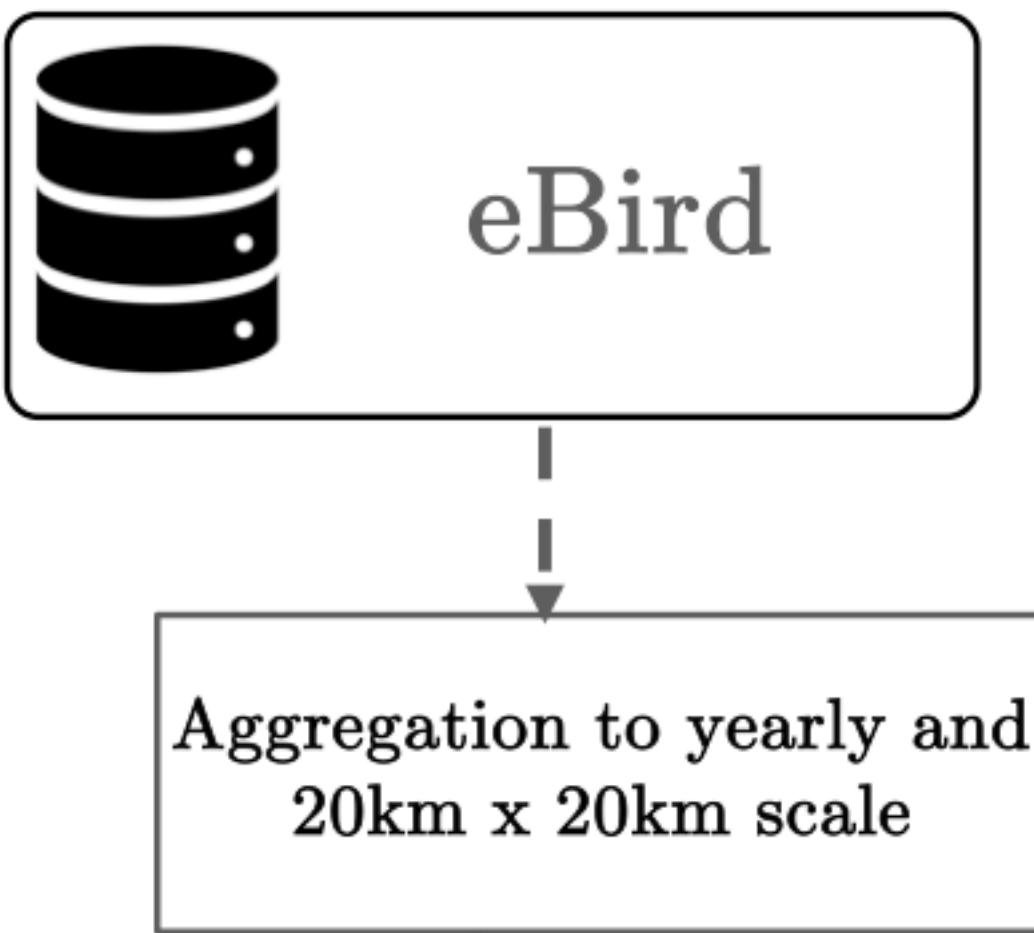
eBird data processing

Year 2020

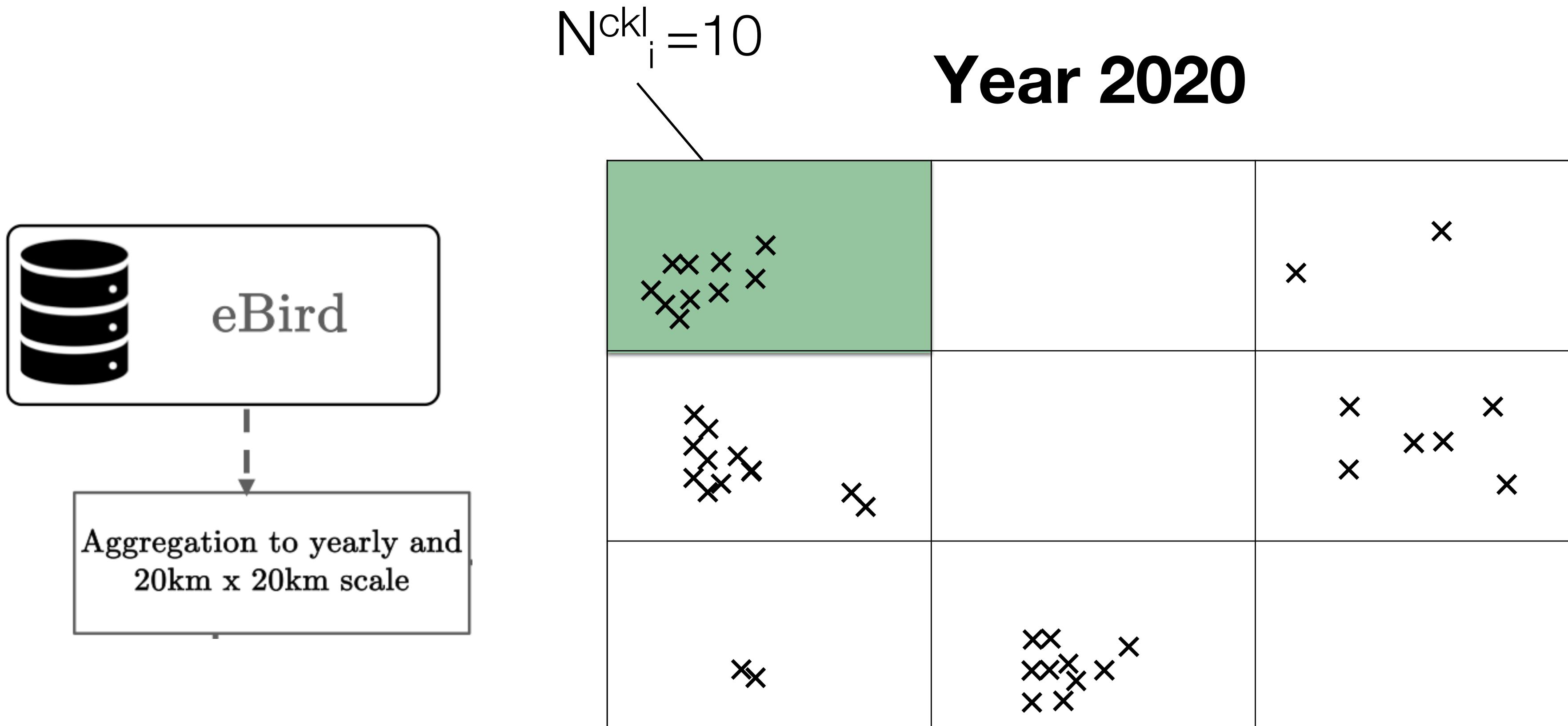


eBird data processing

Year 2020



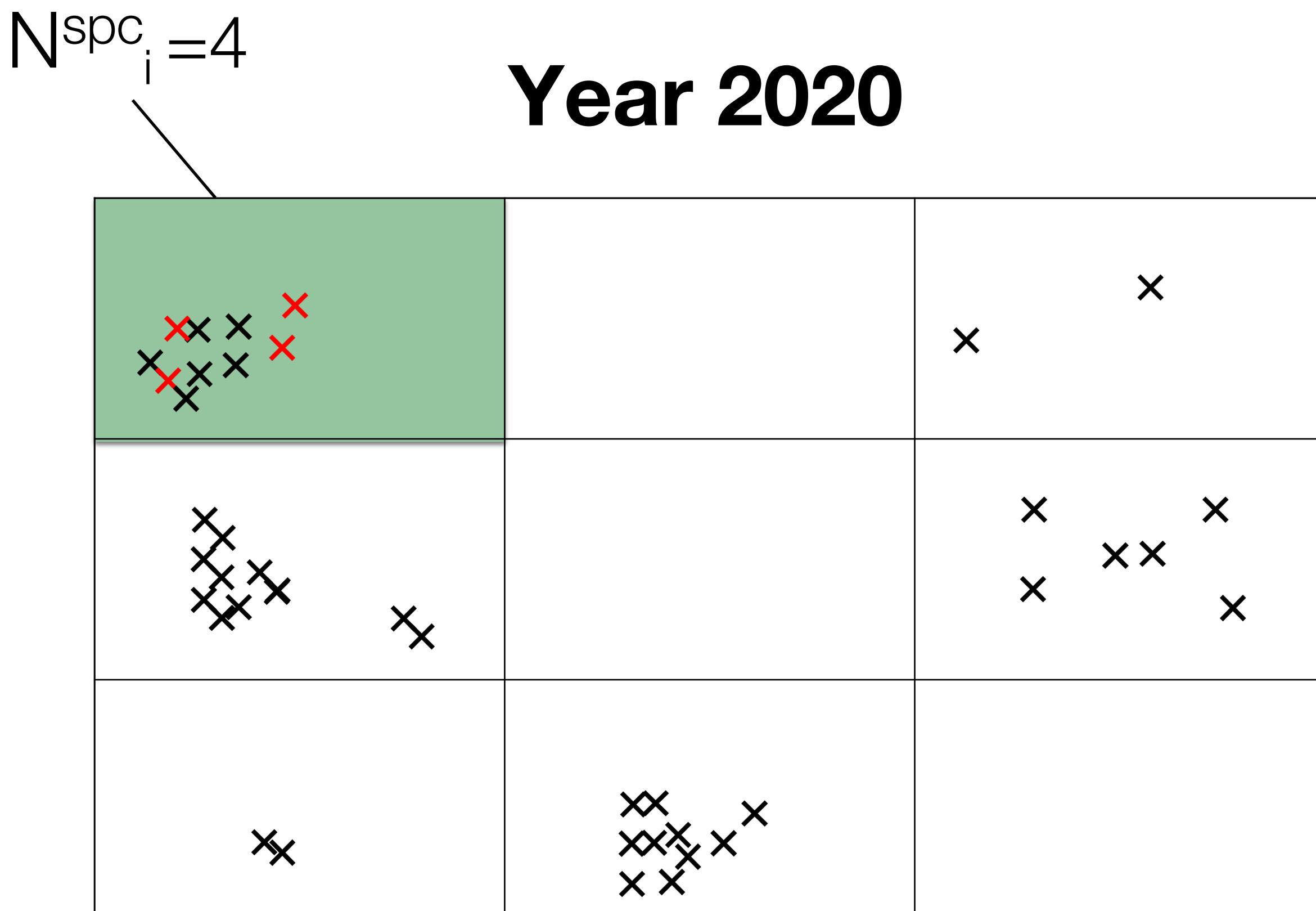
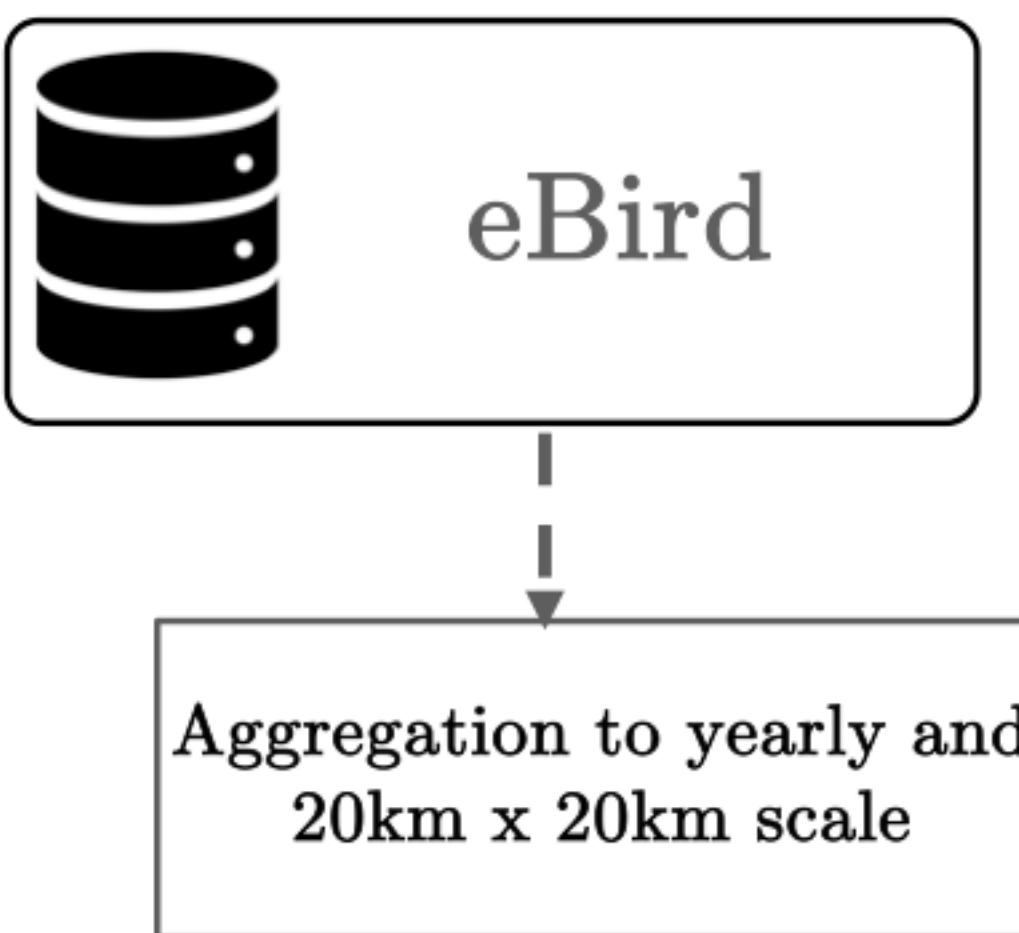
eBird data processing



