ICES Journal of Marine Science, 2023, **00**, 1–12
DOI: 10.1093/icesjms/fsad069
Original Article

Combining fishery data through integrated species distribution models

Iosu Paradinas [1,2,*], Janine B. Illian [3], Alexandre Alonso-Fernández [4], Maria Grazia Pennino [5], and Sophie Smout [1]

[1]Scottish Oceans Institute. University of St Andrews, East Sands, St Andrews, KY16 8LB, UK
[2]AZTI, Txatxarramendi Ugartea z/g, 48395 Sukarrieta, Bizkaia, Spain
[3]School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK
[4]Instituto de Investigaciones Marinas (IIM-CSIC), Eduardo Cabello 6, 36208 Vigo, Pontevedra, Spain
[5]Instituto Español de Oceanografía (IEO-CSIC), Centro Oceanográfico de Vigo, Subida a Radio Faro 50-52, 36390 Vigo, Pontevedra, Spain
[*] *Corresponding author: tel: +34 611044315;* e-mail: paradinas.iosu@gmail.com.

Species Distribution Models are pivotal for fisheries management. There has been an increasing number of fishery data sources available, making data integration an attractive way to improve model predictions. A wide range of methods have been applied to integrate different datasets in different disciplines. We focus on the use of Integrated Species Distribution Models (ISDMs) due to their capacity to formally accommodate different types of data and scale proportional gear efficiencies. ISDMs use joint modelling to integrate information from different data sources to improve parameter estimation by fitting shared environmental, temporal and spatial effects. We illustrate this method first using a simulated example, and then apply it to a case study that combines data coming from a fishery-independent trawl survey and a fishery-dependent trammel net observations on Solea solea. We explore the sensitivity of model outputs to several weightings for the commercial data and also compare integrated model results with ensemble modelling to combine population trends in the case study. We obtain similar results but discuss that ensemble modelling requires both response variables and link functions to be the same across models. We conclude by discussing the flexibility and requirements of ISDMs to formally combine different fishery datasets.

Keywords: essential fish habitat, fish distribution modelling, fisheries management, integrated species distribution modelling, spatial modelling.

Introduction

Species Distribution Models (SDM) use spatially georeferenced data to quantify the relationship between species occurrence or abundance with biotic and abiotic factors in order to gain ecological and evolutionary insight (Elith and Leathwick, 2009). Fish SDMs play a key role at identifying essential fish habitats (EFHs) (Laman *et al.*, 2018; Paradinas *et al.*, 2020; Tolimieri *et al.*, 2020) and producing unbiased population trends (Maunder and Punt, 2004; Thorson and Ward, 2013; Thorson *et al.*, 2015) that may later be used for stock assessment.

Fishery-independent (FI) surveys collect high-quality data that are often used to estimate population trends but have limited coverage in space and time due to their high economic cost. The establishment of on-board sampling programs has produced extensive new fishery-dependent (FD) spatial data that could complement the spatial and temporal coverage of FI data (Pennino *et al.*, 2016; Paradinas *et al.*, 2021; Rufener *et al.*, 2021). FD data are generally regarded as lower quality data since sampling locations are not randomly selected, species are often identified to a higher taxonomic level, and fish are caught using different commercial vessels and different gear types. The integration of FI and FD data may be a way to improve our understanding and prediction of EFHs and population trends, but integrating different fish distribution data may not be straightforward.

In isolation, each dataset provides estimates of species' relative abundance and can provide a characterization of their habitat and spatial distribution. However, when we seek to integrate two or more datasets, we may need to combine different species-at-length catchability coefficients (gear efficiency in what follows) and different types of data (e.g. biomass, abundance, and occurrence). A number of SDM studies have proposed different approaches to integrate data (Fletcher *et al.*, 2019; Alglave *et al.*, 2022).

Data pooling

The most simple approach pools data together without explicitly considering the different sources and their specific sampling issues. Data pooling is depicted on the left panel of Figure 1. This method assumes that the nature of the response variable is the same across all sources. If the two data sources are not of the same type, one of the datasets needs to be transformed, which may result in some loss of information (Isaac *et al.*, 2020) (e.g. degrading abundance data to presence–absence or presence only data).

