



Statistiques pour les
Sciences Participatives



Risques, Extrêmes et
Statistique Spatio-Temporelle



Integrating causal methods into spatio-temporal Species Distribution Models : application to gbif data on french butterflies

NICOLAS FERMON

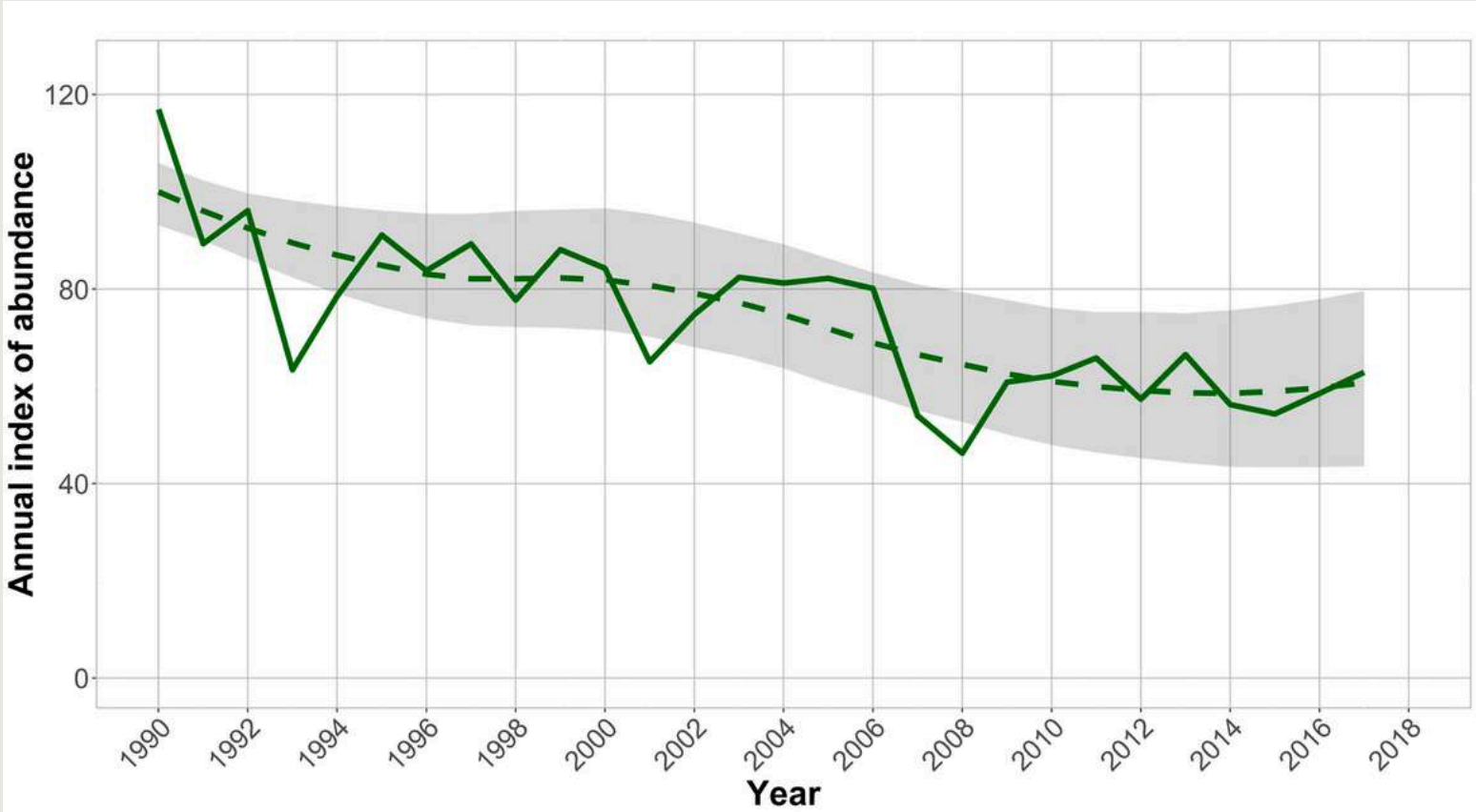
Supervised by
Marianne Tzivanopoulos,
Joaquim Estopinan,
Sara Si-Moussi et
Wilfried Thuiller

02/02/2026



Understanding changes in biodiversity

Biodiversity is declining...



The Grassland Butterfly Indicator for EU countries.
The smoothed line starts at 100, and the shaded areas represent the 95% confidence limits surrounding the smoothed trend¹

But how and why?

- Need to quantify both the changes and the drivers of these changes

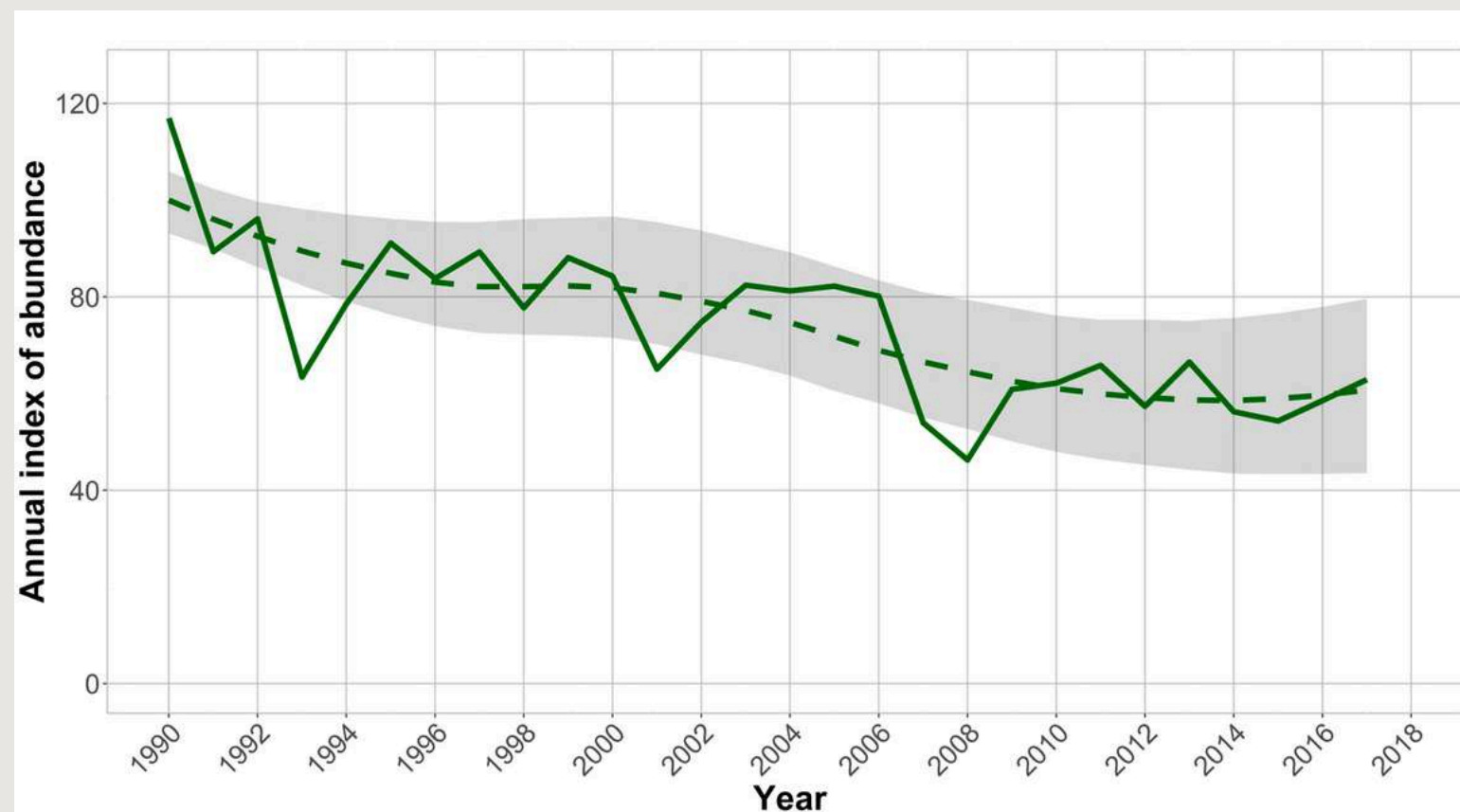
1. Warren and al., 2021

2. Schrödt and al., 2024,

3. Grace, 2024,

Understanding changes in biodiversity

Biodiversity is declining...



The Grassland Butterfly Indicator for EU countries.

The smoothed line starts at 100, and the shaded areas represent the 95% confidence limits surrounding the smoothed trend¹

But how and why?

