

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

Jonathan Koh, Thomas Opitz

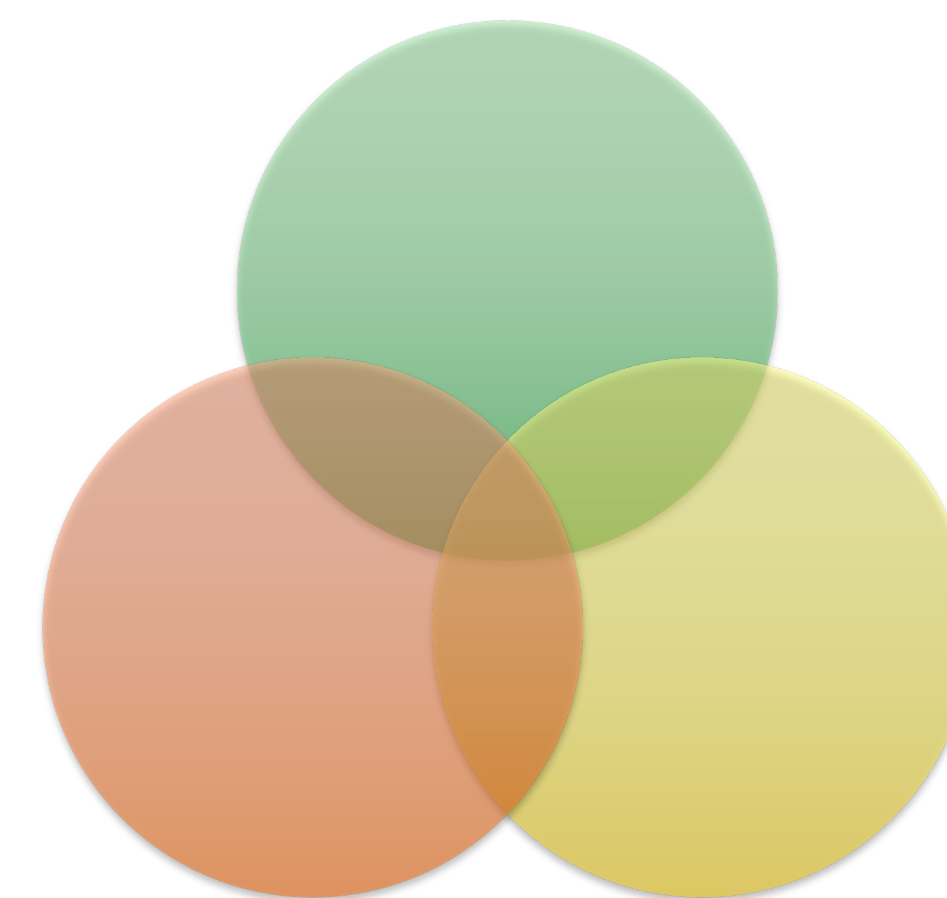
Workshop CisStats & RESSTE
Approaches to modeling heterogeneous data
2–3 February 2026
Avignon



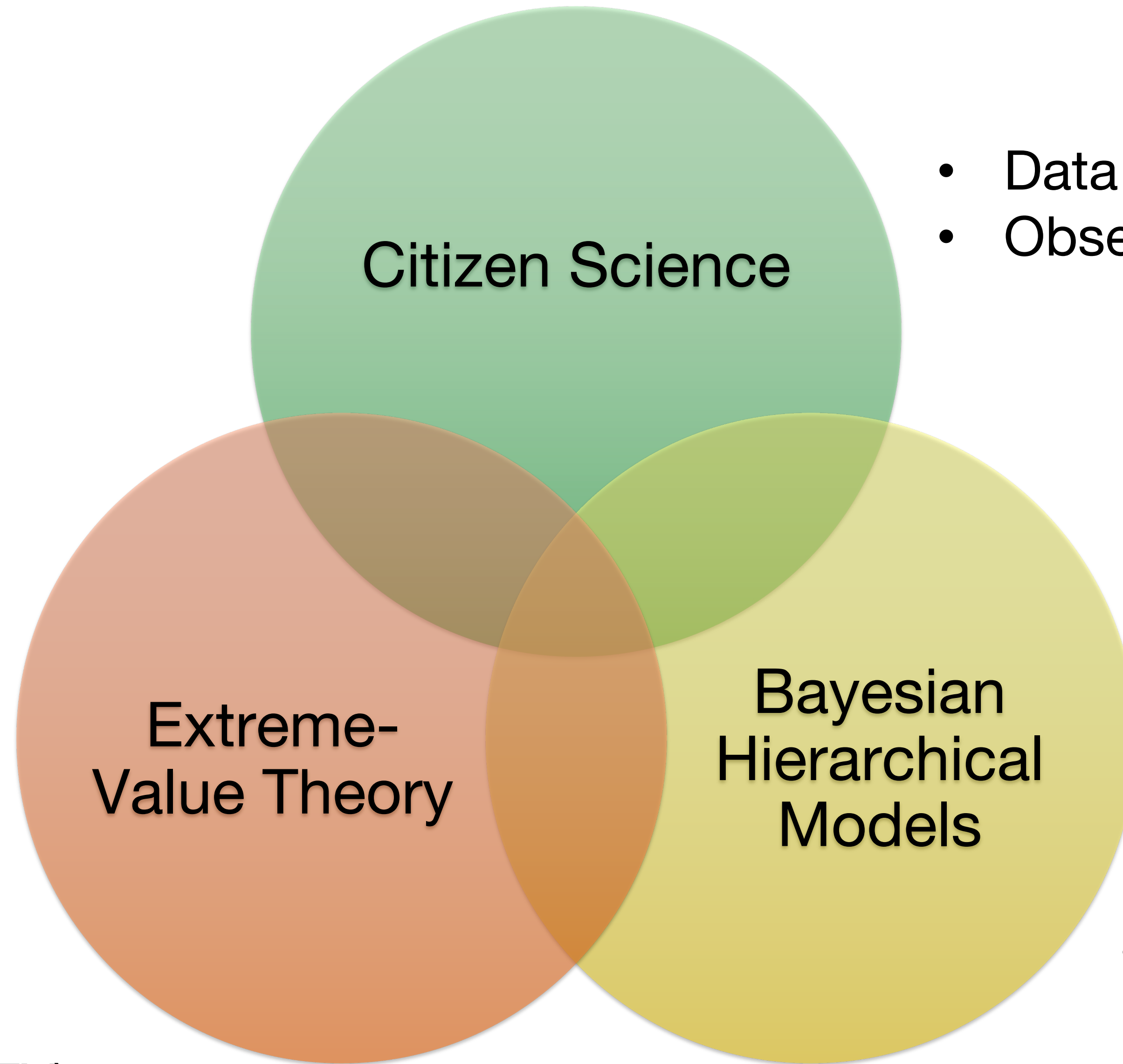
u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



