



HAL
open science

Combining scientific survey and commercial catch data to map fish distribution

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

► **To cite this version:**

Baptiste Alglave, Etienne Rivot, Marie-Pierre Etienne, Mathieu Woillez, James Thorson, et al.. Combining scientific survey and commercial catch data to map fish distribution. *ICES Journal of Marine Science*, 2022, 79 (4), pp.1133-1149. 10.1093/icesjms/fsac032 . hal-03614713

HAL Id: hal-03614713

<https://hal.science/hal-03614713v1>

Submitted on 4 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining scientific survey and commercial catch data to map fish distribution

Alglave Baptiste ^{1,2,*}, Rivot Etienne ², Etienne Marie-Pierre ³, Woillez Mathieu ⁴, Thorson James T ⁵, Vernard Youen ¹

¹ DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Nantes 44980, France

² DECOD (Ecosystem Dynamics and Sustainability), Institut Agro, IFREMER, INRAE, Rennes 35042, France

³ Mathematical Research Institute of Rennes IRMAR, Rennes University, Rennes 35042, France

⁴ DECOD (Ecosystem Dynamics and Sustainability), IFREMER, Institut Agro, INRAE, Brest 29280, France

⁵ Habitat and Ecological Processes Research Program, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98112, USA

* Corresponding author : Baptiste Alglave, email address : baptiste.alglave@agrocampus-ouest.fr

Abstract :

Developing Species Distribution Models (SDM) for marine exploited species is a major challenge in fisheries ecology. Classical modelling approaches typically rely on fish research survey data. They benefit from a standardized sampling design and a controlled catchability, but they usually occur once or twice a year and they may sample a relatively small number of spatial locations. Spatial monitoring of commercial data (based on logbooks crossed with Vessel Monitoring Systems) can provide an additional extensive data source to inform fish spatial distribution. We propose a spatial hierarchical framework integrating both data sources while accounting for preferential sampling (PS) of commercial data. From simulations, we demonstrate that PS should be accounted for in estimation when PS is actually strong. When commercial data far exceed scientific data, the latter bring little information to spatial predictions in the areas sampled by commercial data, but bring information in areas with low fishing intensity and provide a validation dataset to assess the integrated model consistency. We applied the framework to three demersal species (hake, sole, and squids) in the Bay of Biscay that emphasize contrasted PS intensity and we demonstrate that the framework can account for several fleets with varying catchabilities and PS behaviours.

Keywords : hierarchical model, integrated modelling, species distribution model, survey data, Template Model Builder (TMB), VMS and logbook data

33 **1 INTRODUCTION**

34 Developing species distribution models (SDM) is critical in marine and fisheries ecology
35 for assessing the relationship between species and their habitat (Guisan and
36 Zimmermann, 2000), identifying essential habitats (Paradinas *et al.*, 2015), forecasting
37 population and ecosystems response to environmental changes (Cheung *et al.*, 2009).
38 The development of statistical models to predict fishery resources distribution has
39 received considerable attention (Planque *et al.*, 2011; Thorson *et al.*, 2015a, 2015b;
40 Martínez-Minaya *et al.*, 2018; Moriarty *et al.*, 2020). Recent developments have
41 generalized SDM to analyze biological data representing condition, stomach contents,
42 size structure, and other demography and population dynamics features (Thorson, 2015;
43 Grüss *et al.*, 2020). Ongoing research also seek to integrate individual movement, growth,
44 species interactions into SDM (Kristensen *et al.*, 2014; Thorson *et al.*, 2017a, 2019),
45 although these approaches are “data hungry” and therefore require integrating different
46 sources of data within a single model.

47 Scientific survey and commercial catch data consist in two potentially complementary data
48 sources to estimate harvested fish spatial distribution (Pennino *et al.*, 2016). Scientific
49 surveys are key data sources in fisheries ecology. They most often benefit from a
50 standardized sampling plan and a constant catchability (Hilborn and Walters, 1992; Ocean
51 Studies Board and National Research Council, 2000; ICES, 2005; Nielsen, 2015). They
52 are generally designed to cover the full geographical extent of specific populations
53 including areas of low or null abundance, and are thus adapted to develop unbiased
54 abundance indices and spatial predictions of species distribution (Rivoirard *et al.*, 2008;
55 ICES, 2012). In addition, they often seek to minimize selectivity in order to sample as
56 many species, size groups and life stages as possible. However, the related expansive

57 charges generally comes at the cost of a relatively low sampling density in space and/or
58 time. For instance, trawl survey can sample a limited number of spatial locations, and
59 most often occur once or twice a year. Thus, they may provide poor information regarding
60 intra-annual variability (Pennino *et al.*, 2016; Rufener *et al.*, 2021) and imprecise estimates
61 of species abundance and distribution (ICES, 2005).

62 Commercial catch declarations (logbooks) data constitute a complementary data source
63 that may benefit of a higher sampling effort than scientific survey. In Europe, catch
64 declarations must be reported in logbooks data for all fishing vessels; besides, geolocation
65 through Vessel Monitoring System (VMS) is mandatory for all fishing boat above 12m long
66 (Hintzen, 2021). Hence, logbook data combined with VMS data can provide high
67 resolution maps of Catch Per Unit Effort (CPUE - Gerritsen and Lordan, 2010; Murray *et al.*,
68 2013) with a relatively dense spatio-temporal sampling within the range of the
69 commercial fleets. However, inferring SDM with commercial data can be challenging as
70 they generally arise from a preferential sampling (PS) behavior, i.e. a sampling that
71 directly or indirectly depends upon the biomass of the target species. Indeed, fishermen
72 tend to target areas with high biomass or may also favor fishing zones based on other
73 criteria (like bottom substrate or distance to the coast for instance - Hintzen *et al.*, 2021)
74 that are indirectly related to the target species abundance. When not properly considered
75 in statistical models, PS associated with commercial data may lead to biased estimates
76 of fish distribution and biomass (Trenkel *et al.*, 2013; Pennino *et al.*, 2019). In particular,
77 when the biomass is spatially heterogeneous, ignoring PS may overestimate the spatial
78 predictions and the overall biomass estimates.

79 Recent research has tackled this challenge and developed methods to account for PS in
80 statistical inferences. Model based PS was first introduced by Diggle *et al.* (2010) who

81 proposed a base framework for estimating PS and applied it to led pollution data in Galicia.
82 The authors extended a standard geostatistical approach within a hierarchical framework
83 where the variable of interest is jointly modelled with the spatial intensity of the sampling
84 effort which also contributes to the inference and accounts for PS towards the variable of
85 interest. This approach was extended by Pati *et al.* (2011) who introduced covariates and
86 random effects in the model. Conn *et al.* (2017) followed the same ideas and developed
87 a more generic model for ecological applications, which they applied to aerial seal count
88 data. Pennino *et al.* (2019) applied similar ideas to infer the distribution of shrimps from
89 onboard fishery data.

90 Provided PS is accounted for, integrated models (IM) appear as an attractive tool to
91 combine fishery-independent and fishery-dependent data to infer harvested fish spatial
92 distribution. IM have received considerable attention in the ecological literature (Schaub
93 and Abadi, 2011; Parent and Rivot, 2012; Gimenez *et al.*, 2014). By sharing the
94 information between different data types, IM may provide more accurate estimates and
95 predictions compared with separate analysis of different data types. Recently, Rufener *et*
96 *al.* (2021) demonstrated the potential of IM to integrate scientific data and onboard
97 observer count data to improve SDM of fishery resources. However, although onboard
98 observer data provide useful complementary information to scientific survey, they
99 generally only represent a small proportion of all sea trips (1% in average for the French
100 observer programs - Cornou *et al.*, 2021). By contrast, the combination of commercial
101 catch declarations in logbooks with VMS data provides a more extensive data source to
102 map fish spatial distribution. Furthermore, the potential of embedding PS within a
103 hierarchical SDM to integrate catch declaration data and scientific survey is still an open
104 challenge and new methodology are required to handle PS behaviors of commercial fleets

105 while accounting for all the complexity related to fishing locational choice (Salas and
106 Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017).

107 In this paper, we develop an IM model to infer fish spatial distribution by combining both
108 scientific and commercial catch declaration data while taking into account the PS induced
109 by fishing targeting behavior.

110 To assess the challenges, the benefits and also the limits of the approach, we evaluate
111 the performance of our IM based on simulated data. Simulations are primarily designed
112 to assess the respective contribution of each data source to inference for different model
113 configurations. We first evaluate how the balance between the commercial and scientific
114 sample sizes affect the model outputs. Because the commercial data may often only
115 partially cover the distribution area of a targeted species, we assess how this issue may
116 affect the quality of estimation and how scientific data may contribute to reduce the effect
117 of this gap in the commercial data. Introducing PS within an IM framework involves adding
118 new parameters, complexifying the model structure and then increasing the computational
119 cost. We therefore assess how perform a more parsimonious model that would ignore PS.
120 Last, in addition to the PS, the fishing locations can be controlled by other factors
121 independent from the species distribution (e.g. logistical constraints, management
122 regulations – see Girardin *et al.*, 2017; Ducharme-Barth *et al.*, 2022). We therefore assess
123 how such process blurring strict PS may affect the quality of inferences.

124 We demonstrate the flexibility of the approach by fitting the model to three different
125 important European demersal fishery resources in the Bay of Biscay: common sole (*Solea*
126 *solea*, Linnaeus, 1758), hake (*Merluccius merluccius*, Linnaeus, 1758) and squids
127 (*Loliginidae* family). With these contrasted examples, we illustrate the capacity of the

128 framework to handle multiple commercial fleets with potentially distinct PS intensities and
129 different fishing behaviors.

130 **2 MATERIAL AND METHODS**

131 **2.1 Spatial integrated model**

132 Below we provide the core elements of the modelling approach. Additional details are
133 provided in supplementary material (SM 1). The model is structured in four layers:
134 observations (here commercial and scientific CPUE in weight per unit of effort), sampling
135 process, latent field (here fish biomass relative density) and parameters (Figure 1 - all
136 notations are available in SM 1.1, Table S1). Sampling process is usually ignored in
137 hierarchical models as it is mostly considered independent of the quantity of interest, and
138 then has no consequence on the estimation procedure (Diggle *et al.*, 2010). Here, the
139 spatial distribution of commercial fishing is explicitly modelled as a inhomogenous Poisson
140 point process whose intensity may depend on the biomass field and contributes to the
141 likelihood. The observation processes of scientific and commercial data are conditional
142 upon the biomass latent field and the sampled locations.

143 All processes are considered to occur in a discrete fine grid (see for instance SM 2.1,
144 Figure S2.1 or SM 3.1, Figure S3.1). We assume the density of the point process is
145 piecewise constant in each cell grid which brings simplification in the expression of the
146 likelihood of the point process (Diggle, 2013 - see SM 1.2). The time component is omitted
147 and both commercial and scientific data are assumed to occur at the same time step.

148 The IM is designed to assimilate the scientific data of several surveys and/or the
149 commercial data of several fleets. In the following, the subscript j refers to the different
150 data sources either scientific or commercial. For instance, in a model with one scientific
151 survey and two commercial fleets, j will take the values $j = 1,2,3$, with $j = 1$ for the
152 scientific data and $j = 2,3$ for the two commercial fleets.

153 **2.1.1 Latent field of relative biomass**

154 The fish biomass relative density S (eq. (1) – (2)) is modeled through a latent log Gaussian
155 spatial field defined on the same discrete spatial domain as the point process. The mean
156 of the Gaussian field depends on environmental covariates through a log link where the
157 linear predictor combines an intercept α_S , the linear effect of environmental covariates
158 $\Gamma_S(x)$ (effects captured by the corresponding fixed parameters β_S representing the
159 species-habitat relationship). The remaining spatial variation is accounted for through a
160 zero-mean Gaussian random field (GRF) denoted $\delta(x)$ parameterized with a Matérn
161 correlation function $M(x, x'; \kappa, \phi)$, characterized by the shape κ and the scale ϕ (Cressie,
162 1993; Gelfand *et al.*, 2010; Lindgren *et al.*, 2011 and Banerjee *et al.*, (2014)). The shape
163 can be expressed in term of range $\rho = \frac{\sqrt{8}}{\kappa}$ where ρ is the distance for which the correlation
164 between points is near 0.1.