Ensemble modelling

The middle panel of Figure 1 illustrates the use of ensemble modelling to combine predictions based on FI and FD data sets. Independent models are fitted for each dataset and predictions are then combined. This procedure is frequently applied

Received: 1 August 2022; **Revised:** 1 April 2023; **Accepted:** 3 April 2023

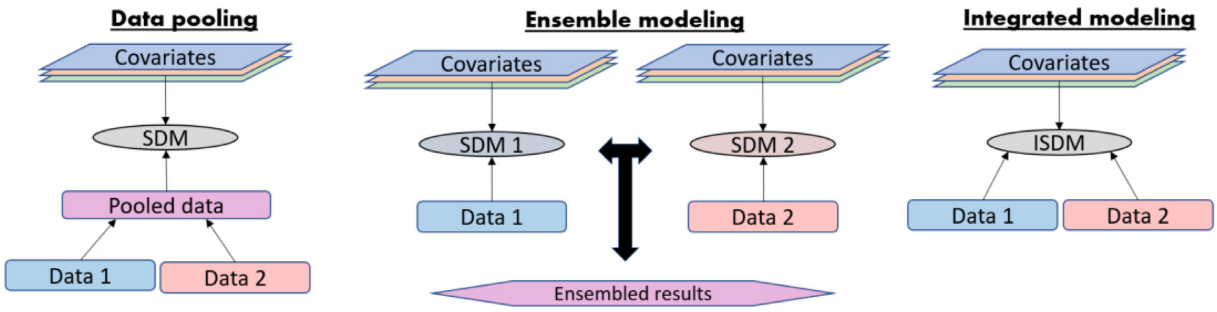

Figure 1. Three major approaches to combining different datasets. Data pooling occurs when different data sources are combined and a single model is fit. Ensembled modelling fits separate models to each dataset and results are then combined. ISDMs use joint modelling to integrate different datasets by explicitly accounting for the differences in the sampling process.

in order to combine predictions or fitted effects from different modelling algorithms (Araújo and New, 2007). However, it may be difficult to formally combine parameter estimates unless both datasets are sampled at a similar spatial resolution (Fletcher *et al.*, 2019) and follow the same type of response variable (and link functions if we were to combine model effects, e.g. temporal trends).

Integrated modelling

The right panel of Figure 1 illustrates integrated modelling. It refers to modelling different datasets together by explicitly describing the differences in the sampling process. The strength of ISDMs lies in combining information from different datasets to estimate shared parameters across models using joint-likelihood procedures. In other words, it allows combining different data sources, each using a different response variable, while estimating the common model parameters. This allows a better informed parameter estimation than using individual models and, as opposed to data pooling and ensembled modelling, allows a formal combination of different types of data (i.e. response variables) and link functions.

This article focuses on the use of ISDMs for integrating fishery data in order to accommodate different gear efficiencies and types of response variables. The "Integrating fishery data through ISDMs" section describes the use of ISDMs to integrate different datasets, sampling methods and types of data together. The "Data weighting" section follows by describing data weighting and the reasons for using this approach. Then the "Simulation study" section presents a simulation study that tests the performance of ISDMs to integrate three different data sources that collect different types of data. Next, the "Case study" section presents a common sole (*Solea solea*) case study in northern Iberian Atlantic waters, followed by a "Results" section and we finish by providing a discussion in the "Discussion" section.

Integrating fishery data through ISDMs

Integrating different fish distribution data implies combining data collected using different gears, and therefore different gear efficiencies. Gear efficiency determines the fraction of the actual number of fish in the ocean that the gear will fish, for a given fish species and length. Generally gear efficiency constitutes a non-linear function [$f(s, l)$] that depends on the fish species caught (s) and the length of the fish (l) (Fraser *et al.*, 2007).

Gear efficiency is hard to quantify (Zhou *et al.*, 2014) and is most often unknown, which makes the integration of fish distribution data a difficult task. However, certain gears could be assumed to have proportional gear efficiency curves, i.e. their quotient at different fish-lengths is a constant value. In such situations, ISDMs provide a way to formally integrate different types of data. Assuming that gear efficiencies are proportional across two different samplers,

$$f_1(s, l) = \gamma_s \times f_2(s, l), \tag{1}$$

the catch per unit effort (CPUE) is also proportional. Note as well that as we narrow the size group of the population under study, the proportionality assumption becomes more robust. For example, if we narrowed down the population study to single length, gear efficiencies should be proportional given that we observe the same species $f_1(s) = \gamma_s \times f_2(s)$.