eBird data processing



Purple Martin



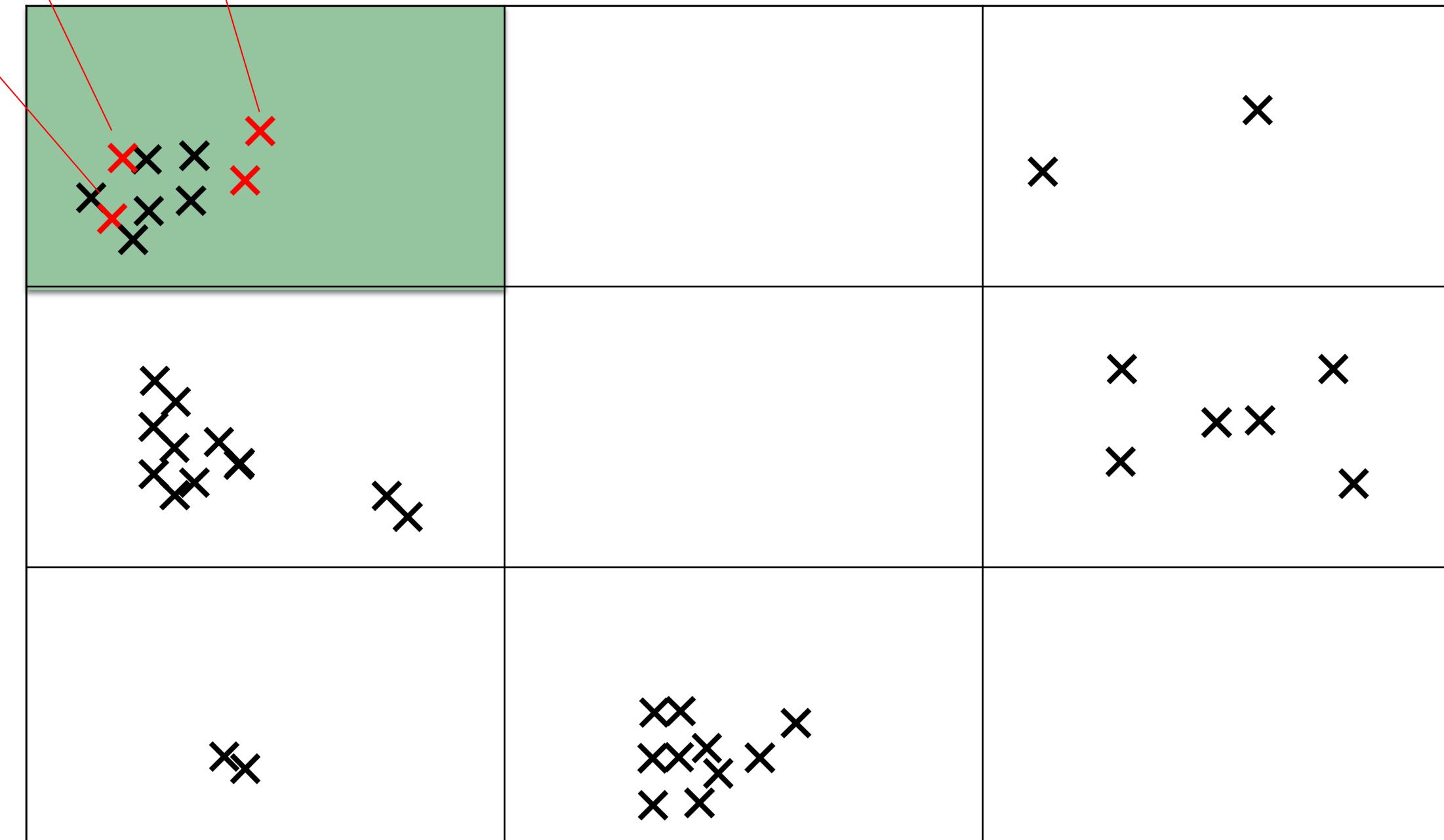
eBird data processing



Purple Martin



eBird

Aggregation to yearly and
20km x 20km scale $Y_{i,2} = 02/04/2020$ $Y_{i,1} = 28/03/2020$ $Y_{i,3} = 05/04/2020$ **Year 2020**

eBird data processing



Z'_i = First arrival date: 28/03/20

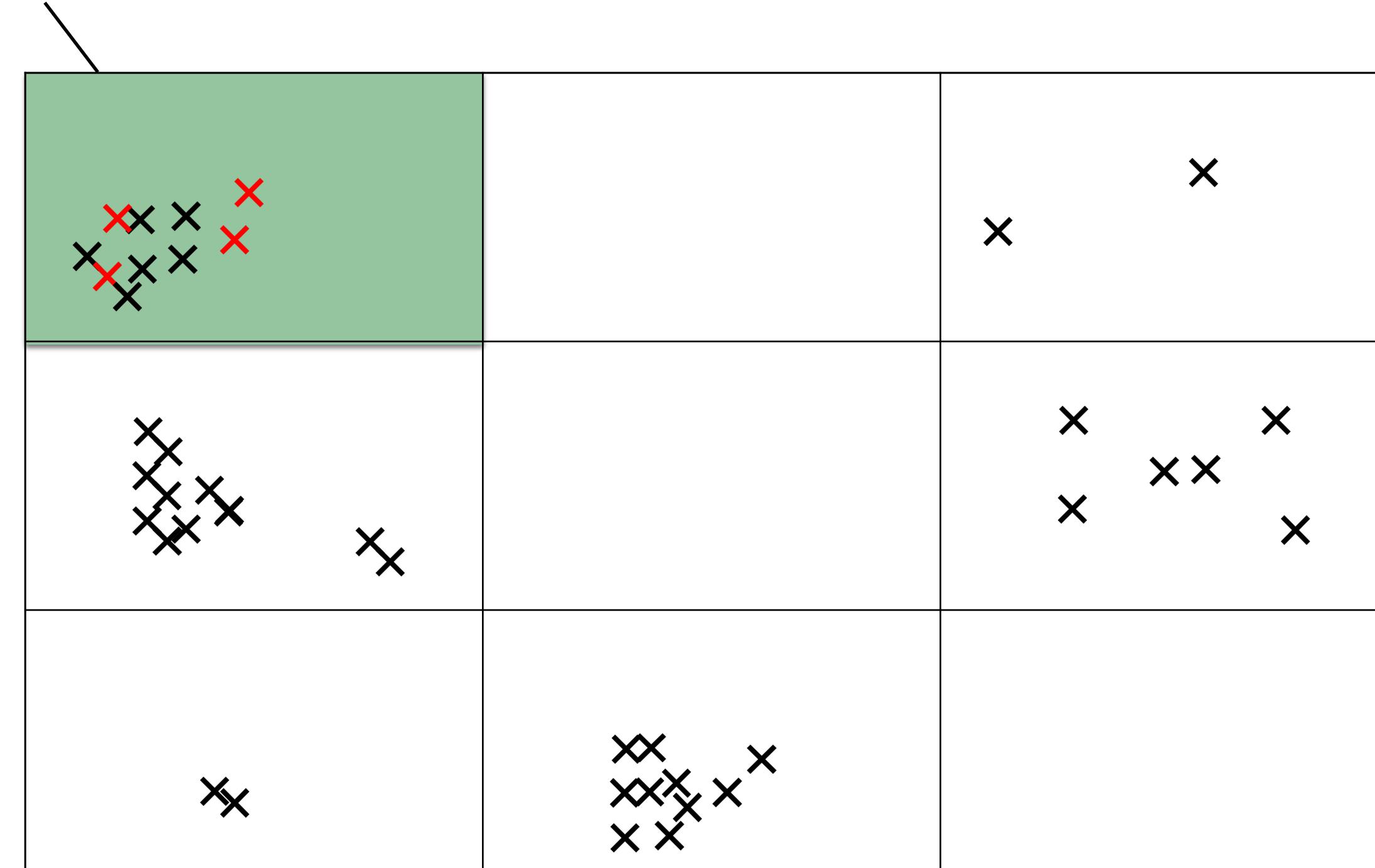
Purple Martin



eBird

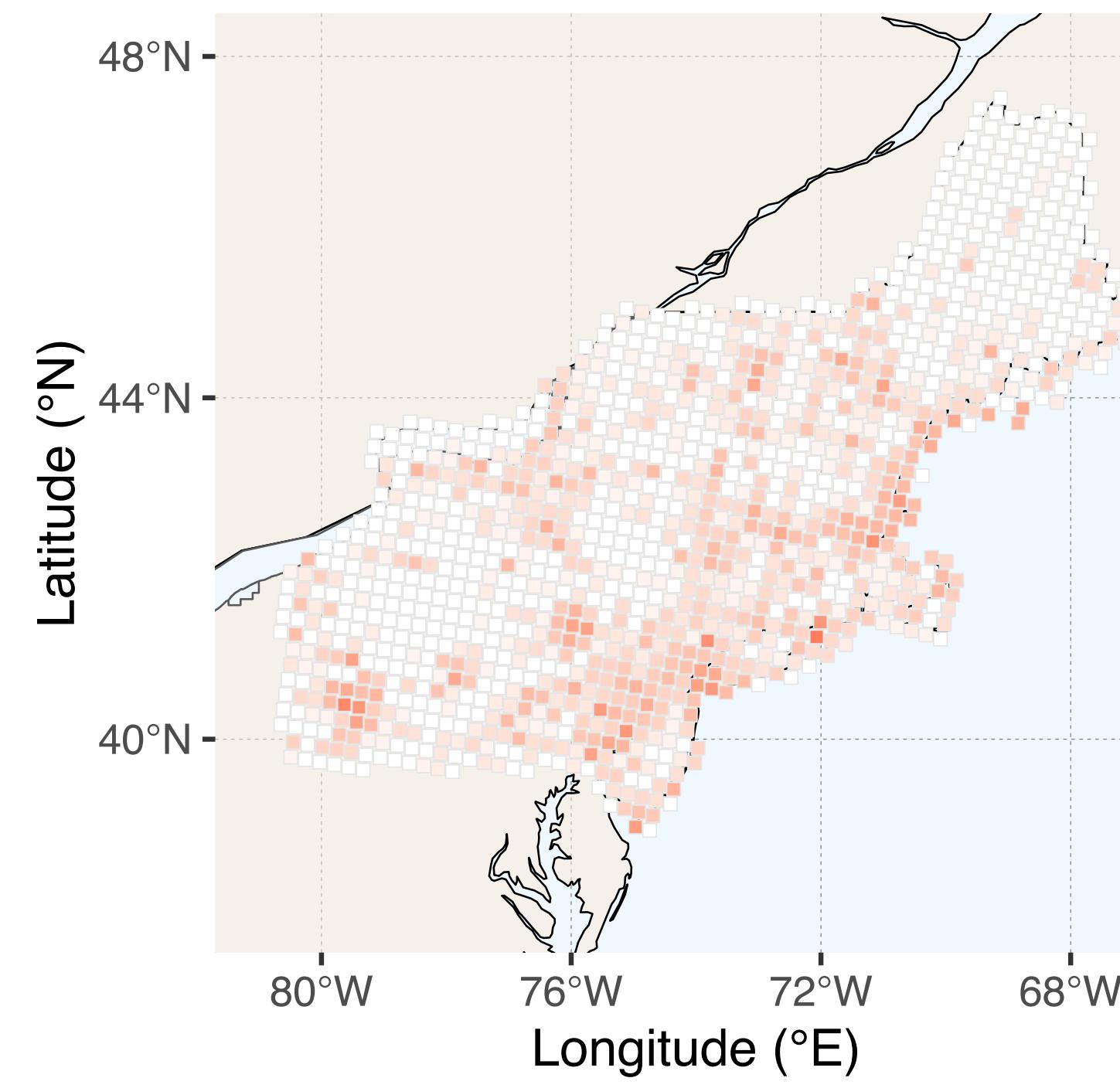
Aggregation to yearly and
20km x 20km scale