165
$$\log(S(x)) = \alpha_S + \Gamma_S(x)^T \cdot \beta_S + \delta(x) \quad (1)$$

166
$$\delta(x) \sim GRF(0, M(x, x'; \kappa, \phi)) \quad (2)$$

167 **2.1.2 Sampling process**

168 Recent literature has emphasized the complexity of the targeting behavior processes
169 (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Abbott *et al.*, 2015; Girardin *et al.*, 2017;
170 Hintzen, 2021). In this paper, we did not attempt to model explicitly all those processes
171 (e.g. resource distribution, logistical constraints, tradition, management regulations) and
172 opted for a simplified representation where the spatial targeting directly depends on the
173 biomass field S and on an additional spatially structured random term.

174 Let us denote X_{comj} the spatial point process where commercial vessels of fleet j are
175 identified as fishing. In the following, all vessels in the same commercial fleet are assumed

176 to have homogeneous behaviors. Following Diggle *et al.* (2010), the set of fishing locations
177 are modeled conditionally on S , as a inhomogeneous Poisson point process with
178 piecewise constant intensity $\lambda_j(x)$ (eq. (3) - (4)).

$$179 \quad X_{comj} \sim \mathcal{JPP}(\lambda_j(x)) \quad (3)$$

$$180 \quad \log(\lambda_j(x)) = \alpha_{Xj} + b_j \cdot \log(S(x)) + \eta_j(x) \quad (4)$$

181 For any fleet j , intensity $\lambda_j(\cdot)$ of the Poisson point process is modeled as a log-linear
182 combination of the intercept α_{Xj} , the logarithm of the relative biomass $S(\cdot)$ scaled by a
183 parameter b_j , and a residual spatial effect $\eta_j(\cdot)$ with the same structure as $\delta(\cdot)$ but with
184 specific parameters κ and ϕ . All parameters α_{Xj} , b_j and the spatial random effect $\eta_j(x)$
185 are specific to each fleet.

186 The parameter b_j quantifies the strength of PS by scaling the relationship between the
187 local value of the resource field and the local fishing intensity.

188 Fishing locations potentially depend on many other factors than fish distribution such as
189 distance to harbor, logistical constraints, management regulations - spatial closures,
190 quotas – or fishing habits/tradition (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin
191 *et al.*, 2017). The spatial random effect $\eta_j(\cdot)$ is needed to capture any remaining additional
192 effect not captured by the dependence to $S(\cdot)$.

193 In that sense, a zero value for b_j indicates that the choice of the sampling locations does
194 not depend on the fish biomass relative density but only on the spatial random effect.

195 In addition to b_j , a dimensionless spatial metric was developed to quantify the strength of
196 PS (SM 1.3).

197 **2.1.3 Observation process**

198 Both scientific and commercial observations are considered as proportional to the
199 underlying biomass through a zero-inflated observation process. In our applications,
200 observations are expressed as CPUE (in weights / unit effort), with high proportion of
201 zeros (zeros represent on average 30% of the commercial data and 10 to 50% of scientific
202 data).

203 Observations are modelled through a zero-inflated lognormal model conditionally on
204 biomass $S(x)$ in cell x (eq. (5-6)). The model is derived from Thorson *et al.* (2016) or
205 Thorson (2018). We assume that the expected catch $\mu_j(x)$ for any fleet/data source j in
206 the cell x depends on the latent field value $S(x)$ and a catchability coefficient q_j (eq. (5)).

207 A zero catch ($y = 0$) is modeled as a Bernoulli random variable with parameter $\exp(-e^{\xi_j} \cdot$
208 $\mu_j(x))$, where ξ_j is the parameter controlling the intensity of zeros relatively to the
209 expected catch (eq. (6)). Then, $\mu_j(x)$ being fixed, the higher (resp., the lower) ξ_j , the lower
210 (resp. the higher) the probability of obtaining a zero-catch.

211 The distribution of a positive catch $y > 0$ at a given x is defined as the combination of the
212 probability of obtaining a non-zero catch $(1 - \exp(-e^{\xi_j} \cdot \mu_j(x)))$ times a positive
213 continuous distribution L (here a lognormal distribution) with expected value
214 $\frac{\mu_j(x)}{(1 - \exp(-e^{\xi_j} \cdot \mu_j(x)))}$ and standard deviation σ_j . This formulation allows to represent the zero
215 catch while assuring that the expected catch still equals $\mu_j(x)$.

216
$$\mu_j(x) = q_j \cdot S(x) \quad (5)$$

217 $P(Y = y|x, S(x)) =$

218
$$\begin{cases} \exp(-e^{\xi_j} \cdot \mu_j(x)) & \text{if } y = 0 \\ \left(1 - \exp(-e^{\xi_j} \cdot \mu_j(x))\right) \cdot L\left(y, \frac{\mu_j(x)}{(1 - \exp(-e^{\xi_j} \cdot \mu_j(x)))}, \sigma_j^2\right) & \text{if } y > 0 \end{cases} \quad (6)$$

219 Per se, catchability q_j are not identifiable as there is no information in the model to
 220 estimate the absolute scale of S . Commercial catches and/or scientific surveys will be only
 221 informative about fish biomass relative density and additional information must be
 222 provided to ensure statistical identifiability. If only one data type feeds the model (only
 223 scientific or commercial data), relative catchability is fixed to 1 and the spatial random field
 224 values is in the same scale as the data. If two data types (or more) are used to feed the
 225 model, one of the relative catchability (denoted q_{ref}) has to be fixed, the other ones being
 226 estimated relatively to the first one through a scaling factor k_j (eq. (7)).

227
$$q_j = k_j * q_{ref} \quad (7)$$

228 As it is illustrated further in the simulation-estimation study (see section 3.1.1), the choice
 229 of the reference level can have important consequences on the precision of estimation.

230 **2.1.4 Maximum likelihood estimation**

231 The estimation of the model is performed with TMB (Template Model Builder - Kristensen
 232 *et al.* (2016)) and the spatial random effects are estimated through the SPDE approach
 233 (Lindgren *et al.*, 2011) within the R software (R Core Team, 2020). More details on
 234 estimation are available in the supplementary material (SM 1.4).

235 **2.1.5 Integrated model validation**

236 A key issue with IM is whether the different data sources provide consistent or conflicting
 237 information (Saunders *et al.*, 2019; Zipkin *et al.*, 2019; Peterson *et al.*, 2021). In our

238 framework, the key question is whether integrating commercial data in addition to scientific
239 data will complement or will disrupt the inferences obtained from the scientific data,
240 considered as a reference source of information. To address this issue, we propose a
241 validation procedure based on the consistency check initially developed by Rufener *et al.*
242 (2021) and designed to check whether estimates obtained from the IM are consistent with
243 those obtained from the model fitted to scientific data only. The procedure would reject
244 consistency if the parameters estimates from the IM fall outside the 95% confidence region
245 of parameters estimates from scientific data only (see SM 1.5 for more details on the
246 procedure). This validation step is applied to both simulations and case studies.

247 **2.2 Simulation-estimation experiments**

248 We conducted simulation-estimation experiments to assess the performance of the
249 method for different data/model configurations (Table 1, see also SM 2 for extended
250 details on simulations). For all scenarios, simulations of data, covariates and GRF were
251 parameterized to tailor the case studies described hereafter. All scenarios and
252 configurations are repeated 100 times so as to capture the variability between replicates.
253 Simulation-estimation experiments were specifically designed to address four questions
254 detailed below. In all cases, commercial data were simulated with various levels of PS
255 ($b = 0$ for uniform sampling, $b = 1$ for moderate PS, $b = 3$ for strong PS) to assess the
256 effect of PS on model's performance (Figure 2).

257 *(Q1) How does each data source contribute to inferences?*

258 In real case study, commercial data sample size may be far superior to scientific data
259 (specifically when using landings data) which might result in commercial data that
260 dominate inferences. To assess how the balance between the scientific and commercial

261 sample sizes drives the relative contribution of each data source, simulations were
262 conducted with few scientific samples (50 each) with increasing commercial samples
263 (50=small, 400=medium and 3000=large), and with a large commercial sample size
264 (3000) with increasing scientific sample size (50=small, 400=medium, 3000=large). No
265 scenario with more scientific samples than commercial samples is presented here as it is
266 a very unlikely configuration when using logbook catch data.

267 For each combination of commercial and scientific sample size, we fitted four different
268 models: a model fitted to scientific data only, a model fitted to commercial data only, and
269 two IM fitted to both commercial and scientific data, one with the scientific data used as
270 reference level and another one using the commercial data as reference level (Cf. eq. (7)).

271 For questions Q2, Q3 and Q4, all simulations were conducted using $n_{scientific} = 50$ and
272 $n_{commercial} = 3000$ to tailor the case studies. Commercial data are used as the reference
273 for catchability in the IM.

274 *(Q2) How does a partial coverage of the study area by the commercial data affect*
275 *the quality of the estimation?*

276 While scientific surveys are supposed to cover the full population distribution area, partial
277 coverage of the area by commercial fishing boats may arise from different sources like
278 spatial management closures (e.g. box closure) or too expensive travels from the coast.
279 To assess how a partial coverage by commercial data can affect estimates, we simulated
280 data with the commercial sampling intensity arbitrarily fixed to 0 in a fixed 9x9 box (15%
281 of the domain) while some biomass and some scientific samples are still simulated in this
282 area. We compared estimates of the biomass in the entire area with those obtained with
283 commercial data available on the whole domain.

284 **(Q3)** *What is the cost of ignoring PS in estimation when sampling is preferential?*

285 Modelling preferential sampling involves conditioning results upon a specified structural
286 assumption about sampling as well as increased computational cost. Here, we assess
287 how ignoring PS would affect the quality of inferences when sampling is actually
288 preferential. We voluntarily introduce misspecification between the model used for
289 simulating the data (with various levels of PS intensity) and the one used in the estimation
290 procedure (b is alternatively estimated or arbitrarily fixed at 0).

291 **(Q4)** *How does the estimation perform when additional processes other than PS
292 drive the fishing locations?*

293 Fishing locations potentially depend on many other factors independent from the species
294 distribution (Salas and Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017). To
295 assess how such process blurring strict PS may affect the quality of inferences, we
296 simulate data with a sampling intensity that depends on both the biomass distribution (PS)
297 and an additional spatial random terms $\eta_f(\cdot)$ independent from the biomass distribution
298 (eq. (4); see Table 1 for more details on $\eta_f(\cdot)$ parameterization), and compare the
299 inferences obtained from a data set simulated with strict PS ($\eta_f(\cdot) = 0$ on the full domain).

300 Note that for questions Q1, Q2 and Q3, the random effect η was fixed to 0 in simulations
301 (but it is still estimated in the estimation model), so that the sampling process only
302 depends on the distribution of biomass.

303 **2.2.1 Performance metrics**

304 The performance of the estimation method was assessed using different metrics on key
305 model outputs such as the total biomass, the PS parameter b and the spatial biomass
306 predictions.

307 The quality of the total biomass estimation (the sum over all grid cells, $B = \sum_x S(x)$) was
308 explored through the relative bias $\frac{(B-\hat{B})}{B}$, that quantifies how much the total biomass is over
309 or under-estimated.

310 The quality of the estimation of the parameter b is assessed through the relative bias
311 defined as $\frac{b-\hat{b}}{b}$ (except for $b = 0$, where only the absolute bias is considered). We also
312 assessed the relative bias of the species-habitat relationship estimate $\hat{\beta}_S$ and range
313 parameter ρ as these parameters are meaningful for understanding species distribution.

314 The precision of the spatial predictions was studied with the mean squared prediction error
315 between the simulated and the estimated latent field values $\frac{1}{n} \sum_x (S(x) - \widehat{S}(x))^2$ (MSPE –
316 n stands for the number of grid cells).

317 **2.3 Case studies**

318 We applied the approach on three case studies of demersal fisheries in the Bay of Biscay:
319 the common sole (*Solea solea*, Linnaeus, 1758), the hake (*Merluccius merluccius*,
320 Linnaeus, 1758) and the squids (Loliginidae family). These case studies were selected
321 because they emphasize different intensities of preferential sampling. Further details on
322 case studies and data are provided in SM 3.

323 To compare models on the same spatial domain for the three species, we limited the
324 analysis to scientific and commercial data available on the Bay of Biscay only (SM 3.1,
325 Figure S3.1 for the spatial grids). Besides, to get some replicates of the analysis, we
326 applied the approach on 2 years for each case study (2017 and 2018 for common sole –
327 2014 and 2015 for hake and squid). To keep it synthetic, only the data and the results of
328 the models for hake in 2014, sole in 2017 and squids in 2015 are presented in this

329 manuscript as the related IM pass the consistency check and they emphasize contrasted
330 level of PS.

