Going from coarse landings data to fine scale species distribution

First Author1 & Ernst-August Doelle1,2

1 Wilhelm-Wundt-University

2 Konstanz Business School

Author note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: [my@email.com](mailto:my@email.com)

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

*Keywords:* keywords

*Word count:* X

Going from coarse landings data to fine scale species distribution

**To do:**

* **email to Kasper**
* **redaction of M&M**
* **run single square simulations**
* **run multiple square simulations**

References for change of support issues in integrated modelling:

<https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1002/ecy.2710>

<https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12793>

<https://www.sciencedirect.com/science/article/pii/S0169534719302551#b0010>

<https://www.sciencedirect.com/science/article/pii/S0006320721001993>

# Material and methods

We base our approach on the model developped by Alglave et al. (2022) and build on these developpments to provide a solution to overcome the issues related to change of support in our specific context.

## Defining the problem from a modelling point of view

Let’s define the latent field the punctual observations . is a spatial Gaussian Field (GF) defined in space such as . is the covariate term with the matrix of deisgn of the coavariates and the effect of the covariates. is GF capturing spatial correlation ().

Punctual observations are conditionnal on the latent field values at fishing location for observation . Observations are supposed to follow some distribution . In our case, data are positive continuous zero inflated data and they are modelled through a zero-inflated lognormal model described in supplementary material (SM) previously introduced by Thorson (2018) and already used in Alglave et al. (2022).

All observations belong to a catch declaration D\_j such as:

is the declaration index with the number of declarations. is the vector of all fishing positions related to the declaration. The punctual observations are indexed through with the number of fishing positions belonging to the declaration.

In standard processing, we know through logbooks data, through VMS data and are reallocated uniformly on related so that derived punctual observations are computed as . The likelihood is then directly computed on reallocated assuming these are the exact punctual observation. This strongly simplifies the actual process of observation and have most likely repercussion on model performance.

To overcome such limitation, an alternative is to consider that only the catch declarations are observed, but they can inform of the punctual observations which are not-observed (i.e. they are latent variables).

Such way, we define some distribution as the distribution of the catch declarations with .

Our approach is to match the 2 first moments of (obtained from the moments of i.e.  probability distribution) and .

First, as at the punctual level, the declarations can be decomposed in 2 components: (1) the probability to obtain a zero value and, if the declaration is positive, (2) the probability to obtain a certain declaration value.

1. We derive the probability to obtain a zero-declaration by simply multiplying the probability to obtain a zero-punctual declarations to all fishing points as .
2. If the declarations is positive, based on the moments of , we derive and (see Table …). Assuming is lognormal (let’s denote it ), we make the two first moments of the coincides with and . This is done through the formulas for the mean component of and for its variance component.

Formulas of and are in the table … and all calculations to obtain these formulas are available in SM.

See 3\_formulas.rmd.

## Simulation-estimation

### Single-square simulations.

How many simulations? 100 or more?

How do we simulate the covariate?

Still some lack of convergence.

Number of pings | nu: 1|10 10|100 100|1000

Size within a fishing sequence: radius = 3

Number of fishing zones: 1|3|5

### Multiple-squares simulations.

How do we simulate the covariate?

## Case-study: sole of the Bay of Biscay

# Results

## Single square analysis

Table 1: Percentage of convergence per simulation/model configuration at the level of a single rectangle

Nb samples

Fishing sequence

Reallocation

Likelihood level

Convergence (%)