- Need to quantify both the changes and the drivers of these changes
- Need for understanding **causal relationships**²
 - $X \rightarrow Y$ can be considered a causal relationship if there is reason to think that variations induced in X can propagate to subsequent variations in Y ³
- Correlative methods can lead to biased explanation
 - *Use of stronger assumptions*

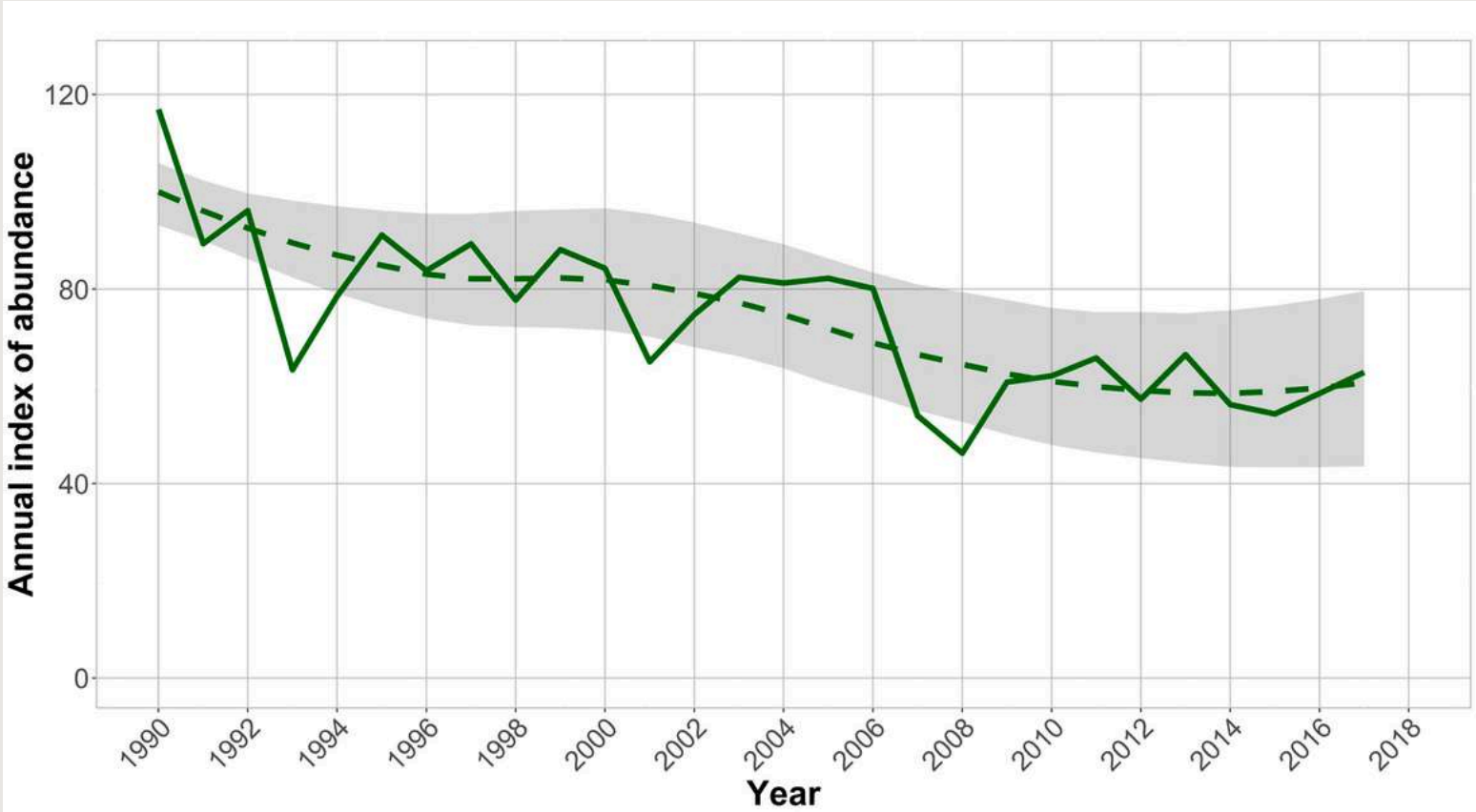
1. Warren and al., 2021

2. Schrödt and al., 2024,

3. Grace, 2024,

Understanding changes in biodiversity

Biodiversity is declining...



The Grassland Butterfly Indicator for EU countries.
The smoothed line starts at 100, and the shaded areas represent the 95% confidence limits surrounding the smoothed trend¹

But how and why?

- Need to quantify both the changes and the drivers of these changes
- Need for understanding **causal relationships**²
 - $X \rightarrow Y$ can be considered a causal relationship if there is reason to think that variations induced in X can propagate to subsequent variations in Y ³
- Correlative methods can lead to biased explanation
 - Use of stronger assumptions
- Causal methods for *drivers attributions*
 - Building causal models with directed acyclic graphs (DAGs)
 - Estimations of causal effects (covariate adjustment, SEM, matching,...)
 - Evaluation of models (sensitivity, interpretability)