331 **2.3.1 Survey data**

332 Scientific data (CPUE, in kg/hour - Figure 3) were derived from the Orhago survey for
333 common sole and EVHOE survey for hake and squids (ICES, 2020a). The sampling
334 density (number of data points / km²) of those two surveys revealed representative of the
335 sampling density of the main European trawl surveys from the DATRAS database (see
336 SM 3.2). In comparison, commercial data used in the case studies are denser by 2 orders
337 of magnitude. Scientific data was aligned on commercial data by filtering only individuals
338 above the minimum landing size when available (24 cm for sole, 27 cm for hake - ICES,
339 2020). The Orhago survey provides 49 samples for 2017 and 2018 and the EVHOE survey
340 provides 86 samples for 2014 and 2015.

341 **2.3.2 Commercial data**

342 For each species, we filtered commercial data for 'bottom trawlers' as they cover a wide
343 part of the study area (Figure 3) and provide easy to compute and reliable CPUE.
344 Commercial data were standardized by the fishing effort in (kg/hour). For hake and sole,
345 we filtered the métier targeting demersal fish (called OTB_DEF) and for squids, the métier
346 targeting cephalopods (called OTB_CEP).

347 In comparison with scientific data, the orders of magnitude of commercial sample size is
348 much larger. For hake (i.e. OTB_DEF), there are 6852 commercial samples in 2014 and
349 5000 in 2015. For squids (i.e. OTB_CEP), there are 7486 commercial samples in 2014
350 and 9611 in 2015. For sole (i.e. OTB_DEF), there are 2401 samples in 2017 and 3325 in
351 2018.

352 **2.3.3 Habitat covariates**

353 Two covariates classically used to describe benthic species distribution were selected:
354 depth and sediment type (Le Pape *et al.*, 2003; Witman and Roy, 2009; Rochette *et al.*,
355 2010). Depth was separated into several categories and was considered (as sediment)
356 as a categorical variable (SM 3.7, 3.8).

357 **2.3.4 Model configurations**

358 As for the simulation-estimation experiments, the models of the case studies were fitted
359 under different configurations. To assess the information brought by each dataset, we
360 compared the model fitted to scientific data only, to commercial data only and to both
361 scientific and commercial data. To assess the effect of PS on model outputs, we compared
362 the IM accounting for PS (b is estimated) with the IM where PS is ignored (b is fixed to 0).
363 For the sole case study, we compared results obtained from the IM by considering one
364 homogeneous or two distinct fleets with specific catchability and targeting parameters.
365 Note that splitting one fleet in 2 distinct fleets is performed through a PCA coupled with a
366 HCPC analysis on vessels characteristics data derived from both logbooks and VMS data.
367 All the clustering analysis is described in SM 3.9.

368 **2.3.5 Model evaluation**

369 Uncertainty of the predictions are quantified through the coefficient of variation and all
370 estimates (e.g. fixed parameters, total biomass) are represented with related 95 %
371 confidence intervals. We assess the consistency of the IM through the statistical tests
372 described in section 2.1.5 and in SM 1.5. Finally, the different IM are compared through a
373 5-fold cross validation, and model performance was quantified based on two metrics: the
374 $MSPE_{fit}$ that measures goodness of fit (MSPE – mean squared prediction error), and the
375 PCV that measures predictive capacity (see SM 3.10 for more details on the metrics and

376 guidelines for interpretation). For both metrics, the lower the values, the better the model
377 fits/predicts the data.

378 **3 RESULTS**

379 **3.1 Simulations**

380 We summarize the main results of the simulation-estimation experiments below.

381 Additional results are provided in SM 4.

382 **3.1.1 Contribution of each data source in the integrated model**

383 Models fitted on scientific data only provide systematically unbiased estimates of total
384 biomass (the mean bias is close to 0 for all sample size - Figure 4, 1st row), and the
385 variance of estimations decreases with scientific sample size. Note that the species-
386 habitat relationship estimates $\hat{\beta}_s$ are also unbiased (see SM 4.1).

387 Overall, inferences from the IM revealed consistent with those obtained from scientific
388 data only (SM 4.2.1). Even when the commercial sample size is large and the scientific
389 sample size is small, only 3% of the p-values fall below the 0.05 threshold for the fixed
390 effect test (the test wrongly rejects consistency). For the random effect test, the results
391 are more contrasted as 10% of the p-values fall below the 0.05 threshold when data size
392 are very unbalanced (low scientific sample – high commercial sample).

393 In almost all configurations, the IM provide unbiased and more precise estimates for total
394 biomass and spatial biomass predictions compared to the model fitted to scientific data
395 only (Figure 4). As expected, the larger the commercial and the scientific sample size, the
396 more accurate the spatial predictions, the PS parameter b and total biomass estimates.
397 Estimates of b are unbiased in most cases except when commercial sample size is small
398 and PS is strong (Figure 4, 2nd row).

399 As expected, the contribution of each data sources in the IM directly depends on the
400 balance in the sample size. When sample size is balanced between the data sources,
401 then integrating the two data sources in the model systematically improves the inferences
402 with regards to situations where only one data source is analyzed. For instance, for large
403 commercial and scientific sample size (com.L_sci.L) and no PS, the precision is 1.5 higher
404 (i.e. the MSPE is 1.5 lower) for the IM compared to single-data models (either scientific or
405 commercial - Figure 4, 3rd row, 1st column). However, when the sample sizes are
406 unbalanced, the data source with the larger sample size (here commercial data)
407 dominates inference and integrating another data source with a smaller sample size (here
408 scientific data) contributes to a much lesser extent to inference. See for instance the
409 situation where commercial sample size is large and scientific sample size is small
410 (com.L_sci.S - Figure 4, 3rd row, 1st column). In this case, the performances of the model
411 fitted to commercial data alone – with reference level fixed to commercial data - are very
412 close to those of the IM whatever the intensity of PS.

413 Interestingly, the higher the intensity of PS, the higher the benefits of fitting commercial
414 data in the model (Figure 4, 3rd row); for instance, when both datasets have large sample
415 sizes (com.L_sci.L), increasing PS reduces error predictions (i.e. increases accuracy) by
416 2 each time (i.e. for $b = 0$, $E(MSPE) = 20$; for $b = 1$, $E(MSPE) = 10$; for $b = 3$,
417 $E(MSPE) = 5$).

418 Still, the simulations also reveal some limits in the inferences. First, the range parameter
419 might be poorly estimated and slightly biased when the sample size is small while being
420 better estimated when increasing the sample size or integrating additional data in the
421 analysis (see SM 4.3).

422 Also, in unbalanced cases the accuracy of total biomass estimates from the IM revealed
423 highly sensitive to the choice of the reference level (Figure 4, 1st row). When the
424 commercial sample size far exceeds the scientific sample size, setting the reference level
425 to the commercial data produces more precise estimates than setting the reference level
426 to scientific data. When defining scientific data as reference level, the intercept of the
427 latent field of relative biomass is estimated from the few scientific samples and resulting
428 estimates are less precise than when defining the reference level with a more numerous
429 data source (here commercial data). This is also true - to a lesser extent - for spatial
430 predictions (Figure 4, 3rd row).

431 In the following, only the case where commercial samples exceed scientific samples and
432 the reference level is fixed with commercial data is explored further as it is the closest to
433 the case studies configuration (Table 1).

434 **3.1.2 Impact of a partial coverage of the study area by the commercial data**

435 When commercial data only partially cover the distribution area, commercial data still
436 provide valuable information to predict biomass spatial distribution whatever the PS
437 intensity is (Figure 5, 2nd column). When sampling is not preferential (data simulated with
438 $b = 0$), a partial coverage of the distribution area produces on average 1.5 less precise
439 spatial predictions but estimates remain unbiased (Figure 5, 3rd row, comparing 1st and
440 2nd column). When sampling is preferential (either moderate or high), biomass estimates
441 are slightly underestimated. Integrating scientific data in the analysis does not correct this
442 bias.

443 Finally, all model configurations allow for unbiased and precise estimation of the species-
444 habitat parameters $\hat{\beta}_S$ whether or not there is a partial coverage of the domain (see SM
445 4.1) and overall almost all IM are consistent with scientific-based model (SM 4.2.2).

446 **3.1.3 How does ignoring PS impact inferences?**

447 As expected, the impact of ignoring PS in the estimation model is negligible when data is
448 simulated with no PS, and becomes more and more detrimental when the intensity of PS
449 increases in the truth (Figure 5, 3rd column). With no surprise, when data are generated
450 with no PS ($b = 0$), ignoring PS in the estimation procedure has no effect on the estimation
451 performance. When PS is moderate, total biomass estimates are 5 % overestimated ($b =$
452 1). In the case of strong PS ($b = 3$), ignoring PS in the estimation strongly deteriorates the
453 quality of inferences regarding total biomass estimates (Figure 5, 1st row, 3rd column).
454 Total biomass estimates are overestimated by 50% on average. However, the main spatial
455 patterns are well identified with or without consideration of PS, even though more precise
456 when accounting for PS (Figure 5, 3rd row, 1st column). SM 4.4 (Figure S4.4.1) presents
457 maps comparing a simulated biomass field and model predictions obtained by considering
458 or ignoring PS when $b = 3$. The areas with high biomass values (i.e. where commercial
459 sampling is dense) are well predicted by the models accounting for PS or not. The main
460 differences are localized in poorly sampled areas where biomass is low. Accounting for
461 PS in estimation allows to interpret the low sampling intensity areas as low-density areas,
462 and therefore to reduce the bias in those areas (SM 4.4, Figure S4.4.2).

463 Finally, from a computational point of view, accounting for PS on average multiplies by 4
464 the computational time (see SM 4.5).

465 **3.1.4 Effect of other spatially structured processes affecting fishing locations**

466 As expected, precision of estimates are deteriorated when fishing locations actually
467 depend upon a combination of biomass distribution (PS) and other mechanisms (here
468 captured by a spatially structured random term - Figure 5, 4th column). In this case, the
469 IM still provides valuable inferences on fish distribution, fish total biomass and estimates
470 of b , although estimations are less accurate than the base case. For instance, MSPE are
471 5 times lower when nothing else than PS affects sampling locations compared with a case
472 where sampling locations depend on both PS and other independent spatial processes
473 (Figure 5, 3rd row, 1st and 4th column). But interestingly, the weight of scientific data
474 increases when the sampling distribution of commercial data is blurred by spatial
475 processes independent from biomass spatial distribution. MSPE and relative bias
476 provided by the IM are both 1.4 smaller compared to those obtained when the model is
477 fitted to commercial data only.

478 **3.2 Case studies**

479 Below we summarize the main results obtained from the application of the framework to
480 the three case studies. Additional results and maps are provided in SM 5.

481 **3.2.1 Contribution of each dataset to the inferences**

482 Almost all the case studies successfully passed the consistency test between the IM and
483 the model fitted to scientific data only (see SM 5.1).

484 Models based on scientific data provide different spatial predictions compared with the
485 IM. Predictions for sole and squids from the scientific-based model are mainly shaped by
486 the covariate effects (Figure 6; for further analysis see SM 5.2, SM 5.3 and SM 5.4). On
487 the other hand, predictions from the IM are mainly shaped by the spatial random effect as
488 commercial data allow to better capture the local spatial correlation structures.

489 Consistently with simulations, inferences from the IM are mainly driven by the commercial
490 data (Figure 6). This logically arise from the much larger sample size of commercial data
491 compared with scientific data, combined with the good coverage of commercial data in
492 high-density areas (Figure 3). As commercial data is denser than scientific data, they will
493 better capture local spatial correlation structures than scientific data. SM 5.5 provide some
494 additional analysis of the information brought by commercial data in the IM.

495 In this configuration, scientific data bring information to model predictions in areas poorly
496 covered by the commercial data (SM 5.6 - e.g. for squids, the offshore predictions are
497 downscaled by scientific data).

498 **3.2.2 Preferential sampling and other processes affecting fishing locations**

499 In this section and related SM (SM 5.7 to SM 5.10), we focus on results from the IM only.

500 For the three case studies, estimates of b are positive, suggesting sampling by fishermen
501 is preferential towards high biomass density areas. The hake case study has the lowest
502 PS parameter ($\hat{b} = 0.88$, $sd(\hat{b}) = 0.107$), followed by sole ($\hat{b} = 2.4$, $sd(\hat{b}) = 0.046$) and
503 squids ($\hat{b} = 3.5$, $sd(\hat{b}) = 0.025$). For more intuition concerning the strength of PS and
504 how it varies in space, refer to SM 5.7. In all case studies, the spatial random term η in
505 the sampling process turned out to be spatially structured (SM 5.8) and captures 25% to
506 97% of the spatial variability of fishing locations (SM 5.9). This highlights the importance
507 of other spatial mechanisms in the choice of fishing locations compared to strict PS
508 towards biomass distribution.