10

1

No

Yi

99.668

10

1

Yes

Yi

0.333

10

1

Yes

Dj

0.000

100

10

No

Yi

100.000

100

10

Yes

Yi

100.000

100

10

Yes

Dj

92.000

1000

100

No

Yi

100.000

1000

100

Yes

Yi

100.000

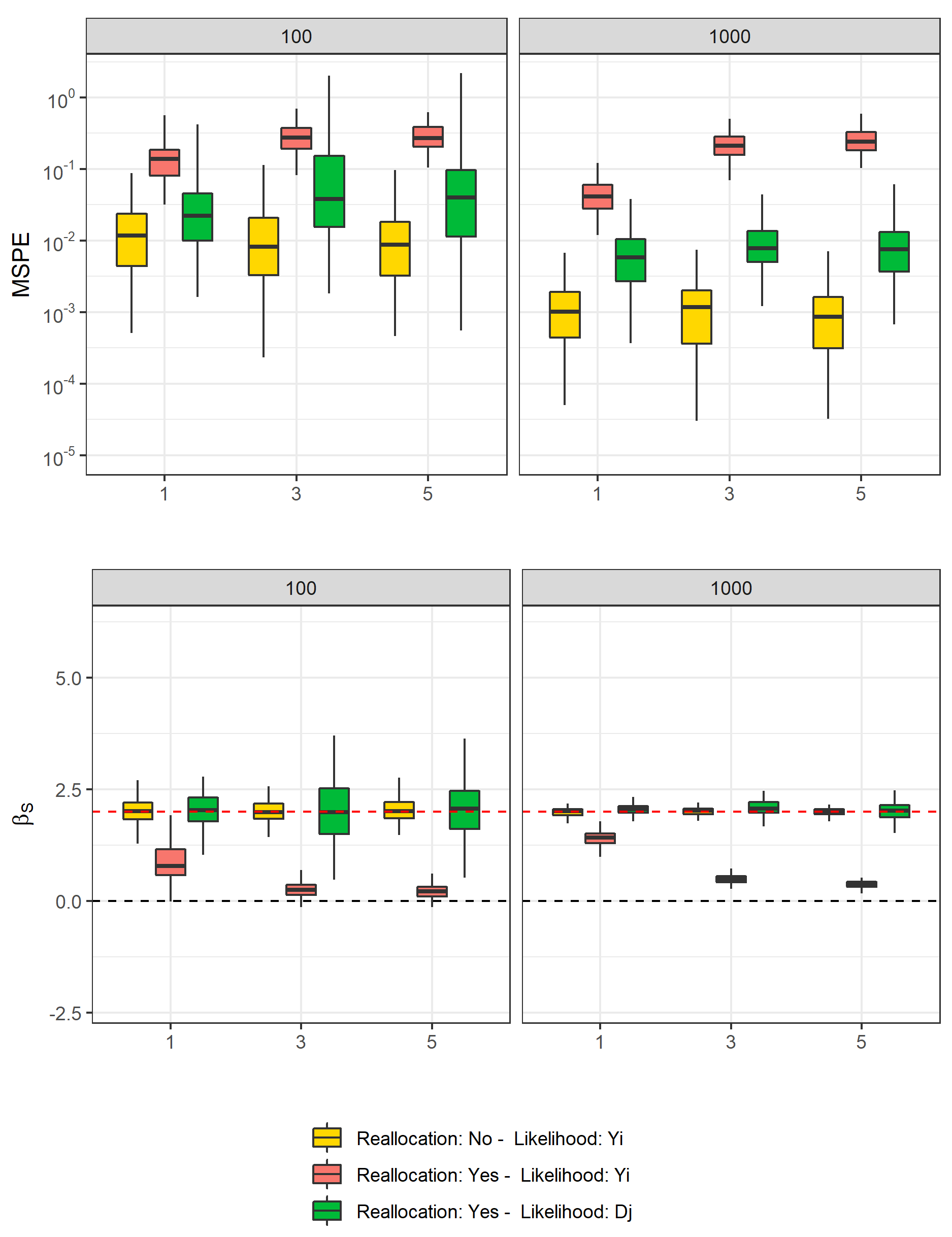
1000

100

Yes

Dj

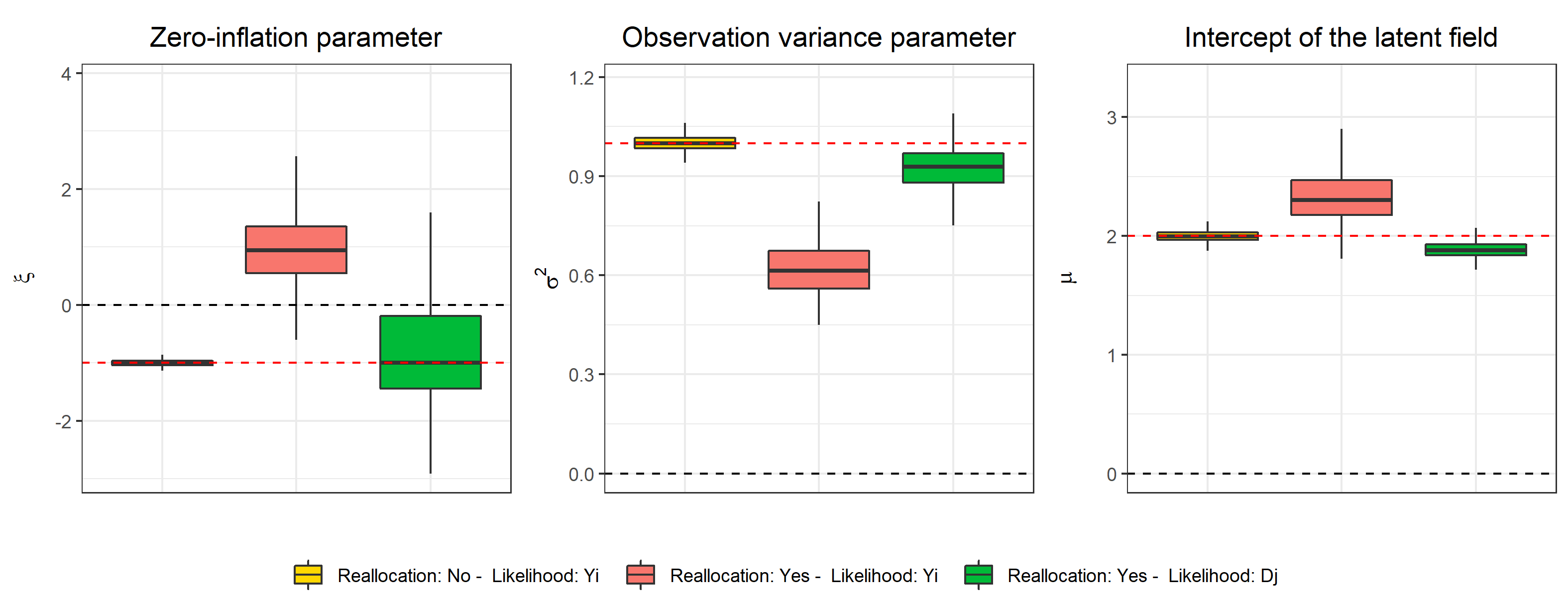
97.333



*Figure* *1.*  Performance metric for single-square simulations. columns: number of commercial pings. x-axis: number of zones visited within each declaration. ‘Reallocation:’, data are or are not reallocated in simulations. ‘Likelihood:’, the likelihood is computed on reallocated observations or on catch declarations . Gold: golden standard. Red: actual situation ( level). Green: alternative model ( level).

Reallocation has a major effect on predictions and estimates accuracy (Figure 1). Reallocating data conduct to a 10 to 200 times decrease of predictions accuracy when working at punctual observations level (MSPE gold compared to red boxplots) and decreases as the number of visited zones within a declaration increases. It also leads to the loss of the species habitat relationship; is biased an tends towards 0 as the number of fishing zones within a declaration increases. Increasing the number of samples only increases the accuracy of the estimates and spatial predictions but overall spatial predictions remain inaccurate and remain biased. Regarding other parameters, the zero-inflation parameter is over-estimated (i.e. when reallocating, the quantity of data is under-estimated), the observation variance is underestimated (i.e. the data is estimated to be less noisy than they actually are) and the interecept of the latent field is slightly over-estimated (Figure 2).