INRAE



- Data fusion
- Observational bias

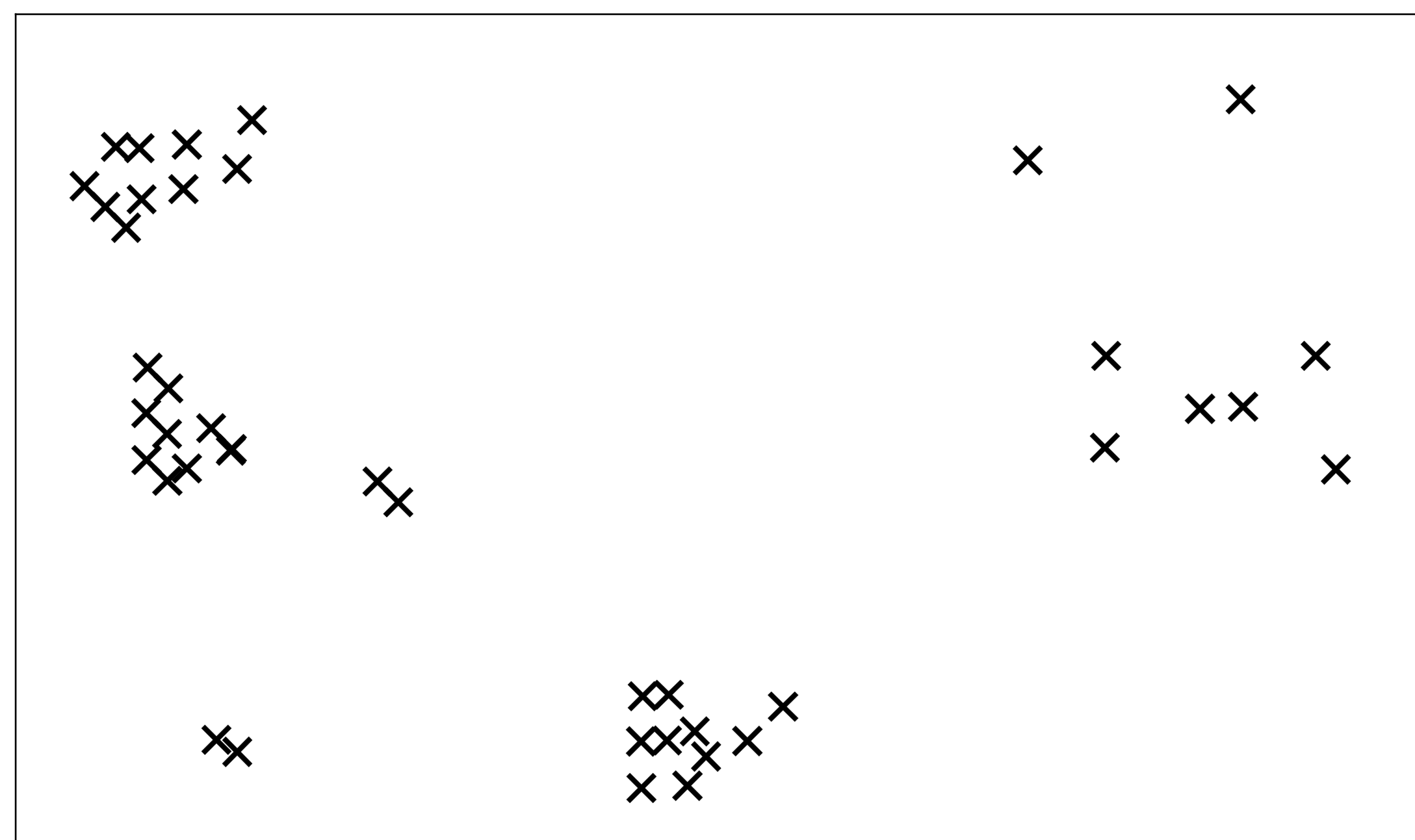
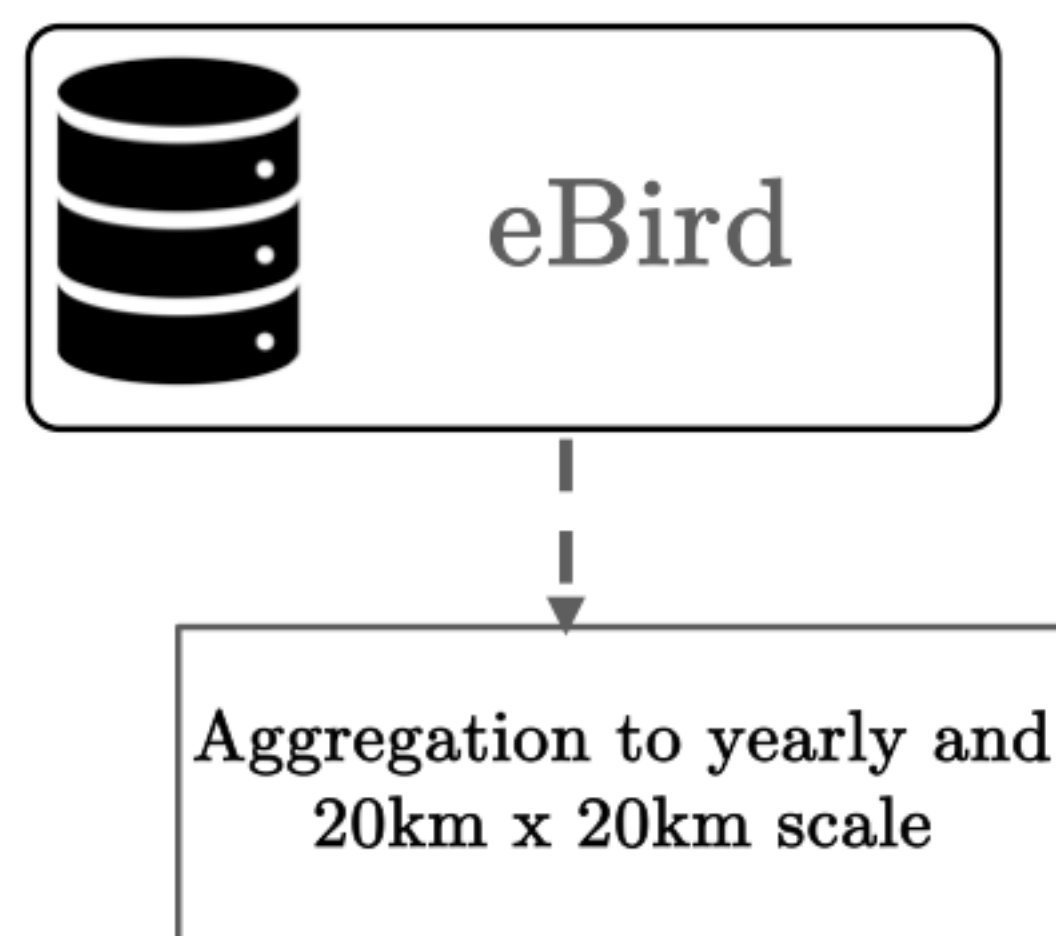
- Extremes of phenological events
- Generalized Extreme Value Distribution (GEV)