• Transparency regarding assumptions

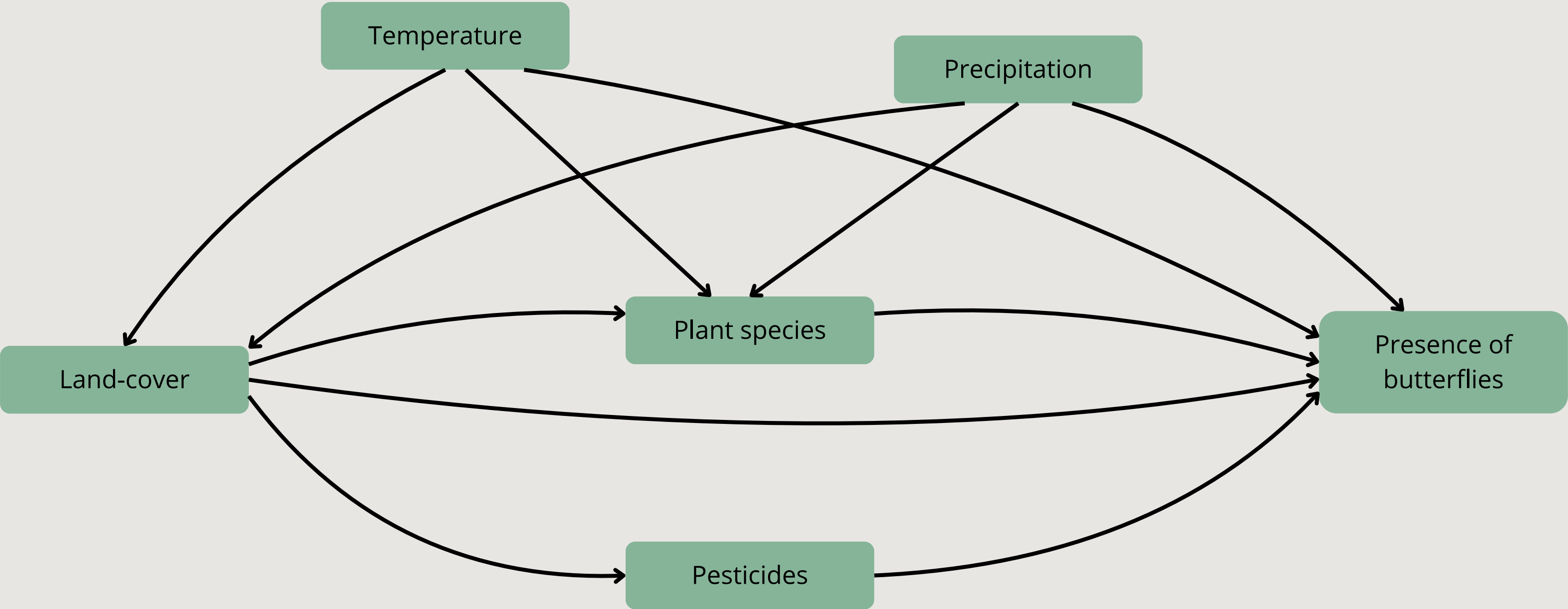
1. Warren and al., 2021

2. Schrödt and al., 2024,

3. Grace, 2024,

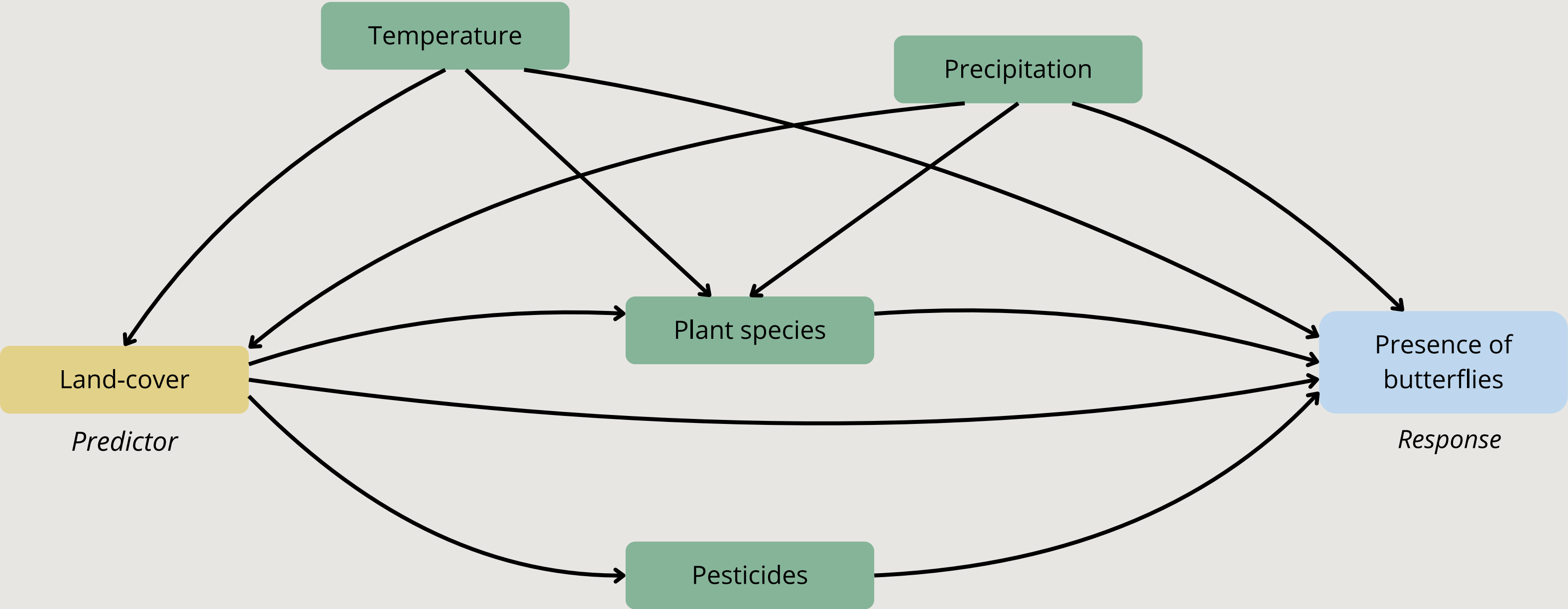
Directed Acyclic Graph

A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹



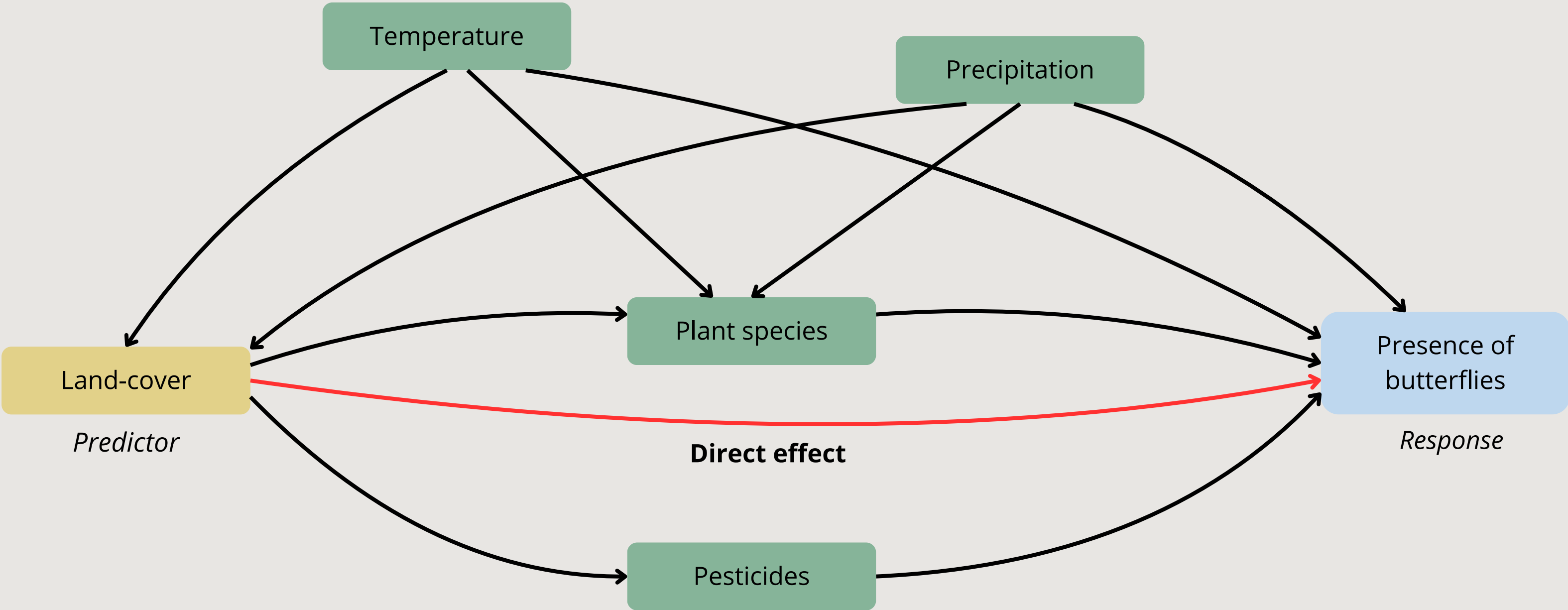
Directed Acyclic Graph

A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹



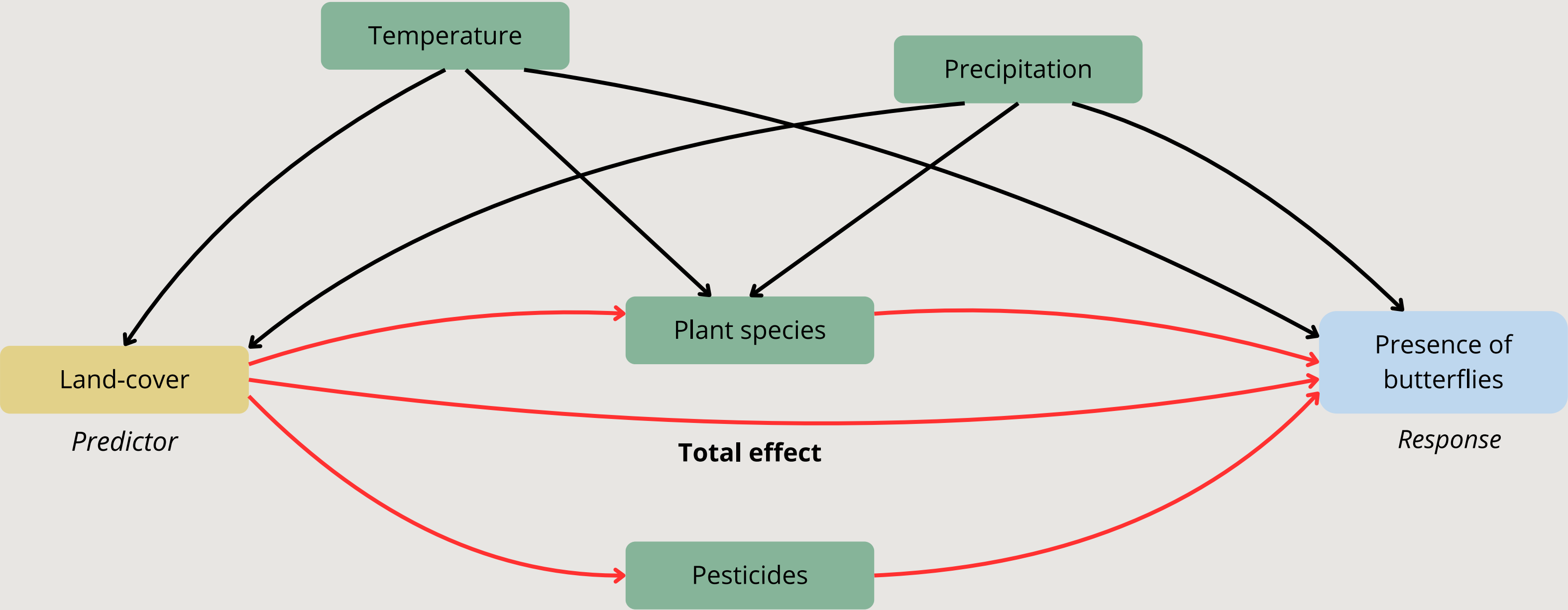
Directed Acyclic Graph

A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹



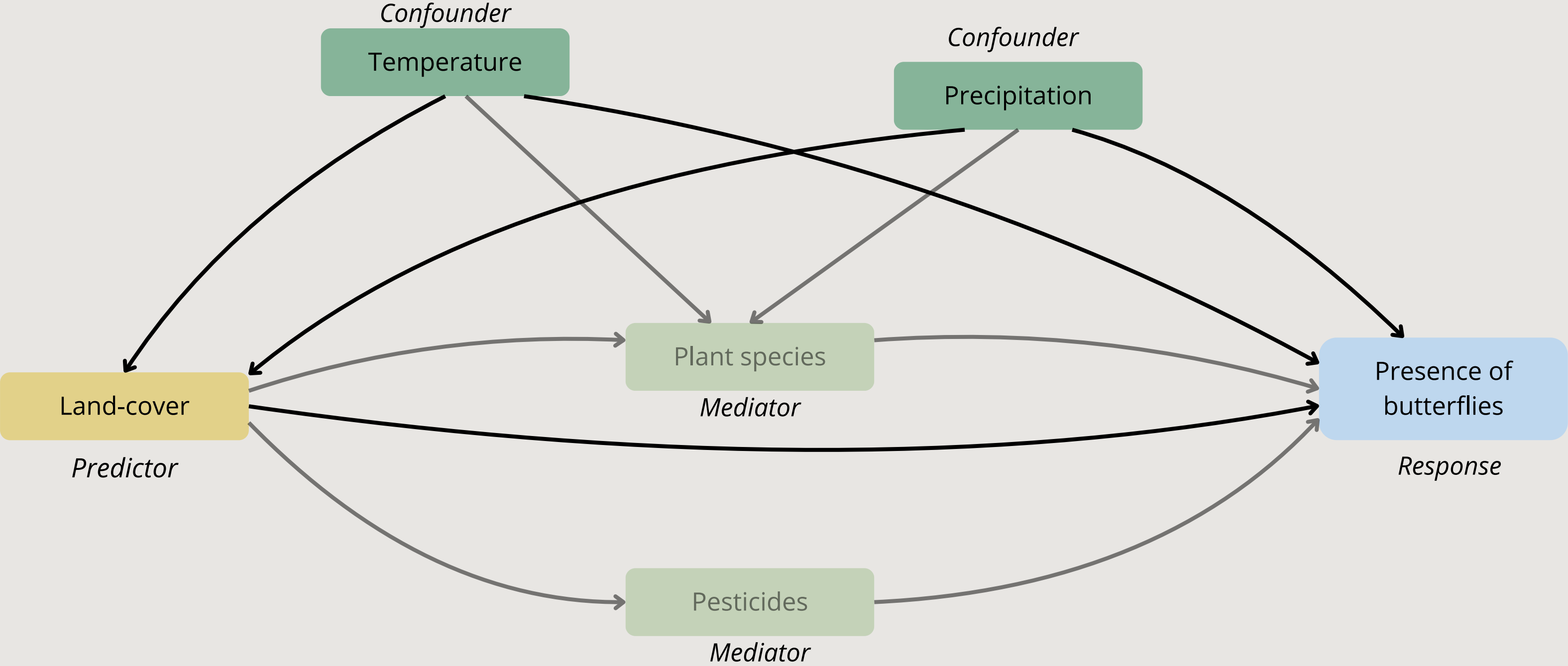
Directed Acyclic Graph

A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹



Directed Acyclic Graph

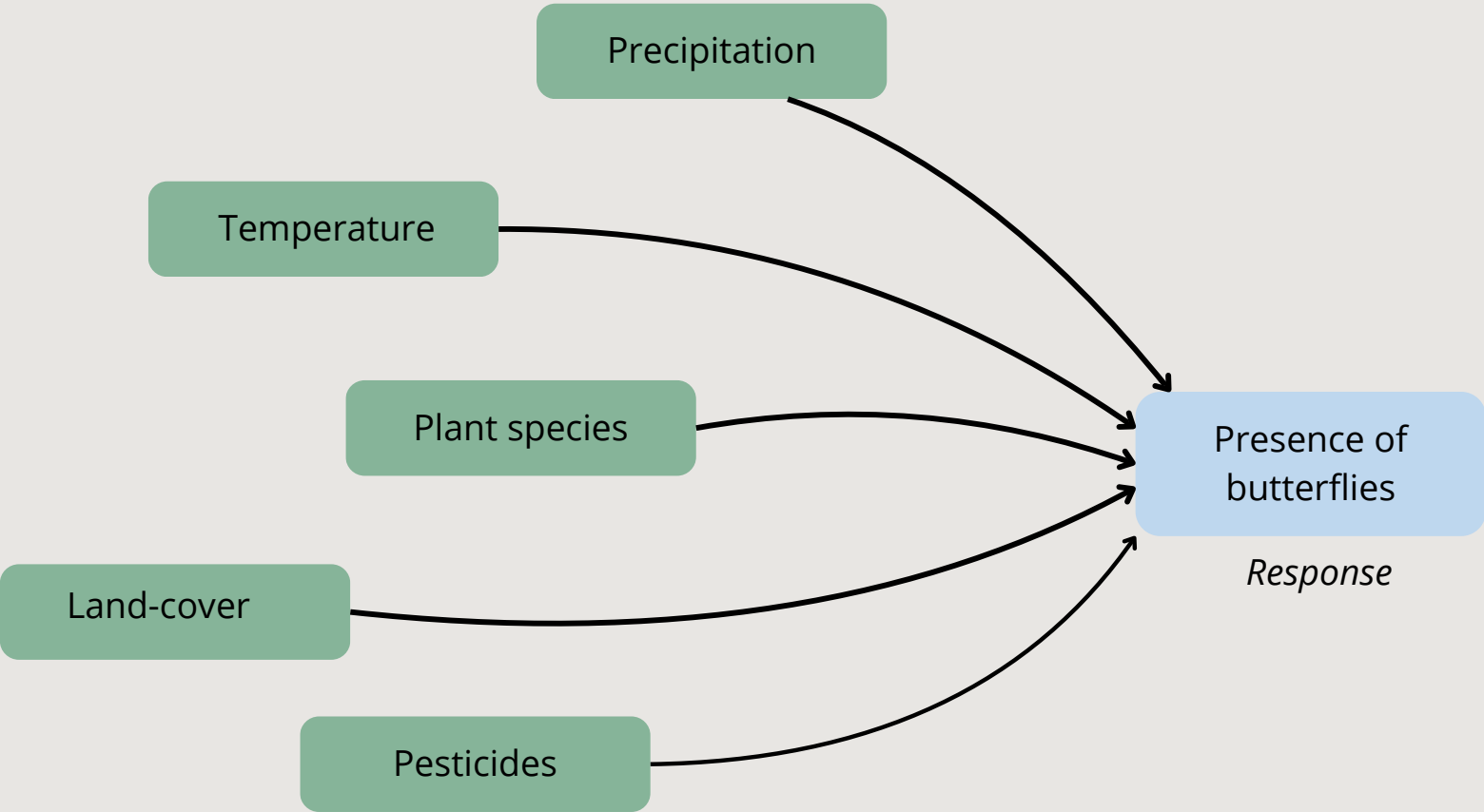
A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹



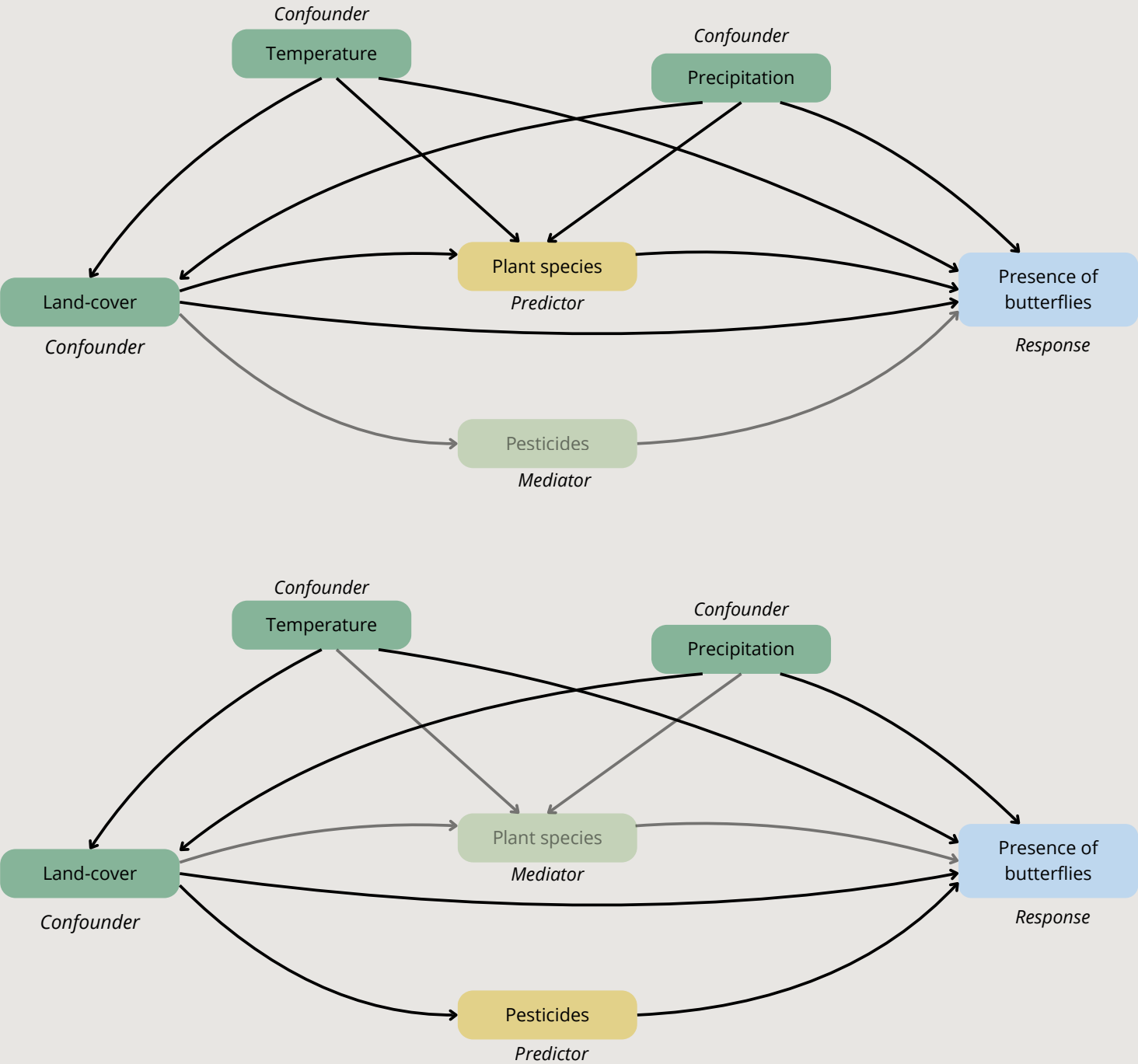
Directed Acyclic Graph

A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹

Traditional “causal salad” approach¹



- Better fit to the data¹
- Less effective generalization²

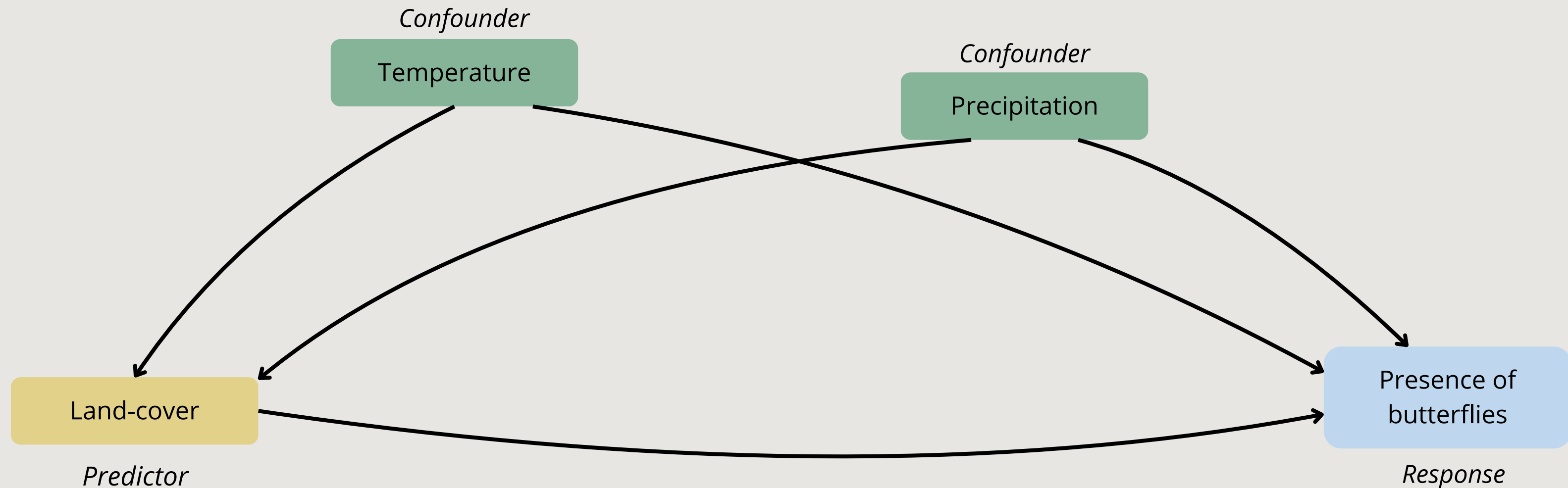


1.Arif and al., 2022

2.Pichler and al., 2023

Directed Acyclic Graph

A set of connected nodes (variables) representing ***hypothetized causal relationships*** within a given system¹



➡ Spatio-temporal modelisation to both detect changes and estimate causal effects

- ***Counterfactual scenarii*** : artificially modify the predictor to estimate the impact on the response
- What if :
 - we had converted 20% of crops to grassland for few years?
 - we had stopped the urbanization ten years ago?

Spatio-temporal SDM

Integrating temporal dimension

- Use spatially and temporally matching data¹
- Add a time effect in a GLM
 - *Captures global trends*
- Use hierarchical bayesian models^{3,4}
 - *Captures spatio-temporal correlations*

INLA approach¹

- Adapted to relatively large datasets
- Construction of a SPDE latent field
- Opportunistic data with presence-only

➡ **Point-process model of Poisson**

Spatio-temporal SDM

Integrating temporal dimension

- Use spatially and temporally matching data¹
- Add a time effect in a GLM
 - Captures global trends
- Use hierarchical bayesian models^{2,3}
 - Captures spatio-temporal correlations

INLA approach¹

- Adapted to relatively large datasets
- Construction of a SPDE latent field
- Opportunistic data with presence-only

➡ Point-process model of Poisson

Mean number of observations

$$Y(s, m, y) | \mu(s, m, y) \sim \text{Poisson}(\mu(s, m, y))$$

$$\log(\mu(s, m, y)) = \log(y_{TG}(s, m, y)) + \beta_0 + \sum_i \beta_i x_i(s, m, y) + W^{(m)}(s) + f(y)$$

Space

Month

Year

Target-group observations

As proxy for sampling effort

Covariates

- Precipitation
- Temperature
- Land-cover

Latent field

- Spatio-temporal Gaussian field
- Intra-annual spatial variations
- Auto-correlation AR1

Inter-annual effect

- Non-spatialized

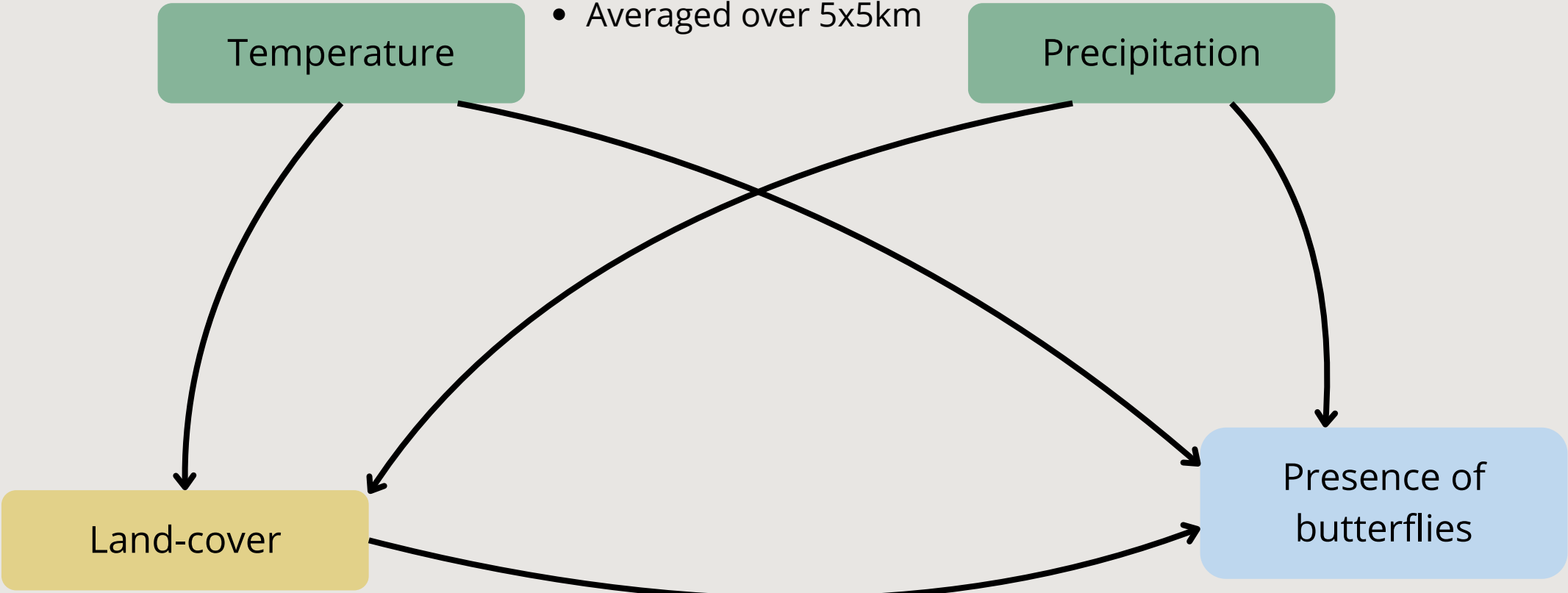
1. Milanesi and al., 2020
2. Lasgorceux and al., 2024
3. Shirey and al., 2026

Datasets

Needs for temporal data

CHELSA-monthly data

- Mean temperature and precipitation
- Monthly from 2001 to 2020
- Averaged over 5x5km