509 Consistently with simulations, the higher the PS intensity, the higher the differences
510 between inferences obtained with and without considering PS. When comparing biomass
511 field values (Figure 7, left column), ignoring PS increases predictions in poorly sampled

512 areas (all red areas – compare with Figure 3). This effect is particularly marked for the
513 squid case study where the relative difference is the strongest in the offshore areas.
514 However, considering PS or not has relatively little effect in areas where sampling is
515 spatially denser (all white areas). Ignoring PS affects total biomass indices estimates and
516 the relative difference between biomass estimates with or without PS increases with the
517 value of b estimates (Figure 7, right column).

518 When the estimated PS intensity is high (i.e. in the case of squids) accounting for PS can
519 improve model goodness-of-fit and predictive capacity (SM 5.10).

520 **3.2.3 Benefits of considering different fleets in the estimation model**

521 Based on the sole case study, we demonstrate the capacity of the model to integrate
522 multiple commercial fishing fleets, each with specific parameters (catchability and
523 targeting). In the sole case studies, considering two different fleets in the IM (instead of
524 one homogeneous) improves goodness-of-fit towards scientific data (SM 5.11, y-axis) and
525 modifies spatial predictions (SM 5.12).

526

527 **4 DISCUSSION**

528 **Main findings**

529 Combining multiple sources of data to build more informative spatio-temporal models for
530 fish distribution is a major challenge in fishery ecology. Commercial catch per unit effort
531 data have long been recognized as a valuable source of information eventually highly
532 complementary to scientific survey data. But the complexity of the mechanisms driving
533 the way fishermen sample in space and time make the combination of scientific and
534 commercial data challenging.

535 In this paper, we provide a hierarchical framework to integrate scientific surveys and
536 commercial catch declaration data to infer species distribution while considering the effect
537 of PS on fishing points distribution. The new model allows for exploring and questioning
538 the challenges raised by such integration. The benefit but also the limits of the new
539 approach were evaluated using simulations and through the application of the model to
540 three contrasted demersal case studies (sole, hake and squids) of the Biscay Bay fishery.
541 Both simulations and case studies demonstrate that ignoring PS in the inference may be
542 highly detrimental when the intensity of PS is strong. The present framework can serve
543 as a tool to assess the benefit of including PS in analysis, depending on the intensity of
544 PS but also on the modelling objectives. As already shown in previous studies (Conn *et al.*,
545 2017; Pennino *et al.*, 2019), when PS actually occurs in commercial catches, ignoring
546 this process may bias inferences on total biomass estimates. Even if ignoring PS may not
547 hamper the capacity to detect areas of high biomass, the biomass in low-density areas
548 may be overestimated. Therefore, if the objective is to compute biomass indices integrated
549 over a large area, then it might be worth accounting for PS to avoid biased results. By

550 contrast, if the objective is to identify hotspots, the benefits of considering PS may be
551 small with regard to the additional computational time it requires.

552 The three case studies illustrated the potential of the model to handle the variability of PS
553 behavior among species and fleets. Low PS was revealed for hake, while a moderate and
554 strong PS was revealed for sole and squids, respectively, which is consistent with the
555 expert knowledge on the behavior of those bottom trawls fleets (Y. Vermard, *com. pers.*).

556 Results also demonstrate the capacity of the framework to integrate commercial catch
557 data from multiple fleets, and the benefits for the quality of inferences when those fleets
558 have different features such as distinct catchabilities or targeting behaviors. For the sole
559 case study, this approach proves useful to distinguish two segments in the bottom trawl
560 fleet, which improved model outputs. This framework could be extended to more than two
561 fleets and combined with other studies analyzing fleets structure (Pelletier and Ferraris,
562 2000; Ferraris, 2002; Stephens and MacCall, 2004; Deporte *et al.*, 2012; Winker *et al.*,
563 2013; Okamura *et al.*, 2018).

564 **Challenges in modelling PS**

565 Still, modelling the spatial distribution of commercial fishing locations remains highly
566 challenging (Hintzen, 2021; Hintzen *et al.*, 2021). Our framework is shaped to integrate
567 data from homogeneous fishing fleets supposed to share the same fishing behavior, which
568 simplifies the modelling of the non-uniform spatial intensity of fishing for each fleet. We
569 propose a parsimonious model where the dependence of the sampling intensity to the
570 biomass is supposed to be linear in the log scale. This is a strong hypothesis and
571 departure from this hypothesis may obviously exist in the truth. For instance, the intensity
572 of PS could vary in space such as in Conn *et al.* (2017) who considered that the degree

573 of PS could change across the landscape. On the other hand, however, the log-log linear
574 assumption is easy to implement in other software including the VAST R package used
575 for operational assessments in some management regions (Thorson *et al.*, 2019).

576 Of course, many other factors may drive the spatial intensity of fishing, and those were
577 simply captured in our model through an additional spatial random term. For instance,
578 fishers' behavior may depend on prior knowledge of fish spatial distribution, on information
579 sharing within fishing cooperatives, on expected distribution of bycatch species, or
580 logistical constraints (e.g., transit costs) (Salas and Gaertner, 2004; Haynie *et al.*, 2009;
581 Girardin *et al.*, 2017). Targeting behavior may also be directed toward an assemblage of
582 species rather than toward a single species (Bourdaud *et al.*, 2019).

583 The random effect should be able to capture additional variations whenever the departure
584 from a continuous Gaussian random field is not too high. If not, for instance in the case of
585 fishery closures where fishing activity suddenly drops to very low levels (as explored in
586 simulation-estimation), the model may produce biased estimates due to model
587 misspecification. We did not detect such misspecification in our case study, but we
588 recommend that future analyses based on fishery-dependent data present a log-log plot
589 between sampling intensity and predicted biomass density to diagnose strong departure
590 from model hypothesis.

591 Still, some non-spatial targeting has been reported from multi-species catch records
592 (Stephens and MacCall, 2004; Okamura *et al.*, 2018). Efforts to integrate these methods
593 into spatio-temporal models are underway (Thorson *et al.*, 2016), although these methods
594 have not previously been extended to jointly analyzing multi-species fishery and survey
595 data.

596 **Relative contribution of scientific and commercial data**

597 Our analysis exemplifies that a key issue in such integrated modelling exercise is to get a
598 sensible evaluation of the relative contribution of the different sources of data in
599 estimation. In particular, critical issues with the IM are whether the different data sources
600 provide eventually highly unbalanced quantity of information (then the inferences are fully
601 dominated by one of the data sources; Fletcher *et al.*, 2019) and whether they provide
602 complementary or conflicting information to the final inferences (Saunders *et al.*, 2019;
603 Zipkin *et al.*, 2019; Peterson *et al.*, 2021).

604 We implemented a likelihood ratio-test (Rufener *et al.*, 2021) to check for model
605 consistency between the IM and the scientific-based model. In most cases, models
606 passed the consistency check successfully, although it was rejected in some cases. Some
607 further analysis should investigate in detail the reasons of these inconsistencies as they
608 could probably shed light on some new research avenues for model improvement. For
609 instance, some neglected vessel effect (*e.g.*, difference in catchability among vessels -
610 Thorson and Ward, 2014) or some too simplistic representation of the sampling and/or
611 the observation process of commercial data might partly explain these inconsistencies.

612 Simulations revealed that when scientific data and commercial data have balanced
613 sample size, they both contribute to inference and the IM will provide better biomass
614 predictions than models based on single-data set. As expected, when the sample size of
615 commercial data far exceeds scientific data, inference about spatial patterns is mainly
616 driven by the commercial data. In the three case studies, we used commercial data with
617 sample sizes that far exceed the scientific one. In that case, scientific data have relatively
618 limited weight in the final inference. Still, they bring valuable information in areas that are
619 not sampled by the commercial fishery. Also, scientific data remain a critical component

620 in the analysis as they provide some reference data through a standardized sampling plan
621 and a controlled protocol allowing then to assess for the IM consistency. It would be worth
622 applying our framework to other case study that may consist in more balanced data sets,
623 such as models seeking to combine scientific with onboard observer data (Rufener *et al.*
624 2021), or in pelagic fisheries where acoustic surveys can provide continuous observations
625 over the full domain.

626 Our results also point out the importance of setting the reference level for the catchability
627 coefficient with either the scientific or the commercial data. In particular, when the sample
628 size of the commercial data far exceeds the scientific survey, fixing the reference level
629 with scientific surveys generally results in higher imprecision, due to the smaller sample
630 size. But still, in certain cases, the scientific data may provide absolute information on
631 biomass and fixing the catchability factor associated with the survey data can result in an
632 interpretable measure of index scale (Thorson *et al.*, 2021). Hence, the choice of the
633 reference level could be a matter of tradeoffs between precision of inferences and
634 interpretation of the results in terms of scale.

635 **The limits of reallocated catch data**

636 Probably one of the major limits of our approach is that the actual framework ignores the
637 uncertainty that arises from the procedure used to reallocate the catch declarations in
638 space. Obtaining the spatialized CPUE inputs used in the model requires pre-treatment
639 of the commercial catch declaration data to allocate declaration data to VMS positions
640 (Hintzen *et al.*, 2012). Raw data corresponds to fishing operations that are daily
641 aggregated and reported at coarse administrative spatial units (0.5° latitude by 1°
642 longitude rectangles). These declarations are then reallocated uniformly on all GPS

643 locations previously identified as fishing in the vessel path. This procedure has been
644 demonstrated to be robust while being a fast and a pragmatic approach for reallocating
645 landings to VMS pings (Gerritsen and Lordan, 2010; Murray *et al.*, 2013). However, it
646 implies strong hypotheses that may artificially increase or transform the information
647 provided by the data. Typically, the uniform reallocation of catch declarations on all GPS
648 positions identified as fishing may smooth the spatial signal, which could potentially
649 explain the lack of species-habitat relationship obtained from the IM. The effect of such
650 reallocation should be explored in further study to better understand its consequences on
651 model predictions/estimates and further model development should investigate how to
652 mitigate its consequences.

653 **Perspectives**

654 Our work raises some major challenges which all constitutes exciting tracks for future
655 research.

656 Data-weighting approaches could be explored further to better control the contribution of
657 the two sources of data and eventually assess if increasing scientific data weight could
658 improve model predictive capacity. Data-weighting methods intend to modify the relative
659 influence of the data sources by assigning or estimating a weight for each data source
660 (Francis, 2017; Punt, 2017; Wang and Maunder 2017; Punt *et al.*, 2020). Only very few
661 studies have already explored the potential for data weighting in the SDM context
662 (Fletcher *et al.*, 2019). Still, several questions regarding the weight specification remain
663 open or largely debated. For instance, how to rigorously fix/estimate/interpret the weight?
664 Also, when can we consider that a data-weighting approach is relevant or is it only a matter
665 of model misspecification? Some theoretical and modelling development could be highly

666 valuable to provide a generic and rigorous formalization for either data weighting or model
667 correction in the context of SDM (but see for instance the approach provided by Thorson
668 *et al.* (2017b) for composition data in the context of stock assessment models).

669 Another option would consist in developing an alternative observation model for the
670 commercial CPUE in order to better capture the uncertainty associated with the
671 reallocation procedure. As a general idea, an observation model could be developed to
672 explicitly represent that CPUE are available at the scale of the daily fishing activity (the
673 scale that corresponds to the catch declaration), rather than artificially reallocating
674 uniformly catch declarations on related VMS pings. Doing so, the quantity of information
675 provided by commercial data would be more representative of the information they really
676 contain.

677 Future work should also seek to better integrate the discrete-choice and econometric
678 analyses emphasizing the complexity of the processes related to the choice of fishing
679 locations. For instance, the sampling process could account for the pluri-specific nature
680 of fisheries (Bourdaud *et al.*, 2019) and additional factors other than fish distribution could
681 be included to explain the variability of sampling intensity in space and time (Salas and
682 Gaertner, 2004; Haynie *et al.*, 2009; Girardin *et al.*, 2017).