- Inference on important latent processes

eBird data processing

eBird → Checklist data: we use coordinates, date, duration of observation

Year 2020

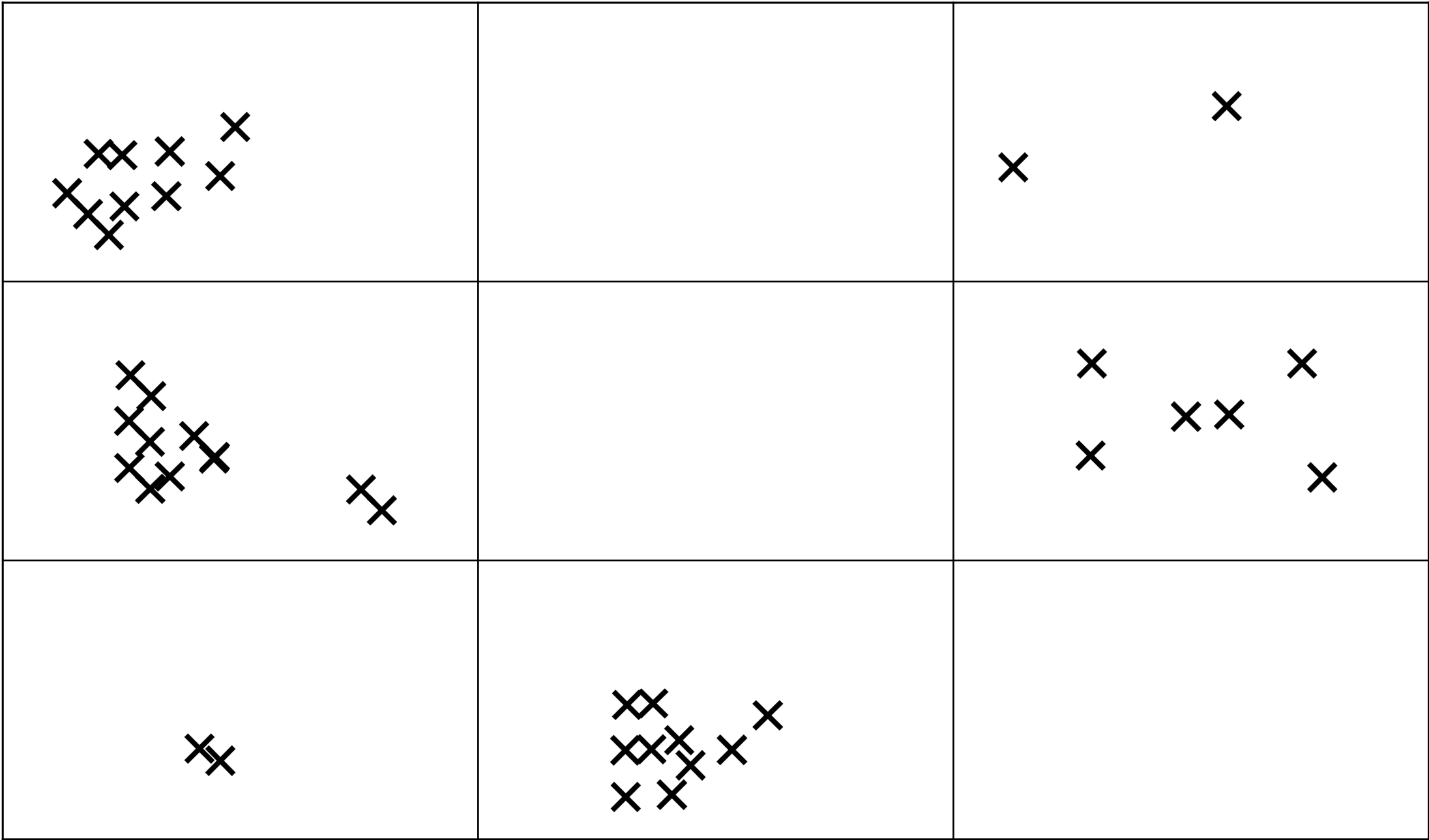


eBird data processing

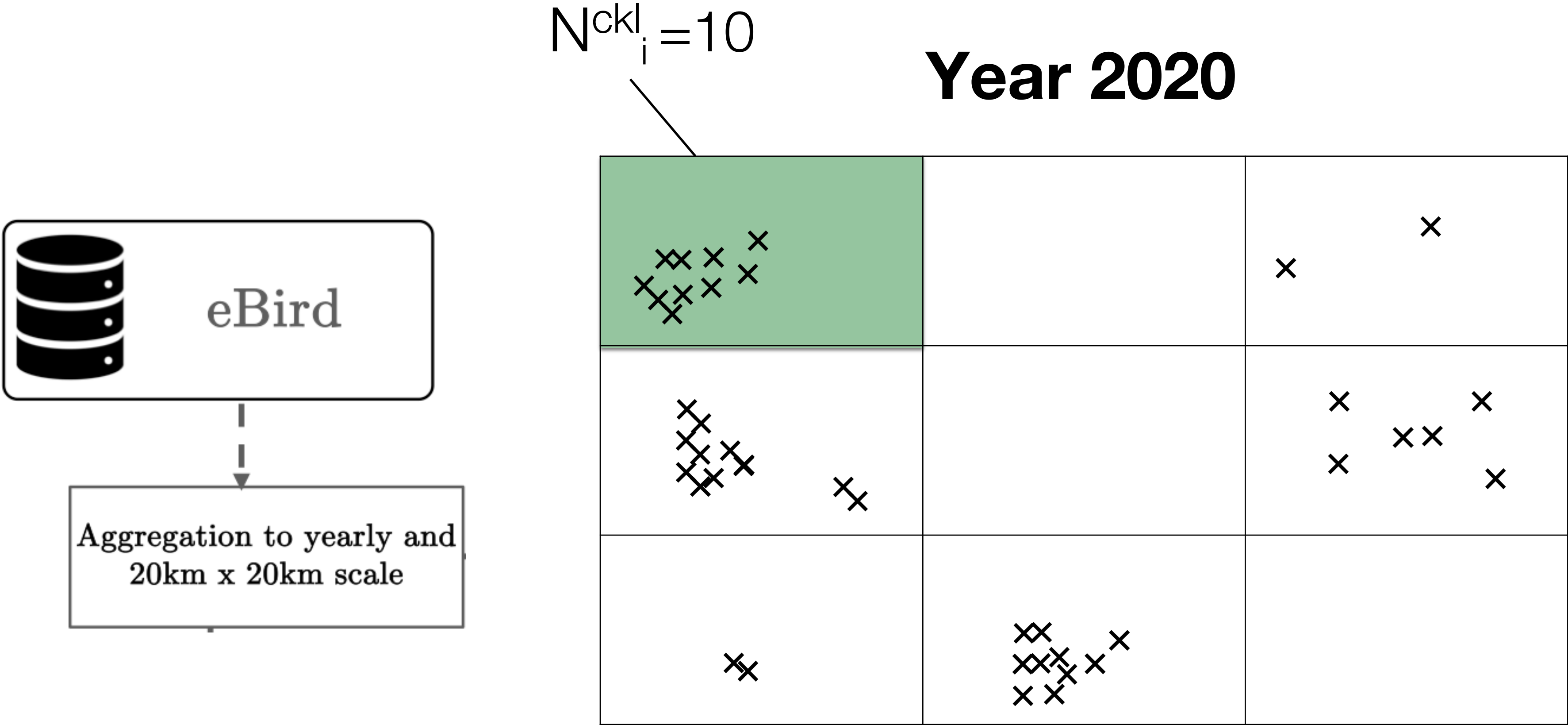
Year 2020



Aggregation to yearly and 20km x 20km scale



eBird data processing



eBird data processing



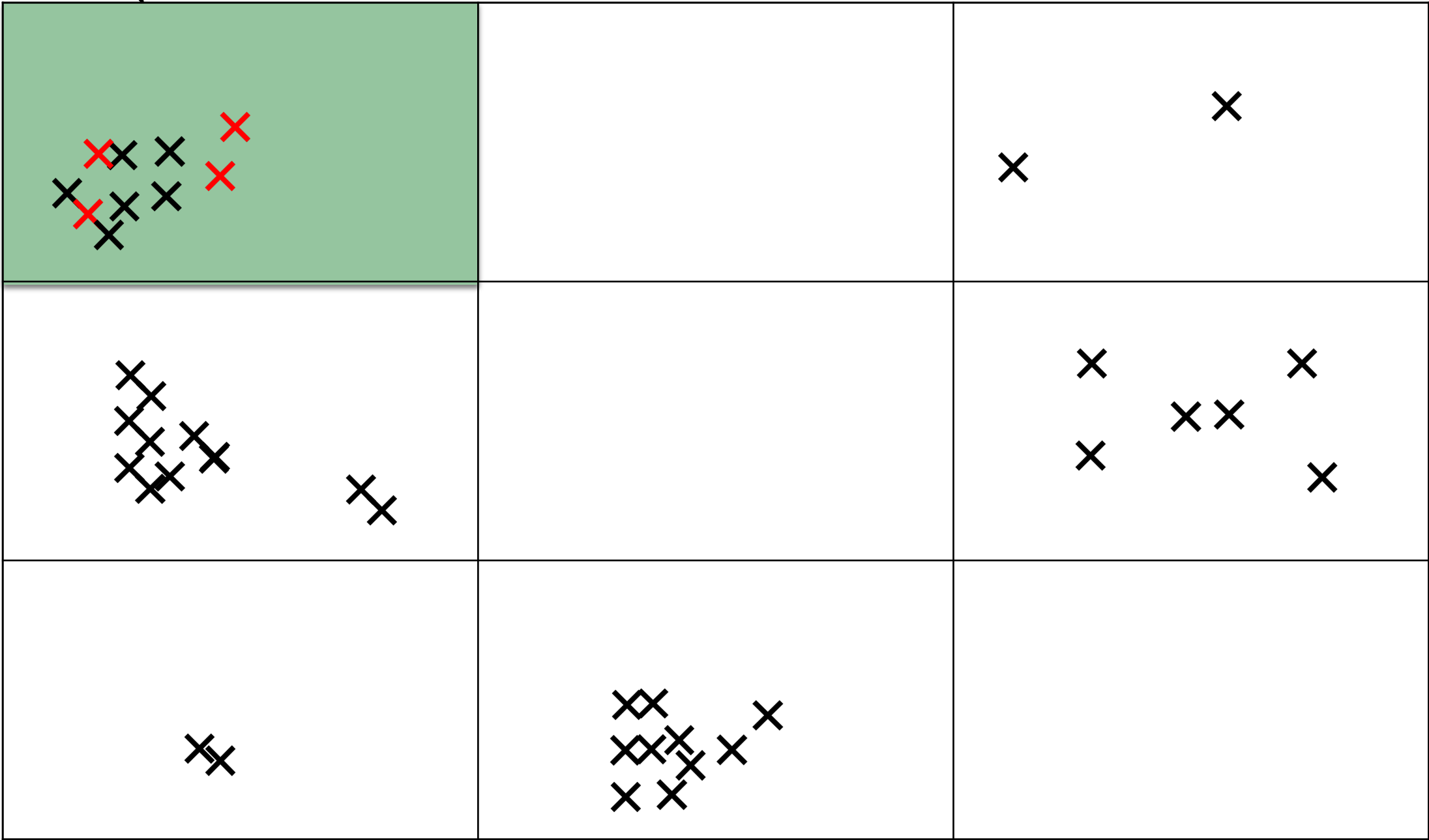
Purple Martin



Aggregation to yearly and 20km x 20km scale

$N^{spc}_i = 4$

Year 2020



eBird data processing