ISDMs use joint-likelihood methods with which, by modelling the regression coefficients hierarchically, can estimate shared parameters across linear predictors (Hogan and Laird, 1997; Knorr-Held and Best, 2001; Paradinas *et al.*, 2017). In other words, it allows combining different data sources, each using a different response variable, while estimating common model parameters. Furthermore, ISDMs can estimate a scaling parameter that accounts for the difference in gear efficiency and/or types of data (e.g. presence–absence, abundance, biomass):

$$\begin{aligned} CPUE_1 &= \gamma_s \times \eta + \epsilon_2, \\ CPUE_2 &= \eta + \epsilon_1, \end{aligned} \tag{2}$$

where $\boldsymbol{\eta}$ refers to any linear predictor that may be suitable to describe the distribution of the species under study and is shared across predictors using joint modelling. γ_s is a scaling parameter that scales the linear predictors and ϵ refers to white noise. The scaling parameter is a fixed parameter that accommodates the differences in the sampling process and gear efficiency. Note that Equation 2 uses the same scaling parameter (γ_s) as Equation 1 to stress the link between the two equations.

Note as well that when using a log or logit link, γ_s could be removed and replaced by including a different intercept to each predictor to do the scaling. These links provide a multiplicative nature to the parameters of the linear predictor, and therefore the intercept performs as a multiplicative scaling parameter itself (Moriarty *et al.*, 2020). Assuming that two datasets observe the same fish population, one collecting abundance data and the other biomass data, are modelled using a Poisson and a gamma distribution respectively, with

parameters λ and μ through logarithmic links:

$$log(\lambda_1) = \beta_{0,1} + \boldsymbol{\beta} X,$$
$$log(\mu_2) = \beta_{0,2} + \boldsymbol{\beta} X, \quad (3)$$

also written as:

$$\lambda_1 = e^{\beta_{0,1}} \times e^{\boldsymbol{\beta} X},$$
$$\mu_2 = e^{\beta_{0,2}} \times e^{\boldsymbol{\beta} X}, \quad (4)$$

where $e^{\beta_{0,1}}$ and $e^{\beta_{0,2}}$ will scale $e^{\boldsymbol{\beta} X}$, which are shared environmental effects characterized in this case as linear. It is important to note that β is the same in both predictors, i.e. shared across the fish abundance and fish biomass predictors using joint modelling.

Data weighting

The joint log-likelihood of an ISDM is the sum of the component log-likelihoods:

$$Log(L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma_s})) = \sum_{d=1}^{D} Log(L_d(\boldsymbol{\beta}, \theta_d, \gamma_{s_{d-1}})), \quad (5)$$

where $\boldsymbol{\beta}$ are shared parameters across D different datasets, $\boldsymbol{\gamma_s}$ are the scaling parameters across linear predictors, and $\boldsymbol{\theta}$ are the parameters associated with the different observational processes. Note that, for identifiability reasons, γ_{s_0} is set to one.

Different data sources may vary in size and quality. Assuming equal weights to every sample may produce results dominated by the larger dataset because the joint log-likelihood function is additive (Equation 5). As a consequence, the larger dataset has a greater contribution to the likelihood, for example, given two datasets, one with 100 samples and the other with 50 samples, the first dataset will contribute two-thirds of the likelihood and the second dataset the remaining one-third. This may not always be desirable, thus Fletcher *et al.* (2019) proposed a weighted likelihood function to fit weighted ISDMs. Assuming the simplest scenario with two datasets:

$$Log(L(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma_s})) = w * Log(L_1(\beta,\theta_1)) + (w-1) * Log(L_2(\beta,\theta_2,\gamma_s)), \quad (6)$$

where $0 < w < 1$, such that $w = 0.5$ would be equivalent weighting between two datasets, which in the default weighting approach (i.e. equal weights) would be as if we had the same sample size in both datasets.

Data weighting is a common practice in stock assessment models. These are complex models that, as opposed to SDMs, require heterogeneous sources of data (compositional data, catch data, survey indexes, etc.) to produce the outputs. These data sources may be inconsistent with one another, thus scientists downweight the influence of certain data source(s) to constrain their influence in the estimates (Punt, 2015; Francis, 2017; Thorson *et al.*, 2017).