683 Finally, including a temporal dimension in the model and fitting a longer time series looks
684 a fruitful research avenue. Moving to spatio-temporal modelling that would consider
685 temporal autocorrelation in the spatial distribution may be methodologically challenging
686 (Cameletti *et al.*, 2013), but represents an exciting step towards a better understanding of
687 the seasonal spatial distribution of fish resources. Indeed, commercial data are often
688 available all along the year, when scientific surveys most often occur once or twice a year.
689 Combining scientific and catch declarations data within an integrated spatio-temporal

690 framework built at an infra-annual time step (e.g., season, month) would allow to
691 complement the gap of information to investigate fish spatio-temporal distribution at a finer
692 temporal scale than what is possible using scientific data only (Bourdaud *et al.*, 2017;
693 Pinto *et al.*, 2019; Rufener *et al.*, 2021). It would offer new opportunities to interpret
694 seasonal patterns of distribution (Kai *et al.*, 2017), identify fish functional habitats such as
695 spawning areas (Paradinas *et al.*, 2015; Delage and Le Pape, 2016), and provide the
696 required knowledge for protecting those habitats (Schmitten, 1999; Erisman *et al.*, 2020).

697 **SUPPLEMENTARY MATERIAL**

698 All the supplementary material documents are available at the ICESJMS online version of
699 the manuscript. They provide additional information on the modelling framework (SM1),
700 material and methods for simulations (SM2) and case studies (SM3), results for
701 simulations (SM4) and case studies (SM5).

702 **ACKNOWLEDGMENTS**

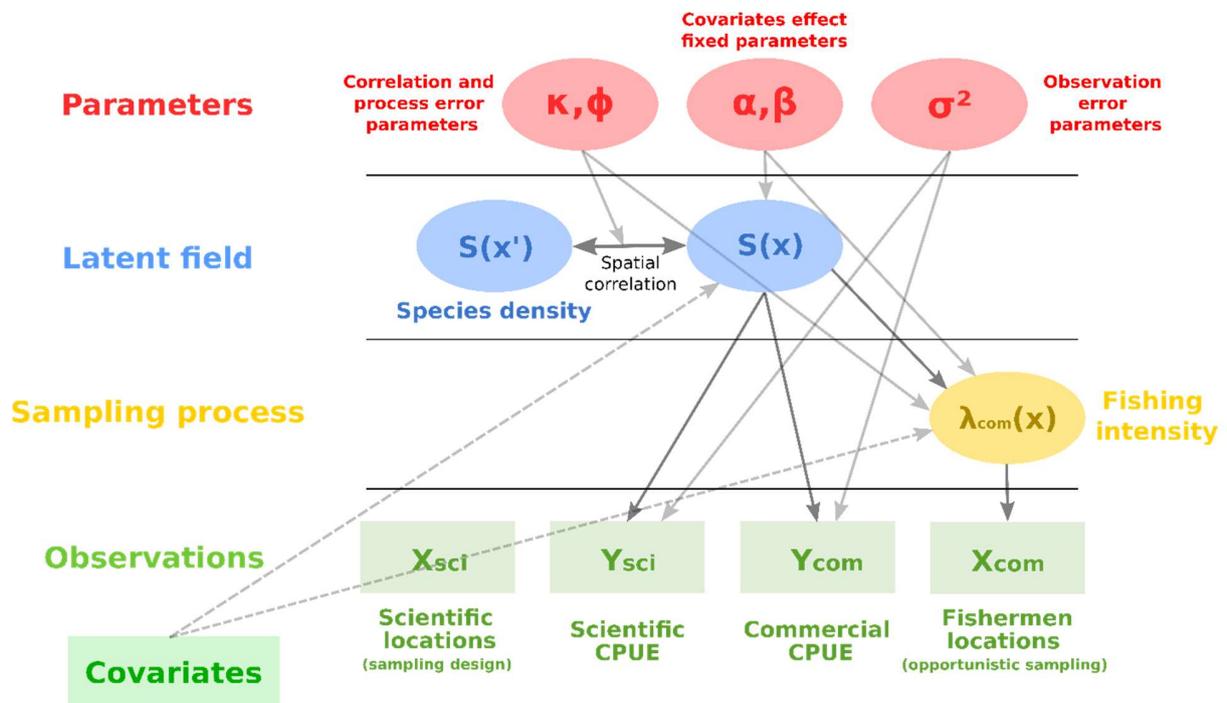
703 The authors are grateful to the Direction des pêches maritimes et de l'aquaculture (DPMA)
704 and Ifremer (Système d'Informations Halieutiques - SIH) who provided the aggregated
705 VMS data. The findings and conclusions of the present paper are those of the authors.

706 The authors thank David Eme who provided tidy environmental covariates data as well as
707 Kasper Kristensen, Jean-Baptiste Lecomte, Louise Day and Pierre-Yves Hervann for
708 their feedbacks and their highly valuable advice. The authors thanks also Maxime Olmos,
709 John Best, and two anonymous reviewers whose feedbacks greatly improved the
710 manuscript.

711 **DATA AVAILABILITY STATEMENT**

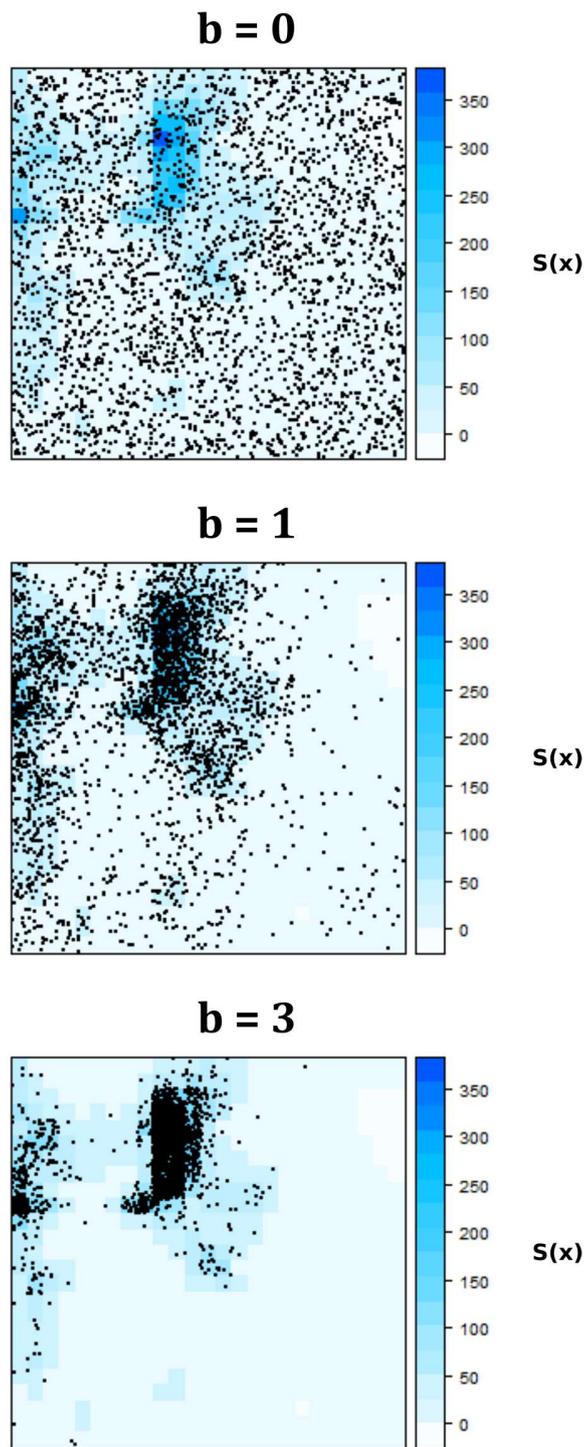
712 Survey data are available through the DATRAS portal ([https://www.ices.dk/data/data-](https://www.ices.dk/data/data-portals/Pages/DATRAS.aspx)
713 [portals/Pages/DATRAS.aspx](https://www.ices.dk/data/data-portals/Pages/DATRAS.aspx)) with the package 'icesDatras' ([https://cran.r-](https://cran.r-project.org/web/packages/icesDatras/index.html)
714 [project.org/web/packages/icesDatras/index.html](https://cran.r-project.org/web/packages/icesDatras/index.html)). Logbooks and VMS data are
715 confidential data and they are available on specific request to DPMA. Codes that support
716 the findings of this study are on gitlab and can be given access on request at the address:
717 baptiste.alglave@agrocampus-ouest.fr.

718



721

722 *Figure 1. Diagram of the spatial integrated model including preferential sampling for*
 723 *commercial data. Locations of scientific trawls do not contribute directly to the likelihood.*

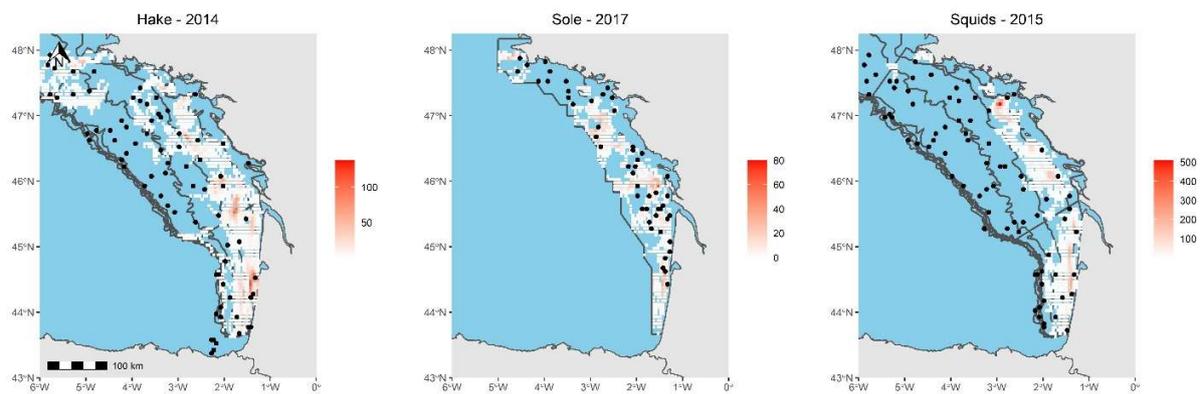


724

725 *Figure 2. Maps of simulated commercial sampling points obtained for three values of*
 726 *preferential sampling ($b=0$, $b=1$, $b=3$). Blue scale: values of the simulated biomass field.*
 727 *Dots: fishing points. For $b = 0$, the targeting metric $T_j(x) = 1$. For $b = 1$,*

728 $\arg \max_x T_j(x) = 12, q_{50\%}\{T_j(x)\} = 0.4$. For $b = 3, \arg \max_x T_j(x) = 80, q_{50\%}\{T(x)\} =$
729 0.002 (SM 1.3).