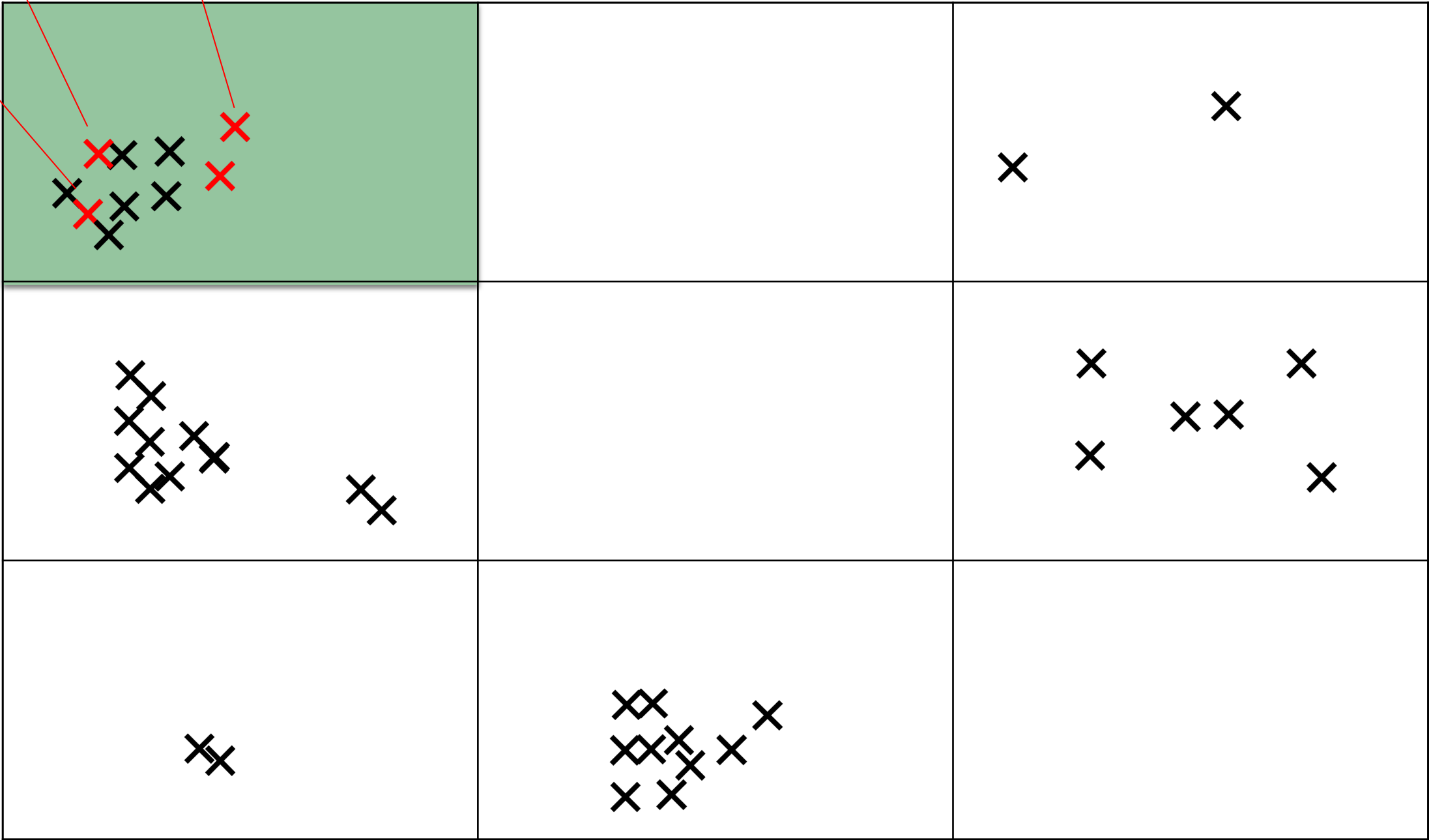


Purple Martin



Aggregation to yearly and 20km x 20km scale

$Y_{i,1}=28/03/2020$ $Y_{i,2}=02/04/2020$ $Y_{i,3}=05/04/2020$ **Year 2020**



eBird data processing



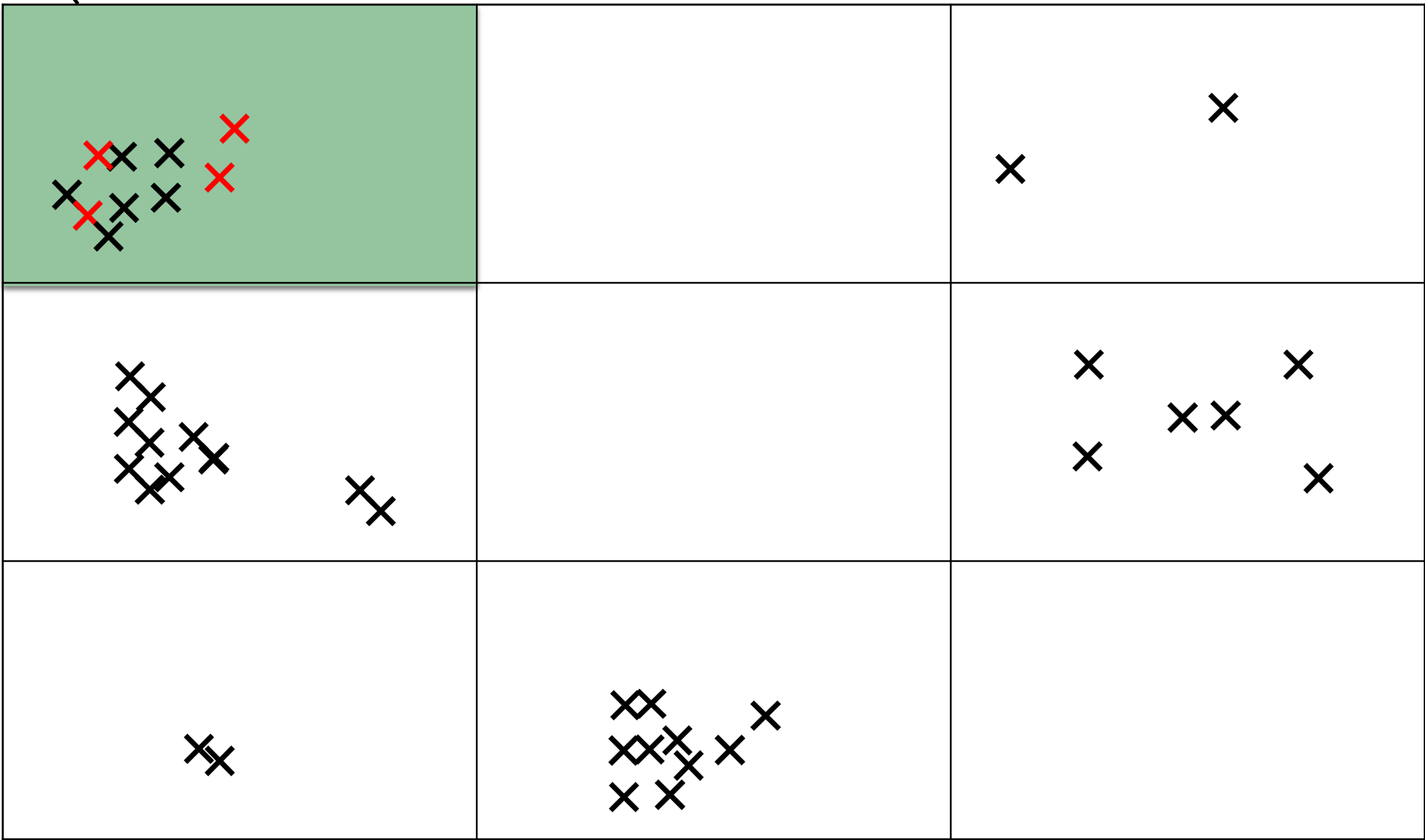
Purple Martin

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

Year 2020

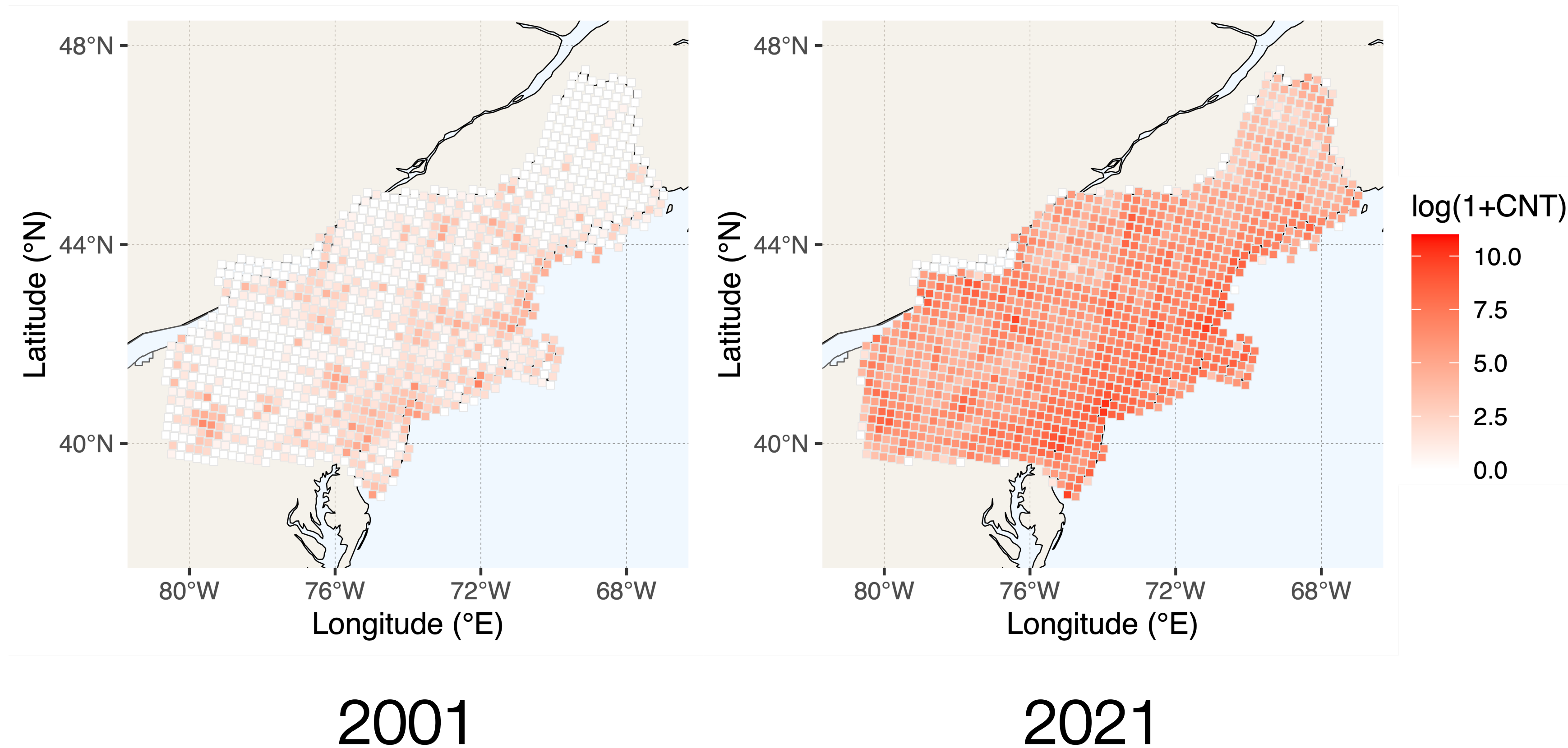


Aggregation to yearly and 20km x 20km scale



→ We will model $-Z_{\text{pixel},\text{year}}$ with a Generalized Extreme Value Distribution (GEV)