Year 2020

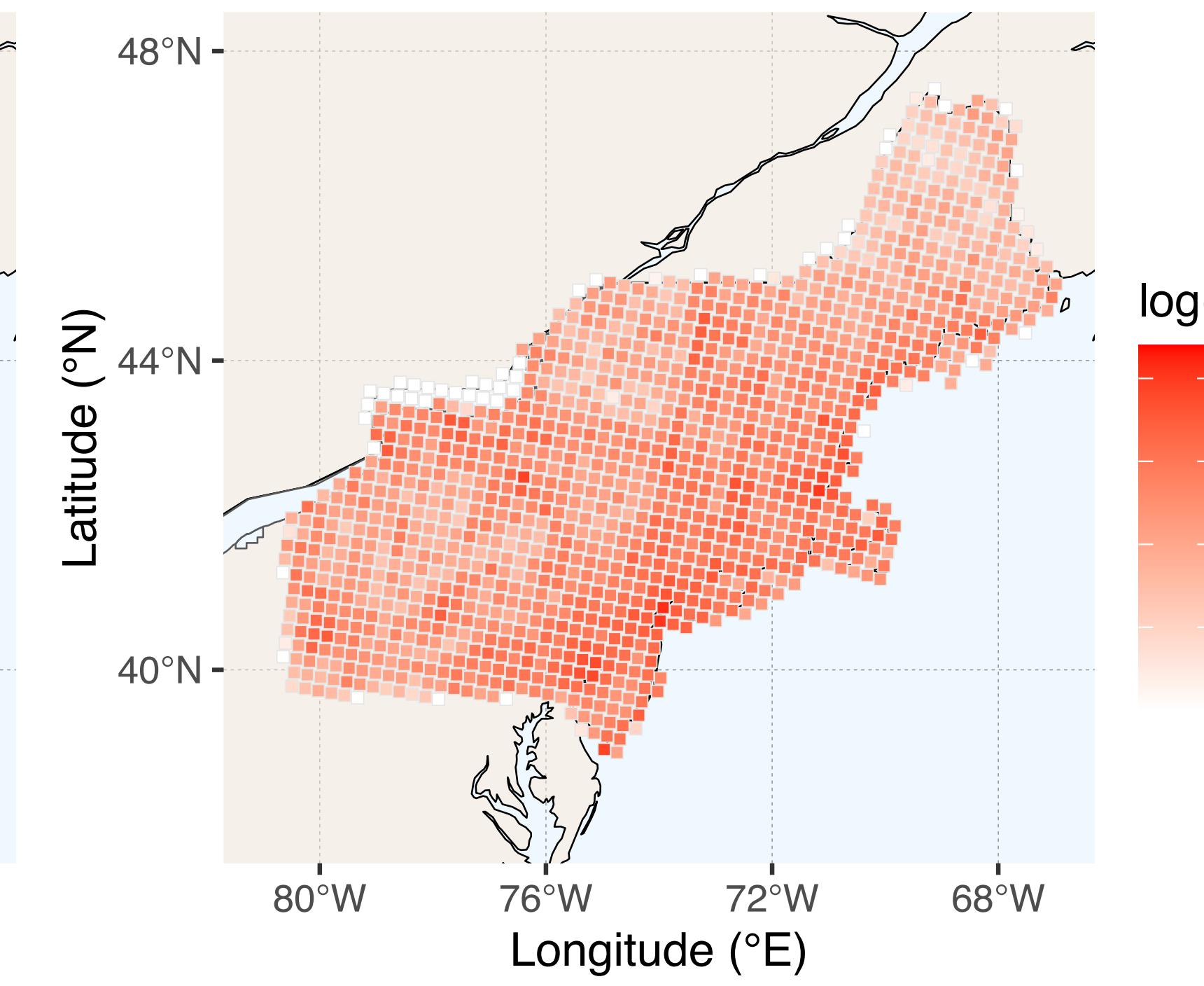


*Wijeyakulasuriya et al. (2024)

Strong temporal trends in reported occurrences

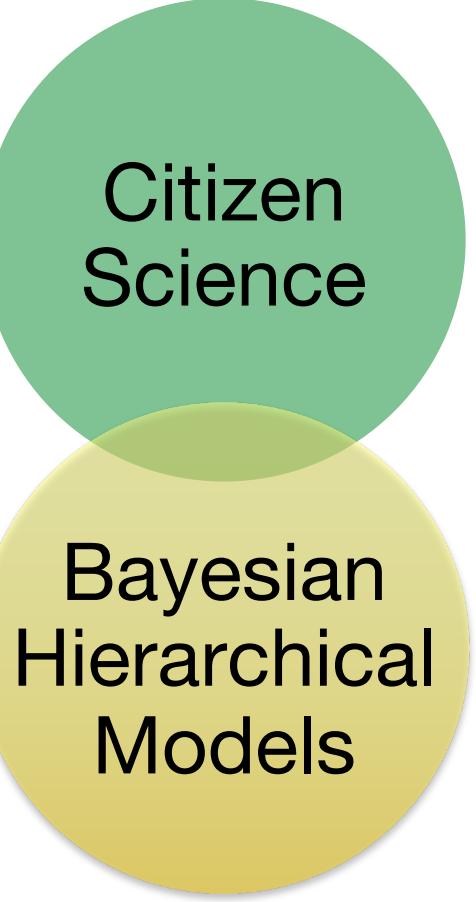


2001

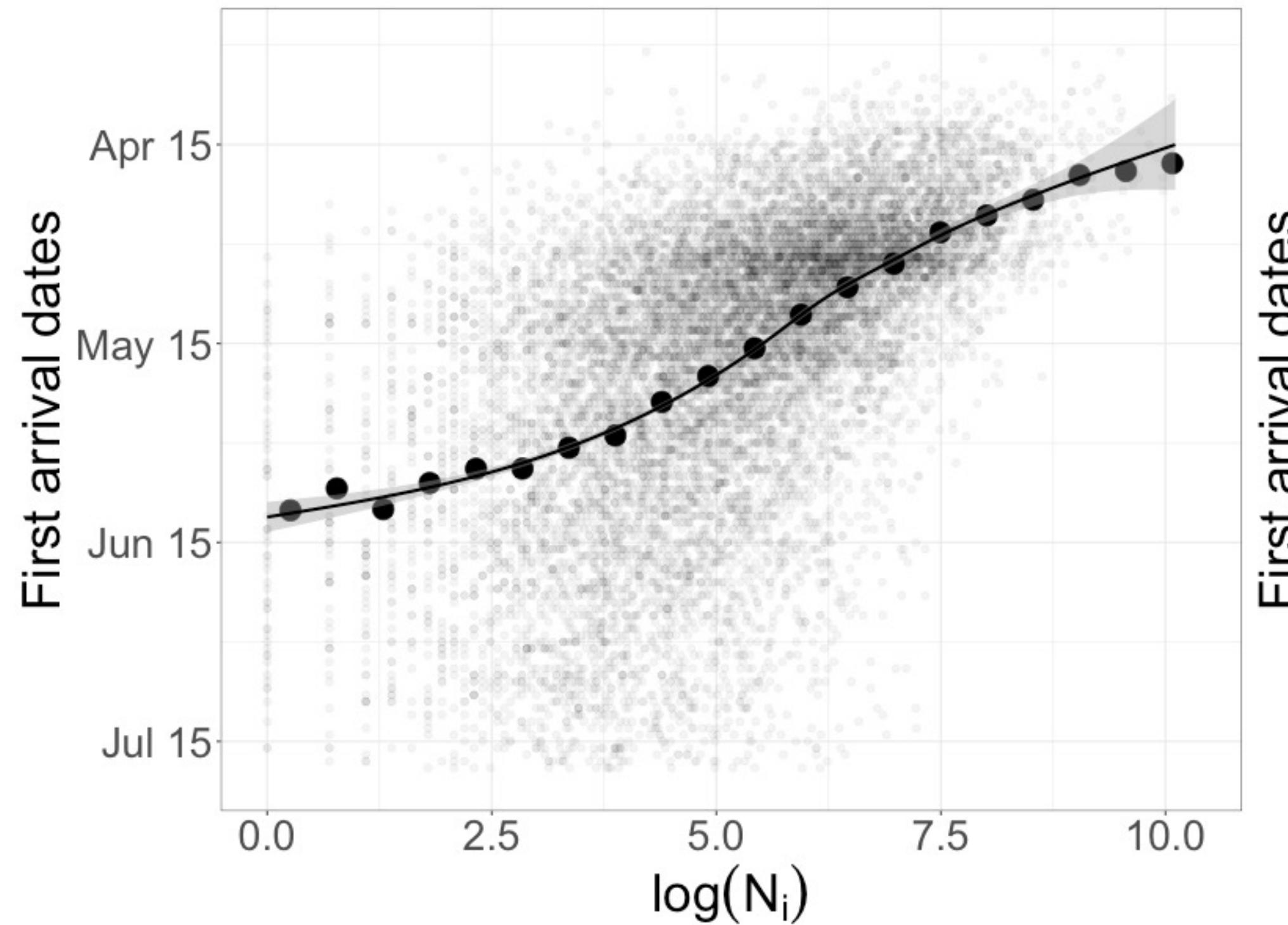


2021

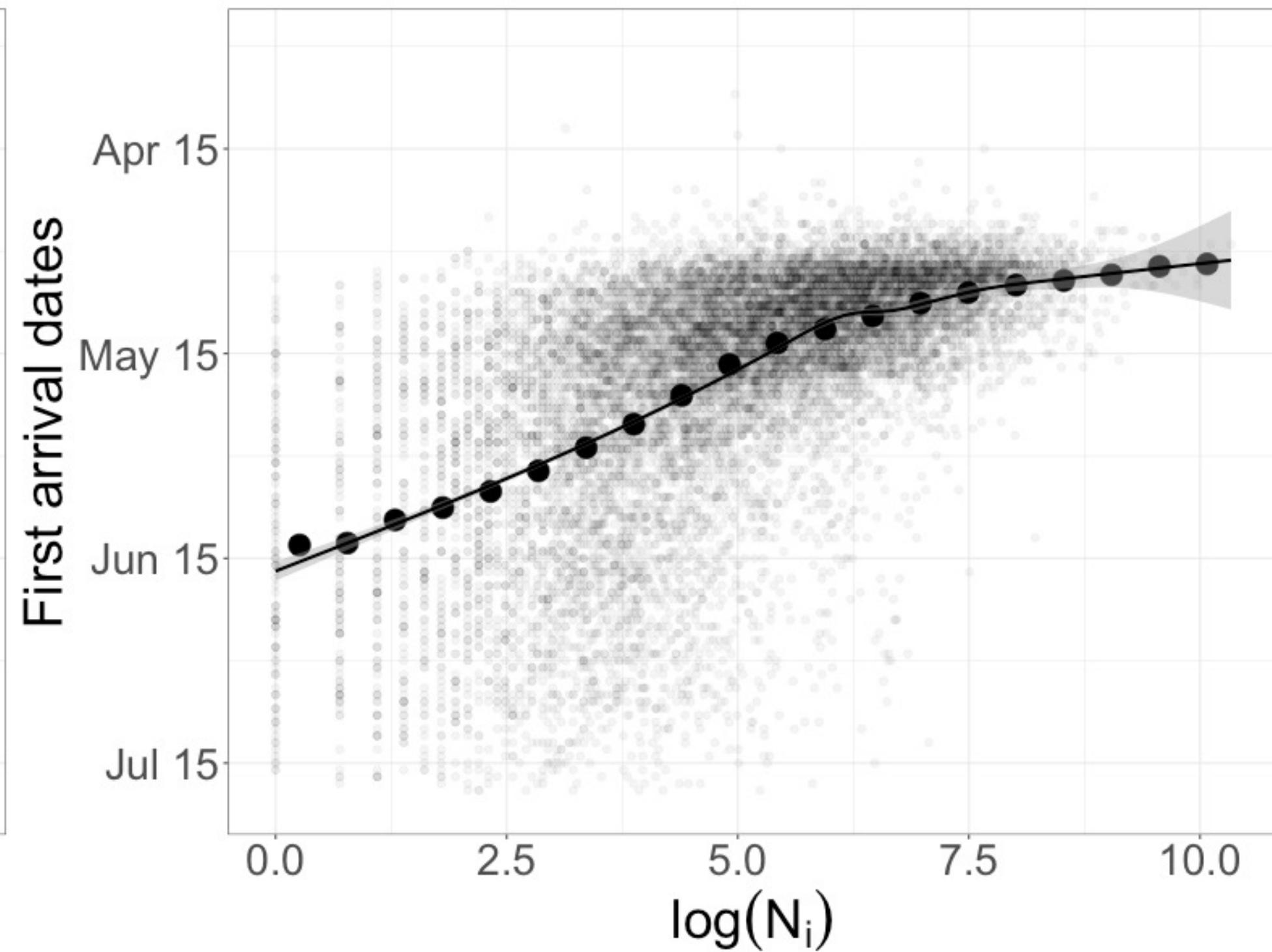
First arrival dates vs. checklist counts



