# Integrating massive and heterogeneous spatio-temporal data in environmental science and ecology

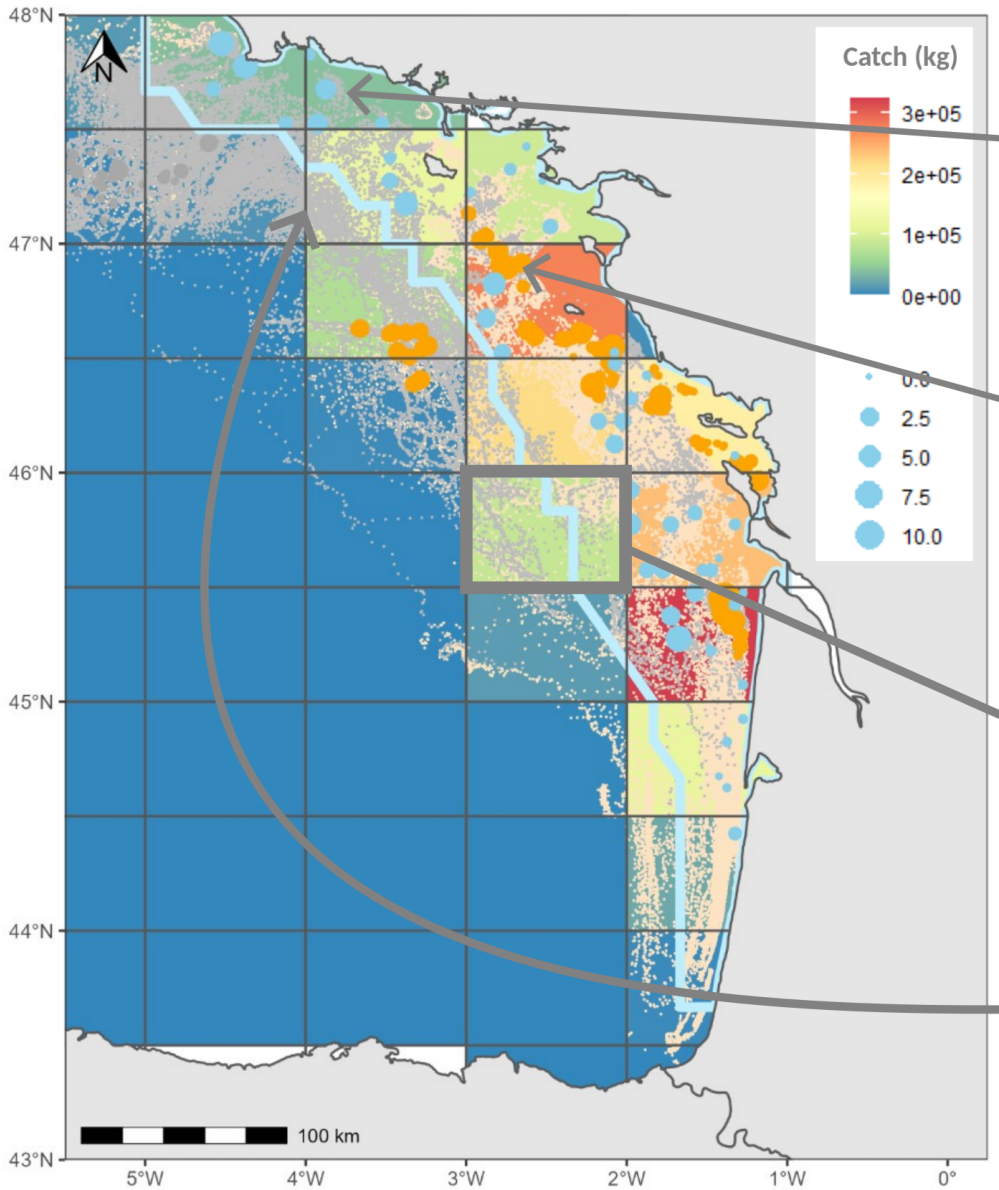
## Fisheries as case of application

Baptiste Alglave, Marie-Pierre Etienne, Youen Vermard,
Mathieu Woillez, Etienne Rivot

June 2024



# Context




**Point-level scientific data**
**(50 obs./year)**

**Point-level onboard observer data**
**(150 obs./semester)**

**Areal-level data**
**(10 000 obs./month)**

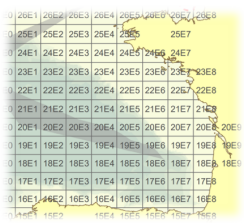
**Fishing locations - VMS**
**(100 000 obs./month)**

# Context

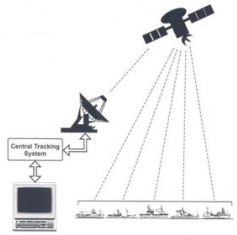
**Scientific survey**


**Onboard observer**


**Catch declarations**

**Vessel Monitoring System (VMS)**




**How to infer spatio-temporal processes from these data?**

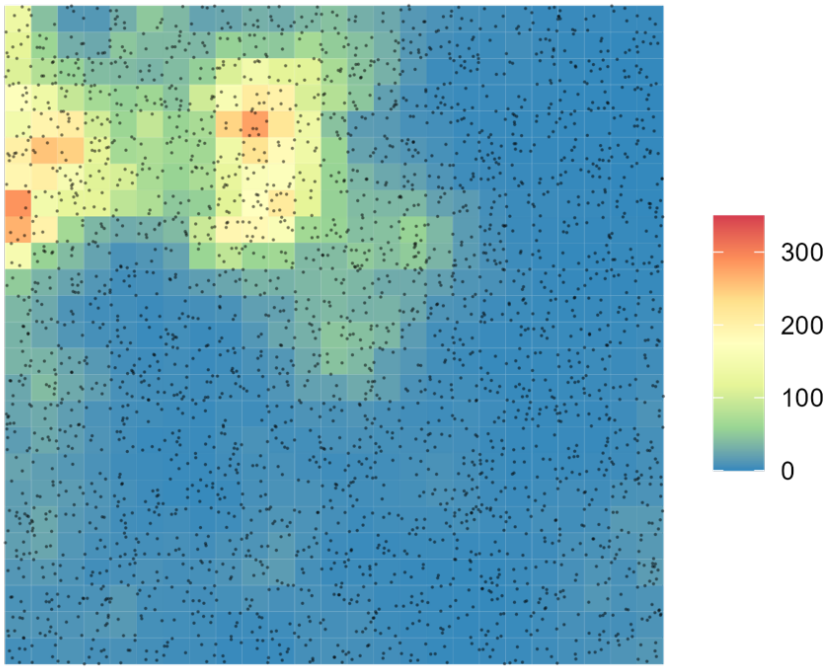
- Handle preferential sampling
- Account for change of support
- Combine these data sources