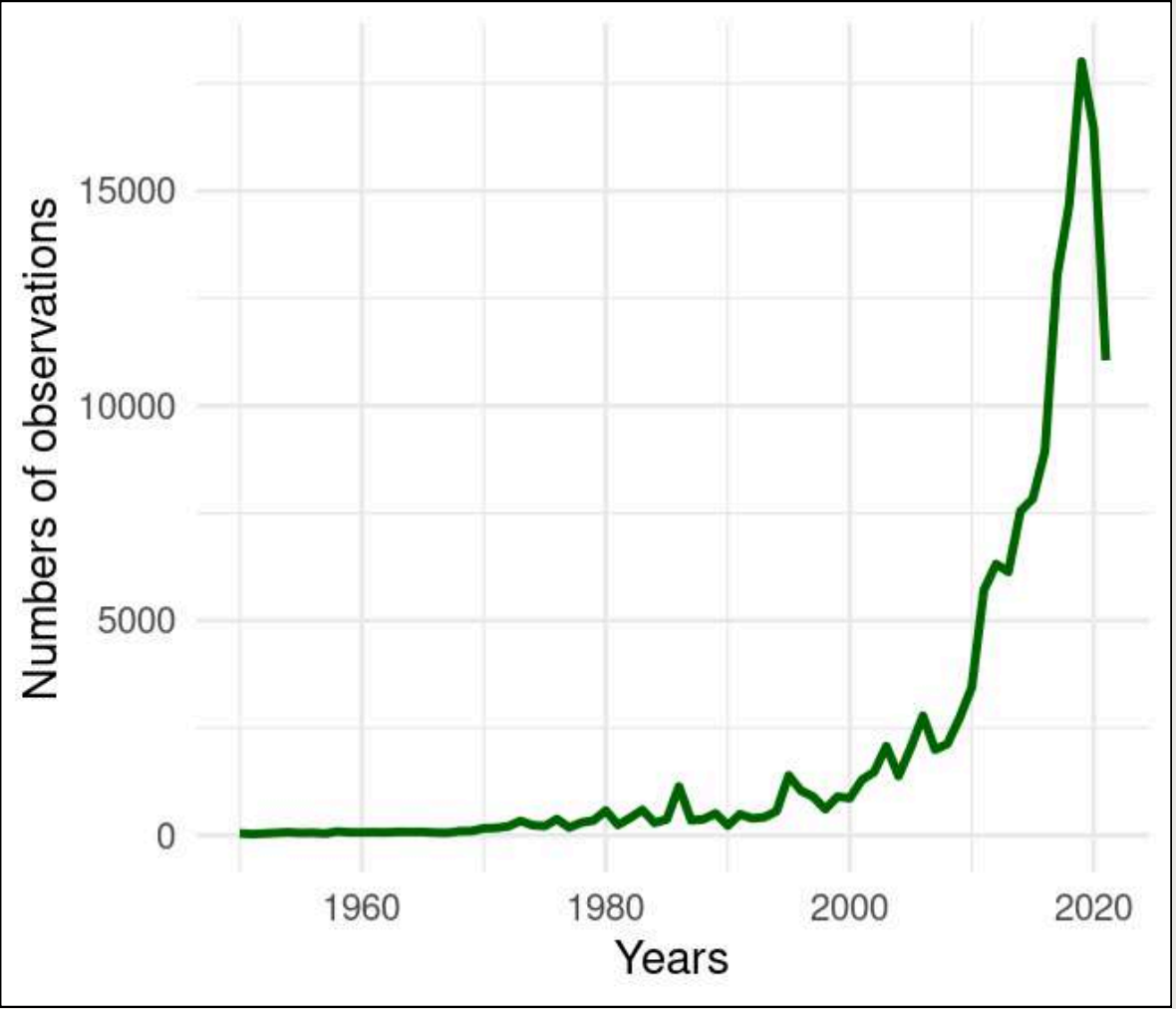


Landsat¹

- 13 classes
- Annual from 2001 to 2020
- Proportion of area for each class on a 5x5 km grid

Gbif

- 30 species of non-generalist butterflies
- Target-group composed of 8 families of butterflies
- From 2001 to 2020
- 5km uncertainty



Trend in the number of observations on GBIF for the 30 studied butterflies species

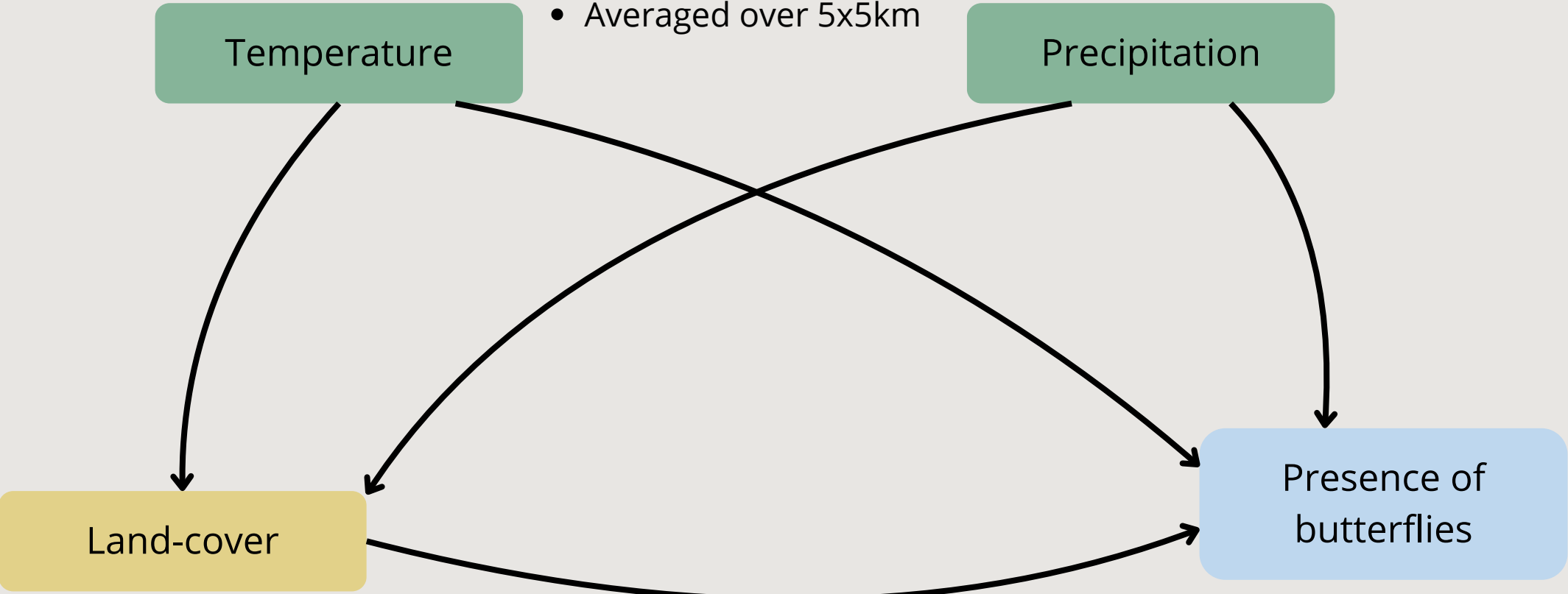
1.Zhang and al., 2024

Datasets

Needs for temporal data

CHELSEA-monthly data

- Mean temperature and precipitation
- Monthly from 2001 to 2020
- Averaged over 5x5km



Landsat¹

- 13 classes
- Annual from 2001 to 2020
- Proportion of area for each class on a 5x5 km grid

Gbif

- 30 species of non-generalist butterflies
- Target-group composed of 8 families of butterflies
- From 2001 to 2020
- 5km uncertainty

Spatial resolution : 5 km

Temporal intra-annual resolution : 6 months

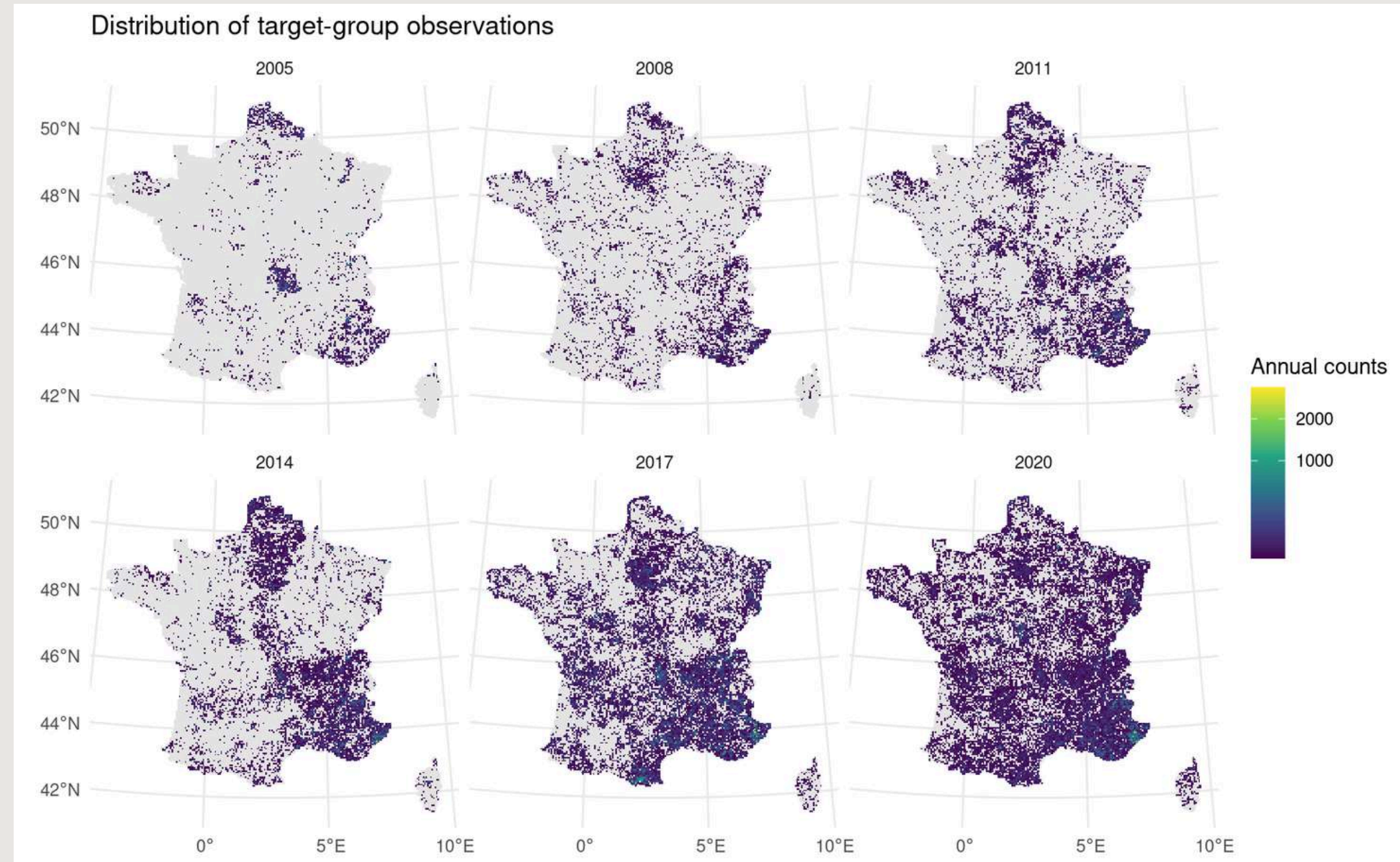
Temporal inter-annual resolution : 1 year