In the contrary, SDMs do not necessarily have to integrate heterogeneous sources of information. In fact, an ISDM would generally respond to the idea of complementing a main data source by other accessory source(s) of data that sample the same population. If the information contained in the accessory data clash with the main dataset, it could suggest that they sample a different process (e.g.,different fractions of the population) or are biased. Different authors agree that sensitivity of results to data weighting is driven by model misspecification (Thorson *et al.*, 2017; Wang and Maunder, 2017), thus ideally results should be invariant to it. Similarly, one may think that results could be sensitive to data quality. However, if data quality is assessed in terms of variance (i.e. the observational process) and not bias, such variance would be absorbed by the dispersion parameter of the probability distributions used in the ISDM and results should remain invariant to data weighting.

In this regard, testing different data weightings may be useful to check whether the different data sources incorporated in the ISDM clash (i.e. outputs sensitive to data weighting), suggesting that they sample different process. In such cases, one should probably not integrate these data sources into an ISDM as it may negatively impact the estimates.

Simulation study

We used a simulation study to test the suitability of ISDMs to integrate different types of data collected over partly overlapping spatial and covariate space of a common target population.

We simulated the distribution of a species over a 100-by-100 grid that was driven by a spatial covariate and an unidentified spatial pattern that displayed different hot-spots in the study area (as seen in the top panels of Figure 2). According to ecological niche theory, the simulated species displayed a non-linear unimodal relationship with the covariate (Hutchinson, 1957).

We then simulated three, partly overlapping, fish distribution datasets, each recording a different type of data: biomass through a gamma distribution; abundance through a Poisson distribution; and presence–absence through a Bernoulli distribution. None of the simulated datasets covered the whole study area, and therefore sampled only fractions of the ecological niche (see Figure 2).

In order to test the performance of ISDMs, we fitted SDMs to each simulated dataset independently, as well as an ISDM to all three datasets together and compared the results. The models included a non-linear fish-covariate relationship (e.g. depth) and a geostatistical effect. All the R code used to perform the simulation study is available in the Supplementary Material.

Case study

Data

This study integrated 529 FI and 1522 FD *Solea solea* (Linnaeus, 1758), abundance samples (i.e. number of specimens) collected between 2013 and 2018 using trawl nets (our FI data) and trammel nets (our FD data) in northern Iberian Atlantic waters (Figure 3).

Common sole is a widely distributed and highly priced flatfish targeted by multi-species and multi-gear fleets (Alonso-Fernández *et al.*, 2019; Pennino *et al.*, 2022b, a). This species' bathymetric niche ranges between 0 and 200 metres in Iberian Atlantic waters (Tanner *et al.*, 2012). The annual trawl scientific survey (i.e. FI data) that collects data on demersal species in this area was not designed to sample common sole, and therefore only samples the deeper part of its bathymetric range, likely resulting in a biased biomass index trend. In contrast, regional FD data provide samples in the shallower part