Strong temporal trends in reported occurrences

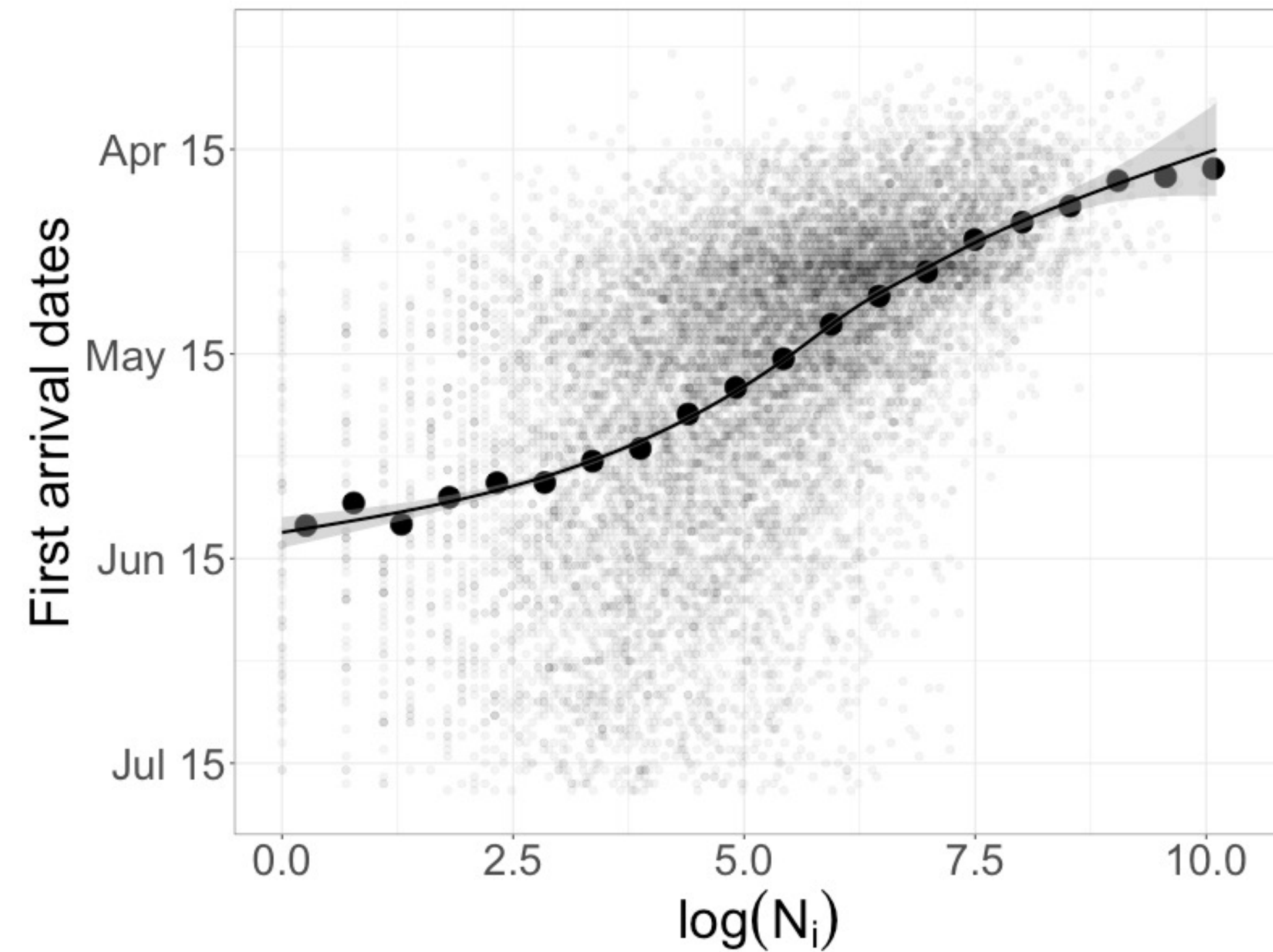


First arrival dates vs. checklist counts

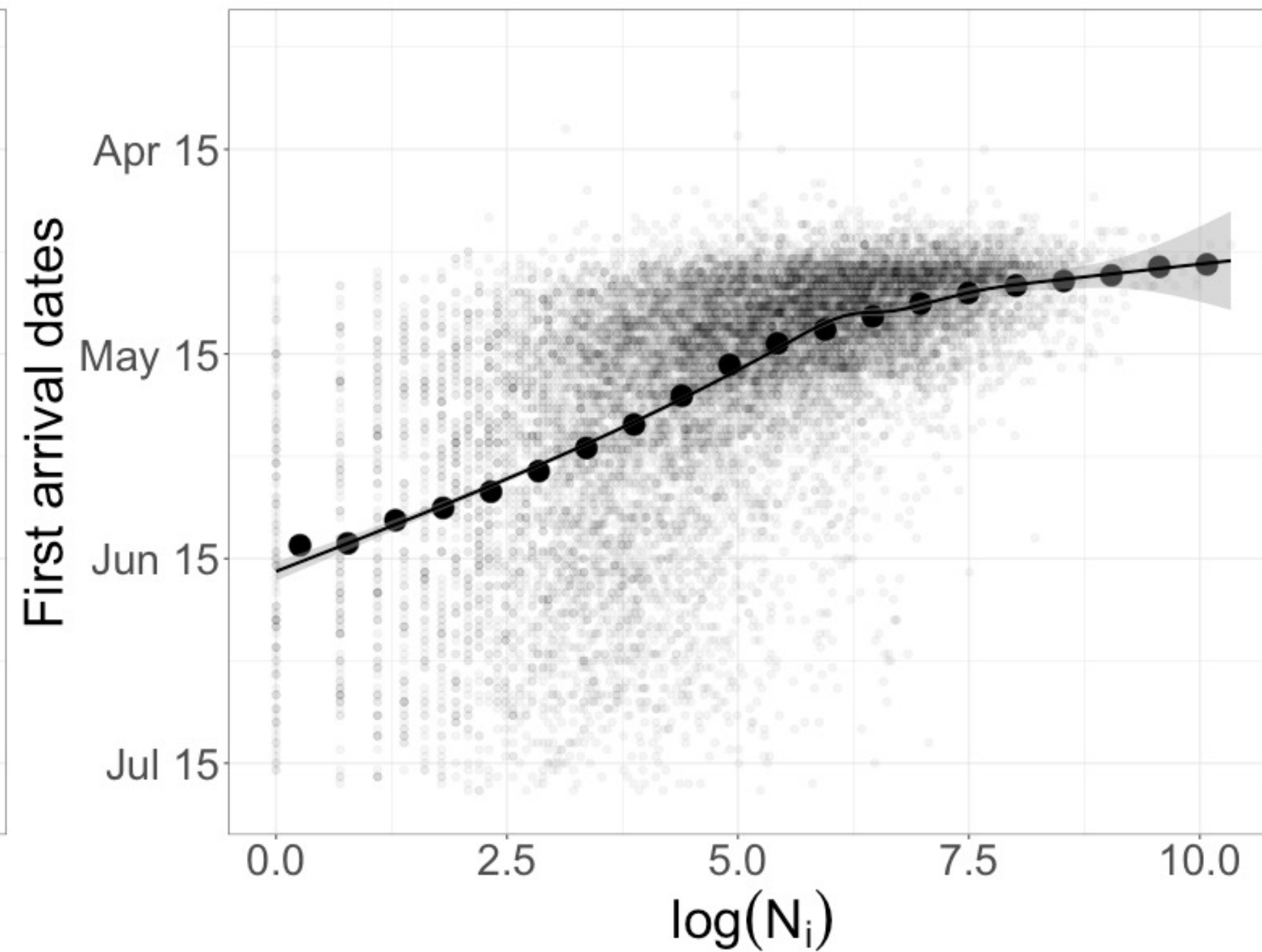
Citizen
Science

Bayesian
Hierarchical
Models

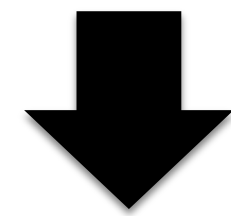
Chimney-Swift



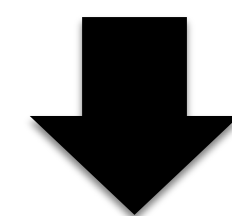
**Chestnut-sided
Warbler**



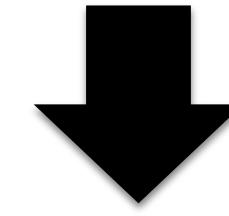
Observational effort = Preference + Activity



space-time
varying



captured by the
sampling intensity
for the checklists

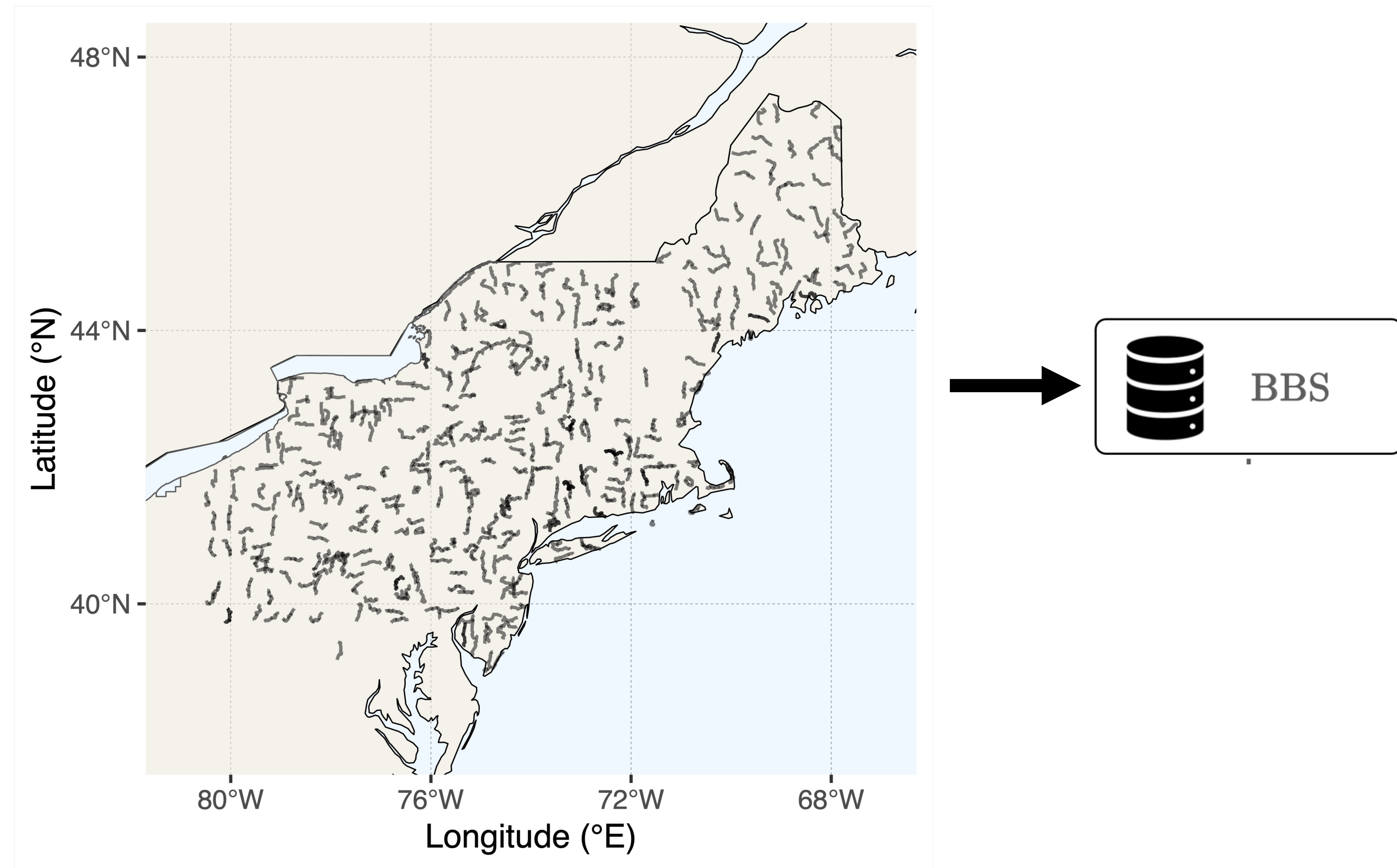


captured by the
(median) time spent
on the checklist

Breeding Bird Survey (BBS) sampling routes

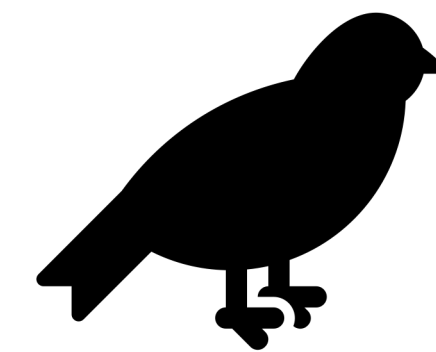
Citizen
Science

- For each route (~40km), bird occurrences are reported at 50 equidistant stops
- 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
- Use the model to make posterior predictions
- Interpolate spatially to locations not visited, in a reasonable way

A multi-response spatial regression system

Bayesian
Hierarchical
Models

→ Estimated for each species

Multi-response spatial regression

Protocol-based data (→ Niche)

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

Checklist count (→ Effort)

$$N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \theta_{\text{ckl}} \sim \text{Pois} \{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \theta_{\text{ckl}}) \},$$

Species PA (→ Niche)

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

First arrival date (→ Phenology)