1.Zhang and al., 2024

Dealing with sampling bias

Estimation of sampling effort

- Observations of a larger dataset as indicators of observers presence
- **Target group** composed of 8 families of butterflies
- Used as an offset in the model

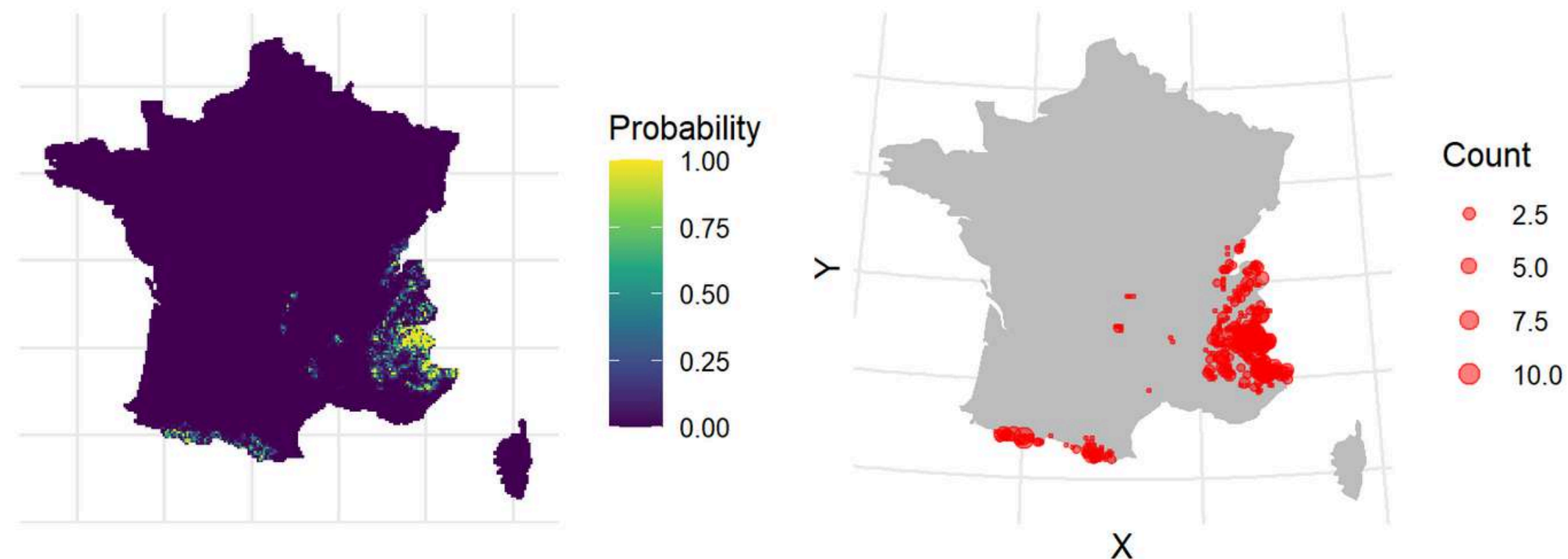


First results

Very new results

Probability of presence – *Parnassius appollo*

Observation counts



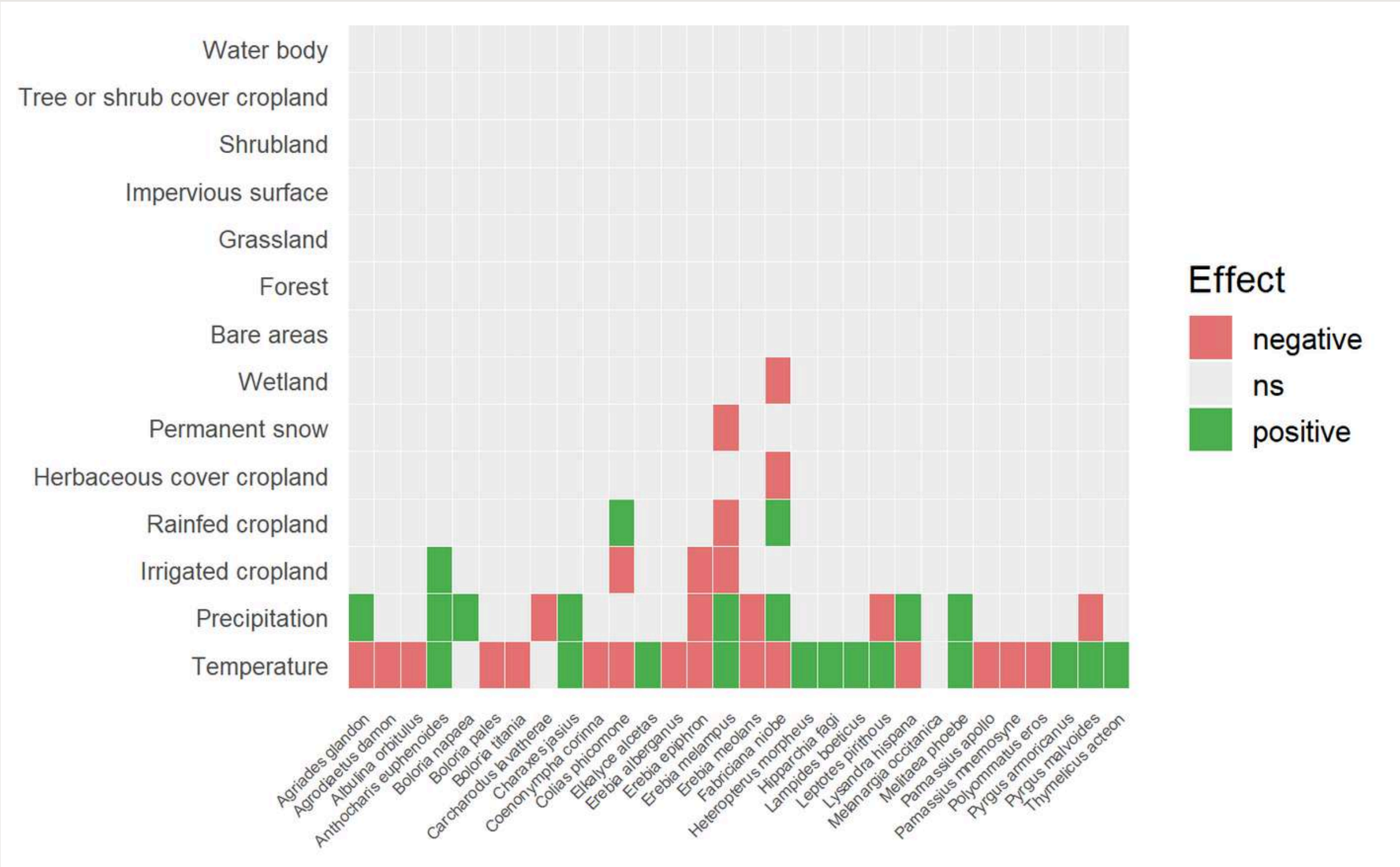
Left : Predicted probability of presence of *Parnassius appollo* in summer 2020;
Right : observation counts in summer 2020

Detect population changes

- Allows us to detect population trends
- Prediction without cross-validation
 - *Spatial, temporal or spatio-temporal?*
- First prediction that seems effective

First results

But not so good results



Estimate causal effects

- Mostly non-significant
- Need to improve the model
 - *Add quadratic covariates*
 - *Variable selection?*
 - *Priors refinement?*

Discussion

Questions to be solved

- Variable selection
 - *Other environmental covariates?*
 - *GDD*
 - *Snow cover*
 - *Previous seasons precipitations/temp?*
 - *Other land-cover covariates?*
 - *Edges*
 - *Human footprint?*
- Model refinement
 - *Priors refinement*
 - *Spatio-temporal K-fold cross validation*
- Interpretability
 - *Weight of latent field in the explanation?*
 - *Interpretability of inter-annual estimate*
- Predictability
 - *Using INLA to predict across space and time*
 - *Construction of scenarii*

From **Lasgorceux and al.**, 2024, *Space-time species distribution modeling for opportunistic presence-only data: a case study of passerines in a protected area*

$$Y_i(s, m, y) | \mu_i(s, m, y) \sim \text{Poisson}(\mu_i(s, m, y))$$

Hypothesis : Species are uniformly sampled

$$\mu_i(s, m, y) = \Lambda_i(s, m, y) \times E_i(s, m, p) = \Lambda_i(s, m, y) \times E(s, m, p)$$