Working at the level of catch declarations allows to recover the species-habitat relationship and to improve the accuracy of spatial predictions even though accuracy is not as accurate as the golden-standard (Figure 1). Furthermore, the zero-inflation parameter is unbiased when the likelihood is built on catch declarations. Other parameters (observation variance, intercept) are also better estimated even though they remain slightly biased (Figure 2). This alternative model have some convergence difficulties (Table 1) as 8% of the model did not converged when sample size is medium (100 pings) and only 3% did not when sample size is large (1000 pings).



*Figure* *2.*  Parameters relative bias for single-square simulations. ‘Reallocation:’, data are or are not reallocated in simulations. ‘Likelihood:’, the likelihood is computed on reallocated observations or on catch declarations . Gold: golden standard. Red: actual situation ( level). Green: alternative model ( level). Only the simulations with 1000 observations are represented. Black line: zero value. Red line: parameter true value.

## Multiple square analysis

Table 2: Percentage of convergence per simulation/model configuration at the level of several ICES rectangles

Model

Likelihood level

Convergence (%)

Commercial model

Yi

100.000

Commercial model

Dj

75.377

Integrated model

Yi

100.000

Integrated model

Dj

76.382

Scientific model

100.000

## Warning: Removed 7 rows containing non-finite values (stat\_boxplot).

The scientific-based model provides species-habitat relationship and range estimates that are unbiased. Whether the model is built on or on , the contribution of either scientific or commercial data can be clearly evidenced from the MSPE plot: the errors related to the integrated model are smaller than the single-data models. Comparing this to the maps obtained from Figure 4, we highlight that the commercial-based models logically miss the hotspot at 2.5°W - 46.75°N as commercial data do not cover this hotspot, while scientific data provide predictions that do not well capture the local hotposts. Integrating scientific and commercial data allows to (1) capture the hotspot missed by commercial data through scientific data and and (2) better capture the local correlation structures through the dense commercial data.

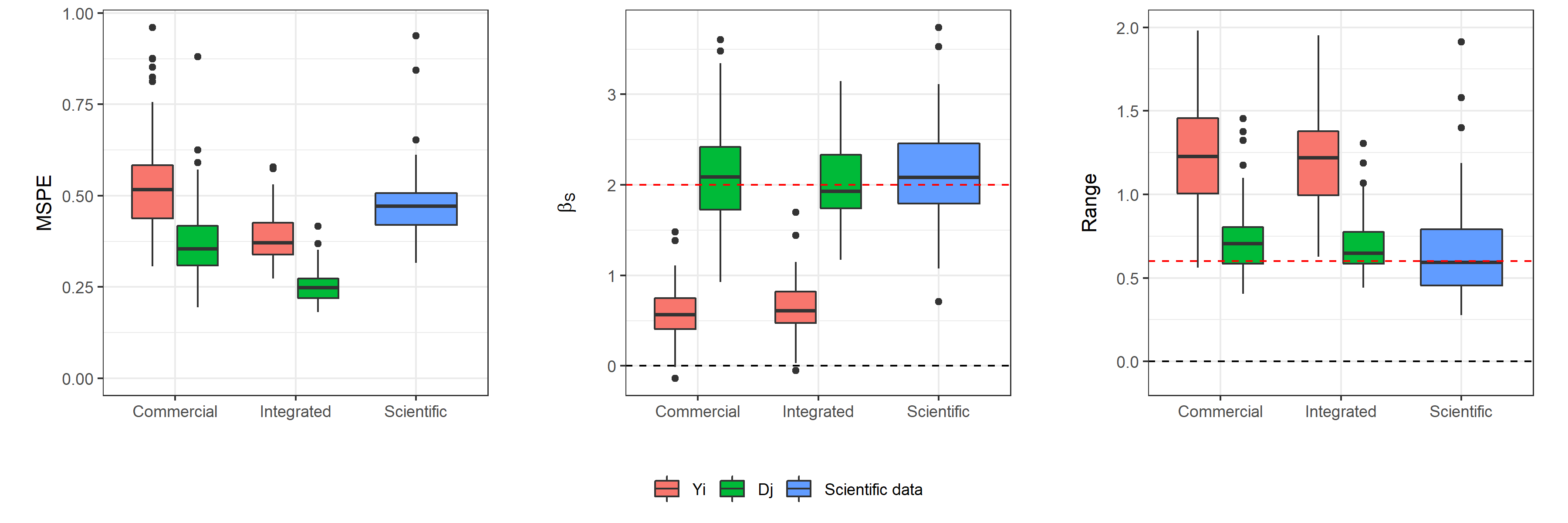
Furtermore, consistently with single-square simulations, working at the level of punctual observations () conduct to a loss in both the species-habitat relationship and predictions accuracy (Figure 3) compared to the model built on commercial declarations ().