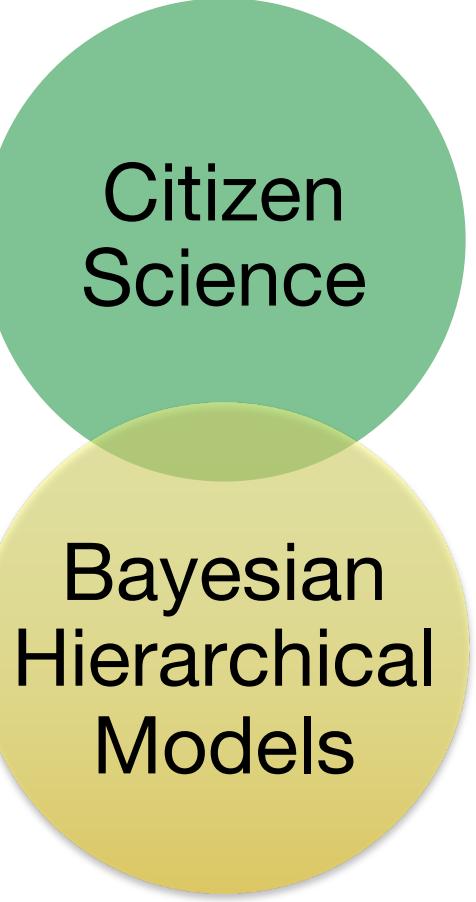
Chimney-Swift



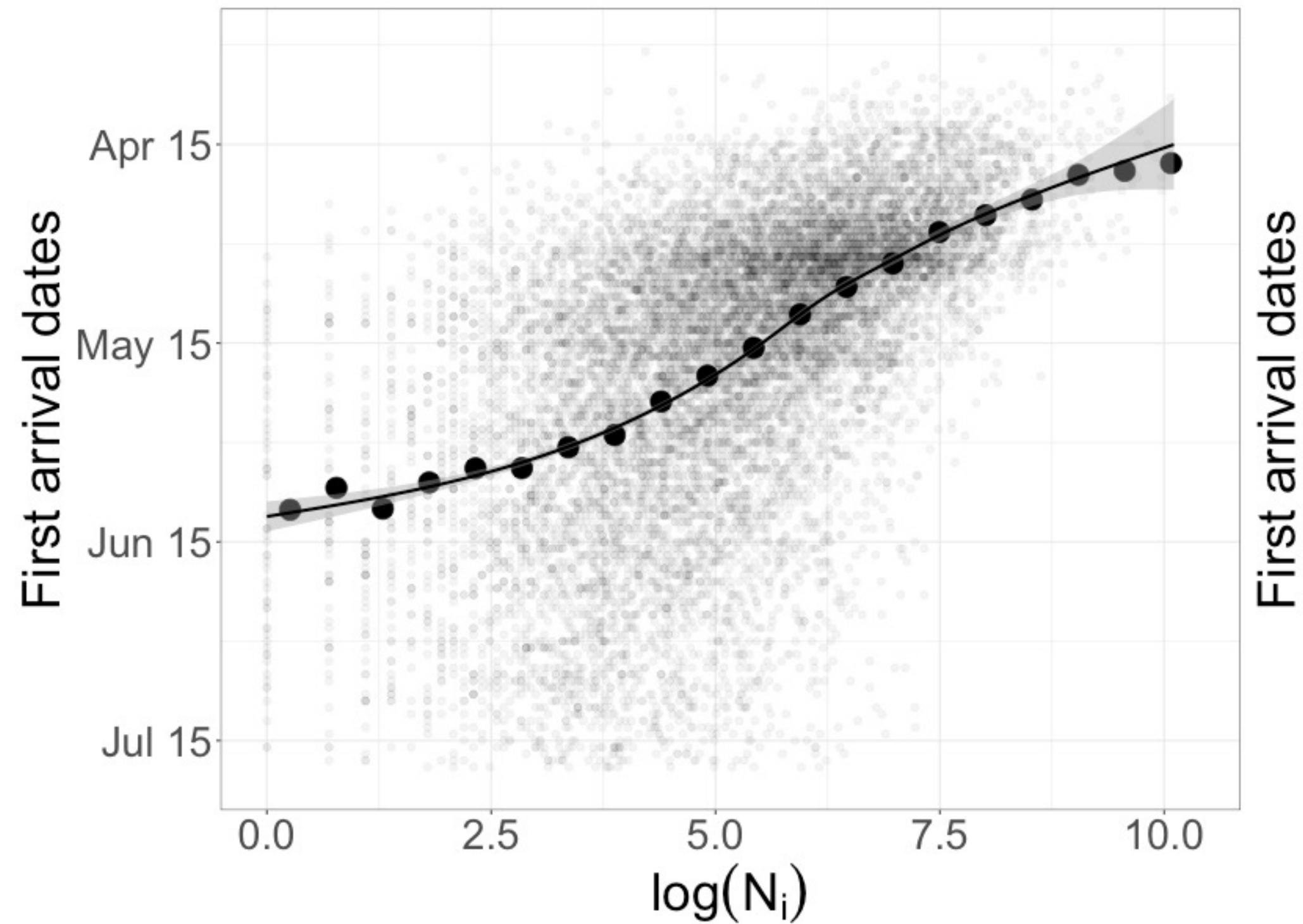
**Chestnut-sided
Warbler**



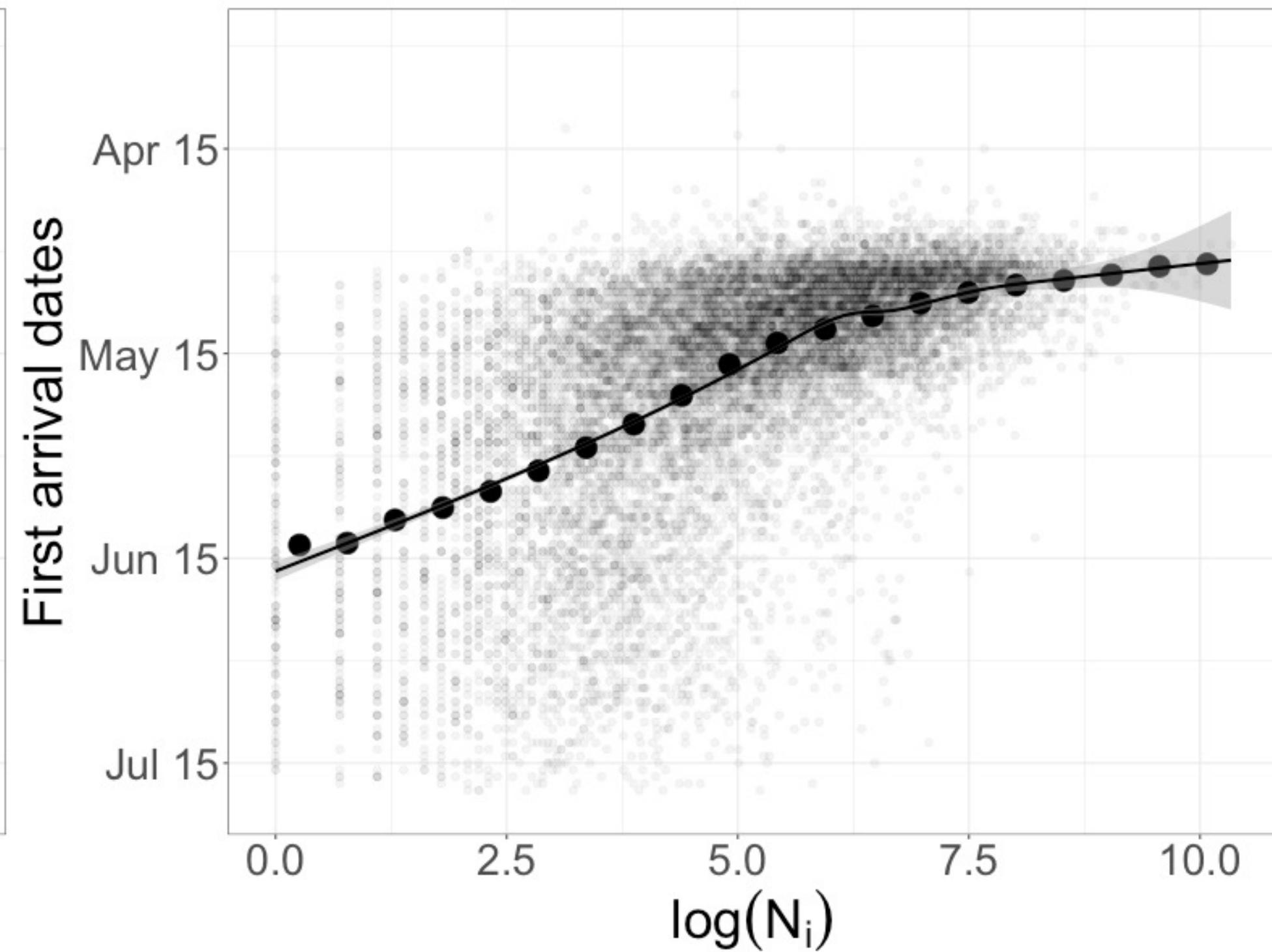
“Observational effort”/ Preference



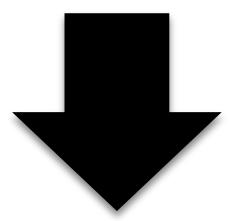
Chimney-Swift



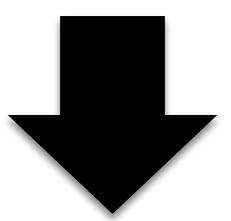
Chestnut-sided Warbler



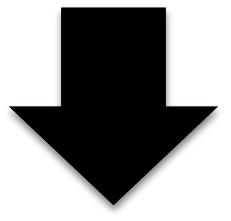
Observational effort = Preference + Activity



space-time
varying



captured by the
sampling intensity
for the checklists

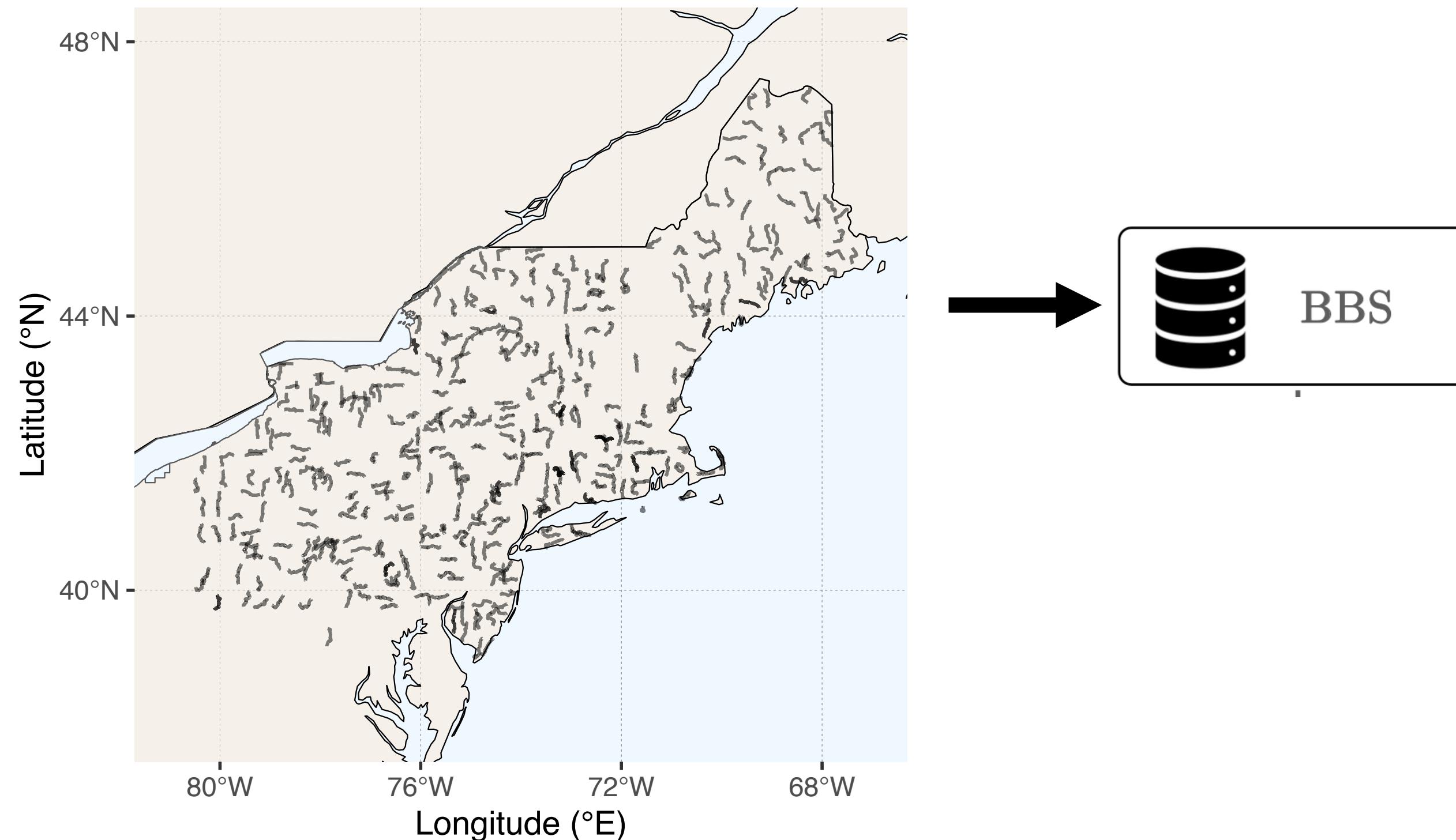


captured by the
(median) time spent
on the checklist

Breeding Bird Survey (BBS) sampling routes

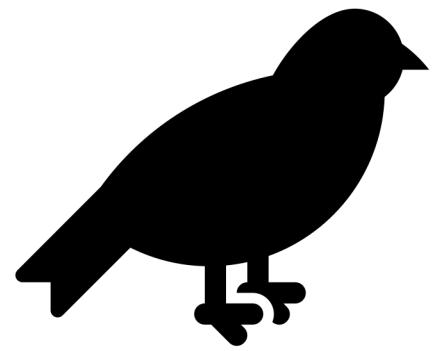


- For each route (~40km), bird occurrences are reported at **50 equidistant stops**
- Only every **summer**
- Complex data preprocessing (missing observations, **missing stop coordinates**, etc.)