Hypothesis : Abundance and sampling effort act in a multiplicative way

$$= \frac{\Lambda_i(s, m, y)}{\Lambda_{TG}(s, m, y)} \times \Lambda_{TG}(s, m, y) E(s, m, p)$$

Hypothesis : Target-group observations y_{TG} as proxy for the second term

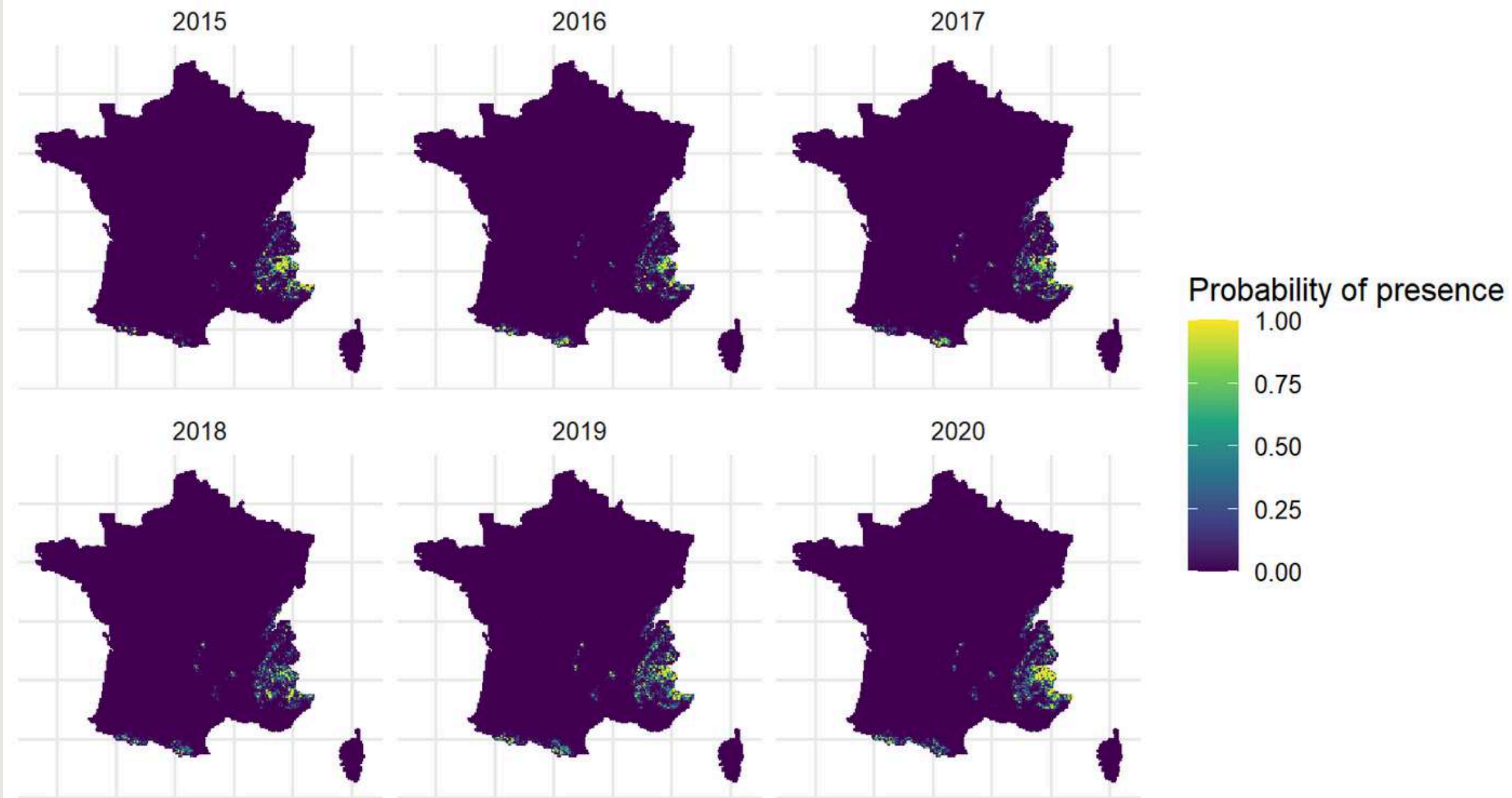
$$\log(\mu_i(s, m, y)) = \log(y_{TG}(s, m, y)) + \log\left(\frac{\Lambda_i(s, m, y)}{\Lambda_{TG}(s, m, y)}\right)$$

$$\log(\mu_i(s, m, y)) = \log(y_{TG}(s, m, y)) + \beta_0 + \sum_j \beta_j x_j(s, m, y) + W^{(m)}(s) + f(y)$$

First results

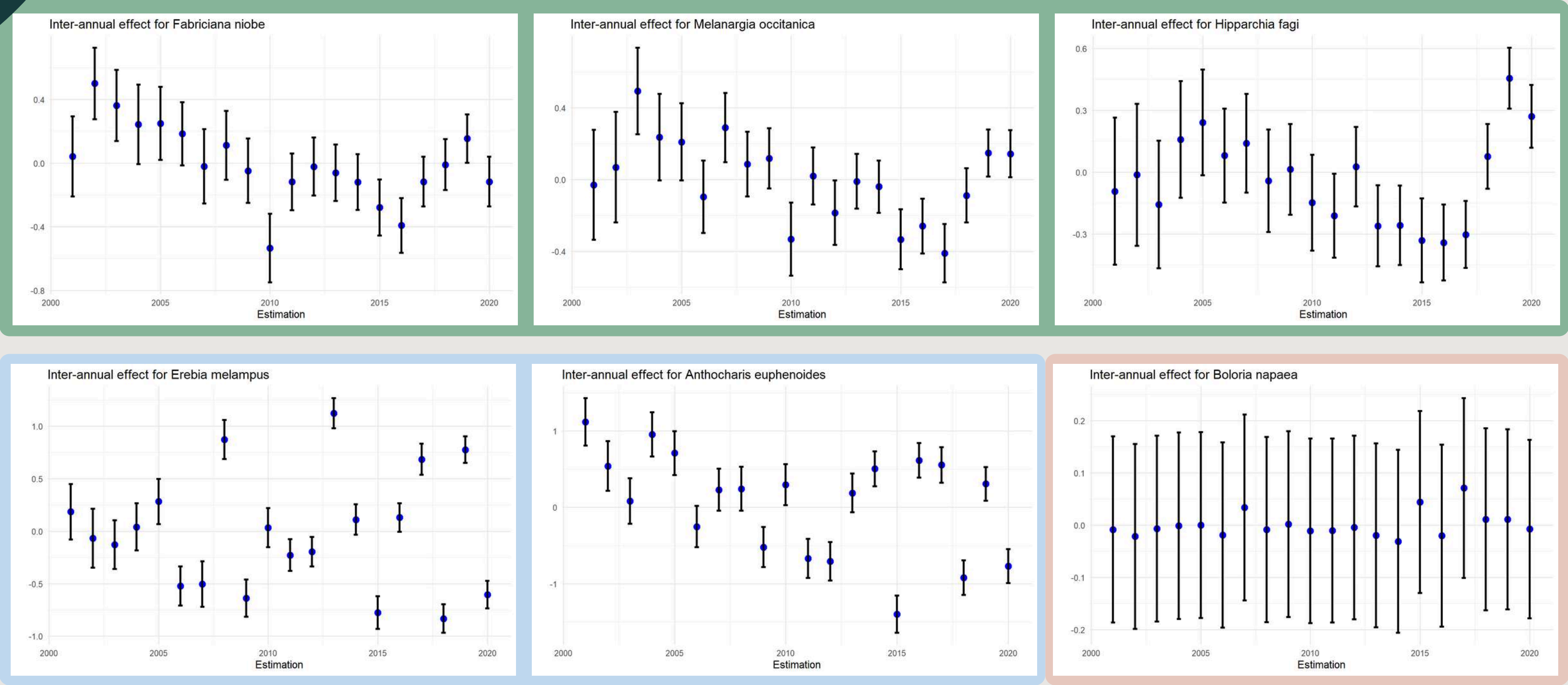
Prediction across space and time

Predictions for different years for *Parnassius apollo*



First results

Time series



SDM as causal tool

Construction of counterfactual scenarii

- Spatio-temporal predictions
- Average treatment effect (ATE) :
$$E(Y_{X=1} - Y_{X=0})$$
- Modelisation of lag-effects or phenological changes

Hypotheses to be tested

- What if :
 - we had converted 20% of crops to grassland for few years?
 - we had stopped the urbanization ten years ago?
 - we let the forest take over the crops? or over the grasslands?

My background

A physicist originally...

- Student of the physics department of the ENS Paris-Saclay

école
normale
supérieure
paris-saclay



... in transition toward ecology...

- **One-year diploma**
“Année de Formation à la recherche en transition écologique”



- **Six-month internship**
in the living-lab VIVALP
“Ecological engineering on mountain slopes : feedback of a nature-based solution”

- Supervised by André Evette, Taina Lemoine and Isabelle Arpin



... still ongoing

- Additional year as student to deepen my skills with internships before applying for a PhD