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

where

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

A multi-response spatial regression system

Multi-response spatial regression



$$\left\{ \begin{array}{l} N_j^{\text{BBS}} \mid \lambda^{\text{BBS}}, \theta_{\text{bbs}} \sim \text{Pois} \left\{ \sum_{k \in \text{route}_j} \omega_k \lambda^{\text{BBS}}(\mathbf{s}_k; \theta_{\text{bbs}}) \right\}, \end{array} \right.$$



$$\left\{ \begin{array}{l} N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \theta_{\text{ckl}} \sim \text{Pois} \{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \theta_{\text{ckl}}) \}, \\ N_i^{\text{spc}} \mid N_i^{\text{ckl}}, p^{\text{spc}}, \theta_{\text{spc}} \sim \text{Bin} \{ N_i^{\text{ckl}}, p^{\text{spc}}(\mathbf{s}_i, t_i; \theta_{\text{spc}}) \}, \\ Z_i \mid \mu, \theta_{\mu}, \sigma, \theta_{\sigma} \sim \text{GEV} \{ \mu(\mathbf{s}_i, t_i; \theta_{\mu}), \sigma(\mathbf{s}_i; \theta_{\sigma}), \xi \}, \end{array} \right.$$

where

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

Sharing random effects

Multi-response spatial regression



$$\begin{aligned}
 & \left\{ \begin{aligned} & N_j^{\text{BBS}} \mid \lambda^{\text{BBS}}, \theta_{\text{bbs}} \sim \text{Pois} \left\{ \sum_{k \in \text{route}_j} \omega_k \lambda^{\text{BBS}}(\mathbf{s}_k; \theta_{\text{bbs}}) \right\}, \\ & N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \theta_{\text{ckl}} \sim \text{Pois} \{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \theta_{\text{ckl}}) \}, \\ & N_i^{\text{spc}} \mid N_i^{\text{ckl}}, p^{\text{spc}}, \theta_{\text{spc}} \sim \text{Bin} \{ N_i^{\text{ckl}}, p^{\text{spc}}(\mathbf{s}_i, t_i; \theta_{\text{spc}}) \}, \\ & Z_i \mid \mu, \theta_{\mu}, \sigma, \theta_{\sigma} \sim \text{GEV} \{ \mu(\mathbf{s}_i, t_i; \theta_{\mu}), \sigma(\mathbf{s}_i; \theta_{\sigma}), \xi \}, \end{aligned} \right.
 \end{aligned}$$

$X^{\text{niche}}(\cdot) \sim \mathcal{GP}(\omega_2),$

where

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

Sharing random effects

Multi-response spatial regression



$$\begin{aligned}
 & \left\{ \begin{aligned} N_j^{\text{BBS}} \mid \lambda^{\text{BBS}}, \theta_{\text{bbs}} &\sim \text{Pois} \left\{ \sum_{k \in \text{route}_j} \omega_k \lambda^{\text{BBS}}(\mathbf{s}_k; \theta_{\text{bbs}}) \right\}, \\ N_i^{\text{ckl}} \mid \lambda^{\text{ckl}}, \theta_{\text{ckl}} &\sim \text{Pois} \{ \lambda^{\text{ckl}}(\mathbf{s}_i, t_i; \theta_{\text{ckl}}) \}, \\ N_i^{\text{spc}} \mid N_i^{\text{ckl}}, p^{\text{spc}}, \theta_{\text{spc}} &\sim \text{Bin} \{ N_i^{\text{ckl}}, p^{\text{spc}}(\mathbf{s}_i, t_i; \theta_{\text{spc}}) \}, \\ Z_i \mid \mu, \theta_{\mu}, \sigma, \theta_{\sigma} &\sim \text{GEV} \{ \mu(\mathbf{s}_i, t_i; \theta_{\mu}), \sigma(\mathbf{s}_i; \theta_{\sigma}), \xi \}, \end{aligned} \right.
 \end{aligned}$$

$X^{\text{pref}}(\cdot) \sim \mathcal{GP}(\omega_1)$

where

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

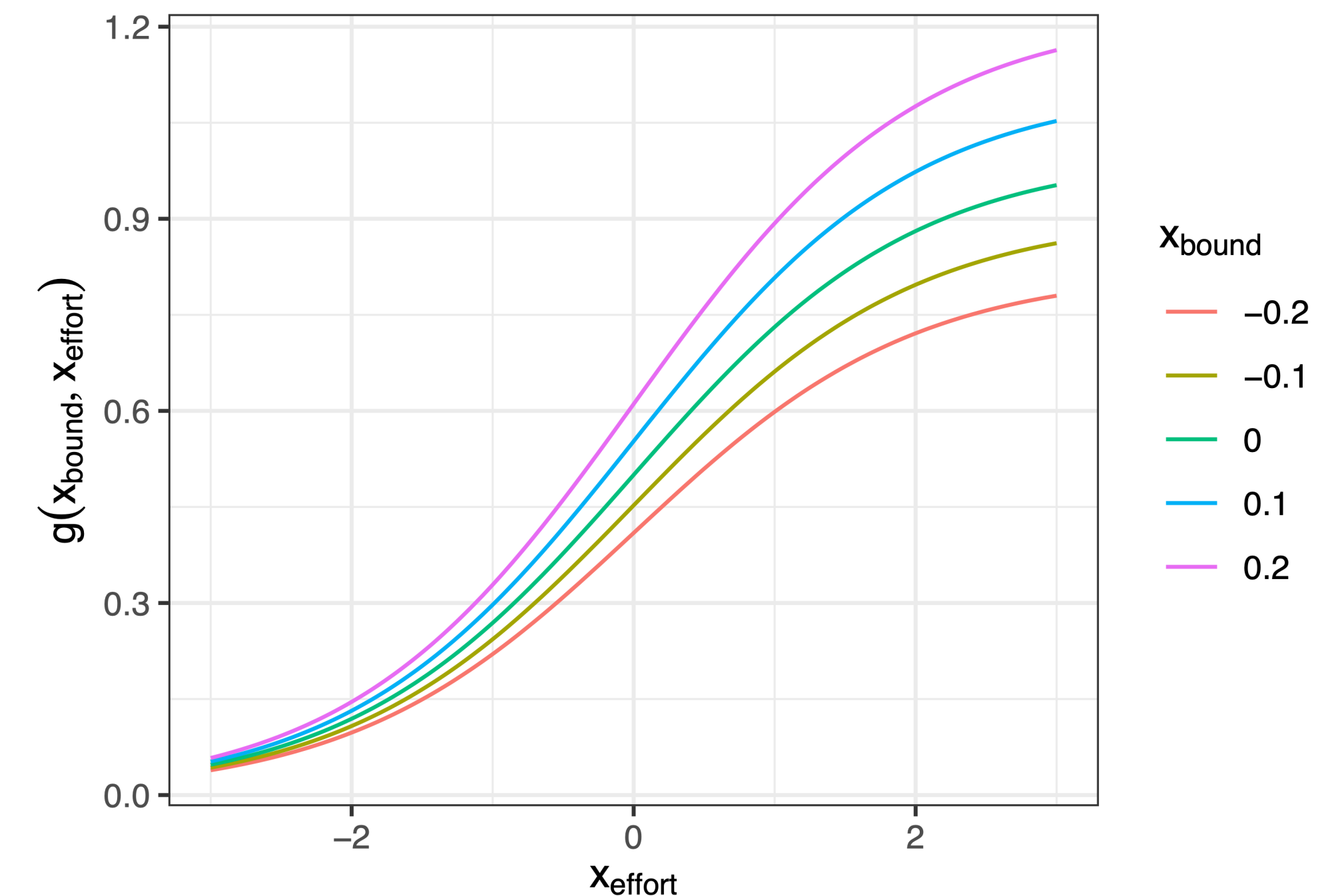
Saturating effect of observational effort

Bayesian
Hierarchical
Models

Extreme-
Value
Theory

- 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



Goodness-of-fit of estimated models

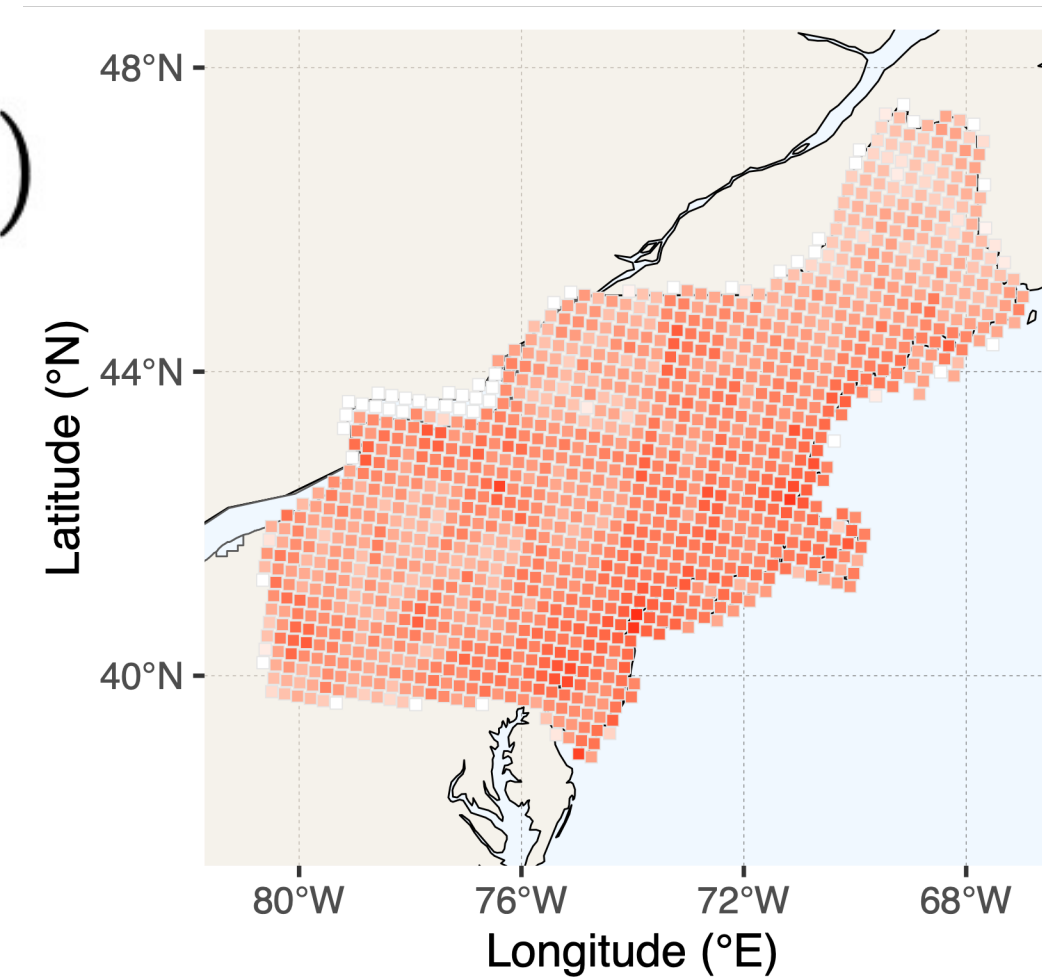
- Generally good match of eBird observations (left maps) with posterior means (right maps)
- Slight differences due to information shared from BBS