MODEL



Modelling goals

- Fit a realistic model to first arrival data, conditional on covariates
- Correct for the observational bias from these datasets

Modelling goals

- Fit a realistic model to first arrival data, conditional on covariates
- Correct for the observational bias from these datasets
- Use the model to make posterior predictions
- Interpolate spatially to locations not visited, in a reasonable way

A multi-response spatial regression system

Multi-response spatial regression

$$N_j^{\text{BBS}} \mid \lambda^{\text{BBS}}, \boldsymbol{\theta}_{\text{bbs}} \sim \text{Pois} \left\{ \sum_{k \in \text{route}_j} \omega_k \lambda^{\text{BBS}}(\mathbf{s}_k; \boldsymbol{\theta}_{\text{bbs}}) \right\},$$

$$N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \boldsymbol{\theta}_{\text{ckl}} \sim \text{Pois} \left\{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \boldsymbol{\theta}_{\text{ckl}}) \right\},$$

$$N_i^{\text{spc}} \mid N_i^{\text{ckl}}, p^{\text{spc}}, \boldsymbol{\theta}_{\text{spc}} \sim \text{Bin}\{N_i^{\text{ckl}}, p^{\text{spc}}(\mathbf{s}_i, t_i; \boldsymbol{\theta}_{\text{spc}})\},$$

$$Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \text{GEV}\{\mu(\mathbf{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\mathbf{s}_i; \boldsymbol{\theta}_\sigma), \xi\},$$

where

$\boldsymbol{\theta}_{\text{bbs}}, \boldsymbol{\theta}_{\text{ckl}}, \boldsymbol{\theta}_{\text{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \text{Hyperpriors}$

A multi-response spatial regression system

Multi-response spatial regression



BBS



eBird

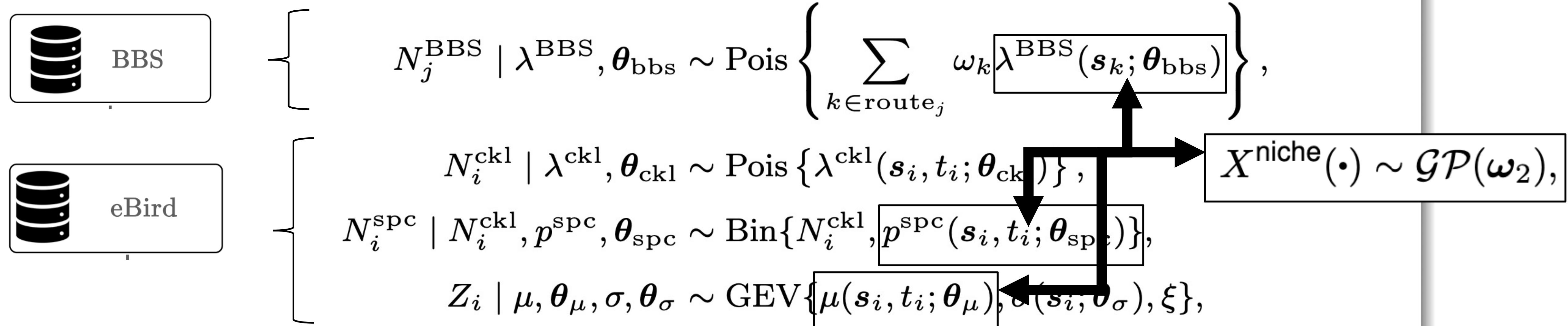
$$\left[\begin{array}{l} N_j^{\text{BBS}} \mid \lambda^{\text{BBS}}, \boldsymbol{\theta}_{\text{bbs}} \sim \text{Pois} \left\{ \sum_{k \in \text{route}_j} \omega_k \lambda^{\text{BBS}}(\mathbf{s}_k; \boldsymbol{\theta}_{\text{bbs}}) \right\}, \\ \\ N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \boldsymbol{\theta}_{\text{ckl}} \sim \text{Pois} \left\{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \boldsymbol{\theta}_{\text{ckl}}) \right\}, \\ N_i^{\text{spc}} \mid N_i^{\text{ckl}}, p^{\text{spc}}, \boldsymbol{\theta}_{\text{spc}} \sim \text{Bin}\{N_i^{\text{ckl}}, p^{\text{spc}}(\mathbf{s}_i, t_i; \boldsymbol{\theta}_{\text{spc}})\}, \\ Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \text{GEV}\{\mu(\mathbf{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\mathbf{s}_i; \boldsymbol{\theta}_\sigma), \xi\}, \end{array} \right]$$

where

$\boldsymbol{\theta}_{\text{bbs}}, \boldsymbol{\theta}_{\text{ckl}}, \boldsymbol{\theta}_{\text{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \text{Hyperpriors}$

Sharing random effects

Multi-response spatial regression



where

$\theta_{\text{bbs}}, \theta_{\text{ckl}}, \theta_{\text{spc}}, \theta_\mu, \theta_\sigma \sim \text{Hyperpriors}$

Sharing random effects

Multi-response spatial regression

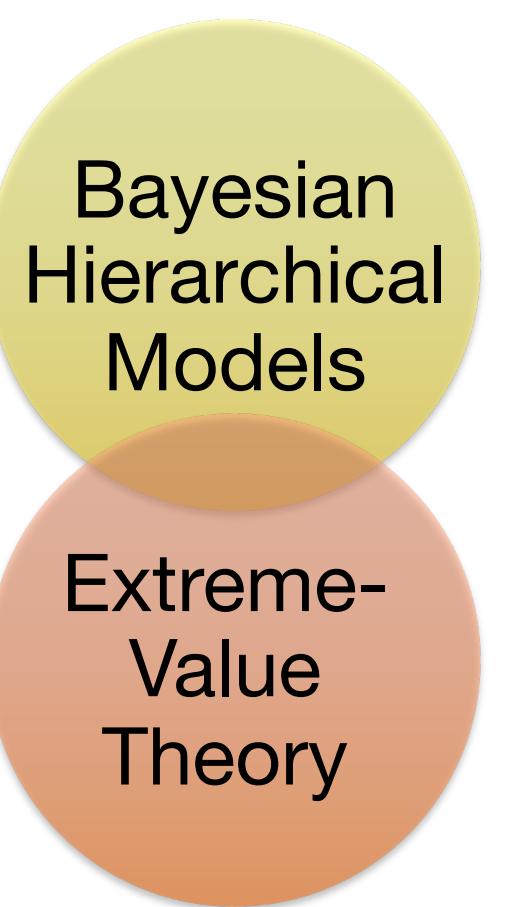


$$\left[\begin{array}{l} N_j^{\text{BBS}} \mid \lambda^{\text{BBS}}, \boldsymbol{\theta}_{\text{bbs}} \sim \text{Pois} \left\{ \sum_{k \in \text{route}_j} \omega_k \lambda^{\text{BBS}}(\mathbf{s}_k; \boldsymbol{\theta}_{\text{bbs}}) \right\}, \\ \\ \left[\begin{array}{l} N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \boldsymbol{\theta}_{\text{ckl}} \sim \text{Pois} \left\{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \boldsymbol{\theta}_{\text{ckl}}) \right\}, \\ N_i^{\text{spc}} \mid N_i^{\text{ckl}}, p^{\text{spc}}, \boldsymbol{\theta}_{\text{spc}} \sim \text{Bin}\{N_i^{\text{ckl}}, p^{\text{spc}}(\mathbf{s}_i, t_i; \boldsymbol{\theta}_{\text{spc}})\}, \\ Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \text{GEV}\{\mu(\mathbf{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\mathbf{s}_i; \boldsymbol{\theta}_\sigma), \xi\}, \end{array} \right] \end{array} \right] \quad X^{\text{pref}}(\cdot) \sim \mathcal{GP}(\boldsymbol{\omega}_1)$$

where

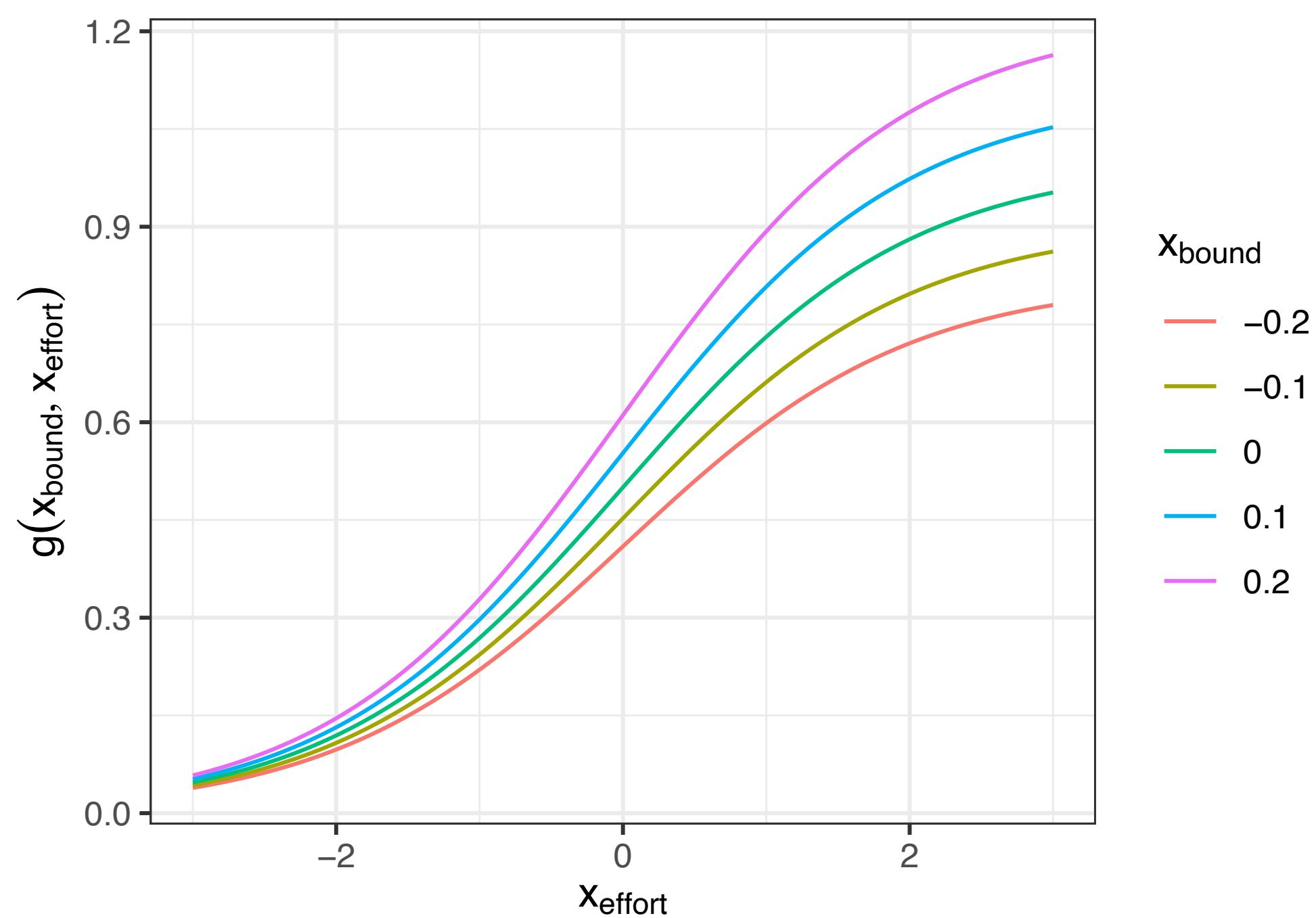
$\boldsymbol{\theta}_{\text{bbs}}, \boldsymbol{\theta}_{\text{ckl}}, \boldsymbol{\theta}_{\text{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \text{Hyperpriors}$

Saturating effect of observational effort



- Observed first arrival is biased towards later dates for low effort, but is the true one for very high effort
- Implementation: $Z_i \sim \text{GEV}(\mu_i, \sigma_i)$ with $\mu_i = g(\text{Predictors}_i, \text{Effort}_i)$
 - Nonlinear function g reaches (unknown) finite upper bound for very high effort
 - Infer g from data
 - Set very high effort for bias-corrected predictions

⚠ Source of high computational complexity



Implementation as a Bayesian hierarchical model

- **MCMC** implementation with Gibbs updates and Metropolis-Adjusted Langevin Algorithm (MALA) updates for the latent Gaussian components to improve mixing of chains. ~3secs per iteration, 80k burn-in, 20k for posterior evaluations with thinning
- **Vecchia** approximations for Gaussian random effects with exponential covariance (penalized complexity priors and sum-to-zero constraints)
- **Simulation** study confirms relevant parameters/ecological processes are identifiable.