730



731

732

733

734

735

736

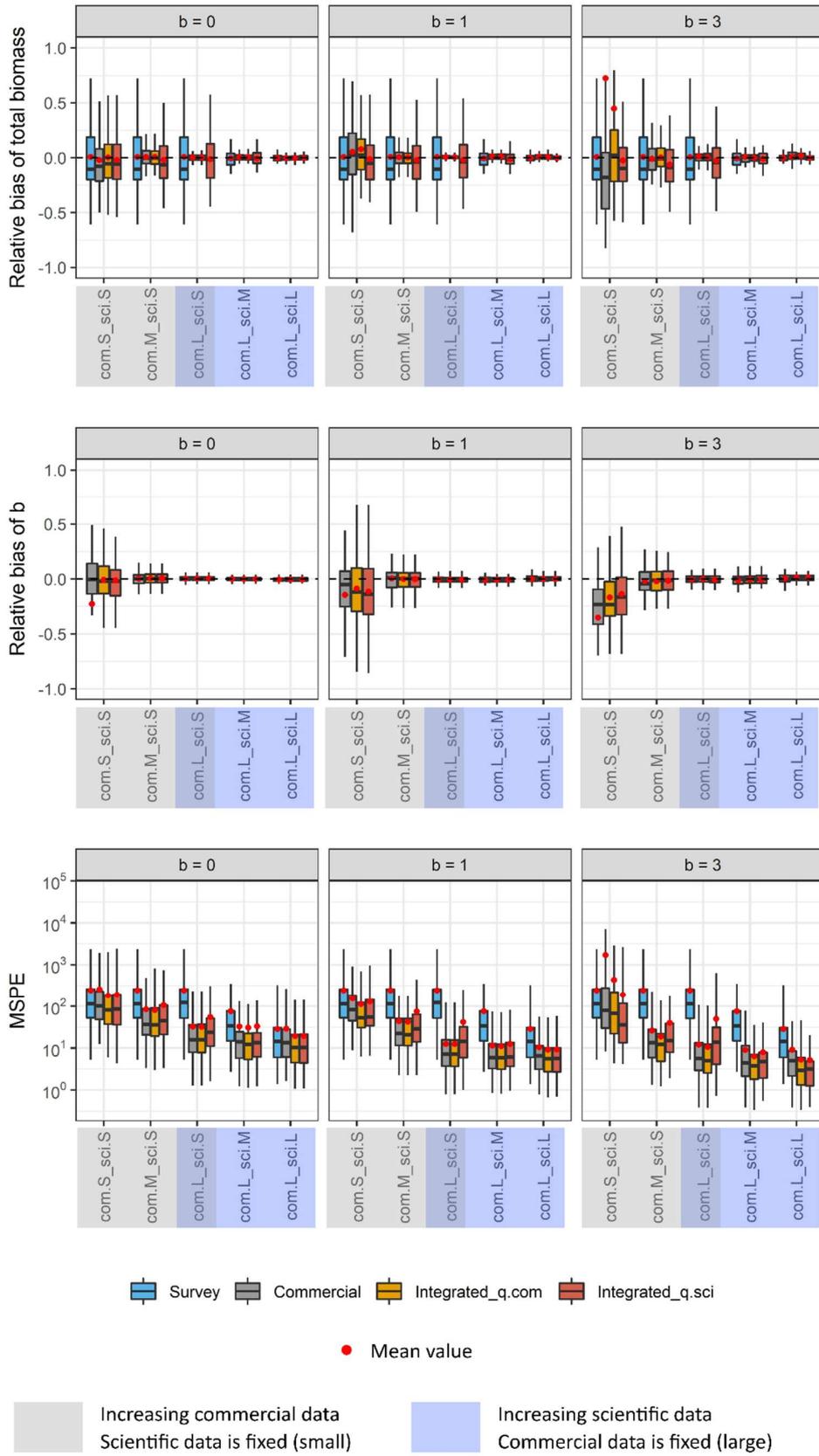
737

738

739

740

Figure 3. Map of scientific samples (black dot) and commercial sampling distribution (red color scale – unit: fishing hours). Note that all scientific hauls last around 30 minutes. Black lines - limits of the spatial domains covered by the scientific survey (Orhago and EVHOE) that delineate the study area. Left – Hake, November 2014 (EVHOE; commercial data from otter bottom trawls targeting demersal species OTB_DEF). Middle – Sole, November 2017 (Orhago; commercial data from otter bottom trawls targeting demersal species OTB_DEF). Right – squid, year 2015 (EVHOE; commercial data from otter bottom trawls targeting cephalopods OTB_CEP).

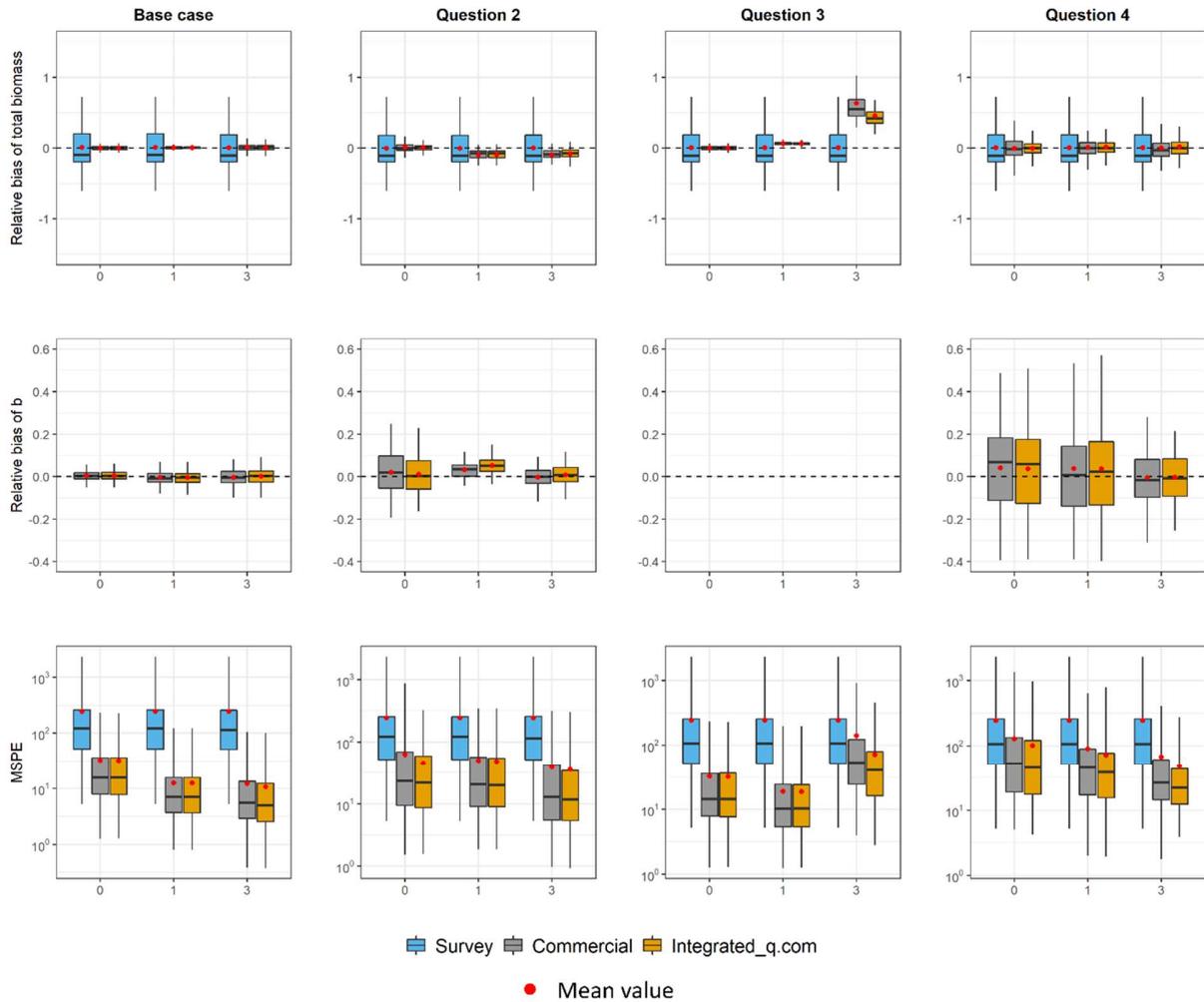


742

743 *Figure 4. Performance metrics obtained for various commercial and scientific data*
744 *sample size. Column: intensity of the preferential sampling in simulated data. x-axis: 5*
745 *combinations of commercial and scientific sample size. 'com' stands for commercial, 'sci'*
746 *stands for scientific, S stands for small sample size (50), M stands for medium sample*
747 *size (400), L stands for large sample size (3000). Colors: model configurations.*
748 *Integrated_q.com: integrated model with catchability fixed to 1 for commercial data;*
749 *Integrated_q.sci: integrated model with catchability fixed to 1 for scientific data. Boxplots*
750 *represent the variability among the 100 replicates.*