# Table of Contents

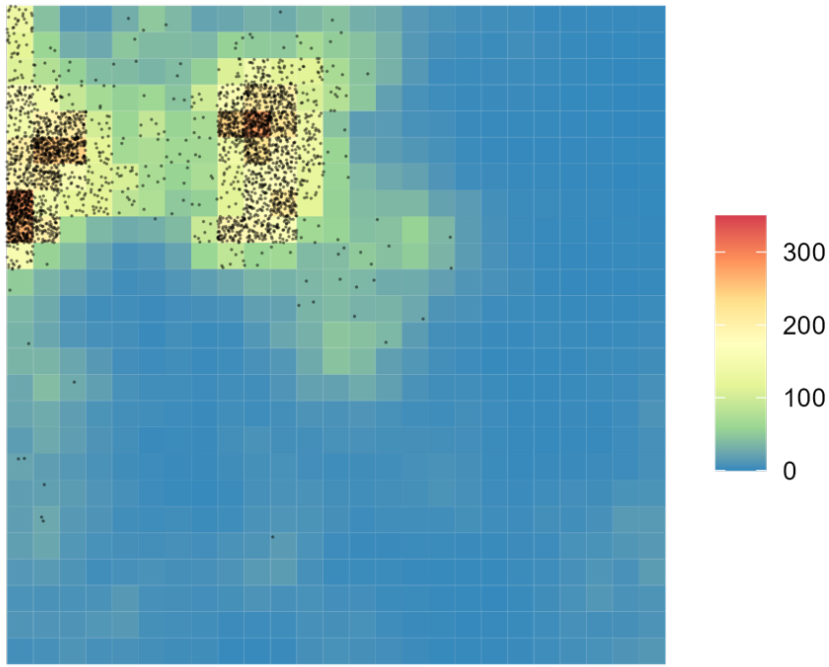
# Table of Contents

# What is preferential sampling?

**Uniform sampling**

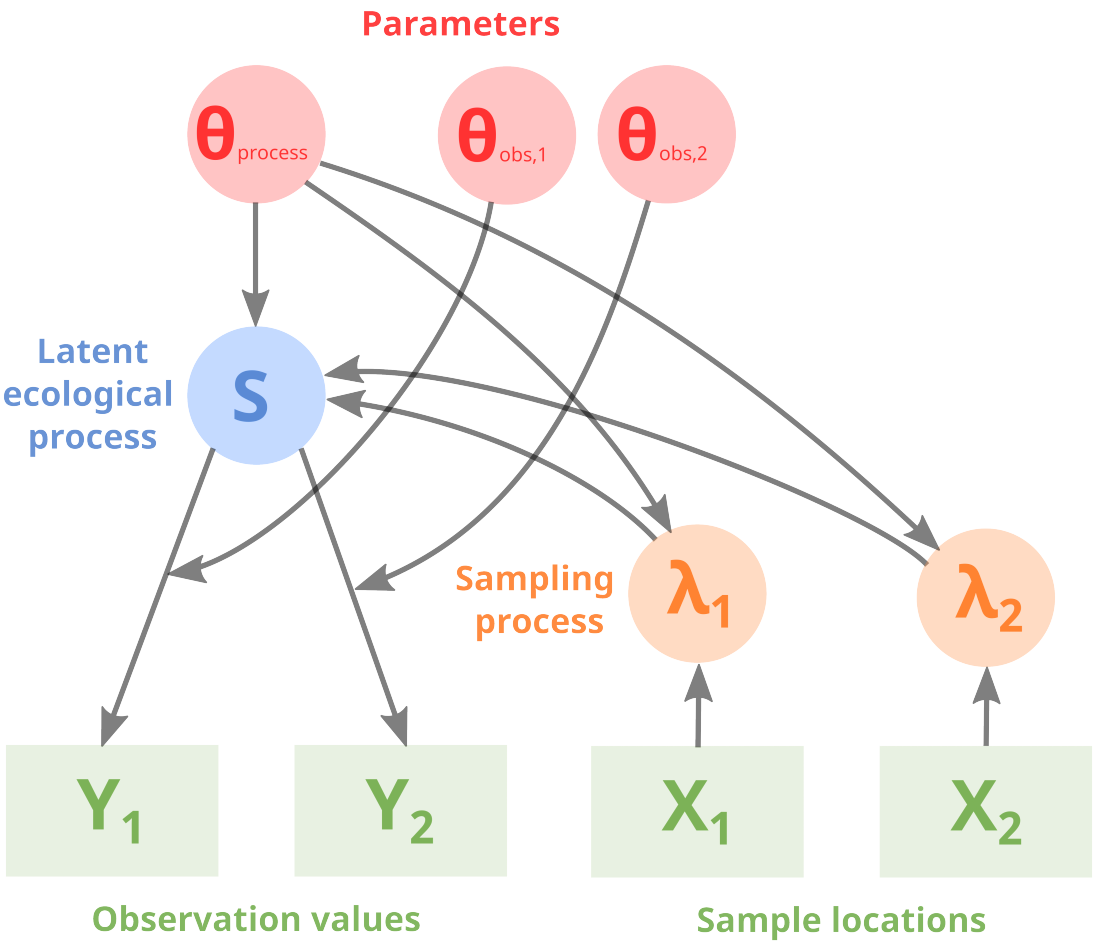



**Strong preferential sampling**




- When sampling locations depend on the process under study

To model the dependence between $X$ and $S$, one solution is to account for $X$ in inference and relate these to $S$ through an extra layer $\lambda$.




**Diggle et al., (2010)**

$$S(x) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

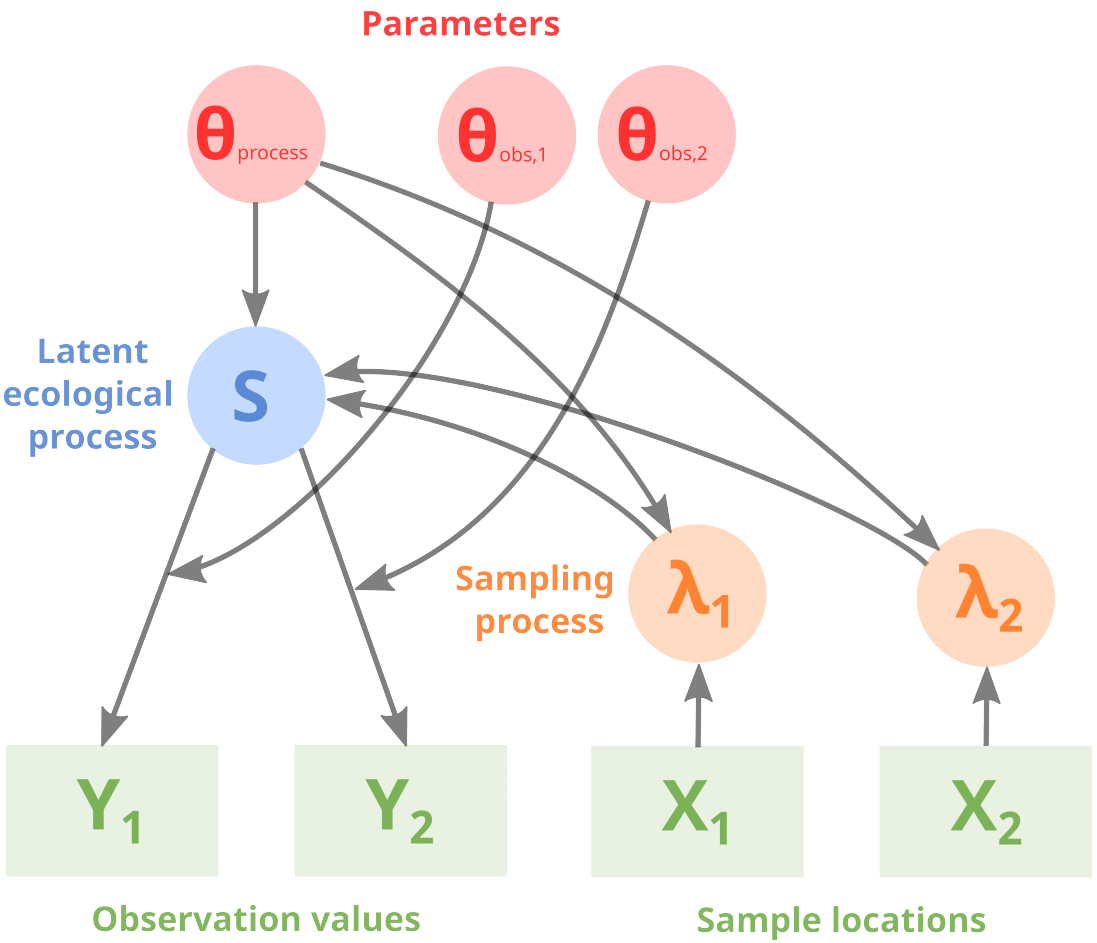
$$X \sim \mathcal{IPP}(\lambda(x))$$
$$\log \lambda(x) = \mu_X + b \cdot S(x)$$

- - - - - - - - - - -

$$Y_i \sim \mathcal{N}(S(x_i), \sigma^2)$$

➠ **Still some limits:** only spatial, only one sampling process, Gaussian observations and latent field, parameterization of preferential sampling (often $X$ is rather a mixture of preferential sampling and other processes).

To model the dependence between $X$ and $S$, one solution is to account for $X$ in inference and relate these to $S$ through an extra layer $\lambda$.




**Diggle et al., (2010)**

$$S(x) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

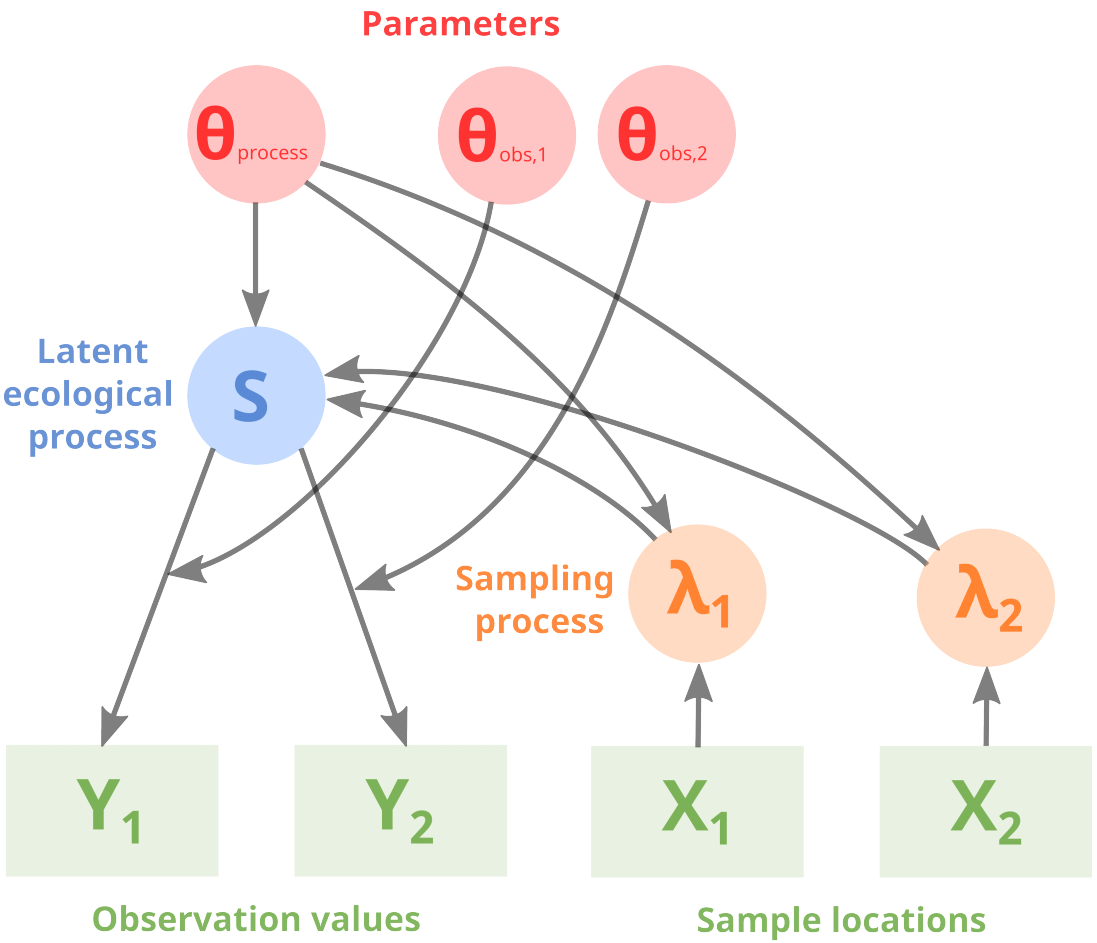
$$X \sim \mathcal{IPP}(\lambda(x))$$
$$\log \lambda(x) = \mu_X + b \cdot S(x)$$

- - - - - - - - - - -

$$Y_i \sim \mathcal{N}(S(x_i), \sigma^2)$$

➠ **Still some limits:** only spatial, only one sampling process, Gaussian observations and latent field, parameterization of preferential sampling (often $X$ is rather a mixture of preferential sampling and other processes).

To model the dependence between $X$ and $S$, one solution is to account for $X$ in inference and relate these to $S$ through an extra layer $\lambda$.