Interestingly, reallocating does not affect only the species-habitat relationship but also the spatial autocorrelation terms such as the range parameter. When working at the level, the range parameter is biased, while it is not when working at the level.

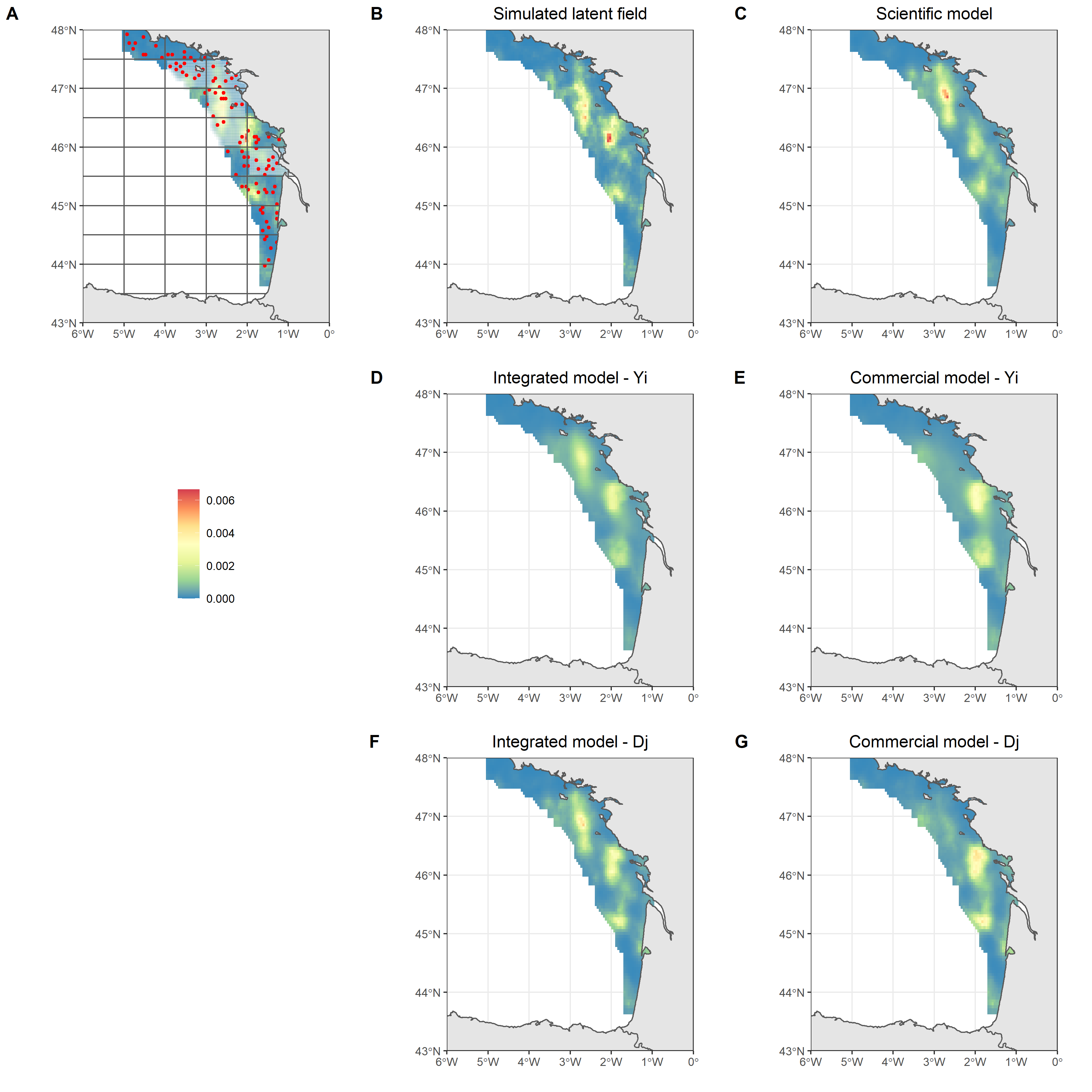
This is a consequence of the loss of the species-habitat relationship: when working at level part of the variability related to the covariate effect is captured by the random effect and then the range parameter captures both autocorrelation related to the actual random effect and to the covariate.