751

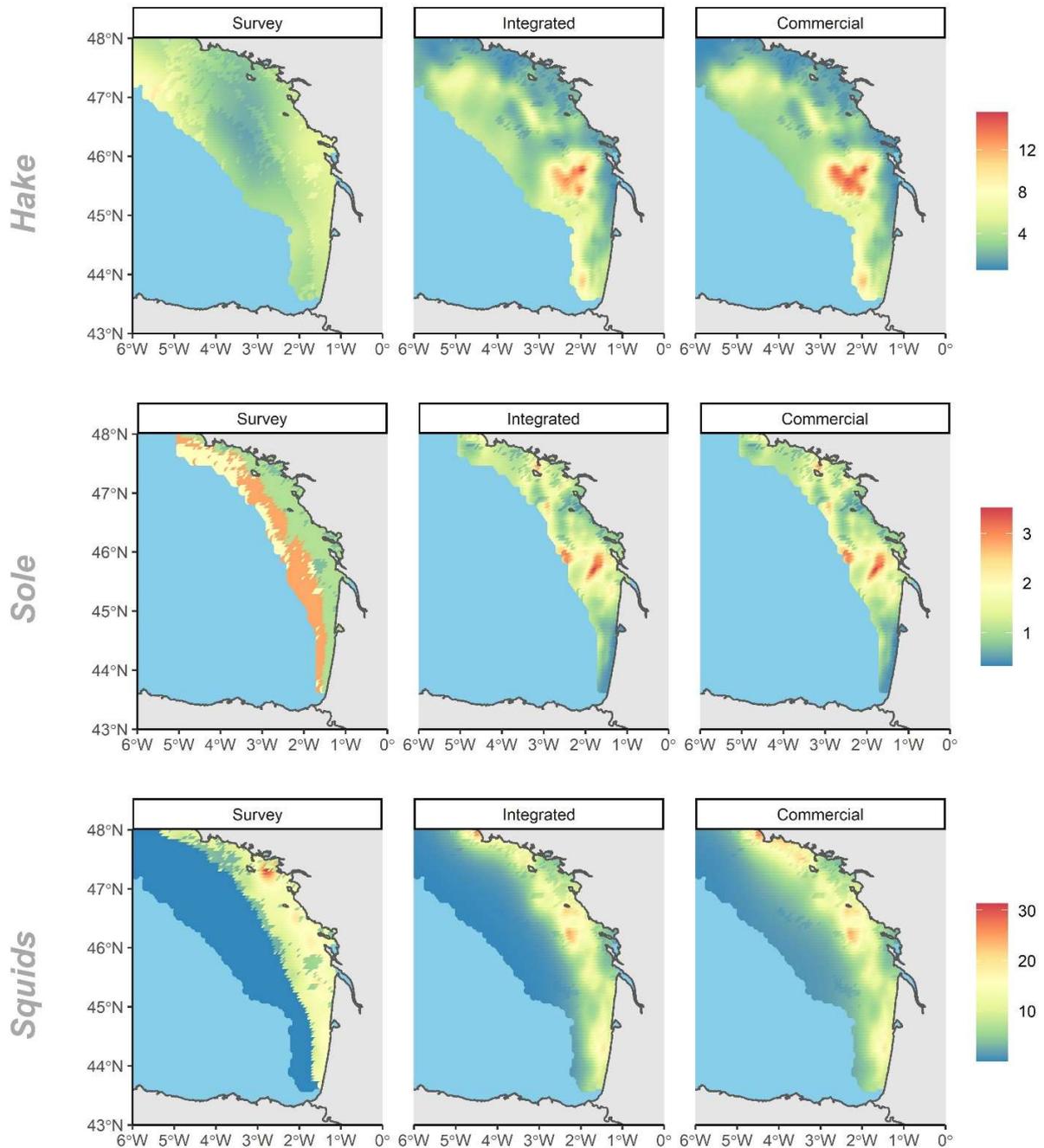
752



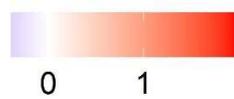
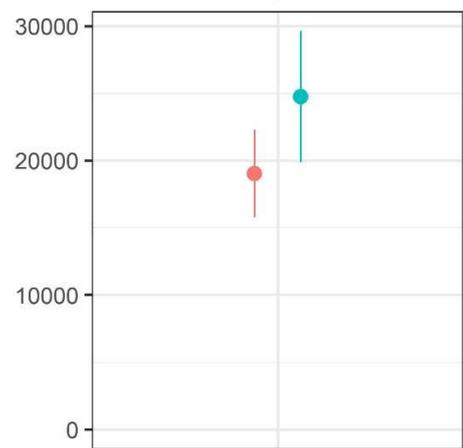
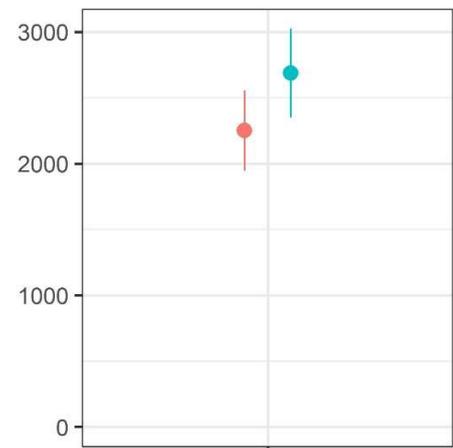
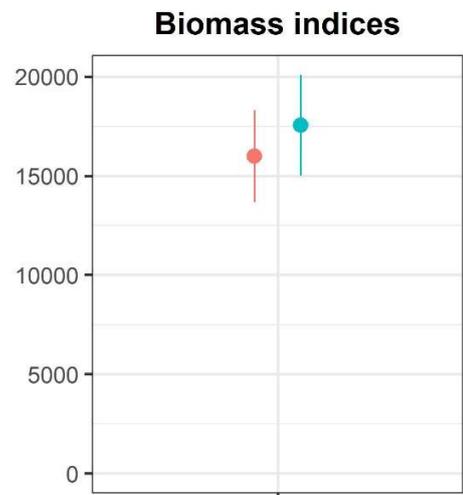
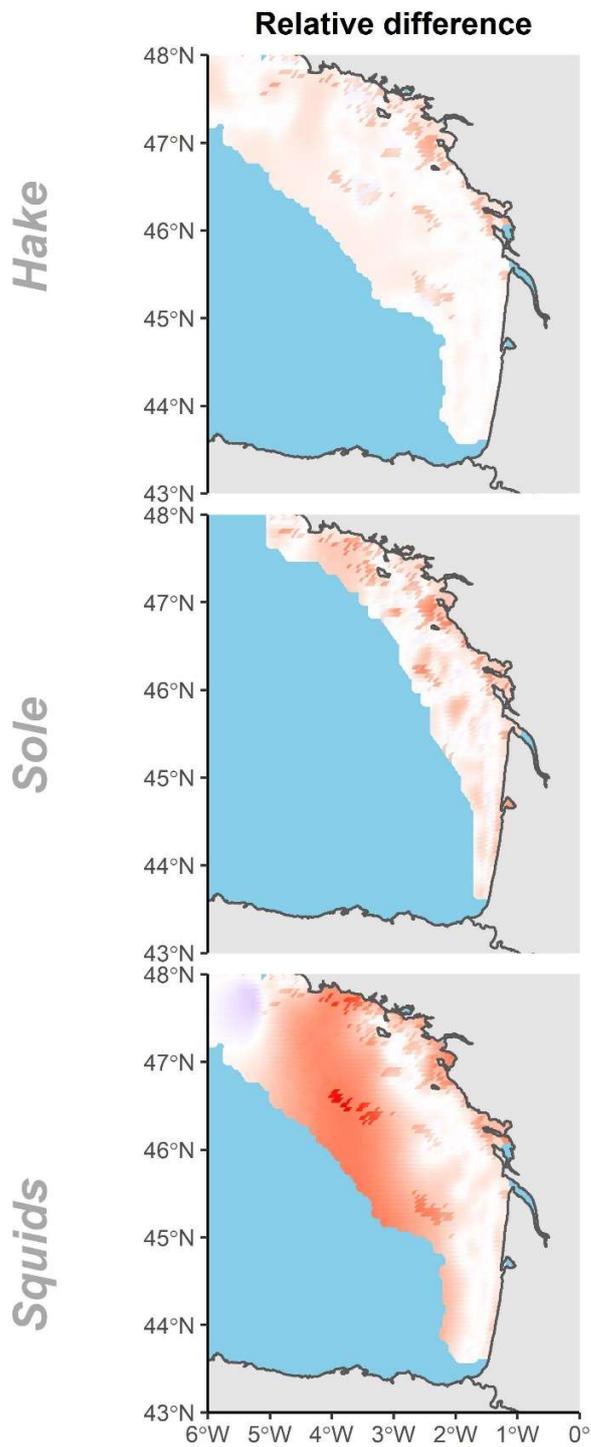
753

754

755 *Figure 5. Performance metrics obtained in different data and model configurations. Red*
 756 *points: mean value. 1st column: no discrepancy between simulation and estimation. 2nd*
 757 *column: commercial data do not cover a 9 x 9 zone of the grid. 3rd column: b is*
 758 *arbitrarily fixed to 0 in the estimation models. 4th column: data simulated with a random*
 759 *effect η in the sampling intensity process. In all configurations, simulations are*
 760 *conducted for three levels of preferential sampling (x-axis: $b = 0$, $b = 1$, $b = 3$). Colors:*
 761 *data sources used in the integrated model for inferences. Integrated_q.com: integrated*
 762 *model with catchability fixed with commercial data. Boxplots represent the variability*
 763 *among the 100 replicates.*
 764



765
 766 *Figure 6. Prediction of the relative biomass for each case study. 1st column: model*
 767 *fitted to scientific data only; 2nd column: integrated model accounting for PS; 3rd*
 768 *column: commercial-based model accounting for PS. When the model is fitted to*
 769 *scientific data only, relative biomass is rescaled with the relative catchability parameter*
 770 *estimated within the integrated model so that all maps are in the same scale.*
 771



● b_est ● b_fix

773 *Figure 7. Comparison of relative difference in biomass spatial predictions (calculated as*
774 *$(S_{b_fix}(x) - S_{b_est}(x))/S_{b_est}(x)$ in space (left) and of total biomass (sum on the spatial*
775 *domain; right) obtained with the integrated models from the 3 case studies when*
776 *accounting or not for preferential sampling. b_est : PS is estimated. b_fix : PS is not*
777 *accounted for.*

778

Table 1: Simulations

General simulations description								
Biomass field	Depends on one continuous covariate (I_s) and one random spatial effect (δ). Both are simulated independently through a GRF with Matérn covariance function. Their range (ρ) and marginal variance are fixed respectively to 10 and 1. <small>n.b. the marginal variance quantifies the variability of the spatial process. For more details on marginal variance parameterization, see Lindgren <i>et al.</i> (2011).</small>					Simulated within a 25 x 25 grid.		
Scientific data	Random stratified plan within 4 strata (see Figure S2.1)			Catchability fixed to 1		Simulated with 10% of zeroes ($\xi_j = 0$)		
Commercial data	<p>Simulated according to three PS levels (i.e. three values for b - see Figure 2).</p> <ul style="list-style-type: none"> - $b = 0$: commercial sampling is not preferential; - $b = 1$: preferential sampling is moderate, commercial vessels mainly target areas where fish biomass is high; - $b = 3$: commercial sampling is highly preferential and vessels strongly target zones where biomass is high. <p>η is set to 0 for Q1, Q2, Q3. For Q4, η is set to tailor the sole case study. The range of η is set to 40 (4 times the range of δ), the marginal variance is set to 5 (5 times the marginal variance of δ). Catchability fixed to 1 Simulated with 30 % of zero when PS is null ($\xi_j = -1$)</p>							
	Simulation scenarios					Model configurations		
	b	Scientific sample size	Commercial samples size	Coverage of the study area	Additional random effect in sampling intensity (η)	Data sources considered in the model	PS estimated	Fixed catchability
Question 1: How do each data source contribute to inferences?	0,1,3	50	50, 400, 3000	Full	No	Scientific only, commercial only, both	yes	Scientific or Commercial
	0,1,3	50, 400, 3000	3000	Full	No	Scientific only, commercial only, both	yes	Scientific or Commercial
Question 2: How does a partial coverage of the study area by the commercial data affect the quality of the estimation?	0,1,3	50	3000	No fishing in a 9x9 cells box	No	Scientific only, commercial only, both	yes	Commercial
Question 3: What is the cost of ignoring PS in estimation when sampling is preferential?	0,1,3	50	3000	Full	No	Scientific only, commercial only, both	no (b fixed at 0)	Commercial
Question 4: How does the estimation perform when additional processes other than PS drive the fishing locations?	0,1,3	50	3000	Full	Yes	Scientific only, commercial only, both	yes	Commercial

782 **REFERENCES**

- 783
- 784 Abbott, J., Haynie, A., and Reimer, M. 2015. Hidden Flexibility: Institutions, Incentives, and
785 the Margins of Selectivity in Fishing. *Land Economics*, 91: 169–195.
- 786 Banerjee, S., Carlin, B. P., and Gelfand, A. E. 2014. Hierarchical modeling and analysis for
787 spatial data. CRC press.
- 788 Bourdaud, P., Travers-Trolet, M., Vermard, Y., Cormon, X., and Marchal, P. 2017. Inferring
789 the annual, seasonal, and spatial distributions of marine species from
790 complementary research and commercial vessels' catch rates. *ICES Journal of Marine*
791 *Science*, 74: 2415–2426.
- 792 Bourdaud, P., Travers-Trolet, M., Vermard, Y., and Marchal, P. 2019. Improving the
793 interpretation of fishing effort and pressures in mixed fisheries using spatial overlap
794 metrics. *Canadian Journal of Fisheries and Aquatic Sciences*, 76: 586–596.
- 795 Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. 2013. Spatio-temporal modeling of
796 particulate matter concentration through the SPDE approach. *AStA Advances in*
797 *Statistical Analysis*, 97: 109–131.
- 798 Cheung, W. W., Lam, V. W., Sarmiento, J. L., Kearney, K., Watson, R., and Pauly, D. 2009.
799 Projecting global marine biodiversity impacts under climate change scenarios. *Fish*
800 *and fisheries*, 10: 235–251. Wiley Online Library.
- 801 Conn, P. B., Thorson, J. T., and Johnson, D. S. 2017. Confronting preferential sampling when
802 analysing population distributions: diagnosis and model-based triage. *Methods in*
803 *Ecology and Evolution*, 8: 1535–1546.
- 804 Cornou, A.-S., Quinio-Scavinner, M., Sagan, J., Cloâtre, T., Dubroca, L., and Billet, N. 2021.
805 Captures et rejets des métiers de pêche français - Résultats des observations à bord
806 des navires de pêche professionnelle en 2019. Ifremer.
- 807 Cressie, N. A. 1993. *Statistics for spatial data*. John Wiley and Sons. Inc., New York.
- 808 Delage, N., and Le Pape, O. 2016. Inventaire des zones fonctionnelles pour les ressources
809 halieutiques dans les eaux sous souveraineté française. Première partie: définitions,
810 critères d'importance et méthode pour déterminer des zones d'importance à
811 protéger en priorité. Rapport de recherche. Pôle halieutique AGROCAMPUS OUEST,
812 Rennes.
- 813 Deporte, N., Ulrich, C., Mahévas, S., Demanèche, S., and Bastardie, F. 2012. Regional métier
814 definition: a comparative investigation of statistical methods using a workflow
815 applied to international otter trawl fisheries in the North Sea. *ICES Journal of Marine*
816 *Science*, 69: 331–342. Oxford University Press.
- 817 Diggle, P. J., Menezes, R., and Su, T. 2010. Geostatistical inference under preferential
818 sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:
819 191–232.
- 820 Diggle, P. J. 2013. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC
821 press.
- 822 Ducharme-Barth, N. D., Grüss, A., Vincent, M. T., Kiyofuji, H., Aoki, Y., Pilling, G., Hampton, J.,
823 *et al.* 2022. Impacts of fisheries-dependent spatial sampling patterns on catch-per-

824 unit-effort standardization: A simulation study and fishery application. *Fisheries*
825 *Research*, 246: 106169.

826 Erisman, B. E., Grüss, A., Mascareñas-Osorio, I., Licon-González, H., Johnson, A. F., and López-
827 Sagástegui, C. 2020. Balancing conservation and utilization in spawning aggregation
828 fisheries: a trade-off analysis of an overexploited marine fish. *ICES Journal of Marine*
829 *Science*, 77: 148–161. Oxford University Press.

830 Ferraris, J. 2002. Fishing fleet profiling methodology. Food & Agriculture Org.

831 Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M.
832 2019. A practical guide for combining data to model species distributions. *Ecology*,
833 100: e02710.

834 Francis, R. C. 2017. Revisiting data weighting in fisheries stock assessment models.
835 *Fisheries Research*, 192: 5–15.

836 Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. 2010. Handbook of spatial statistics.
837 CRC press.

838 Gerritsen, H., and Lordan, C. 2010. Integrating vessel monitoring systems (VMS) data with
839 daily catch data from logbooks to explore the spatial distribution of catch and effort
840 at high resolution. *ICES Journal of Marine Science*, 68: 245–252.

841 Gimenez, O., Buckland, S. T., Morgan, B. J. T., Bez, N., Bertrand, S., Choquet, R., Dray, S., *et al.*
842 2014. Statistical ecology comes of age. *Biology Letters*, 10: 20140698. Royal Society.

843 Girardin, R., Hamon, K. G., Pinnegar, J., Poos, J. J., Thébaud, O., Tidd, A., Vermard, Y., *et al.*
844 2017. Thirty years of fleet dynamics modelling using discrete-choice models: What
845 have we learned? *Fish and Fisheries*, 18: 638–655.

846 Grüss, A., Thorson, J. T., Carroll, G., Ng, E. L., Holsman, K. K., Aydin, K., Kotwicki, S., *et al.*
847 2020. Spatio-temporal analyses of marine predator diets from data-rich and data-
848 limited systems. *Fish and Fisheries*, 21: 718–739.

849 Guisan, A., and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology.
850 *Ecological modelling*, 135: 147–186. Elsevier.

851 Haynie, A. C., Hicks, R. L., and Schnier, K. E. 2009. Common property, information, and
852 cooperation: Commercial fishing in the Bering Sea. *Ecological Economics*, 69: 406–
853 413.