**Alglave et al., (2022)**

$$f(S(x)) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

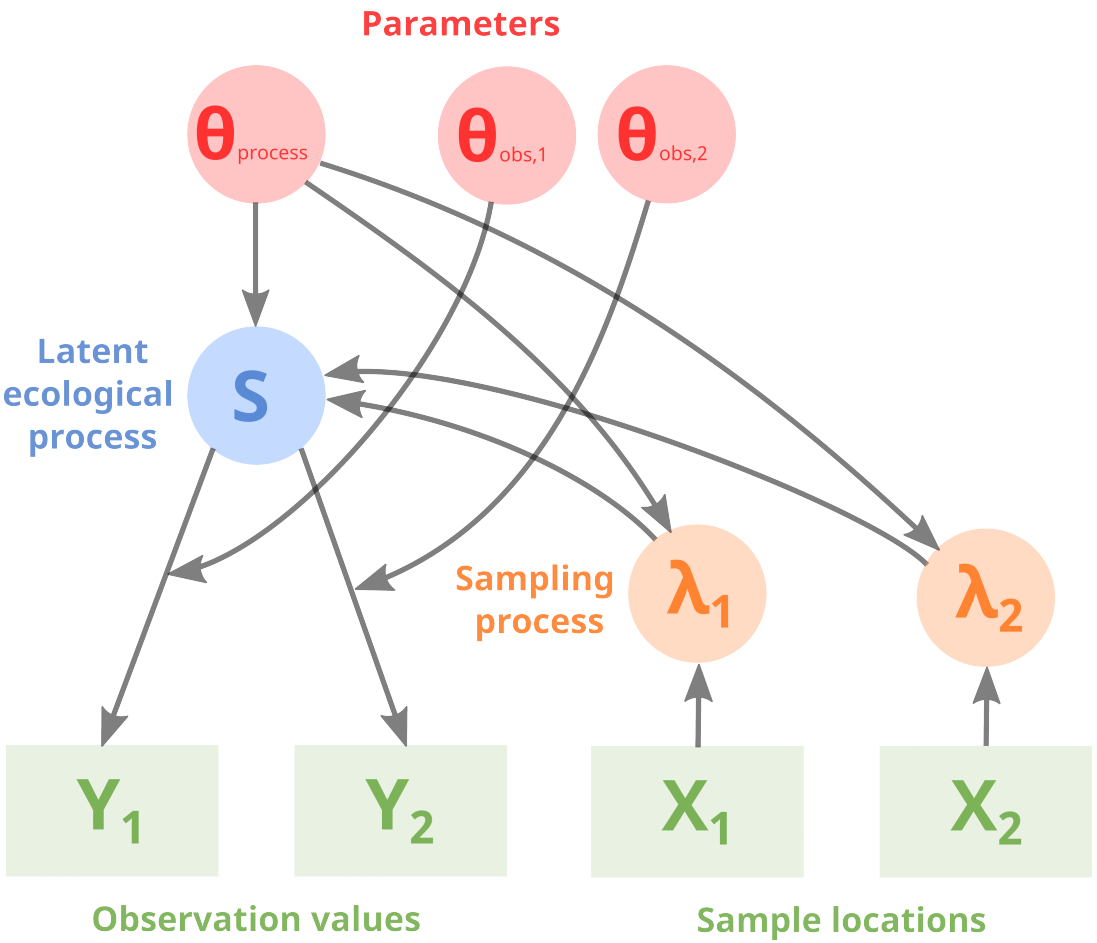
$$X \sim \mathcal{IPP}(\lambda(x))$$
$$\log \lambda(x) = \mu_X + b \cdot f(S(x))$$

- - - - - - - - - - -

$$Y_i \sim \mathcal{L}(S(x_i), \sigma^2)$$

➠ **Still some limits:** only spatial, only one sampling process, **Gaussian observations and latent field**, parameterization of preferential sampling (often $X$ is rather a mixture of preferential sampling and other processes).

To model the dependence between $X$ and $S$, one solution is to account for $X$ in inference and relate these to $S$ through an extra layer $\lambda$.




**Alglave et al., (2022)**

$$f(S(x)) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - - -

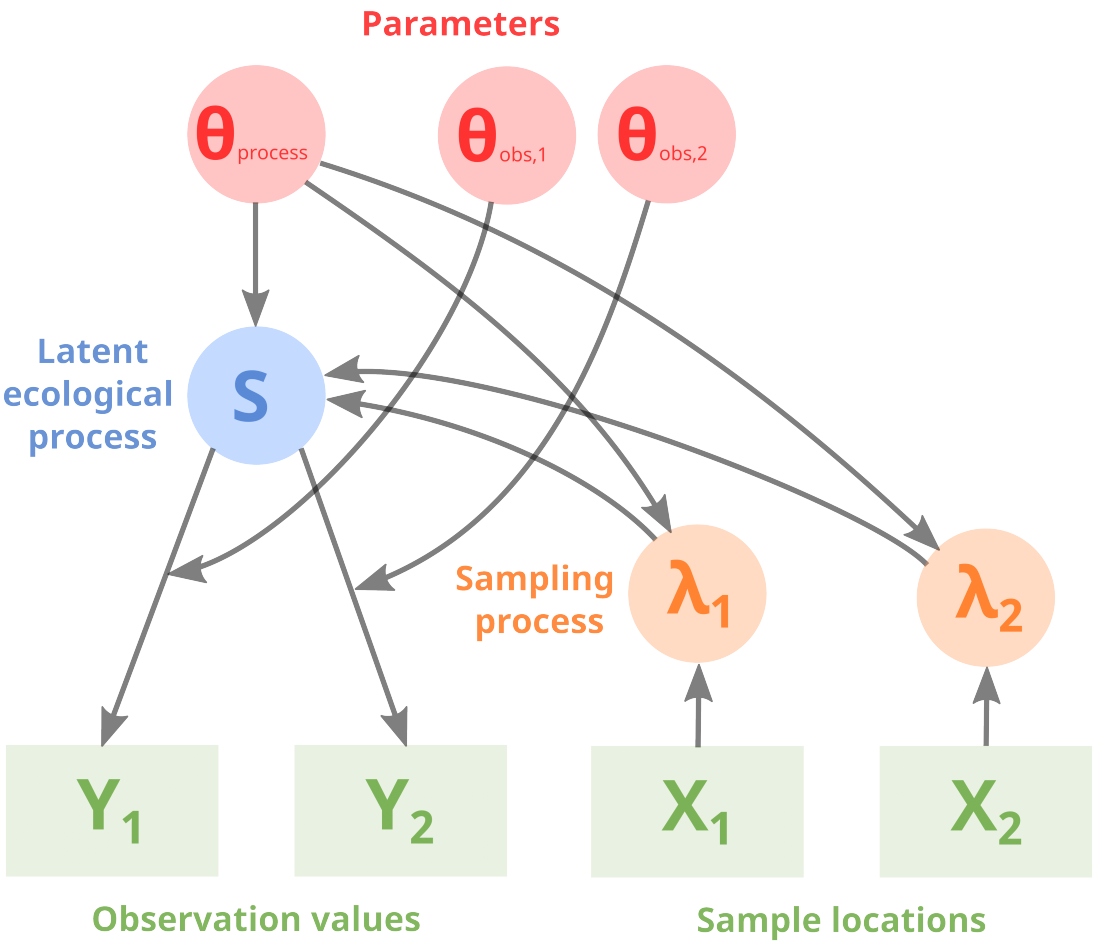
$$X_j \sim \mathcal{IPP}(\lambda_j(x))$$
$$\log \lambda_j(x) = \mu_{X,j} + b_j \cdot f(S(x))$$

- - - - - - - - - - - -

$$Y_i \sim \mathcal{L}(q_j \cdot S(x_i), \sigma_j^2)$$

➠ **Still some limits:** only spatial, **only one sampling process**, Gaussian observations and latent field, parameterization of preferential sampling (often $X$ is rather a mixture of preferential sampling and other processes).

To model the dependence between $X$ and $S$, one solution is to account for $X$ in inference and relate these to $S$ through an extra layer $\lambda$.




**Alglave et al., (2022)**

$$f(S(x)) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

$$X_j \sim \mathcal{IPP}(\lambda_j(x))$$
$$\log \lambda_j(x) = \mu_{X,j} + b_j \cdot f(S(x)) + \Gamma_X(x)^T \cdot \beta_{X,j} + \eta_j(x)$$

- - - - - - - - - - -

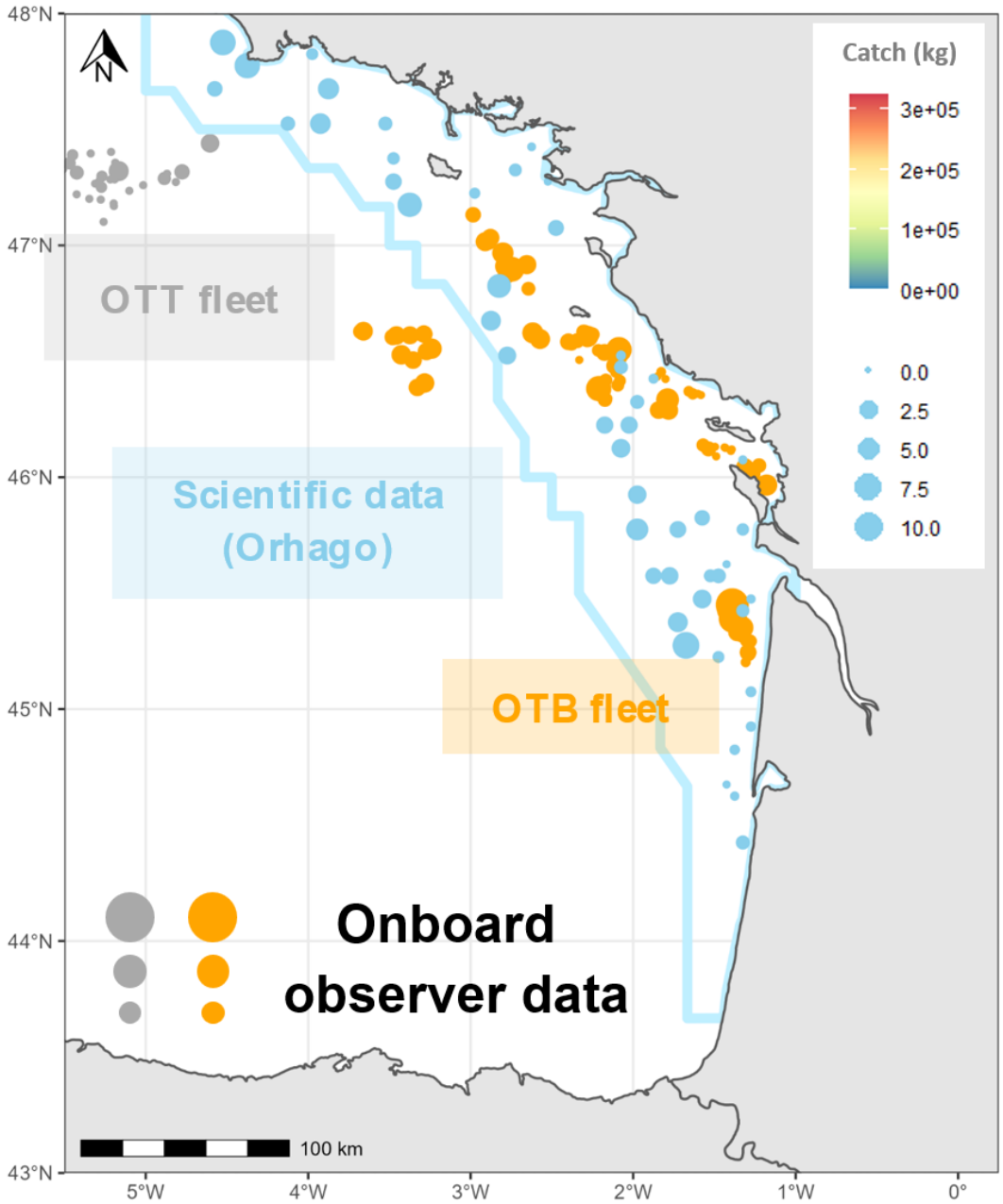
$$Y_i \sim \mathcal{L}(q_j \cdot S(x_i), \sigma_j^2)$$

➠ **Still some limits:** only spatial, only one sampling process, Gaussian observations and latent field, **parameterization of preferential sampling (often $X$ is rather a mixture of preferential sampling and other processes)**.