Goodness-of-fit of estimated models

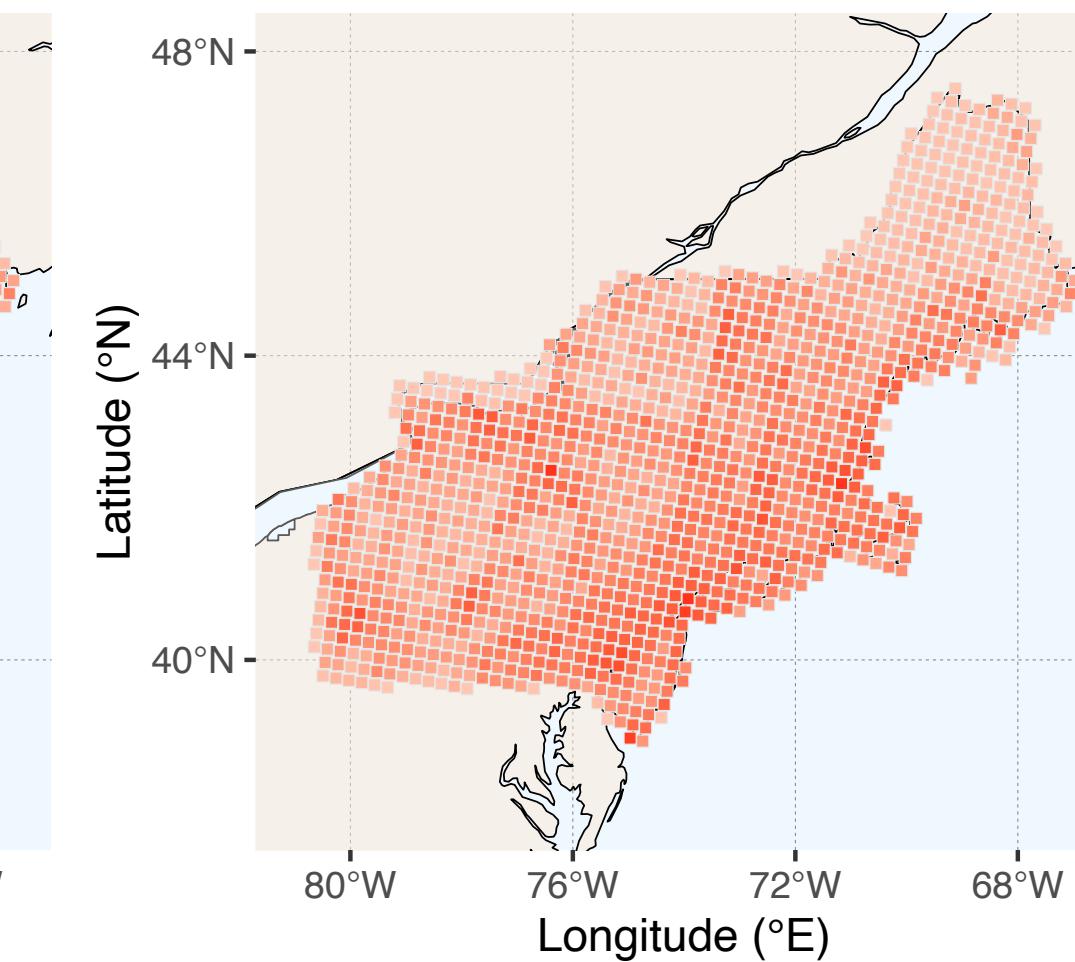
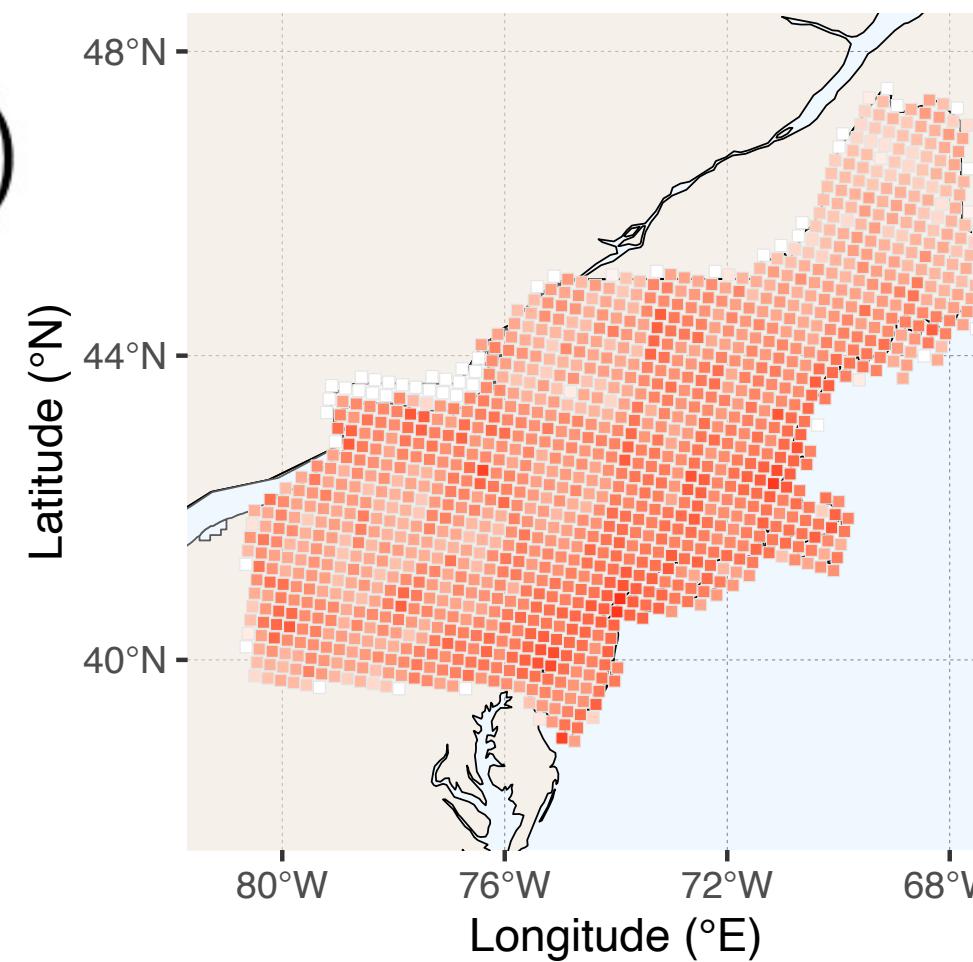
- Generally good match of eBird observations (left maps) with posterior means (right maps)
- Slight differences due to information shared from BBS

Example species:



Great Crested Flycatcher

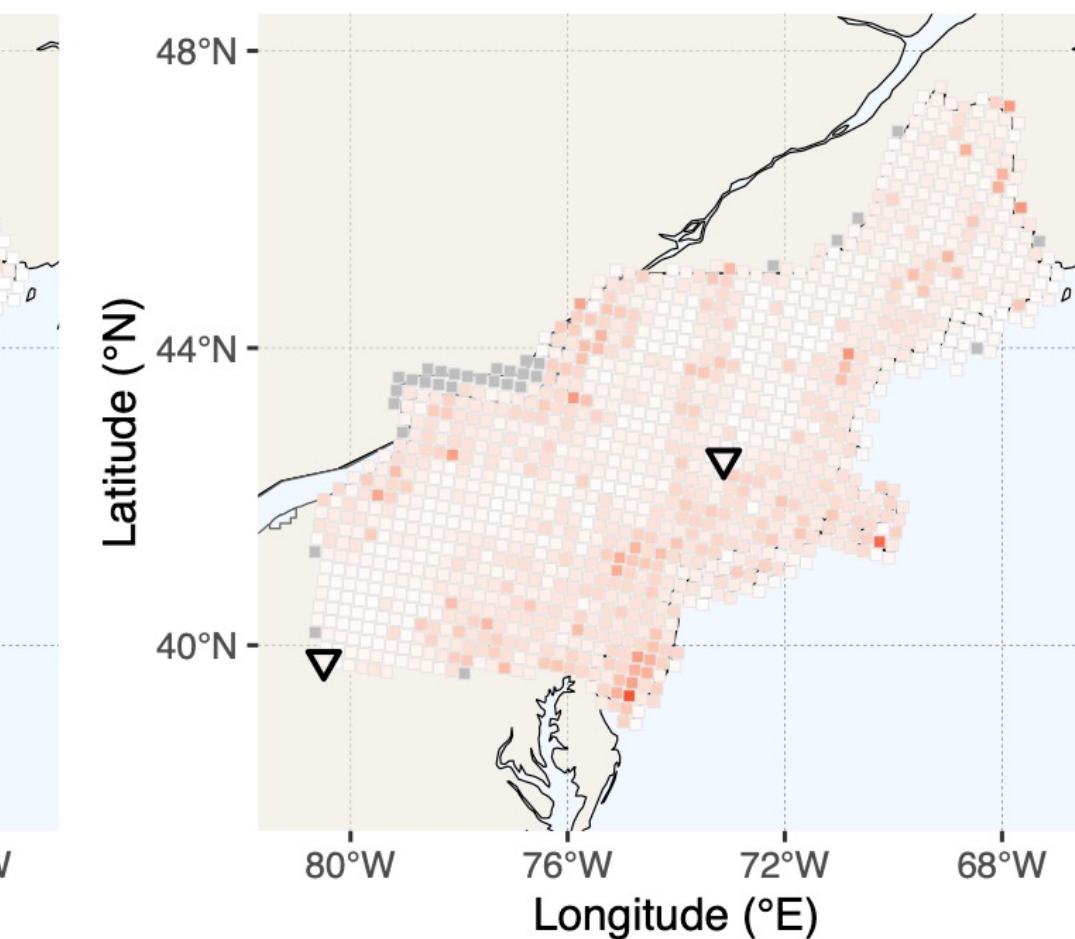
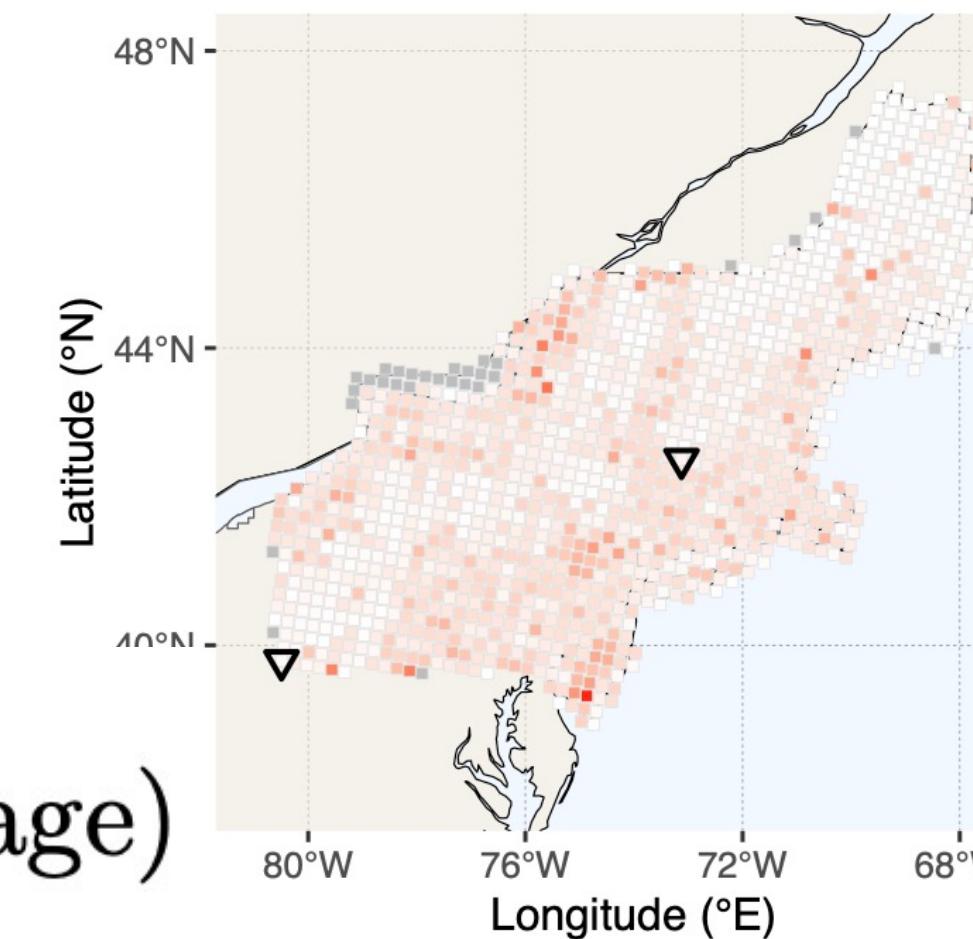
$$N_i^{\text{ckl}}(2021)$$



$$\hat{\lambda}_i^{\text{ckl}}(2021)$$

$\log(1+\text{CNT})$

$$\frac{N_i^{\text{spc}}}{N_i^{\text{ckl}}}(\text{Time-average})$$



$$\hat{p}_i^{\text{spc}}(\text{Time-average})$$

Illustration of bias-corrected prediction of first arrivals (2022)

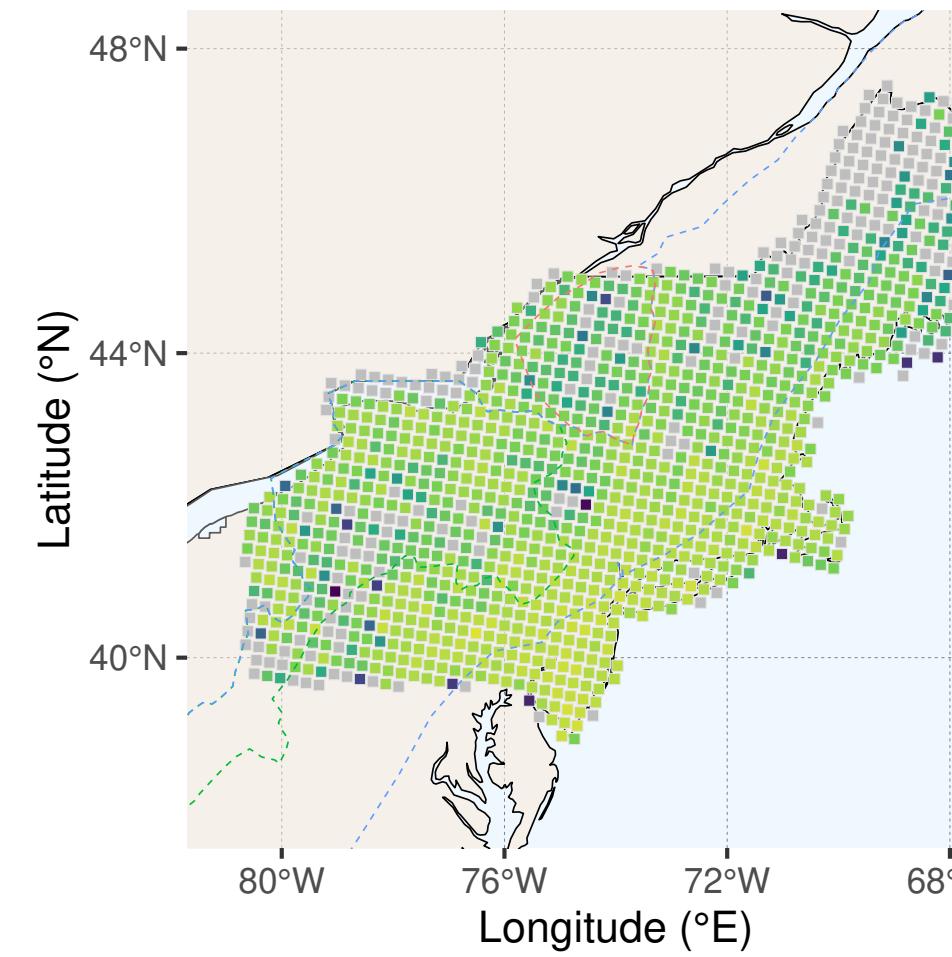
- Based on Generalised Extreme-Value response
- Bias-corrected prediction by fixing saturated observational effort

$$Z_i \mid \mu, \boldsymbol{\theta}^\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \text{GEV}\{\mu(\mathbf{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\mathbf{s}_i; \boldsymbol{\theta}_\sigma), \xi\}$$