Working at the level of catch declarations in estimation allows to recover and disentangle the effect of the species-habitat relationship and of the random effect. This is evidenced in Figure 4 where the model based on provide smoothed maps and do not capture the relatively small scale patterns that are shaped by the covariate effect. On the other hand the model built on (and the scientific-based model too) better capture and disentangle the covariate effect and the spatial random effect and then provide predictions that better emphasize to the small-scale patterns of species distribution.

However, this goes with some difficulty in convergence as only 75% of the model built on catch declarations converge.

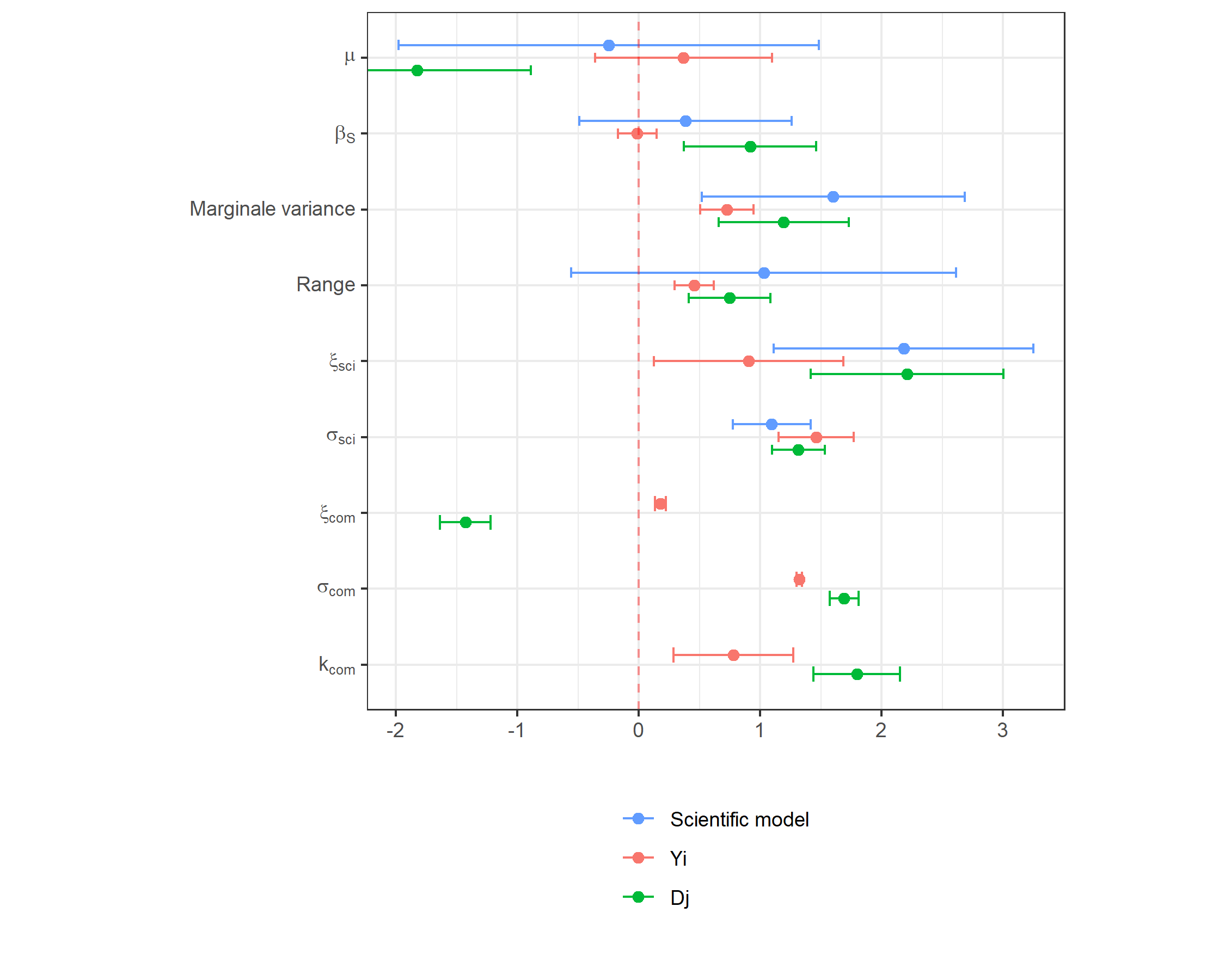


*Figure* *3.*  Performance metric for several rectangles simulations. columns: commercial data coverage. x-axis: likelihood level. row, red line: true value of .