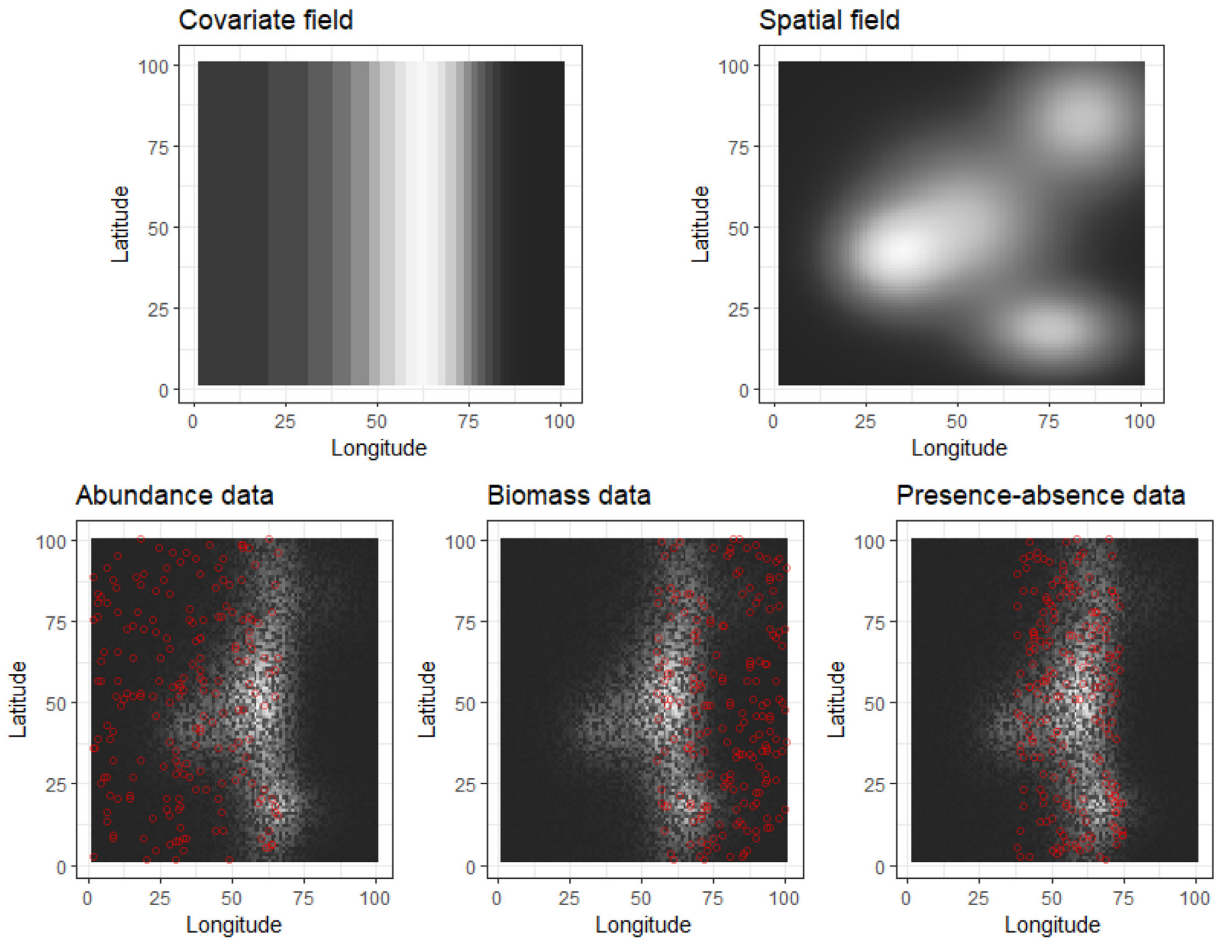

Figure 2. Visualization of the different drivers that affect the simulated process in the top panels, and the resulting process along with the spatial distribution of the samples for the different datasets represented in the bottom panels. Red dots represent the sampling locations of each survey.

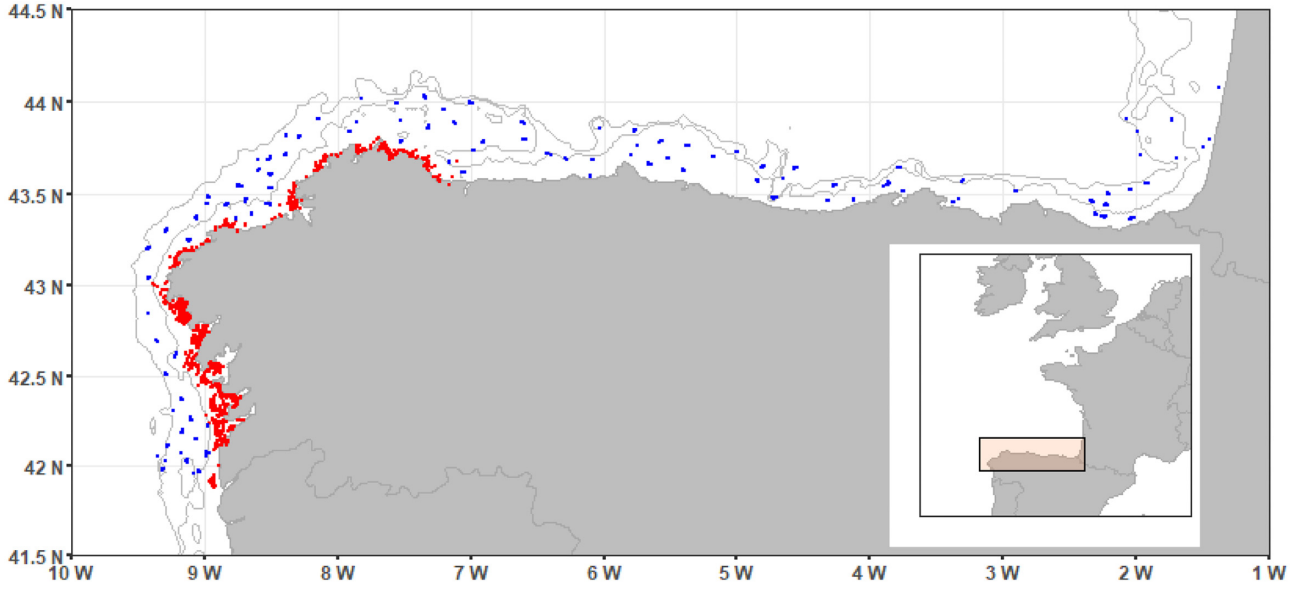

Figure 3. Study area with FD sampling locations (red dots) and FI sampling locations (blue dots). Bathymetric lines indicate the 100 and 200 metre isobaths.

of its bathymetric range, thus by integrating both datasets, we are able to cover the whole bathymetric niche of common sole.