**Onboard observer**

Available over whole year

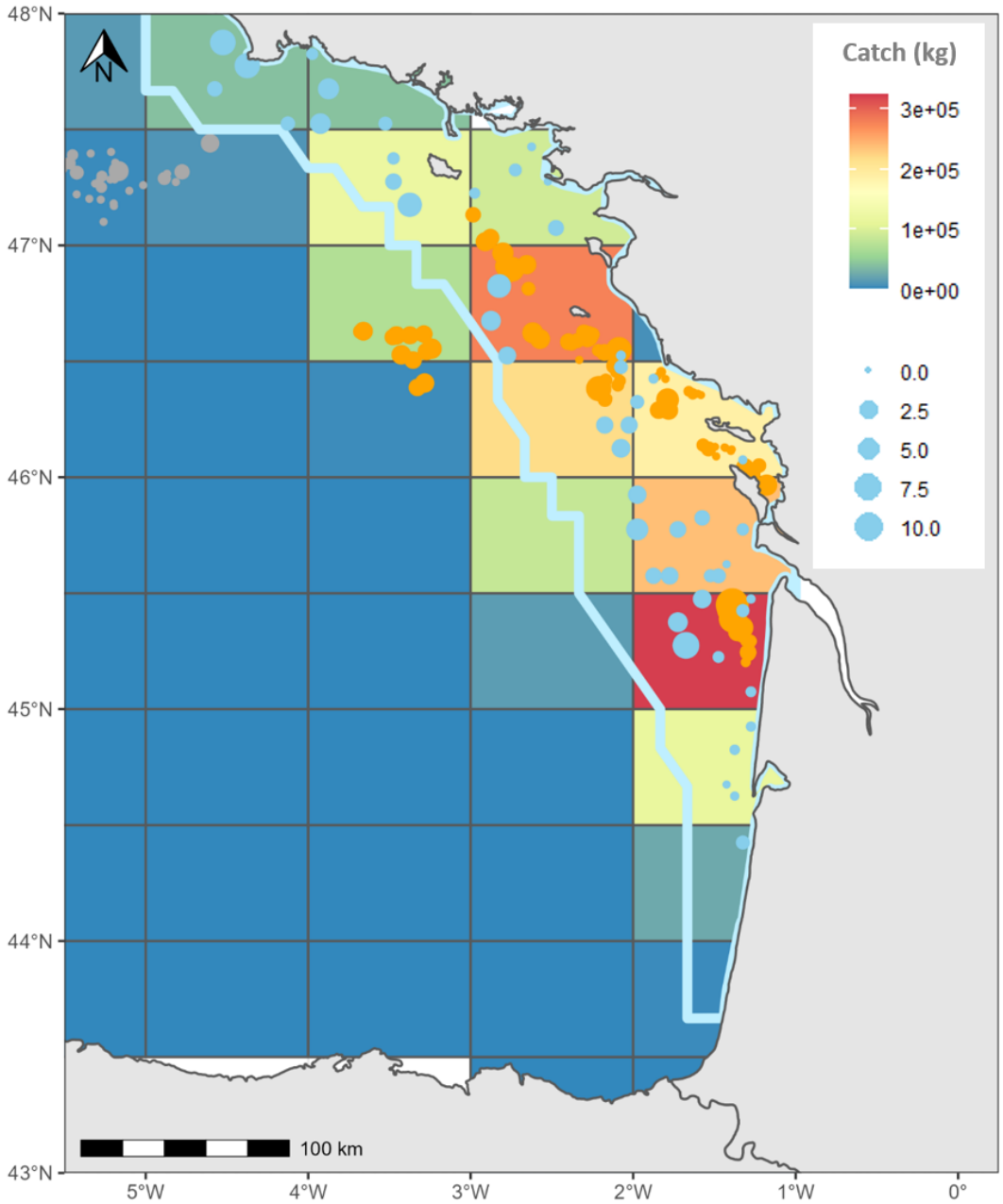
Preferential sampling



## Catch declarations

Large amount of data

Aggregated over rough scale



# Table of Contents

## Big challenge

Combine:

- Punctual observations

$$Y_i | S(x_i), \boldsymbol{\theta}_{obs}$$

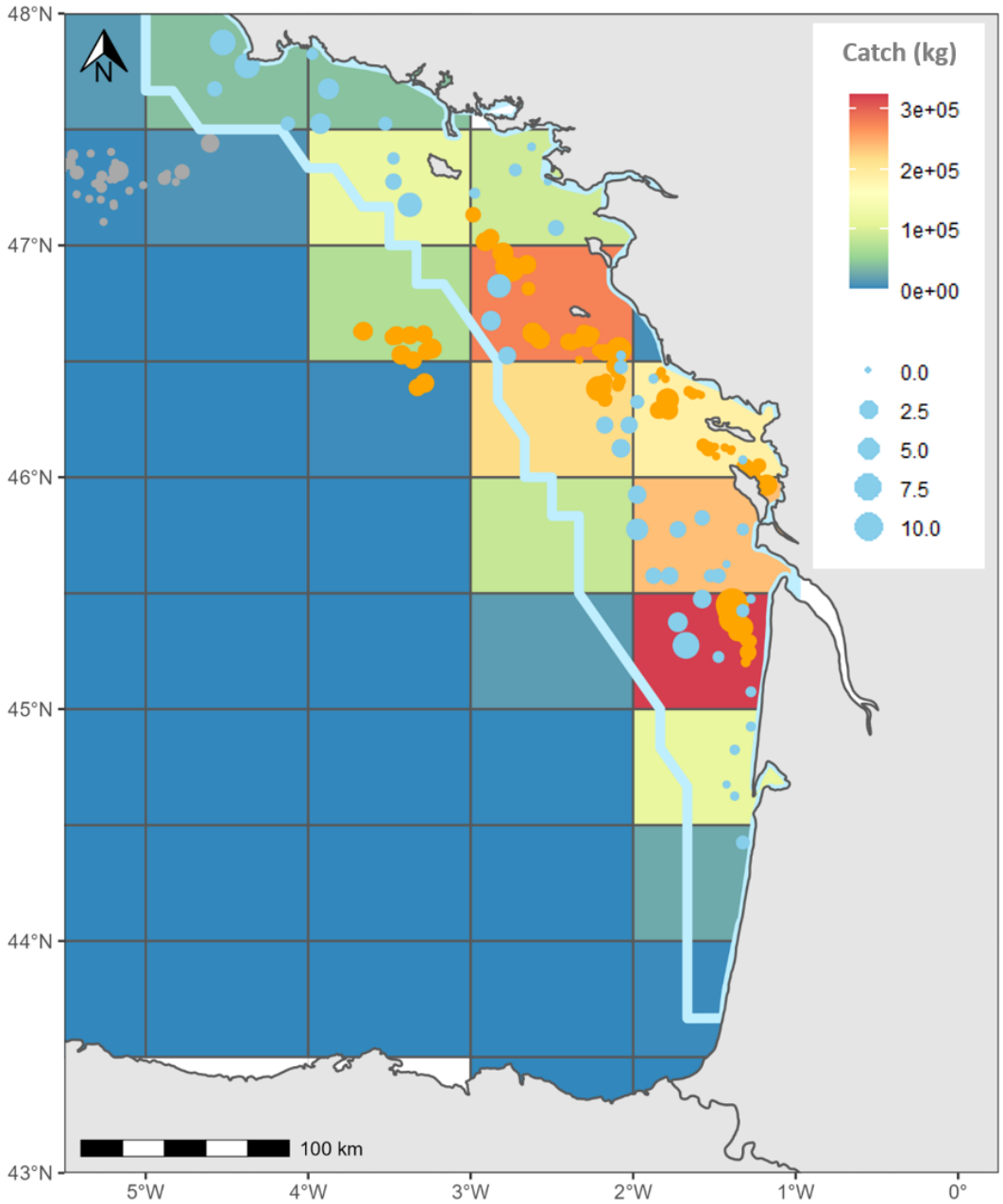
with

- Aggregated observations

$$D_a | S_a, \boldsymbol{\theta}_{obs}$$

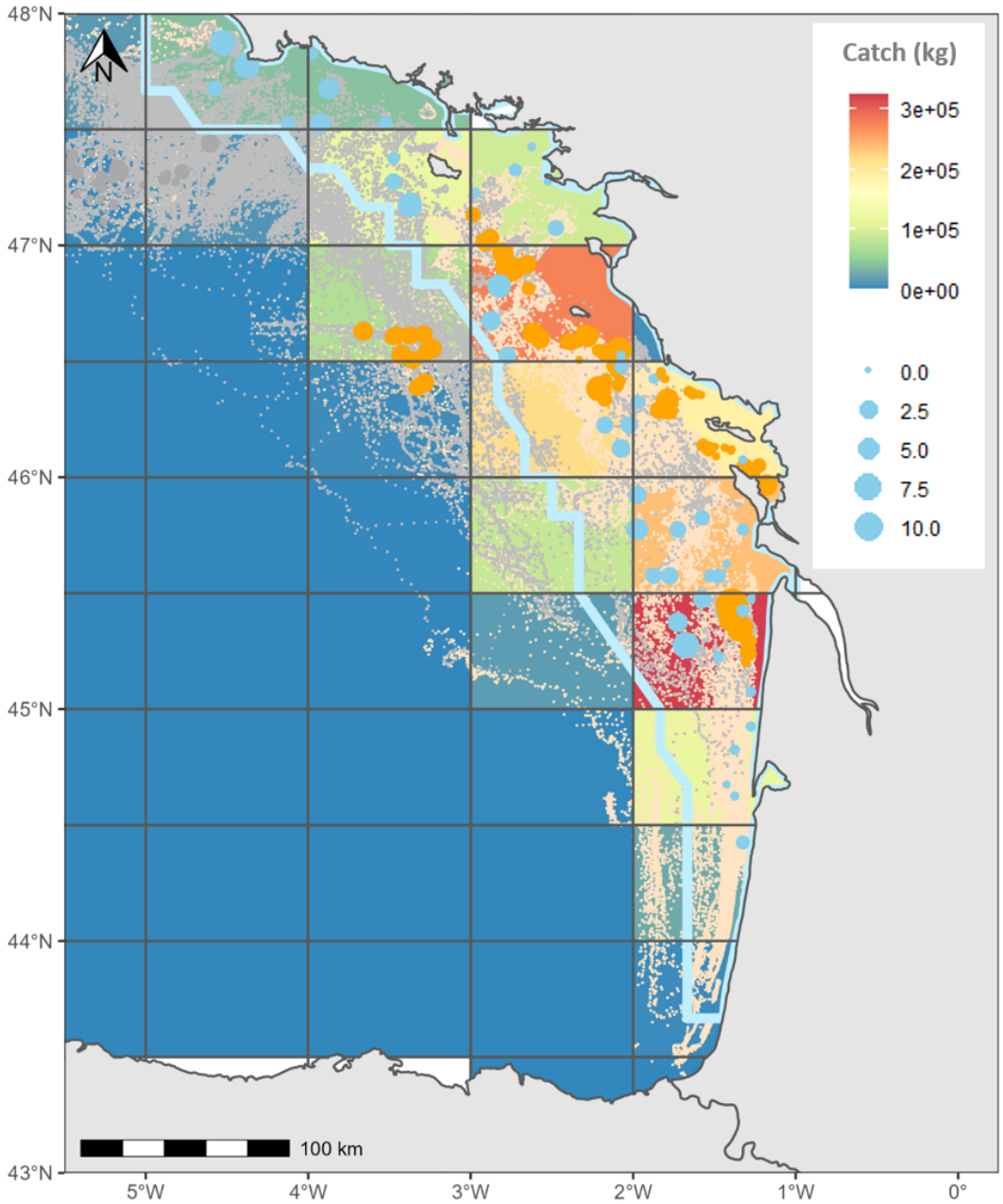
with $a$ being the spatial aggregation unit level

while observations are possibly complex (i.e. zero-inflated positive continuous)



## VMS

Locations of fishing positions



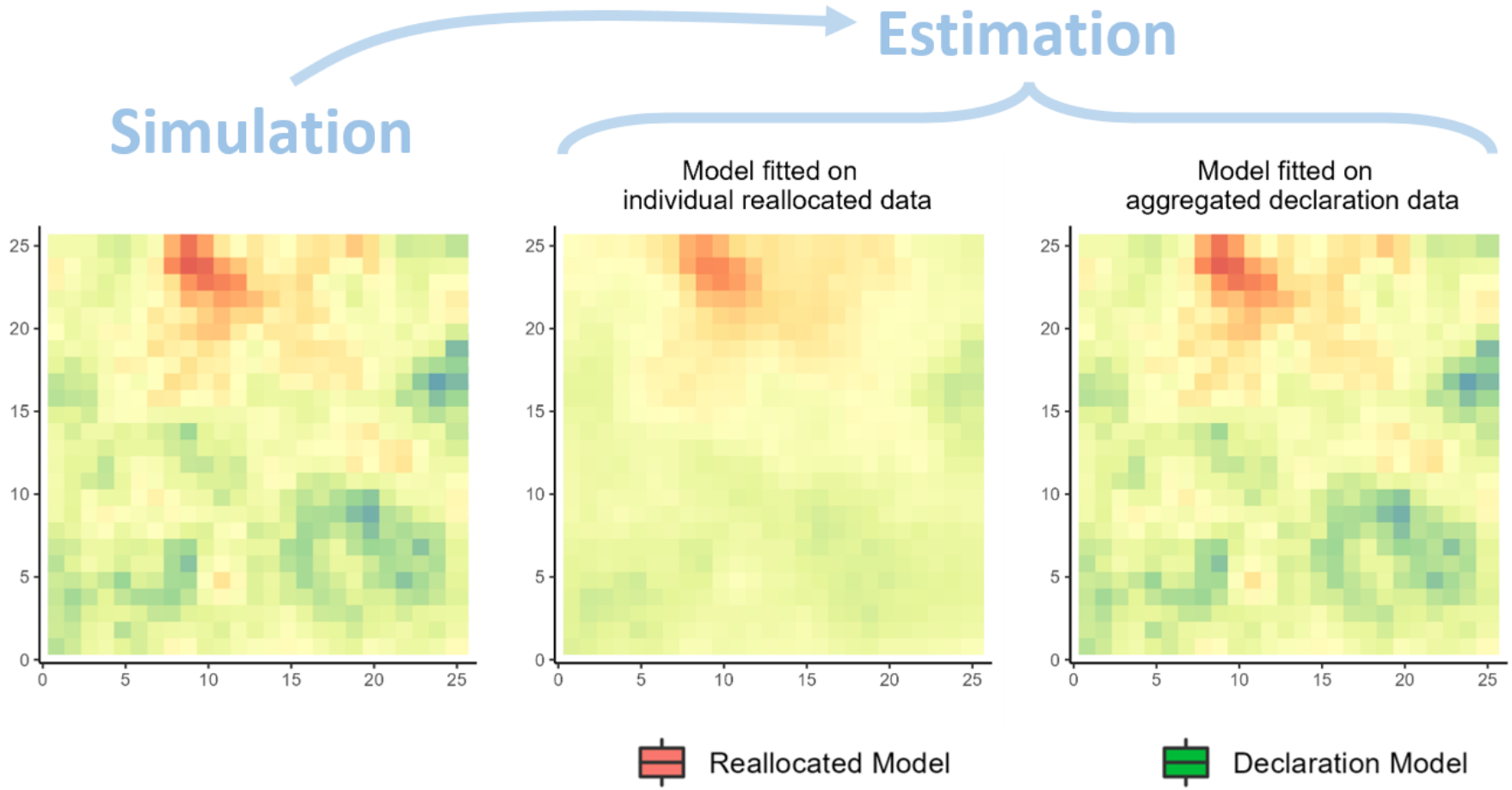
# How to handle change of support?

Standard way is rough (➠ **Reallocated approach**)

Another option (➠ **Declaration model**):

- define $\mathcal{L}_Y$ the probability distribution for (unobserved) punctual observations $Y_i$
- consider $D_a = \sum_{i|x_i \in \mathcal{R}_a} Y_i$ and compute the moment of $D_a$
- Accordingly, define $\mathcal{L}_D$ the distribution of $D_a$ by setting the moments of $\mathcal{L}_D$ as the ones computed for $D_a$

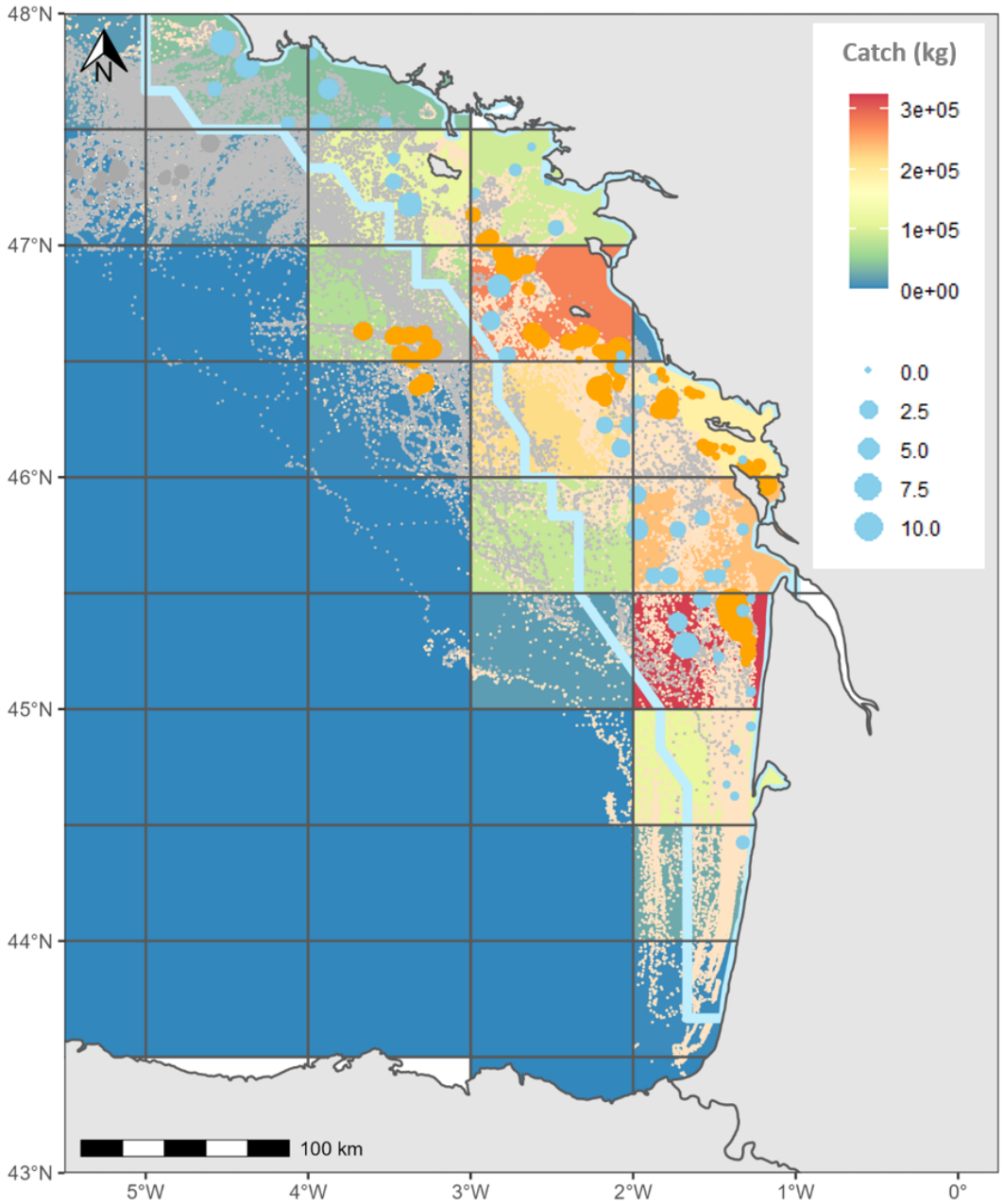
# Simulation testing




➠ **The Reallocated approach** predicts rough smoothed maps, while the **Declaration model** predicts more accurate maps