Example species:

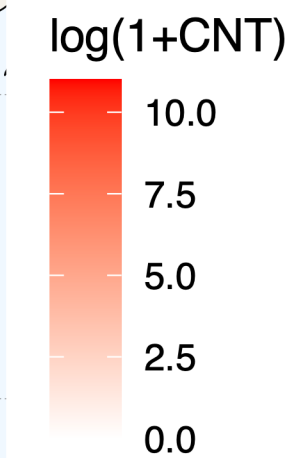
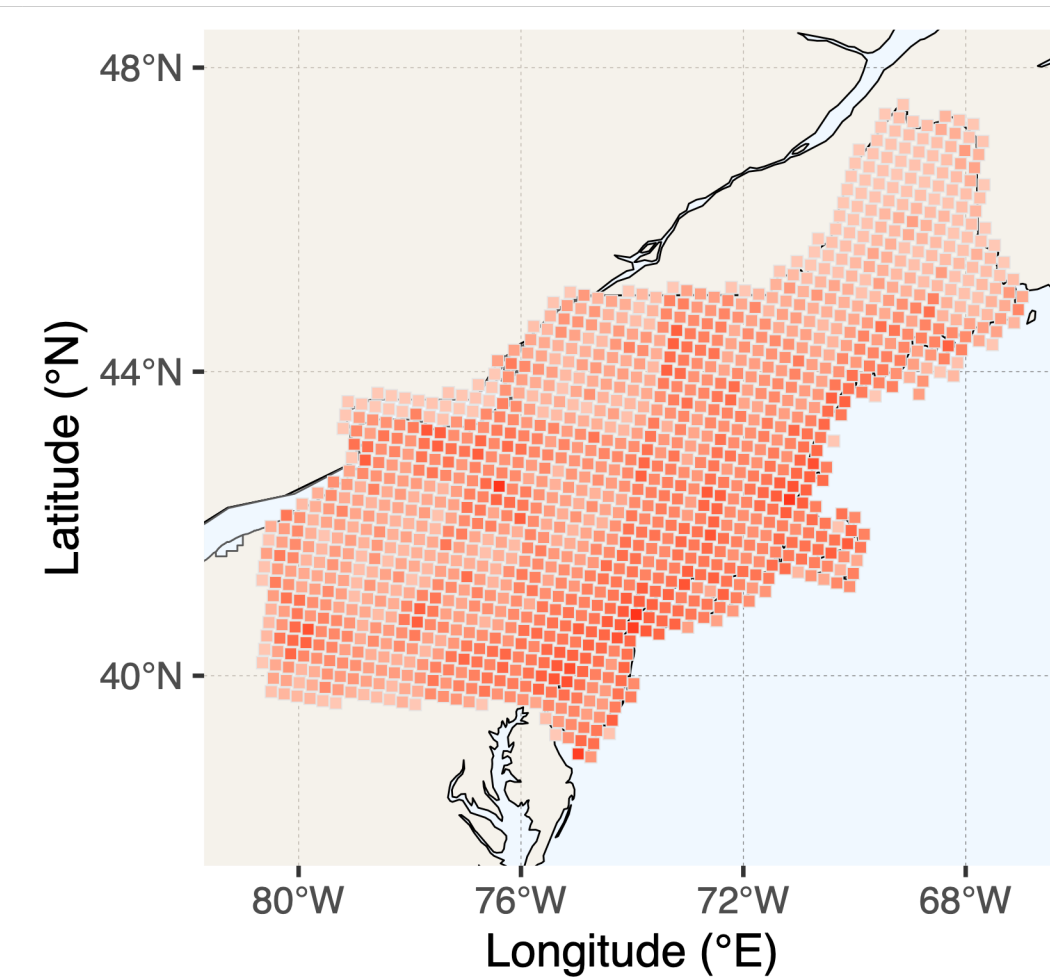


Great Crested Flycatcher

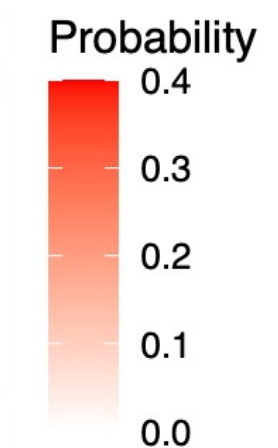
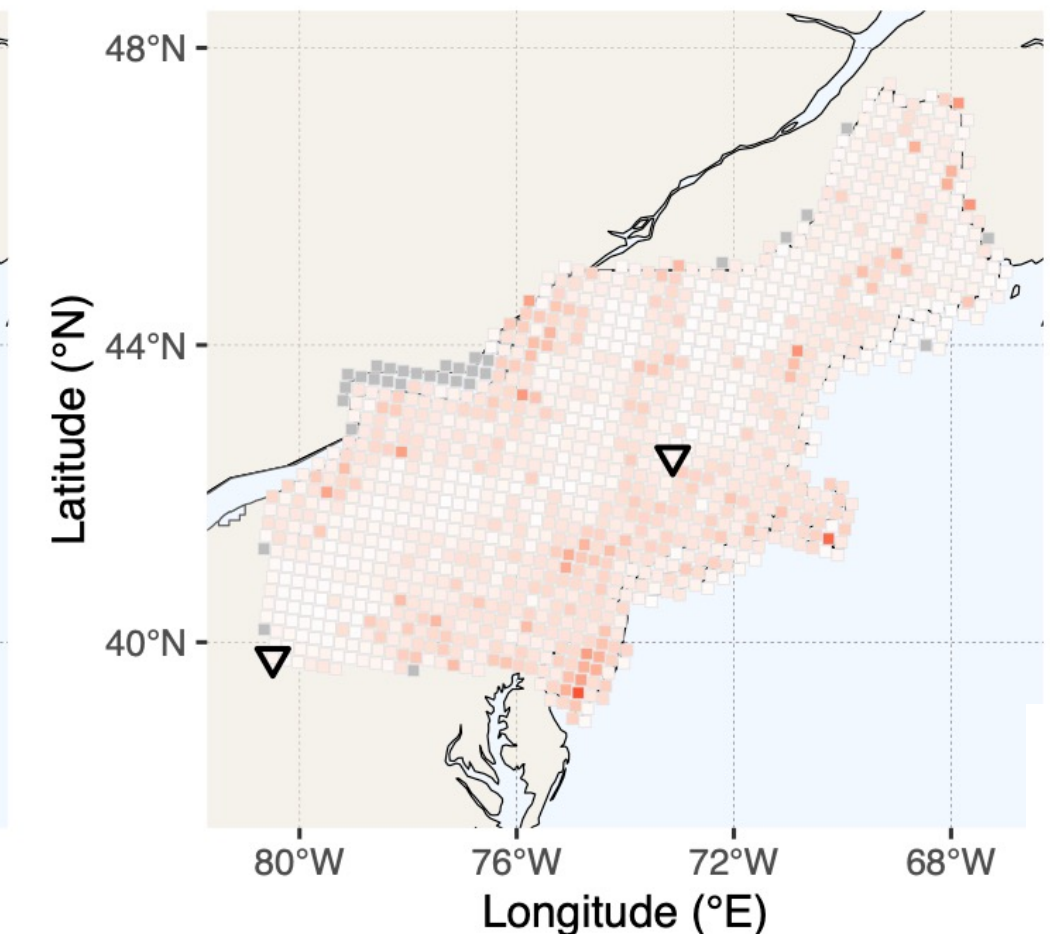
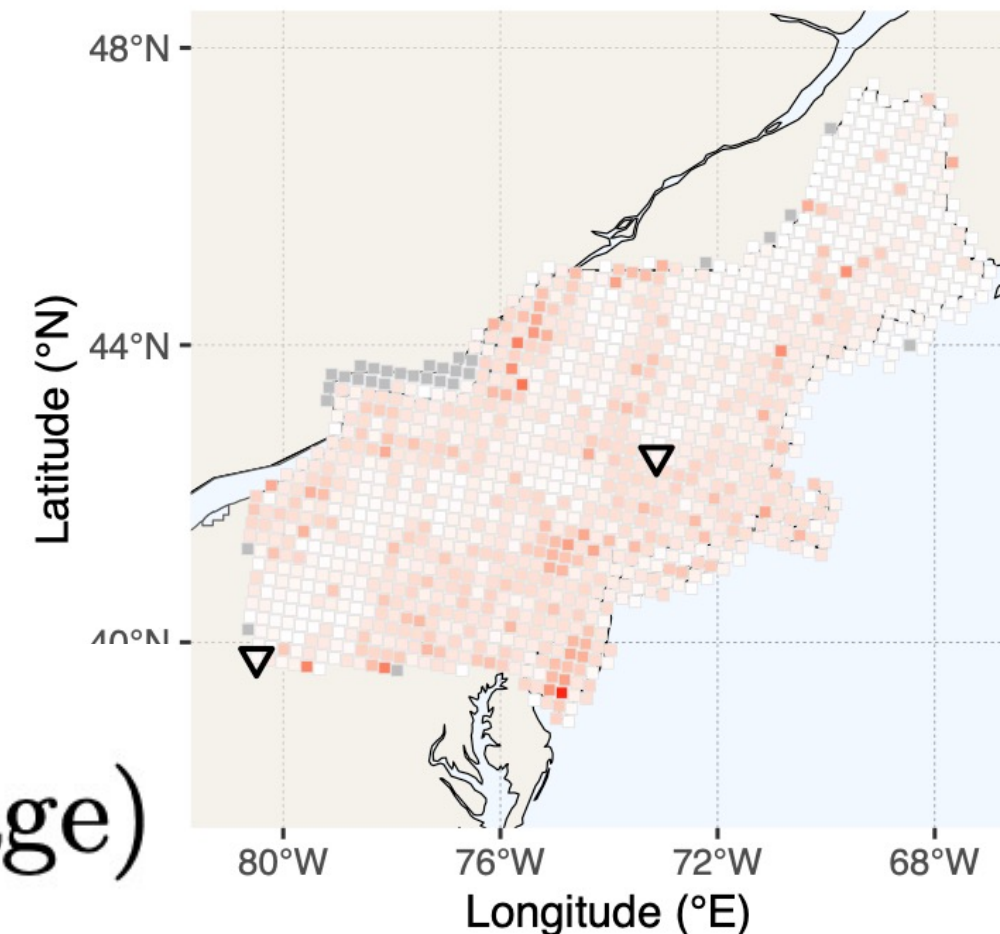
$$N_i^{\text{ckl}}(2021)$$