FI data were collected during the scientific survey series SP-NSGFS, which is performed every year in Iberian Atlantic waters between September and November. Sample locations were selected using a stratified sampling design based on three bathymetric strata: 70–120 m, 121–200 m and 201–500 m. The sampling consisted of 30 min trawling hauls, performing

~115 hauls each year. However, this study only uses the first two bathymetric strata (i.e. 70–120 m, 121–200 m) because common sole does not occur at deeper waters. The dataset also includes 20 samples collected randomly at depths of <70 m. FI data were distributed throughout the study area.

FD data were collected in Galician waters (NW Spain) using on-board observers. Fishing vessels were selected randomly covering the western part of the study area and a wide range of fishing gears throughout the year. Fishing vessels usually perform more than one haul per trip, and at each haul, observers record all basic operational data (i.e. date, position, gear, etc.), the number and weight of all retained and discarded taxa and environmental variables such as the bathymetry and the predominant type of substratum in the haul. We selected trammel net data, which is the better sampler of common sole (Alonso-Fernández *et al.*, 2019) and its gear efficiency should *a priori* be proportional to that of the FI trawl gear. We had two potential effort variables: the number of nets deployed at each location and their soak time.

Response variable

Both sampling schemes observed count data, with a particularly high number of zero observations (i.e. 17 and 32% in FD and FI data, respectively) and some sporadic high abundances. We modelled these processes using a zero-inflated negative binomial (ZINB) model that assumes a positive relationship between the negative binomial (NB) process and the probability of being in the NB process. In other words, the probability of zero inflation is higher at lower abundances and vice-versa.

$$P(y(s)) = p + (1 - p) \times NB(y(s)), \quad s = 1, ..., n,$$

where $y(s)$ are the samples collected in location s, n is the number of samples, p is the probability of zero inflation, and $(1 - p)$ is the probability of being in the NB process, defined as:

$$(1 - p) = \left(\frac{\mu}{1 + \mu}\right)^{\kappa},$$

that depends on μ, i.e. the mean of the NB, and the hyperparameter κ, which shapes the relationship between the probability of presence in the zero-inflation and the mean abundance of the NB.

We selected this particular model for the observation process based on the belief that the occurrence and the abundance processes are linked (Paradinas *et al.*, 2015, 2021; Thorson, 2018). Alternatively, the modelling could also be conducted in two stages (i.e. occurrence and abundance) by fitting a hurdle or delta model (Maunder and Punt, 2004).

Modelling

We applied Bayesian hierarchical modelling using the integrated nested Laplace approximation (INLA) approach through the R-INLA package (Rue *et al.*, 2009). First, we fitted different models to each dataset and performed variable selection independently. The FD model considered non-linear relationships for soak time, bathymetry, month (cyclic effect), and year, as well as a categorical covariate for the type of substratum. The number of nets deployed was used as an offset in the FD model. The FI model only considered bathymetry and year given that soak time does not apply to trawlers, trawlers only sample muddy–sandy bottoms and that the FI survey is performed only once per year. Every FI haul performed the same nominal effort, thus we did not include any offset in it. Both FI and FD models included a geostatistical effect to account for spatial autocorrelation fitted using the stochastic partial differential equations approach (SPDE) (Lindgren *et al.*, 2011).

Table 1. Model comparison of FD and FI SDMs was based on WAIC and LCPO scores.