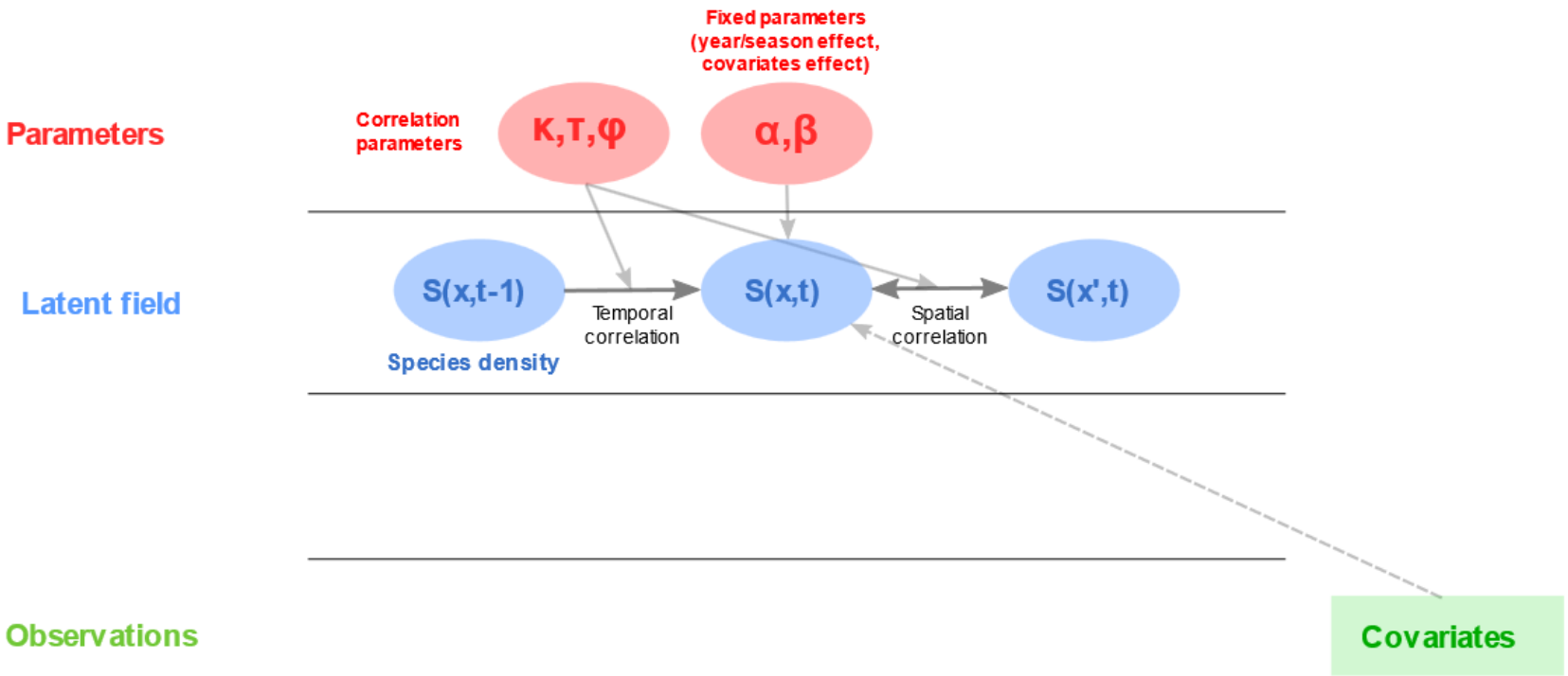
**Let's combine the data sources!**

➠ Extend the framework in time

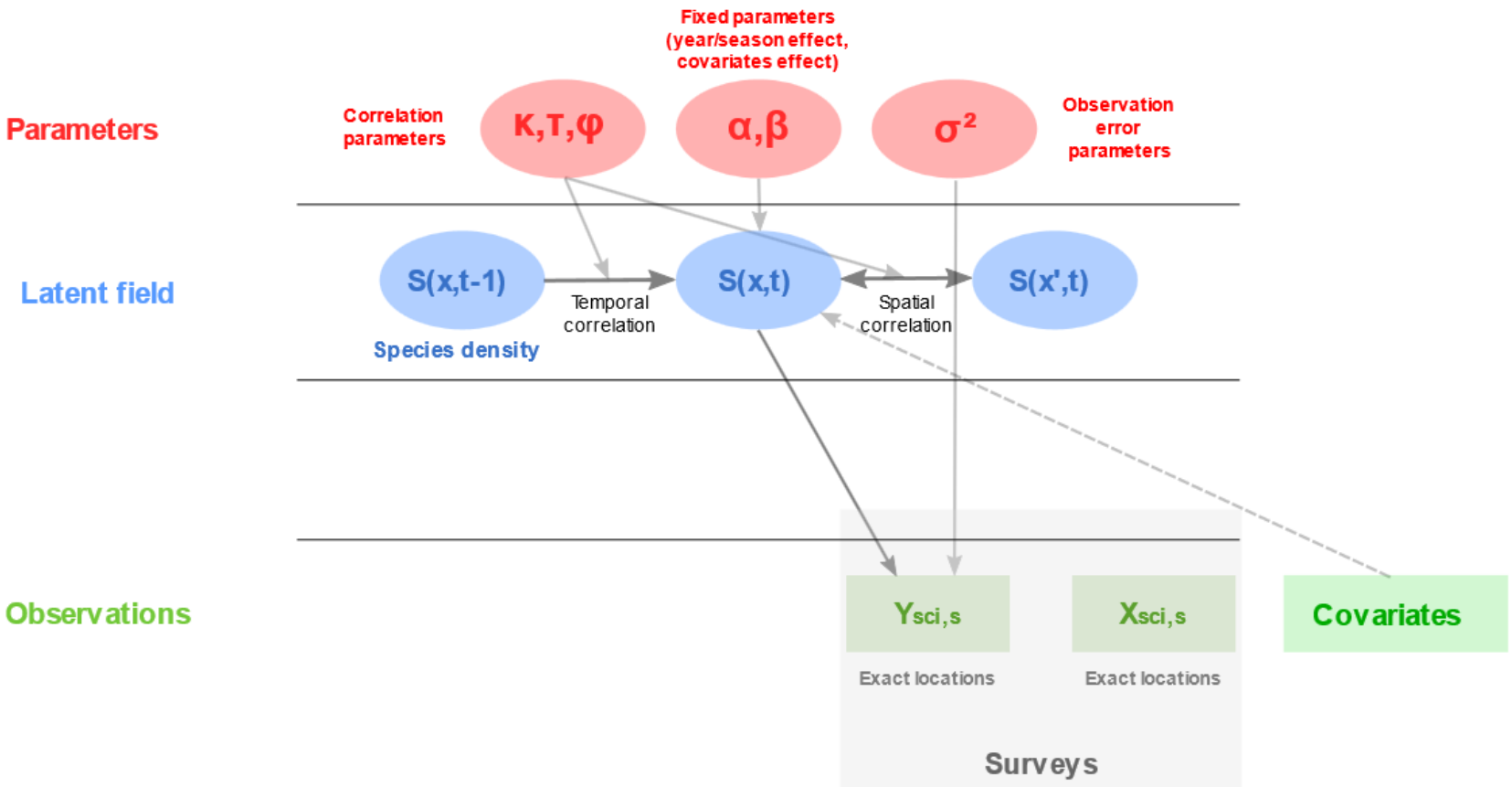


# Table of Contents

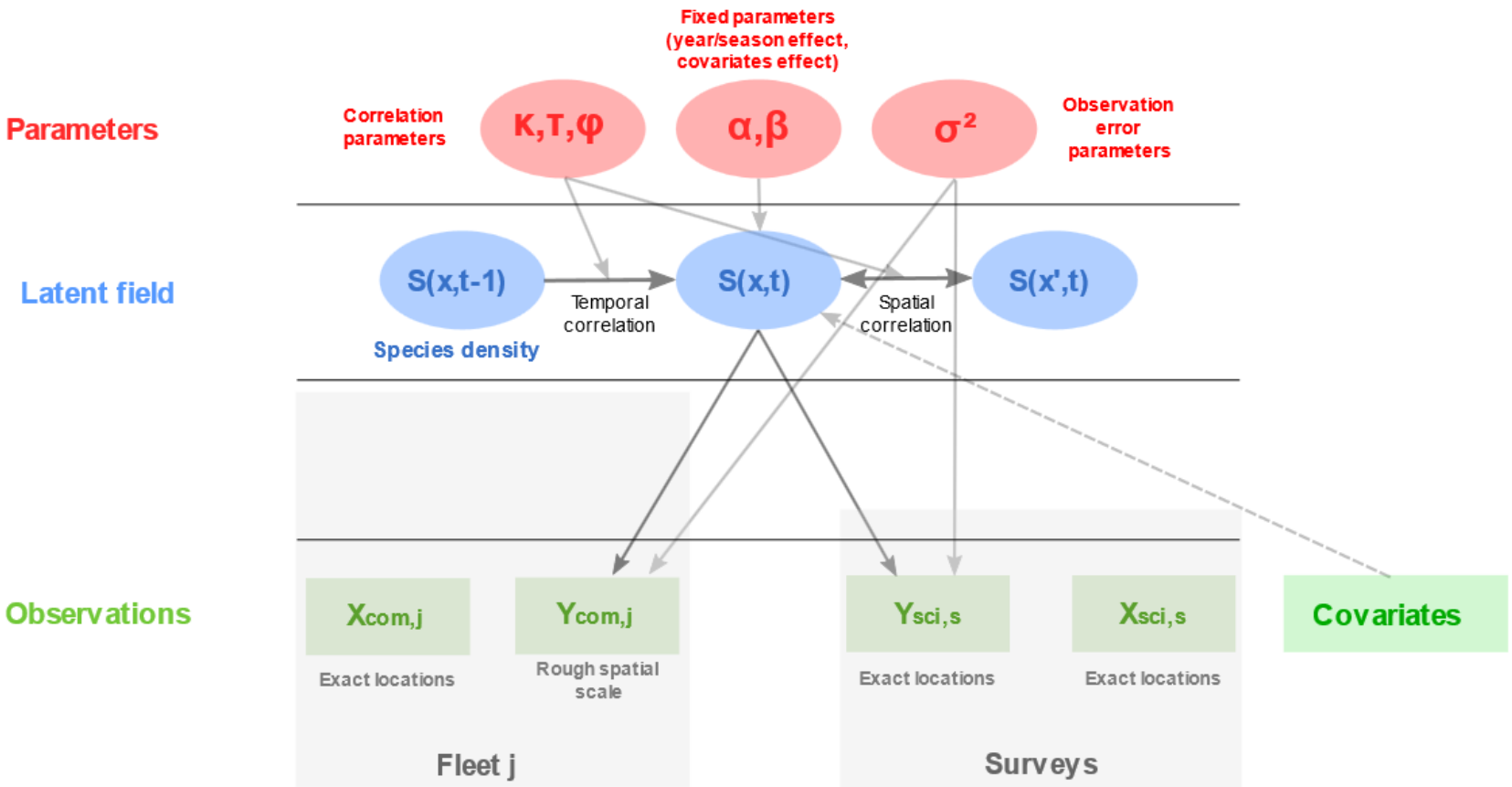
$$log(S(x,t)) = \mu_S(t) + \Gamma_S(x,t)^T \cdot \beta + \delta(x,t)$$
$$\delta \sim GF(0, \mathcal{C}(x,x';t,t'))$$