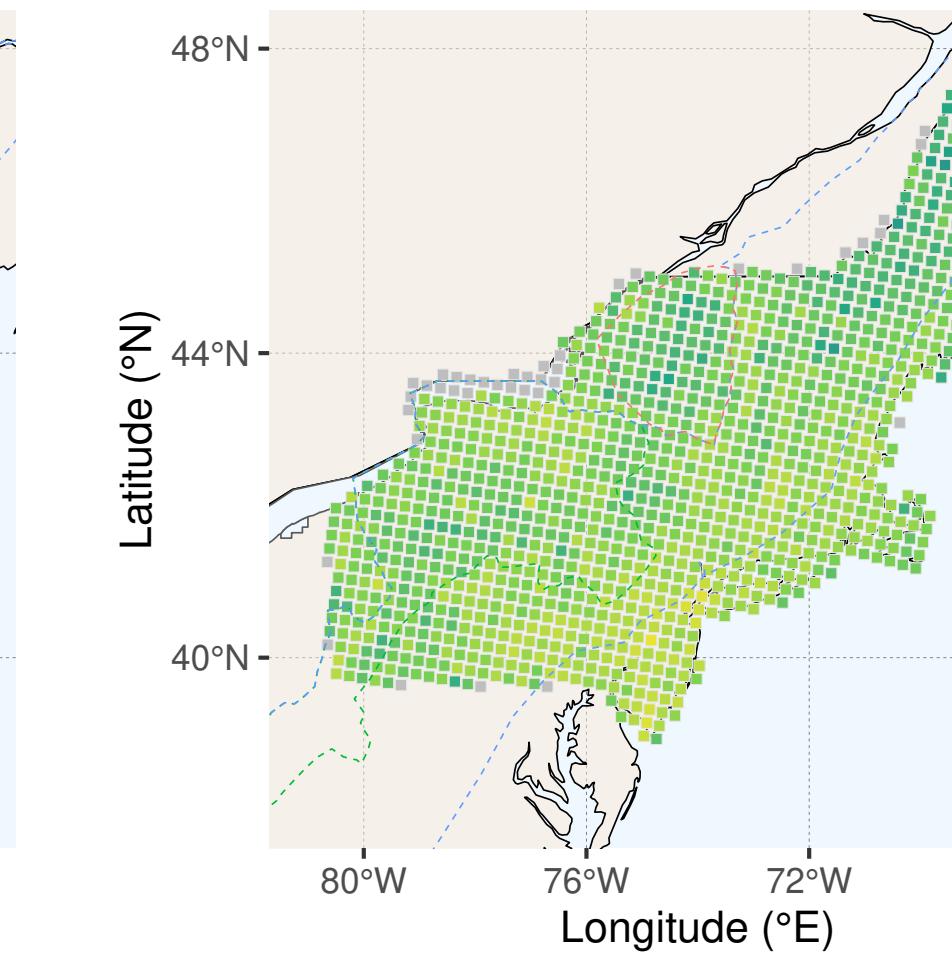
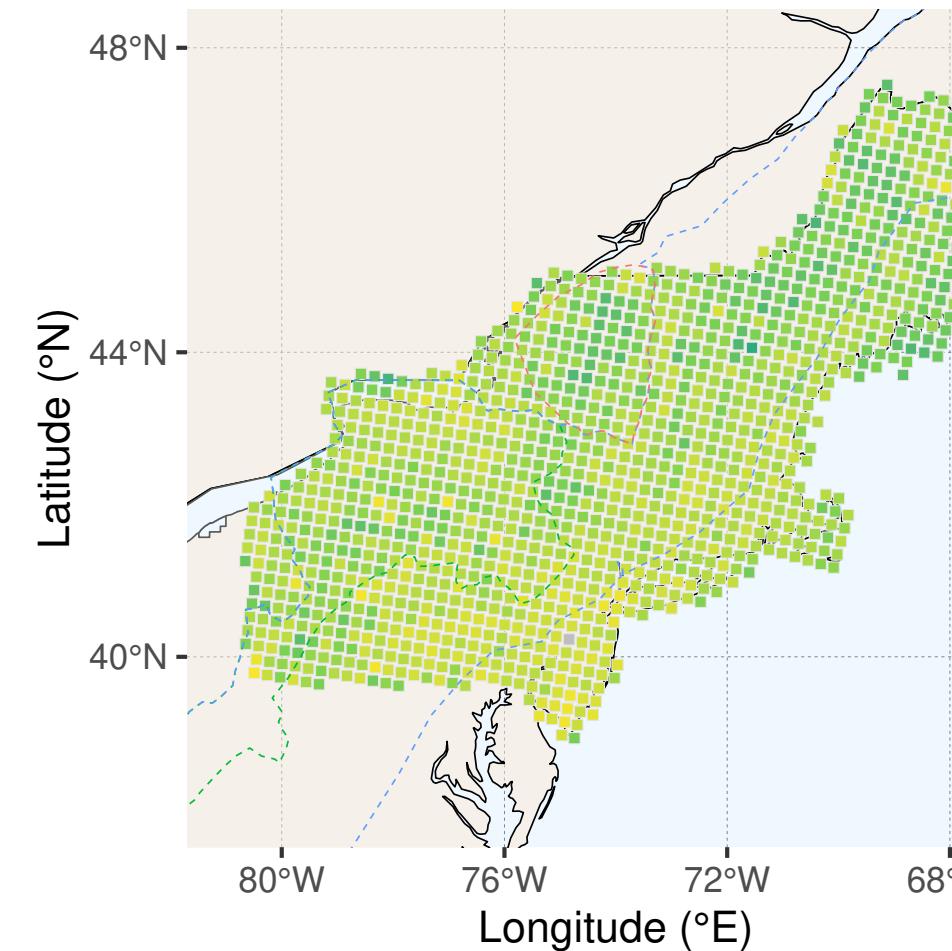


Great Crested
Flycatcher

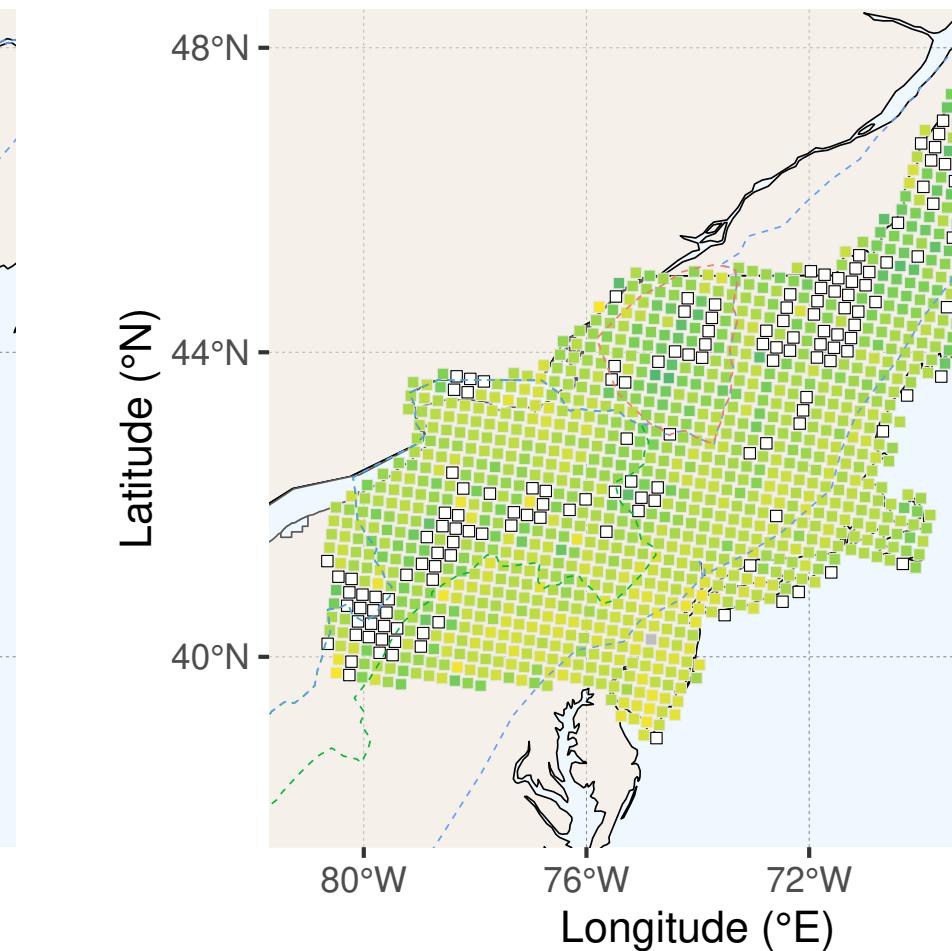
Observed



Posterior predictive
→ Saturated effort



Posterior



Posterior predictive
→ Saturated effort
→ Species niche

Illustration of bias-corrected prediction of first arrivals (2022), cont'd

- Table of estimated key parameters and first arrival dates for two pixels
- Estimated (not bias-corrected) first arrivals tend to occur relatively earlier for
 - higher Preference,
 - higher Activity and
 - in the core area of the niche



Species	Chimney Swift	Great Crested Flycatcher	Chestnut-sided Warbler	Purple Martin
$\hat{\theta}^{\text{pref}}$	0.191 (0.184,0.202)	0.204 (0.199,0.21)	0.187 (0.183,0.191)	0.2 (0.178,0.217)
$\hat{\theta}^{\text{act}}$	-0.15 (-0.217,-0.061)	-0.818 (-0.911,-0.696)	-0.548 (-0.619,-0.454)	-0.03 (-0.269,0.236)
$\hat{\theta}^{\text{niche-GEV}} (\times 10^{-2})$	4.9 (4.664,5.134)	4 (3.894,4.133)	0.2 (0.17,0.278)	6 (5.541,6.443)
Observed	NA	NA	NA	NA
Predicted	09/05	03/05	21/05	07/06
Debiased	03/04	13/04	03/05	28/03
Observed	01/05	04/05	04/05	29/06
Predicted	09/05	15/05	12/05	12/05
Debiased	22/04	05/05	03/05	07/04

Discussion: Ecological data fusion using latent processes



- Incomplete and biased observation of true processes
- Interpretable latent processes for effort and relevant ecological properties
 - Identifiability thanks to shared random effects, but challenging validation
- Towards spatiotemporal, not purely spatial, modelling
 - Improve modelling of temporal dynamics
 - ⚠ Requires disentangling complex observational/ecological dynamics
- Could we implement shared latent processes in other learning algorithms?
(GAMs, ANNs, Random Forests...)
- Multi-species modelling?

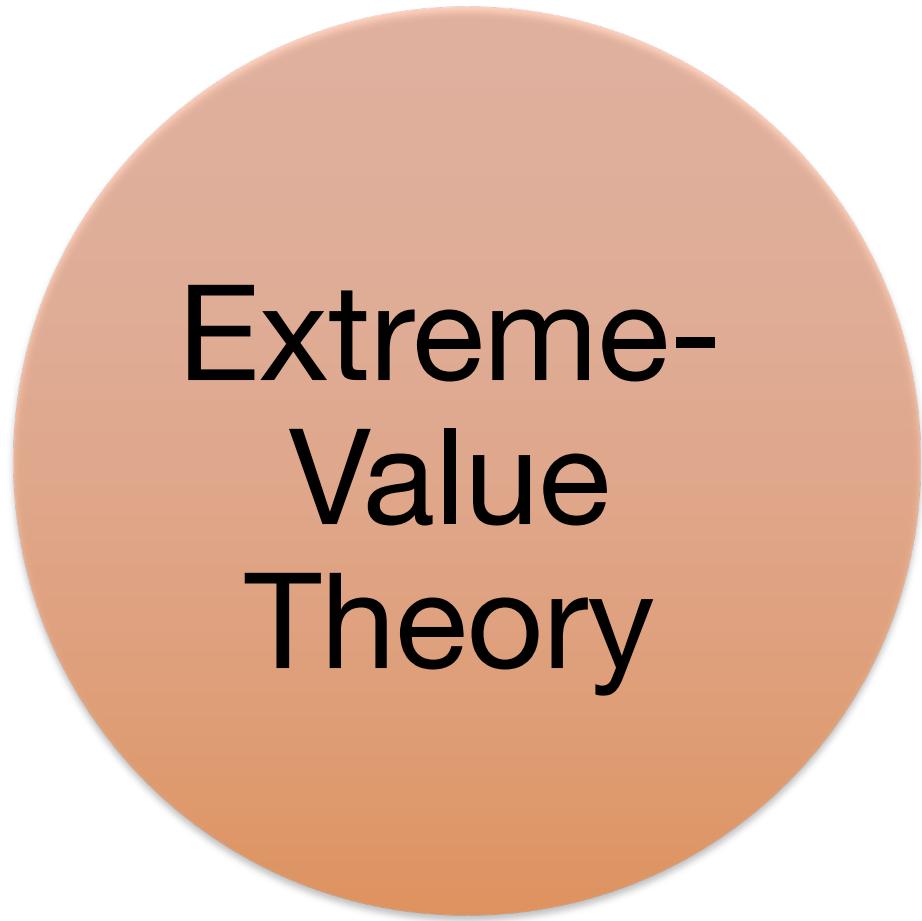
Discussion: Bias and uncertainty reduction



Citizen
Science

- Checklist data, such as eBird, allow generating pseudo-absences, but many opportunistic datasets are less structured
- Could we use these opportunistic datasets in machine learning algorithms?
- Data fusion of opportunistic and structured data in *Integrated Species Distribution Models* is crucial (Fithian et al 2015; Isaac et al 2020)
- Collecting additional exhaustive field data may be necessary
→ Explore optimal sampling design through simulation studies?