Data	Linear predictor	WAIC	LCPO
Fishery independent	B	1 223.72	1.16
	B + T	1 222.91	1.16
	B + W	1 052.54	1.00
	B + T + W	**1 047.70**	**0.99**
Fishery dependent	B + TS	2 071.42	0.68
	B + TS + ST	2 085.55	0.69
	B + TS + M	2 071.34	0.68
	B + TS + T	2 047.42	0.67
	B + TS + W	1 837.75	0.62
	B + TS + T + M	2 047.41	0.67
	B + TS + T + W	**1 827.59**	**0.61**
	B + TS + M + W	1 837.69	0.63
	B + TS + T + M + W	2 047.41	0.67

B refers to non-linear bathymetric effects, *W* refers to geostatistical effects, *T* refers to yearly trend effects, *TS* refers to the type of substratum categorical effect, *M* refers to a cyclic non-linear effect for month, and *ST* refers to soak time. Best models are highlighted in bold.

Finally, based on the individual models selected in the previous step (Table 1), ISDMs included joint bathymetric, spatial, and yearly trend effects, as well as the type of substratum effect applied over the FD data. We used logarithmic links in the modelling, thus intercepts scaled the difference in gear efficiency as mentioned in the end of Section 2.

Let $\mathbf{s} \in \mathbf{S}$ be a location in the study area $\mathbf{S}$, then we have that

$$\begin{aligned} \eta_{\mathrm{FI}}(\mathbf{s}) &= \beta_{0,\mathrm{FI}} + f_D(\tau_D) + f_T(\tau_T) + u(\sigma, r), \\ \eta_{\mathrm{FD}}(\mathbf{s}) &= \beta_{0,\mathrm{FD}} + f_D(\tau_D) + f_T(\tau_T) + u(\sigma, r) + \boldsymbol{\beta}_{\mathrm{TS}}, \end{aligned} \quad (7)$$

where $\eta_{\mathrm{FI}}(\mathbf{s})$ and $\eta_{\mathrm{FD}}(\mathbf{s})$ correspond to the linear predictors for the FI and FD data, respectively. The coefficients $\beta_{0,\mathrm{FI}}$ and $\beta_{0,\mathrm{FD}}$ are intercepts and $\boldsymbol{\beta}_{\mathrm{TS}}$ are fixed effects that quantify the type of substratum effect over the FD data. The functions f_D and f_T are shared random effects over the bathymetry and year, defined as a Gaussian random walks of second order with precisions $\tau_D > 0$ and $\tau_T > 0$. The function u is a shared SPDE object across FI and FD predictors, with standard deviation (σ) and range (r), fitted over a two-dimensional triangulation of the study area (a.k.a. mesh).

The Bayesian approach requires that every hyperparameter of the model is given a prior distribution. Due to the lack of prior knowledge about the behaviour of the fixed effects and intercepts, we chose non-informative priors, set as zero-mean Gaussian priors with a precision of 0.001 for all fixed effects and precision equal to 0 for intercepts (i.e. improper prior). For the remaining parameters, we used Penalized Complexity priors (PC priors; Simpson *et al.*, 2017). The PC prior for σ was set so that the probability of it being bigger than 1 was 0.2 [$\Pr(\sigma > 1 = 0.2)$]. The PC prior for r was set so that its median probability was 80 km [$\Pr(r < 100 = 0.5)$]. The PC priors for τ_D and τ_T were set as uninformative as possible by setting the probability of the precision being smaller than the empirical standard deviation of the FI data is 0.01 [$\Pr(\tau < sd(FI) = 0.01\ )$]

Given that FI data are generally better quality data than FD data, we fitted four ISDMs with different data weightings