Joint likelihood: $[\boldsymbol{Y}, \delta|\theta] = ([\boldsymbol{X}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{sci}|\theta, \delta]) \cdot [\delta|\theta]$

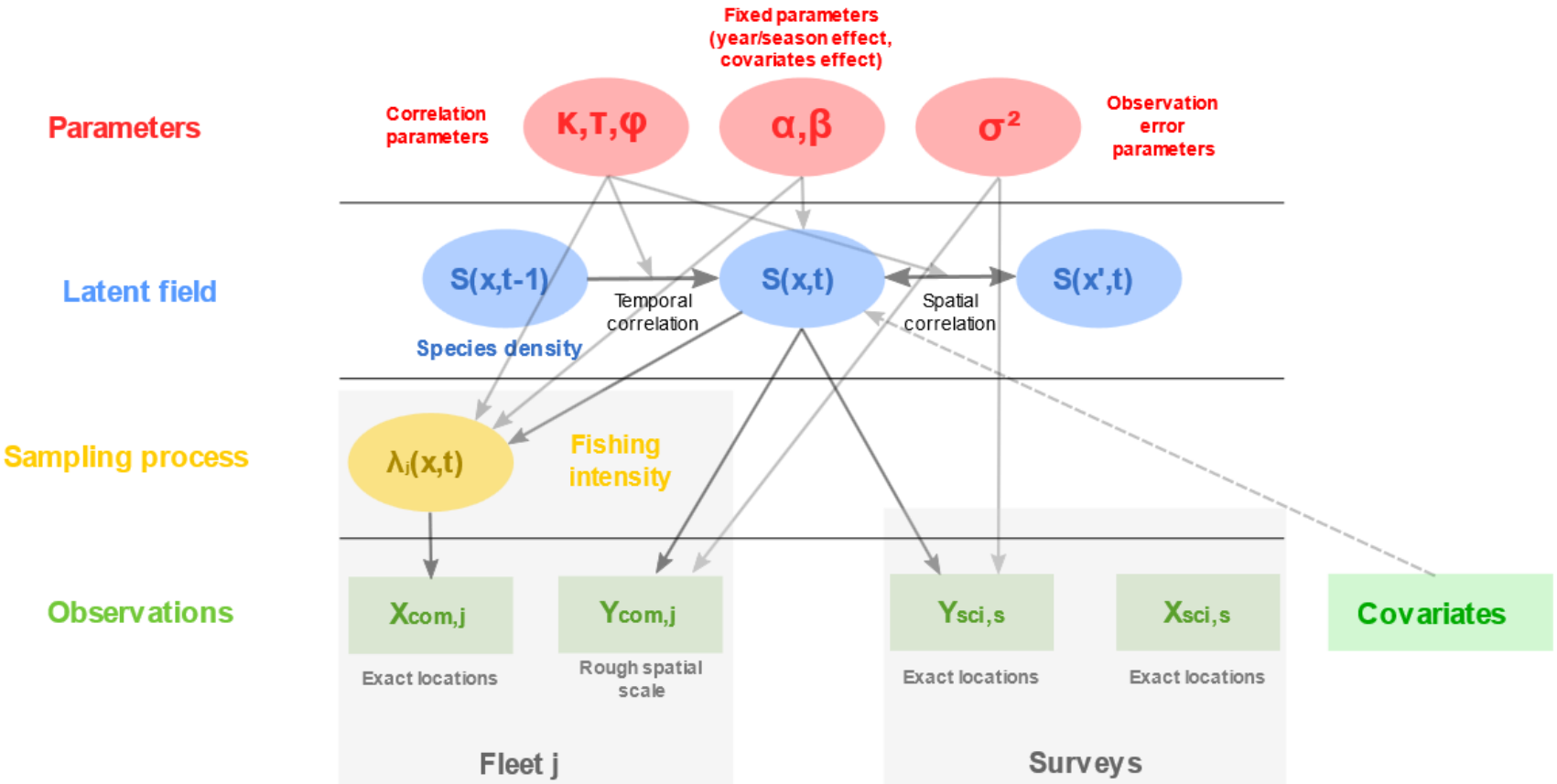


Joint likelihood: $[\boldsymbol{Y}, \delta|\theta] = ([\boldsymbol{X}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{sci}|\theta, \delta]) \cdot [\delta|\theta]$



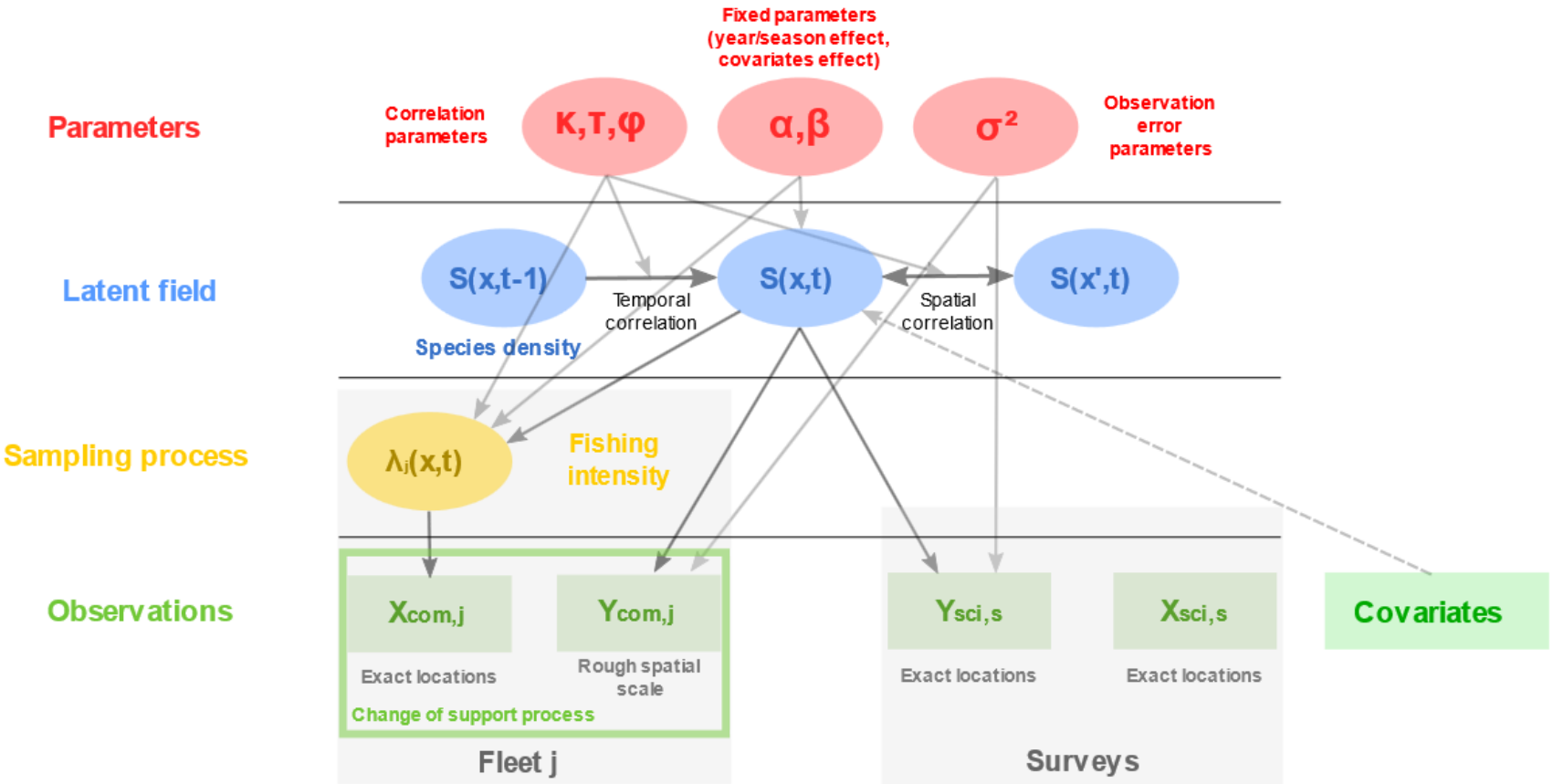
Joint likelihood: $[\boldsymbol{Y}, \boldsymbol{\delta}|\boldsymbol{\theta}] = ([\boldsymbol{X}_{com}|\boldsymbol{\theta}, \boldsymbol{\delta}] \cdot [\boldsymbol{Y}_{com}|\boldsymbol{\theta}, \boldsymbol{\delta}] \cdot [\boldsymbol{Y}_{sci}|\boldsymbol{\theta}, \boldsymbol{\delta}]) \cdot [\boldsymbol{\delta}|\boldsymbol{\theta}]$

$\boldsymbol{X}_{com} \sim \mathcal{IPP}(\lambda(x, t))$



Joint likelihood: $[\boldsymbol{Y}, \delta|\theta] = ([\boldsymbol{X}_{com}|\theta, \delta] \cdot [\boldsymbol{D}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{sci}|\theta, \delta]) \cdot [\delta|\theta]$

$\boldsymbol{D}_{com} = \sum \boldsymbol{Y}_{com}$



# Table of Contents

# Applications

MACCO project

**Mapping data-rich and data-poor species**

CSTEP WG
Closure areas

**Use for marine spatial planning**



Bay of Biscay




Med. Sea

**Identification of potential closure areas**
**+**
**Model parameterization to assess these closure areas**

ISIS-Fish

MARXAN
conservation solutions

# Take home message

- Preferential sampling requires accounting for:
  - non-Gaussian observations,
  - heterogeneous preferential behavior,
  - additional covariates that can affect sampling locations.
- Change of support deals with zero-inflation and data with heavy tails.
- Combining all these data allows to move to a spatio-temporal model with accurate temporal resolution.