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



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



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

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

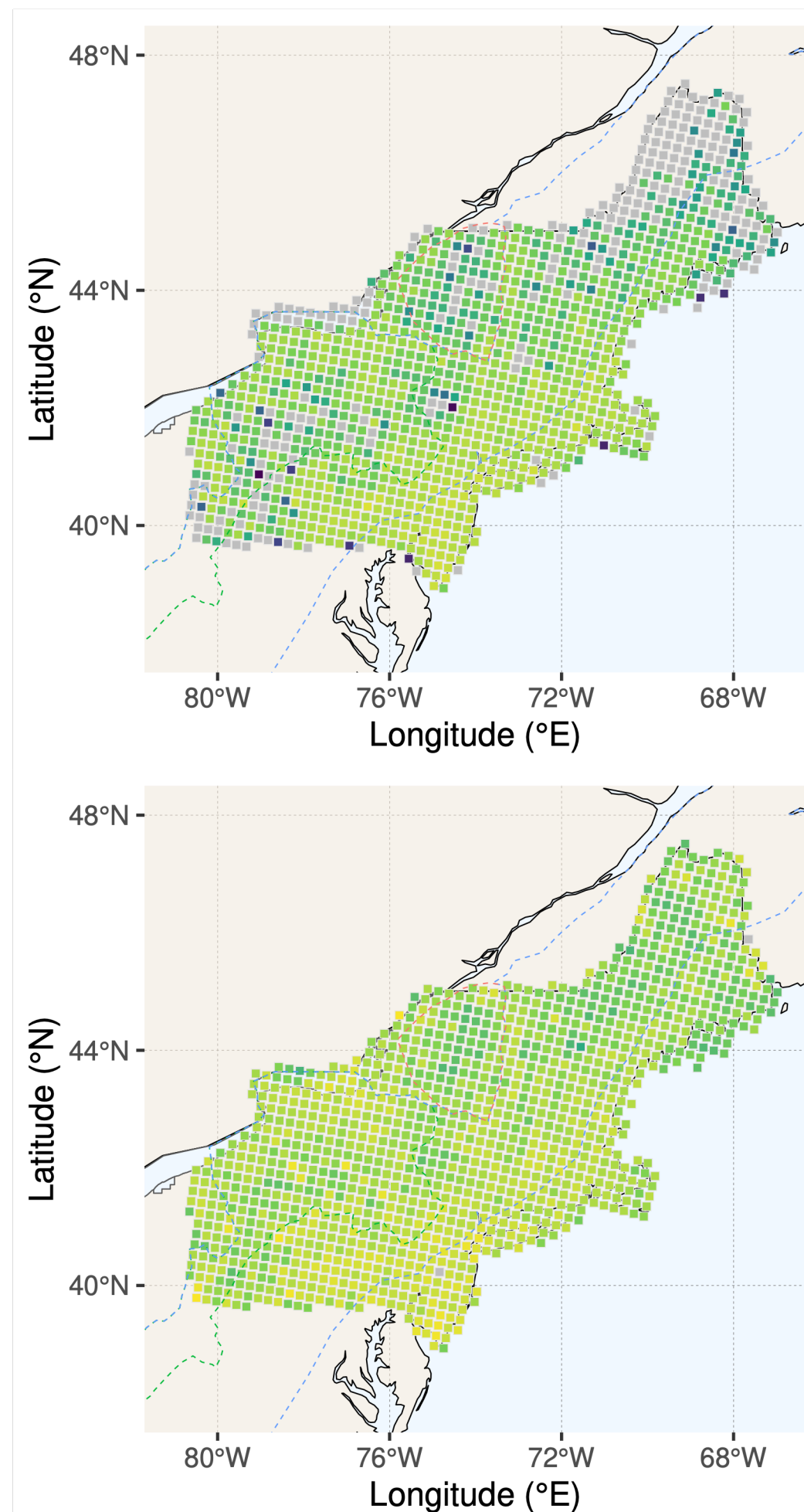
- Based on GEV response distribution
- Bias-corrected prediction by fixing saturated observational effort

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

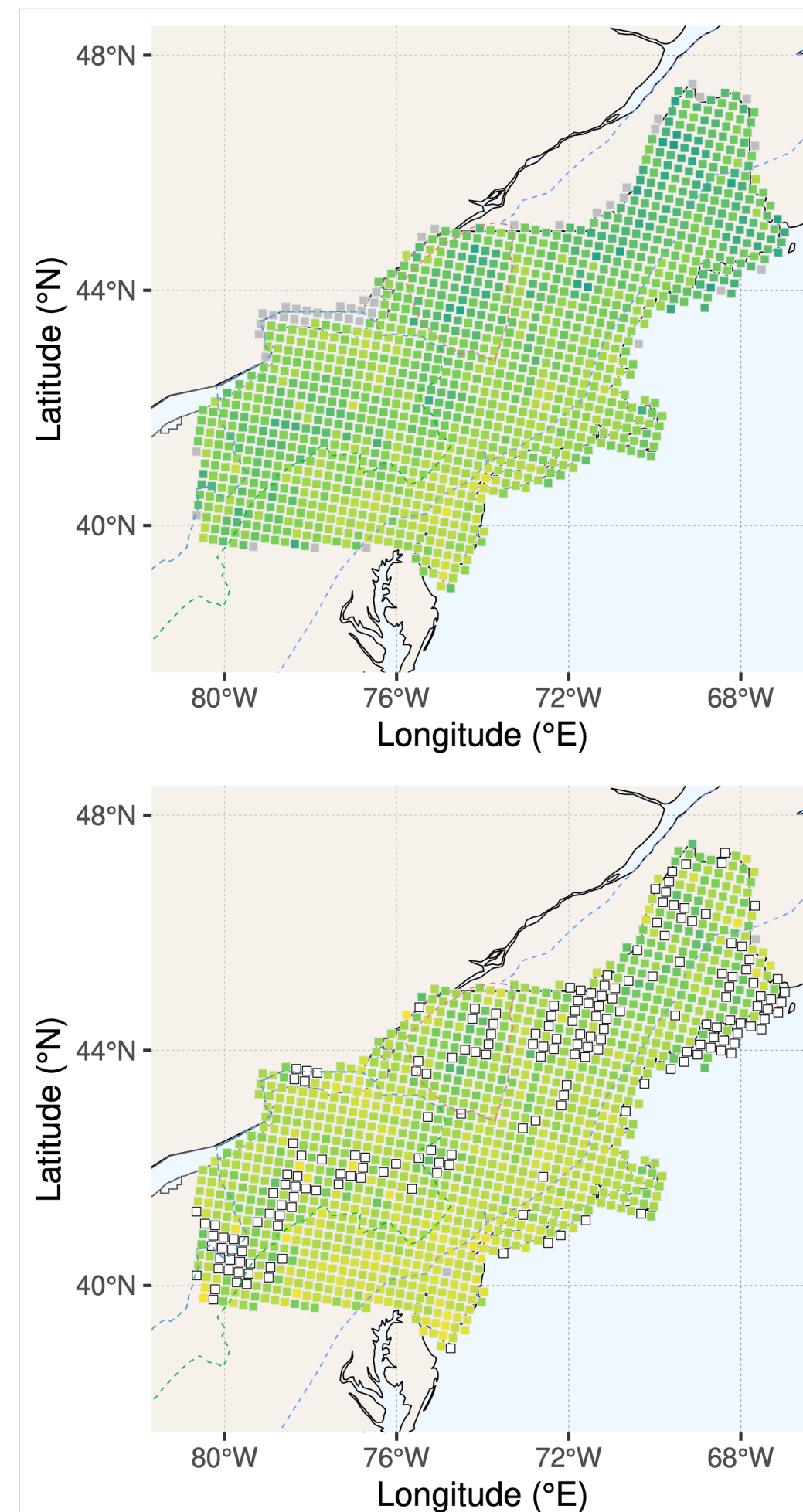


Great Crested
Flycatcher

Observed



Posterior



Posterior predictive
→ Saturated effort

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

Bayesian Hierarchical Models

- 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 speed up inference using **likelihood-free neural estimators**?
- Could we implement shared latent processes in other learning algorithms?
(GAMs, ANNs, Random Forests...)

Discussion: Bias and uncertainty reduction



Citizen Science

- Generating **pseudo-absences** is often crucial
→ Selection algorithm and interpretation of „background“ data is critical!
- Data fusion of opportunistic and structured data in *Integrated Species Distribution Models* is crucial (Fithian et al 2015; Isaac et al 2020)
- Collecting additional locally exhaustive field data may be necessary
→ Explore **optimal sampling design** through simulation studies?
- It remains difficult to publish CS-based scientific findings in top journals
→ Develop theoretical statistical guarantees for CS-based data analysis?
→ Use benchmark datasets to „validate“ methods?

Food for thought

This work:

Koh, Opitz (2025). Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data. Journal of the Royal Statistical Society, Series A (Statistics in Society).

→ Discussion paper + Authors' reply.

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-Infinitely Divisible Process. Journal of Agricultural, Biological and Environmental Statistics.