Discussion: Opportunities for ecological extreme-value analysis

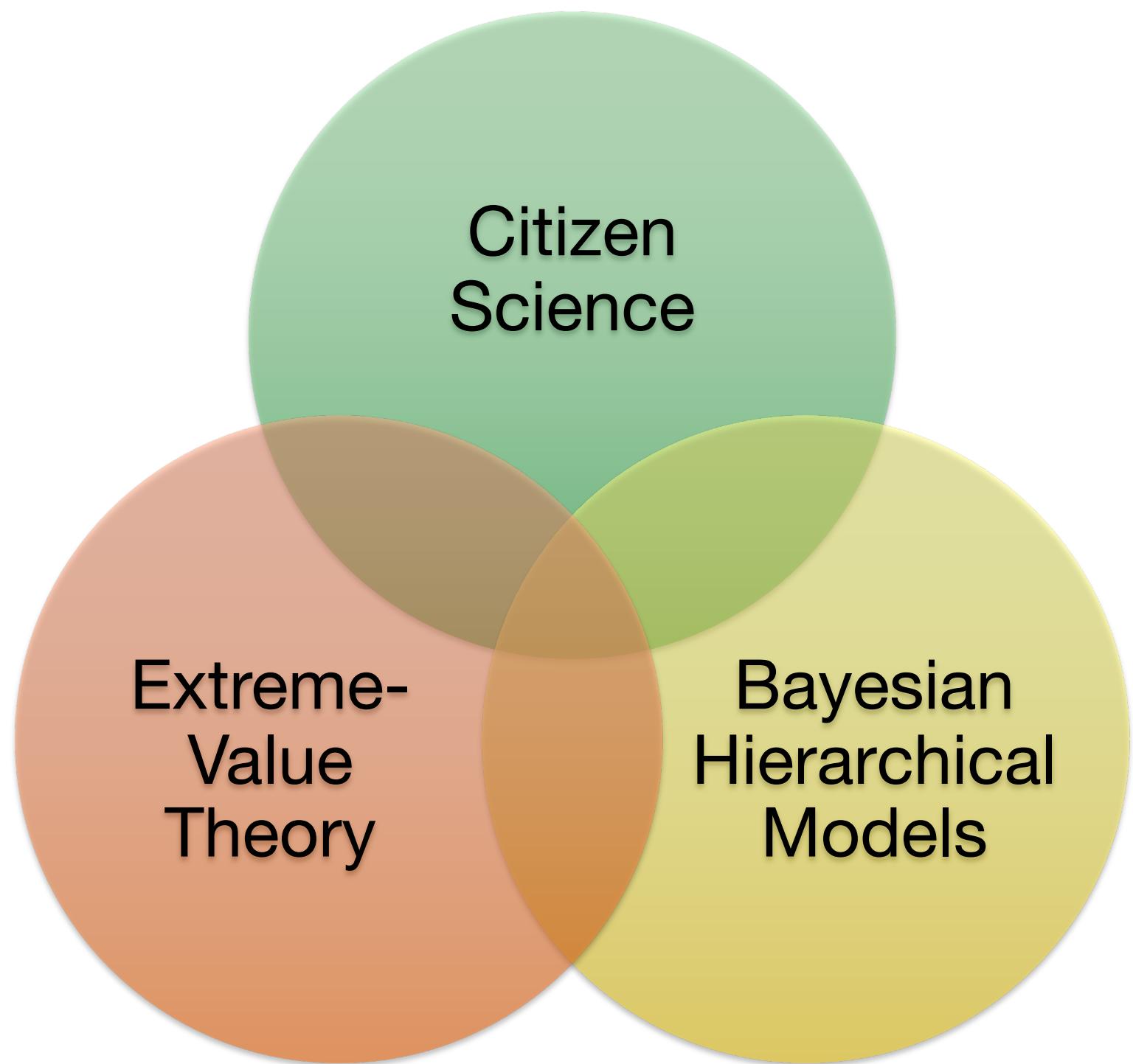


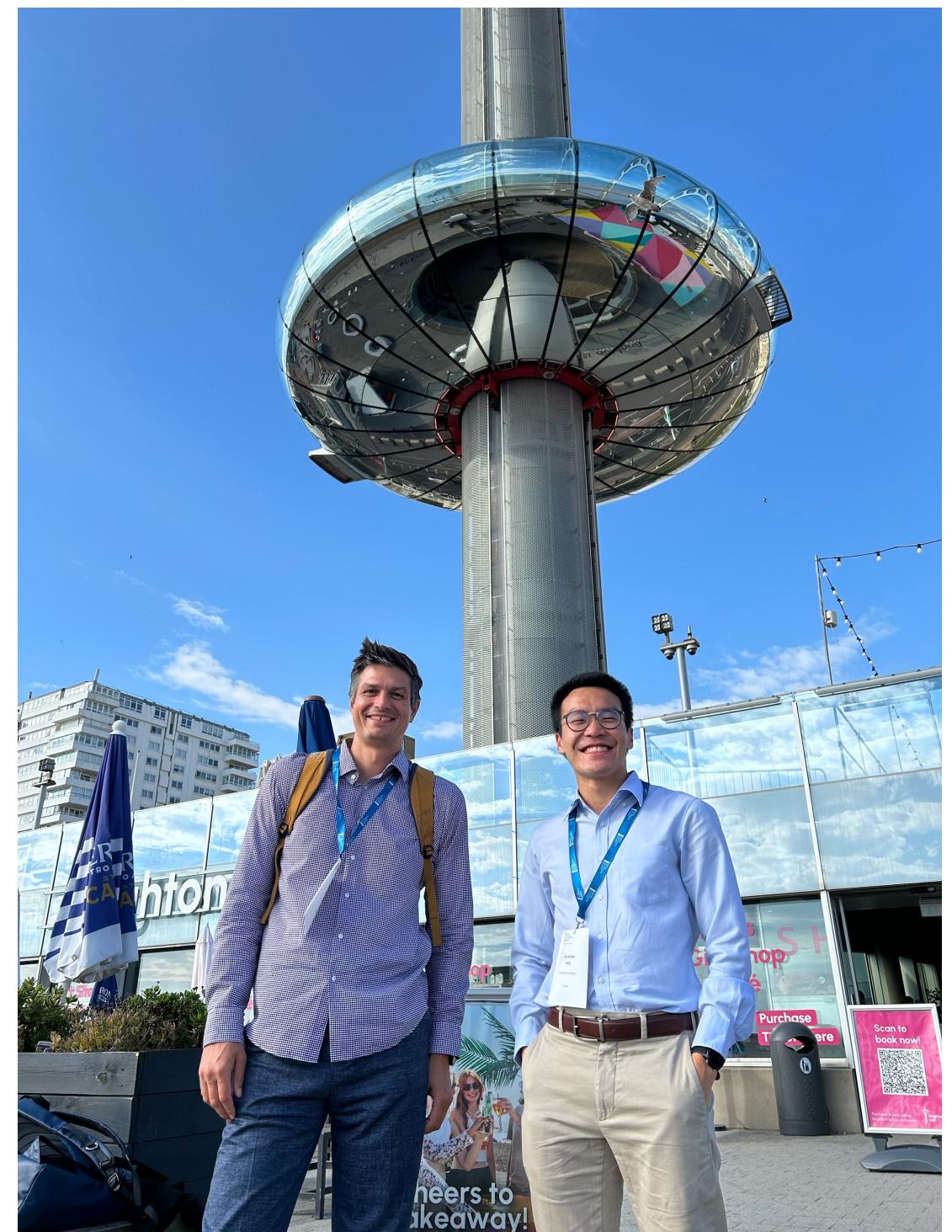
Extreme-
Value
Theory

- EVT generally less relevant for discrete data but promising for modelling extreme phenological events, such as first arrivals
- GEV isn't the only model
- Lots of suggestions/ideas to improve mixing of MCMC
- EVT widely used for extreme climate and environmental events
 - Such events can drive strong species population shifts
 - Focus on specific events, not only long-term climate averages

Outlook

- Citizen science datasets: *Small Data* and *Big Data*, but always complex...
 - Wide opportunities for modelling and decision support
 - An exciting potential playground for statisticians
(and also those in extremes!)





Membership News and Publications Training and events Policy and campaigns Jobs and careers Resources

Sign in Join

About RSS

[Home](#) > [News](#) > [2023](#) > [Member callouts](#) > **Call for discussion papers 2023: Analysis of citizen science data**

Call for discussion papers 2023: Analysis of citizen science data

03.05.23 | Member callouts



Citizen science involves volunteers who participate in scientific research by collecting data, monitoring sites, and even taking part in the whole process of scientific inquiry (Roy et al. 2012, Scyphers et al. 2015). In the past two decades, citizen science (also called participatory or community-based monitoring) has gained tremendous popularity (Bonney et al. 2009, Danielsen et al. 2014; Aceves-Bueno et al., 2017). These opportunistic datasets can be substantially large, numbering hundreds of thousands of units and present many statistical challenges in terms of sparse and missing data, issues with data collection

Special thanks to the discussion contributors!

- Kerrie Mengerson (Seconder)
- Ben Swallow (Proposer)
- Kuldeep Kumar
- Leo Belzile & Rishikesh Yadav
- Andrej Srakar
- Raphael Huser & Andrew Zammit-Mangion
- Abdelaati Daouia & Gilles Stupler
- Allan Reese
- Stefano Rizzelli
- Maozai Tian

Journal of the Royal Statistical Society

Series A: Statistics in Society

Issues More content ▾ Submit ▾ Purchase Alerts About ▾

Journal of the Royal Stats ▾

Article Contents

Abstract

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Authors' reply to the Discussion of 'Extreme-value modelling of migratory bird arrival dates: Insights from citizen-science data' FREE

Jonathan Koh ✉ , Thomas Opitz

Journal of the Royal Statistical Society Series A: Statistics in Society, qnaf058,
<https://doi.org/10.1093/jrsssa/qnaf058>

Published: 14 May 2025 Article history ▾

 PDF  Split View  Cite  Permissions  Share ▾

Abstract

We respond to the discussion comments of the Proposer and Seconder of the vote of thanks, and to the eight other contributions discussing our work.

Issue Section: Discussion Paper Contribution

Food for thought

This work:

Koh, Opitz (2024). Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data. Journal of the Royal Statistical Society, Series A (Statistics in Society).

Other literature:

- Adjei et al. (2023). A structural model for the process of collecting biodiversity data. Authorea Preprints.
- Adjei et al. (2023). The Point Process Framework for Integrated Modelling of Biodiversity Data. arXiv:2311.06755.
- Belmont et al. (2024). Spatio-temporal Occupancy Models with INLA. arXiv:2403.10680.
- Coles (2001). An introduction to statistical modeling of extreme values. Springer.
- Diggle et al. (2010). Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society Series C: Applied Statistics.
- Fithian et al. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution.
- Gelfand & Shirota (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs.
- Isaac et al. (2020). Data integration for large-scale models of species distributions. Trends in Ecology & Evolution.
- Lindgren et al. (2024). *inlabru*: software for fitting latent Gaussian models with non-linear predictors. arXiv:2407.00791.
- Tang et al. (2021). Modeling spatially biased citizen science effort through the eBird database. Environmental and Ecological Statistics.
- Wijeyakulasuriya et al. (2024). Modeling First Arrival of Migratory Birds Using a Hierarchical Max-Ininitely Divisible Process. Journal of Agricultural, Biological and Environmental Statistics.

Upcoming Talks

19 December 2025 - Jonathan Koh (ETH Zürich)

(16:00 UTC) = (11:00 New York) = (17:00 Paris) = (00:00 Shanghai) = (03:00 Sydney)

Tail calibration of probabilistic forecasts

Abstract: Probabilistic forecasts comprehensively describe the uncertainty in the unknown future outcome, making them essential for decision making and risk management. While several methods have been introduced to evaluate probabilistic forecasts, existing evaluation techniques are ill-suited to the evaluation of tail properties of such forecasts. However, these tail properties are often of particular interest to forecast users due to the severe impacts caused by extreme outcomes. In this work, we introduce a general notion of tail calibration for probabilistic forecasts, which allows forecasters to assess the reliability of their predictions for extreme outcomes. We study the relationships between tail calibration and standard notions of forecast calibration, and discuss connections to peaks-over-threshold models in extreme value theory. Diagnostic tools are introduced and applied in a case study on European precipitation forecasts.

22 January 2026 - Axel Bücher (Ruhr University Bochum)

(15:00 UTC) = (10:00 New York) = (16:00 Paris) = (23:00 Shanghai) = (02:00 Sydney)

Latent linear factor models for tail dependence in high dimensions

Abstract: A common object to describe the extremal dependence of a d-variate random vector X is the stable tail dependence function L . Various parametric models have emerged, with a popular sub-class consisting of those stable tail dependence functions that arise for linear and max-linear factor models with heavy tailed factors. We study such models under the assumption that the factors are possibly dependent, which results in a model for L that depends on a $(d \times K)$ loading matrix