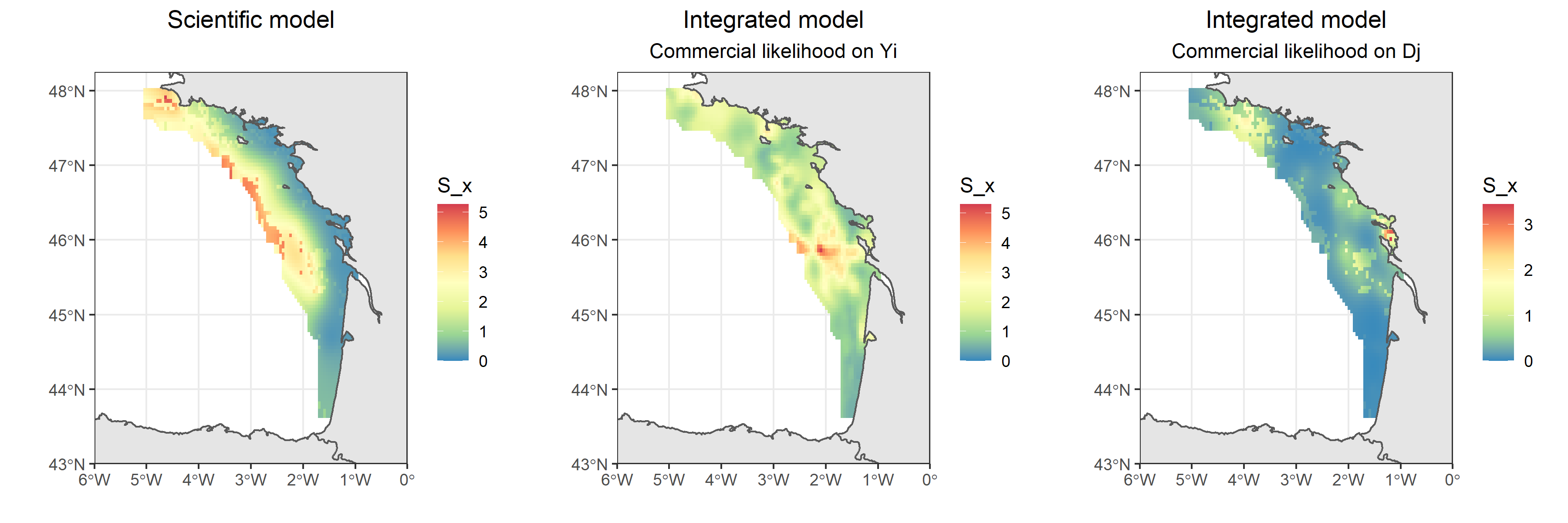


*Figure* *4.*  Relative distribution of simulated/estimated biomass field. A: Simulated biomass field with scientific samples (red) and rectangles that have not been sampled by commercial data (transparent rectangles). B: simulated biomass field. C: biomass field from the scientific-based model. : model fitted at the punctual observation level (D, E). : model fitted at the declaration level (F, G).

## Real case study



*Figure* *5.*  Parameters obtained with scientific-based model, the integrated model built on punctual observations and the integrated model built on catch declarations .



*Figure* *6.*  Maps obtained from (left) the scientific-based model, (center) the integrated model built on punctual observations , (right) the integrated model built on catch declarations .

Passing from punctual observations to catch declarations modifies the overall distribution patterns and some of the parameters estimates and related confidence intervals. Consistently with simulations, model built on tend to provide biased estimates and under-estimate the uncertainty related to other estimates compared with the scientific model and the model buit on . Mainly, the effect of substrate is recovered in the integrated model built on (Figure 6). The zero-inflation parameter is revised downwards (i.e. there are actually more zero-values than in the reallocated data) while the observation variance of commercial data is revised upwards (i.e. the commercial data are more noisy than expected when building likelihood on ). In the model built on , the confidence interval of some parameters (specifically , the marginal variance, the range, , ) are very narrow while in the model built on confidence intervals are larger.