854 Hilborn, R., and Walters, C. J. (Eds). 1992. Quantitative Fisheries Stock Assessment: Choice,
855 Dynamics and Uncertainty. Springer US.
856 <https://www.springer.com/gp/book/9780412022715> (Accessed 14 June 2021).

857 Hintzen, N. T., Bastardie, F., Beare, D., Piet, G. J., Ulrich, C., Deporte, N., Egekvist, J., *et al.* 2012.
858 VMStools: Open-source software for the processing, analysis and visualisation of
859 fisheries logbook and VMS data. *Fisheries Research*, 115: 31–43. Elsevier.

860 Hintzen, N. T. 2021. Zooming into small-scale fishing patterns: The use of vessel monitoring
861 by satellite in fisheries science. Wageningen University.

862 Hintzen, N. T., Aarts, G., Poos, J. J., Van der Reijden, K. J., and Rijnsdorp, A. D. 2021.
863 Quantifying habitat preference of bottom trawling gear. *ICES Journal of Marine*
864 *Science*, 78: 172–184.

865 ICES. 2005. Report of the Workshop on Survey Design and Data Analysis (WKSAD). Sète,
866 France.

867 ICES. 2012. Manual for the international bottom trawl surveys. SISP 1-IBTS Copenhagen,
868 Denmark.

869 ICES. 2020a. International Bottom Trawl Survey Working Group (IBTSWG). ICES Scientific
870 Reports. ICES. <http://www.ices.dk/sites/pub/Publication>
871 [Reports/Forms/DispForm.aspx?ID=37066](http://www.ices.dk/sites/pub/Publication) (Accessed 28 May 2021).

872 ICES. 2020b. Working Group for the Bay of Biscay and the Iberian Waters Ecoregion
873 (WGBIE). ICES Scientific Reports. ICES. <http://www.ices.dk/sites/pub/Publication>
874 [Reports/Forms/DispForm.aspx?ID=36841](http://www.ices.dk/sites/pub/Publication) (Accessed 28 May 2021).

875 Kai, M., Thorson, J. T., Piner, K. R., and Maunder, M. N. 2017. Spatiotemporal variation in
876 size-structured populations using fishery data: an application to shortfin mako
877 (*Isurus oxyrinchus*) in the Pacific Ocean. *Canadian Journal of Fisheries and Aquatic*
878 *Sciences*, 74: 1765–1780.

879 Kristensen, K., Thygesen, U. H., Andersen, K. H., and Beyer, J. E. 2014. Estimating spatio-
880 temporal dynamics of size-structured populations. *Canadian Journal of Fisheries and*
881 *Aquatic Sciences*, 71: 326–336.

882 Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. 2016. TMB: Automatic
883 Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70: 1–21.

884 Le Pape, O., Chauvet, F., Mahévas, S., Lazure, P., Guéroult, D., and Désaunay, Y. 2003.
885 Quantitative description of habitat suitability for the juvenile common sole (*Solea*
886 *solea*, L.) in the Bay of Biscay (France) and the contribution of different habitats to
887 the adult population. *Journal of Sea Research*, 50: 139–149.

888 Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields and
889 Gaussian Markov random fields: the stochastic partial differential equation
890 approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
891 73: 423–498. Wiley Online Library.

892 Martínez-Minaya, J., Cameletti, M., Conesa, D., and Pennino, M. G. 2018. Species distribution
893 modeling: a statistical review with focus in spatio-temporal issues. *Stochastic*
894 *environmental research and risk assessment*, 32: 3227–3244. Springer.

895 Moriarty, M., Sethi, S. A., Pedreschi, D., Smeltz, T. S., McGonigle, C., Harris, B. P., Wolf, N., *et al.*
896 2020. Combining fisheries surveys to inform marine species distribution modelling.
897 *ICES Journal of Marine Science*, 77: 539–552. Oxford University Press.

898 Murray, L. G., Hinz, H., Hold, N., and Kaiser, M. J. 2013. The effectiveness of using CPUE data
899 derived from Vessel Monitoring Systems and fisheries logbooks to estimate scallop
900 biomass. *ICES Journal of Marine Science*, 70: 1330–1340.

901 Nielsen, J. R. 2015. Methods for integrated use of fisheries research survey information in
902 understanding marine fish population ecology and better management advice:
903 improving methods for evaluation of research survey information under
904 consideration of survey fish detection and catch efficiency. Wageningen University.
905 Ocean Studies Board, and National Research Council. 2000. Improving the collection,
906 management, and use of marine fisheries data. National Academies Press.

907 Okamura, H., Morita, S. H., Funamoto, T., Ichinokawa, M., and Eguchi, S. 2018. Target-based
908 catch-per-unit-effort standardization in multispecies fisheries. *Canadian Journal of*
909 *Fisheries and Aquatic Sciences*, 75: 452–463.

910 Paradinas, I., Conesa, D., Pennino, M., Muñoz, F., Fernández, A., López-Quílez, A., and Bellido,
911 J. 2015. Bayesian spatio-temporal approach to identifying fish nurseries by
912 validating persistence areas. *Marine Ecology Progress Series*, 528: 245–255.

913 Parent, E., and Rivot, E. 2012. Introduction to hierarchical Bayesian modeling for ecological
914 data. CRC Press.

- 915 Pati, D., Reich, B. J., and Dunson, D. B. 2011. Bayesian geostatistical modelling with
916 informative sampling locations. *Biometrika*, 98: 35–48.
- 917 Pelletier, D., and Ferraris, J. 2000. A multivariate approach for defining fishing tactics from
918 commercial catch and effort data. *Canadian Journal of Fisheries and Aquatic
919 Sciences*, 57: 51–65. NRC Research Press.
- 920 Pennino, M. G., Conesa, D., Lopez-Quilez, A., Munoz, F., Fernández, A., and Bellido, J. M. 2016.
921 Fishery-dependent and -independent data lead to consistent estimations of essential
922 habitats. *ICES Journal of Marine Science*, 73: 2302–2310. Oxford University Press.
- 923 Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa,
924 D. 2019. Accounting for preferential sampling in species distribution models.
925 *Ecology and evolution*, 9: 653–663.
- 926 Peterson, C. D., Courtney, D. L., Cortés, E., and Latour, R. J. 2021. Reconciling conflicting
927 survey indices of abundance prior to stock assessment. *ICES Journal of Marine
928 Science*, 78: 3101–3120.
- 929 Pinto, C., Travers-Trolet, M., Macdonald, J. I., Rivot, E., and Vermard, Y. 2019. Combining
930 multiple data sets to unravel the spatiotemporal dynamics of a data-limited fish
931 stock. *Canadian Journal of Fisheries and Aquatic Sciences*, 76: 1338–1349. NRC
932 Research Press.
- 933 Planque, B., Loots, C., Petitgas, P., LINDSTRØM, U. L. F., and Vaz, S. 2011. Understanding what
934 controls the spatial distribution of fish populations using a multi-model approach.
935 *Fisheries Oceanography*, 20: 1–17.
- 936 Punt, A. E. 2017. Some insights into data weighting in integrated stock assessments.
937 *Fisheries Research*, 192: 52–65.
- 938 Punt, A. E., Dunn, A., Elvarsson, B. Þ., Hampton, J., Hoyle, S. D., Maunder, M. N., Methot, R. D.,
939 *et al.* 2020. Essential features of the next-generation integrated fisheries stock
940 assessment package: A perspective. *Fisheries Research*, 229: 105617.
- 941 R Core Team. 2020. R: A language and environment for statistical computing. R Foundation
942 for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- 943 Rivoirard, J., Simmonds, J., Foote, K. G., Fernandes, P., and Bez, N. 2008. Geostatistics for
944 estimating fish abundance. John Wiley & Sons.
- 945 Rochette, S., Rivot, E., Morin, J., Mackinson, S., Riou, P., and Le Pape, O. 2010. Effect of
946 nursery habitat degradation on flatfish population: Application to *Solea solea* in the
947 Eastern Channel (Western Europe). *Journal of sea Research*, 64: 34–44. Elsevier.
- 948 Rufener, M.-C., Kristensen, K., Nielsen, J. R., and Bastardie, F. 2021. Bridging the gap between
949 commercial fisheries and survey data to model the spatiotemporal dynamics of
950 marine species. *Ecological Applications*: e02453.
- 951 Salas, S., and Gaertner, D. 2004. The behavioural dynamics of fishers: management
952 implications. *Fish and Fisheries*, 5: 153–167.
- 953 Saunders, S. P., Farr, M. T., Wright, A. D., Bahlai, C. A., Ribeiro Jr., J. W., Rossman, S., Sussman,
954 A. L., *et al.* 2019. Disentangling data discrepancies with integrated population
955 models. *Ecology*, 100: e02714.
- 956 Schaub, M., and Abadi, F. 2011. Integrated population models: a novel analysis framework
957 for deeper insights into population dynamics. *Journal of Ornithology*, 152: 227–237.
958 Springer.
- 959 Schmitt, R. A. 1999. Essential fish habitat: opportunities and challenges for the next
960 millennium. *In American Fisheries Society Symposium*, p. 10.

961 Stephens, A., and MacCall, A. 2004. A multispecies approach to subsetting logbook data for
962 purposes of estimating CPUE. *Fisheries Research*, 70: 299–310.

963 Thorson, J. T., and Ward, E. J. 2014. Accounting for vessel effects when standardizing catch
964 rates from cooperative surveys. *Fisheries Research*, 155: 168–176.

965 Thorson, J. T., Ianelli, J. N., Munch, S. B., Ono, K., and Spencer, P. D. 2015a. Spatial delay-
966 difference models for estimating spatiotemporal variation in juvenile production and
967 population abundance. *Canadian journal of fisheries and aquatic sciences*, 72: 1897–
968 1915.

969 Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K.
970 2015b. Spatial factor analysis: a new tool for estimating joint species distributions
971 and correlations in species range. *Methods in Ecology and Evolution*, 6: 627–637.
972 Wiley Online Library.

973 Thorson, J. T. 2015. Spatio-temporal variation in fish condition is not consistently explained
974 by density, temperature, or season for California Current groundfishes. *Marine
975 Ecology Progress Series*, 526: 101–112.

976 Thorson, J. T., Fonner, R., Haltuch, M. A., Ono, K., and Winker, H. 2016. Accounting for
977 spatiotemporal variation and fisher targeting when estimating abundance from
978 multispecies fishery data. *Canadian Journal of Fisheries and Aquatic Sciences*, 74:
979 1794–1807.

980 Thorson, J. T., Jannot, J., and Somers, K. 2017a. Using spatio-temporal models of population
981 growth and movement to monitor overlap between human impacts and fish
982 populations. *Journal of Applied Ecology*, 54: 577–587.

983 Thorson, J. T., Johnson, K. F., Methot, R. D., and Taylor, I. G. 2017b. Model-based estimates of
984 effective sample size in stock assessment models using the Dirichlet-multinomial
985 distribution. *Fisheries Research*, 192: 84–93.

986 Thorson, J. T. 2018. Three problems with the conventional delta-model for biomass
987 sampling data, and a computationally efficient alternative. *Canadian Journal of
988 Fisheries and Aquatic Sciences*, 75: 1369–1382. NRC Research Press.

989 Thorson, J. T., Adams, G., and Holsman, K. 2019. Spatio-temporal models of intermediate
990 complexity for ecosystem assessments: A new tool for spatial fisheries management.
991 *Fish and Fisheries*, 20: 1083–1099.

992 Thorson, J. T., Cunningham, C. J., Jorgensen, E., Havron, A., Hulson, P.-J. F., Monnahan, C. C.,
993 and von Szalay, P. 2021. The surprising sensitivity of index scale to delta-model
994 assumptions: Recommendations for model-based index standardization. *Fisheries
995 Research*, 233: 105745.

996 Trenkel, V. M., Beecham, J. A., Blanchard, J. L., Edwards, C. T., and Lorange, P. 2013. Testing
997 CPUE-derived spatial occupancy as an indicator for stock abundance: application to
998 deep-sea stocks. *Aquatic living resources*, 26: 319–332. EDP Sciences.

999 Winker, H., Kerwath, S. E., and Attwood, C. G. 2013. Comparison of two approaches to
1000 standardize catch-per-unit-effort for targeting behaviour in a multispecies hand-line
1001 fishery. *Fisheries Research*, 139: 118–131.

1002 Witman, J. D., and Roy, K. 2009. *Marine Macroecology*. University of Chicago Press. 442 pp.

1003 Zipkin, E. F., Inouye, B. D., and Beissinger, S. R. 2019. Innovations in data integration for
1004 modeling populations. *Ecology*, 100: e02713.

1005