**Thank you for your attention!**

# Take home message

- Preferential sampling requires accounting for:
  - non-Gaussian observations,
  - heterogeneous preferential behavior,
  - additional covariates that can affect sampling locations.
- Change of support deals with zero-inflation and data with heavy tails.
- Combining all these data allows to move to a spatio-temporal model with accurate temporal resolution.

**Thank you for your attention!**

# Take home message

- Preferential sampling requires accounting for:
  - non-Gaussian observations,
  - heterogeneous preferential behavior,
  - additional covariates that can affect sampling locations.
- Change of support deals with zero-inflation and data with heavy tails.
- Combining all these data allows to move to a spatio-temporal model with accurate temporal resolution.

Thank you for your attention!

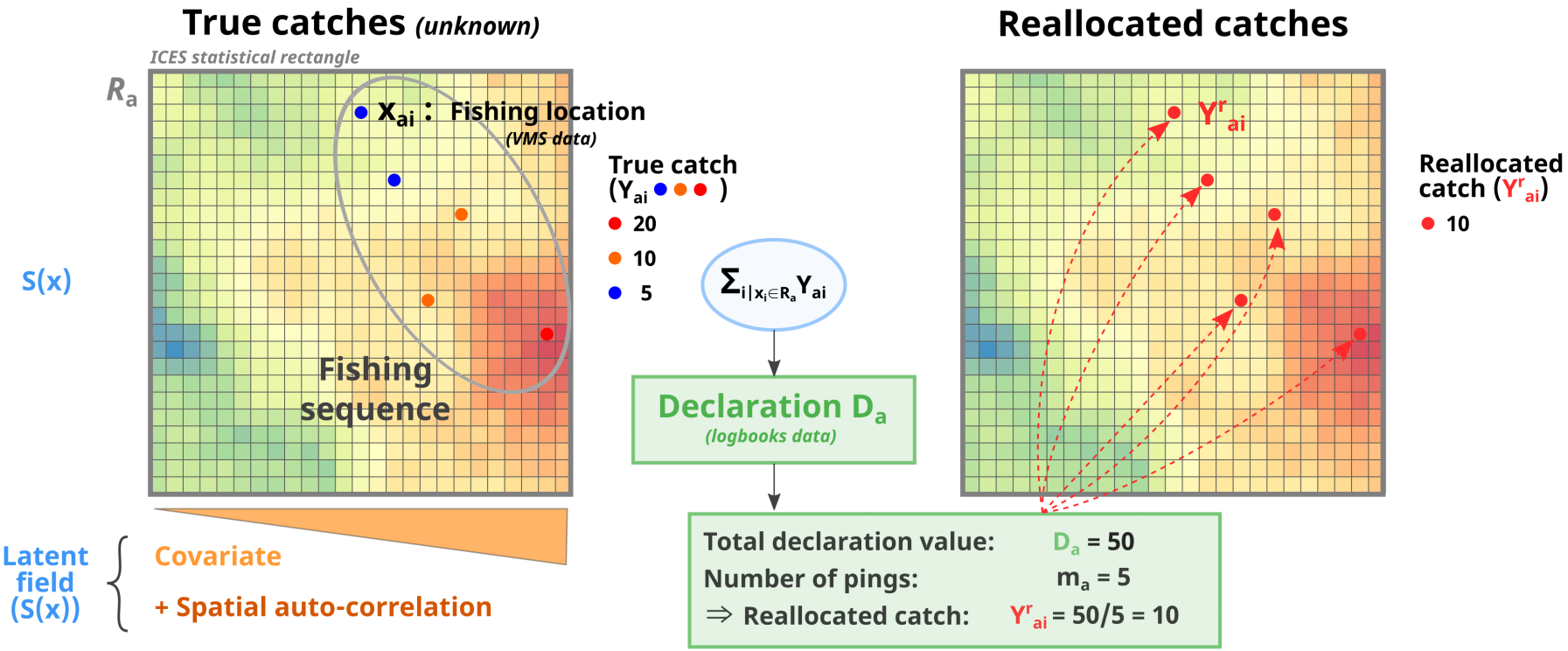
# Take home message

- Preferential sampling requires accounting for:
  - non-Gaussian observations,
  - heterogeneous preferential behavior,
  - additional covariates that can affect sampling locations.
- Change of support deals with zero-inflation and data with heavy tails.
- Combining all these data allows to move to a spatio-temporal model with accurate temporal resolution.

**Thank you for your attention!**

# Any questions?


# How to handle change of support?




Standard way is rough (➠ **Reallocated approach**)

Another option (➠ **Declaration model**):

- define $\mathcal{L}_Y$ the probability distribution for (unobserved) punctual observation $Y_{ai}$
- consider $D_a = \sum_{i|x_i \in \mathcal{R}_a} Y_{ai}$
- Accordingly, define $\mathcal{L}_D$ the distribution of $D_a$ by matching the moments of $D_a$ and $Y_{ai}$

## Punctual observation model $(Y_{ai})$

When making the sum or the product over $i|x_i \in \mathcal{R}_a$, we simply denote $\prod_i$ and $\sum_i$

$\mathrm{L}(y, \mu, \sigma^2)$ is the Lognormal likelihood for observation $y$, mean $\mu$ and variance $\sigma^2$

$Y_{ai}$ and $D_a$ are supposed conditional on $\boldsymbol{S}$ and $\boldsymbol{X}$.

$$\mathrm{P}\left(Y_{ai} = y_{ai}\right) = \begin{cases} p_{ai} & \text{if } y_{ai} = 0 \\ (1 - p_{ai}) \cdot \mathrm{L}\left(y_{ai}, \mu_{ai} = \frac{S(x_{ai})}{(1-p_{ai})}, \sigma^2\right) & \text{if } y_{ai} > 0 \end{cases}$$

$$p_{ai} = \exp(-e^{\xi}.S(x_{ai}))$$

## Declaration model $(D_a = \sum_i Y_{ai})$

$$P(D_a = 0) = \prod_i P(Y_{ai} = 0) = \exp\left\{-\sum_i e^{\xi}.S(x_{ai})\right\} = \pi_a$$

$$\mathrm{P}\left(D_a = d_a | d_a > 0\right) = \quad ?$$

## Punctual observation model $(Y_{ai})$

When making the sum or the product over $i|x_i \in \mathcal{R}_a$, we simply denote $\prod_i$ and $\sum_i$

$\mathrm{L}(y, \mu, \sigma^2)$ is the Lognormal likelihood for observation $y$, mean $\mu$ and variance $\sigma^2$

$Y_{ai}$ and $D_a$ are supposed conditional on $\boldsymbol{S}$ and $\boldsymbol{X}$.

$$\mathrm{P}\left(Y_{ai} = y_{ai}\right) = \begin{cases} p_{ai} & \text{if } y_{ai} = 0 \\ (1 - p_{ai}) \cdot \mathrm{L}\left(y_{ai}, \mu_{ai} = \frac{S(x_{ai})}{(1-p_{ai})}, \sigma^2\right) & \text{if } y_{ai} > 0 \end{cases}$$

$$p_{ai} = \exp(-e^{\xi}.S(x_{ai}))$$

## Declaration model $(D_a = \sum_i Y_{ai})$

$$P(D_a = 0) = \prod_i P(Y_{ai} = 0) = \exp\left\{-\sum_i e^{\xi}.S(x_{ai})\right\} = \pi_a$$

$$\mathrm{P}\left(D_a = d_a | d_a > 0\right) = \quad ?$$

## Compute the moments of $D_a|d_a > 0$

$$E(D_a|d_a > 0) = \frac{\sum_i S(x_{ai})}{1 - \pi_a}$$

$$Var(D_a|d_a > 0) = \frac{\sum_i Var(Y_{ai})}{1 - \pi_a} - \frac{\pi_a}{(1 - \pi_a)^2} E(D_a)^2$$

$$Var(Y_{ai}) = \frac{S(x_{ai})^2}{1 - p_{ai}}(e^{\sigma^2} - (1 - p_{ai}))$$

## Consider $D_a|d_a > 0$ is Lognormal too

$$\mathrm{P}\left(D_a = d_a|d_a > 0\right) =$$

$$\mathrm{L}\left(d_a, \mu_a = E(D_a|d_a > 0), \sigma_a^2 = ln(\frac{Var(D_a|d_a > 0)}{E(D_a|d_a > 0)^2} + 1)\right)$$

## Compute the moments of $D_a|d_a > 0$

$$E(D_a|d_a > 0) = \frac{\sum_i S(x_{ai})}{1 - \pi_a}$$

$$Var(D_a|d_a > 0) = \frac{\sum_i Var(Y_{ai})}{1 - \pi_a} - \frac{\pi_a}{(1 - \pi_a)^2} E(D_a)^2$$

$$Var(Y_{ai}) = \frac{S(x_{ai})^2}{1 - p_{ai}}(e^{\sigma^2} - (1 - p_{ai}))$$

## Consider $D_a|d_a > 0$ is Lognormal too

$$\mathrm{P}\left(D_a = d_a|d_a > 0\right) =$$

$$\mathrm{L}\left(d_a, \mu_a = E(D_a|d_a > 0), \sigma_a^2 = ln(\frac{Var(D_a|d_a > 0)}{E(D_a|d_a > 0)^2} + 1)\right)$$