Furthermore, when comparing the scientific-based model estimates and the integrated model built on , some parameters are better estimated. For instance, while in the scientific-based model the substrate effect was not significant, in the integrated model built on substrate is significant and the confidence interval is smaller. Other parameters are more precisely estimated such as the marginal variance, the range, and .

On the contrary, other parameters do not seems well estimated in either or models. For instance, compared to the scientific-based model, the intercept is revised upwards when building the likelihood on and revised downards when working on . This is consistent with simulations results, see Figure 2.

Regarding the maps of the spatial predictions, working at level strongly modify the model biomass field compared with the model. Mainly, the effect of the covariate have a sharper effect on species distribution. Overall, the strength of the hotspots are revised and the map of the model fitted on is less homogeneous than the map of the model built on .

Finally, the model built on commercial data only does not converge (while the one built on does) emphasizing the model build on catch declarations face difficulties to converge and require punctual observations (here survey data and on-board observer data) to converge on real data. Interestingly, apart from the species-habitat relationship that are similar between the scientific-based model and the integrated model built on , maps are not similar emphasizing commercial data brings information to the spatial predictions and better allow to capture local spatial patches of biomass.

# Discussion

# References

Alglave, B., Rivot, E., Etienne, M.-P., Woillez, M., Thorson, J. T., & Vermard, Y. (2022). Combining scientific survey and commercial catch data to map fish distribution. *ICES Journal of Marine Science*, fsac032. <https://doi.org/10.1093/icesjms/fsac032>

Thorson, J. T. (2018). Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(9), 1369–1382.