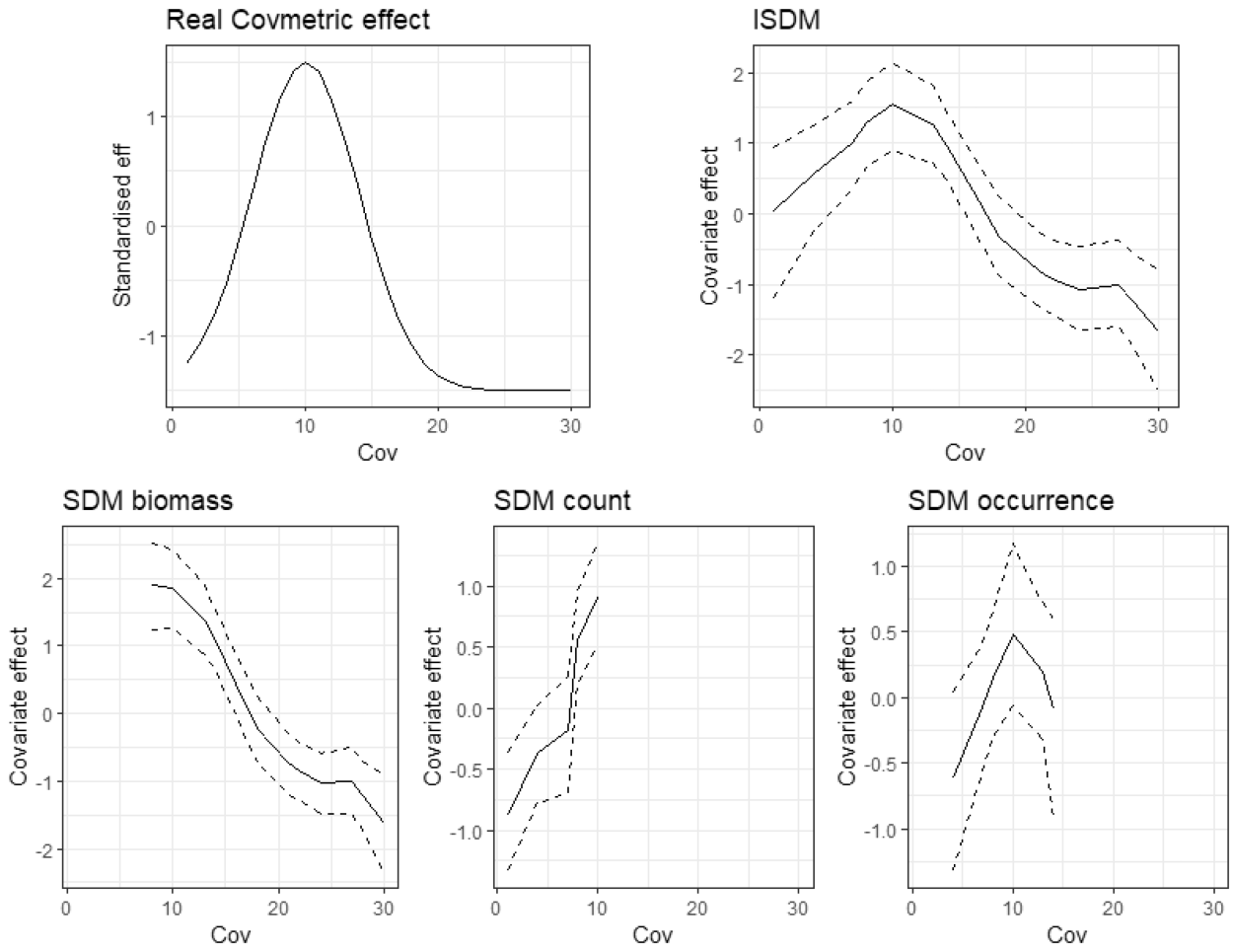

Figure 4. Simulated covariate effect (row 1, left panel) and fitted covariate effects (lower panels) using separate SDMs for each dataset (row 2), or combining all three datasets through ISDMs (row 1, right panel). Solid line represents the mean effect while dashed lines represent credibility intervals.

to assess its impact in the results. One ISDM used a default weighting of equal weights to every sample regardless of the source. Another ISDM assigned the same overall weight to each dataset so that the contribution of each dataset was the same ($w_{FI} = w_{FD}$). The last two weightings, favoured the scientific survey data by assigning the FD dataset a half ($0.5{*}w_{FI} = w_{FD}$) and a quarter ($0.25{*}w_{FI} = w_{FD}$) of the weight of the FI dataset.

Finally, we also used ensemble modelling to estimate joint temporal trend effects based on the independent FI and FD models. This case study was particularly adequate to do so given that both SDMs modelled abundance and used the same link function. The ensembling was performed using the same weights as the ISDMs.

Results

Simulation study

Each simulated survey by itself did not cover the whole covariate range of the species, thus single survey SDMs fitted incomplete process-covariate relationships Figure 4. The ISDM integrated all three datasets to fit a much better characterization of the covariate niche (top right panel in Figure 4). The same occurred with the spatial effect, where each simulated survey did not cover the whole spatial range of the species and therefore single survey ISDMs fitted incomplete spatial effects as compared to the ISDM Figure 5.

Common sole case study

Model selection was performed over the FI and FD data independently, based on the Watanabe Akaike's information criterion (WAIC) (Watanabe, 2010) and the log-conditional predictive ordinate scores (LCPO) (Gneiting and Raftery, 2007) to do so (Table 1). FI and FD SDM results display significant differences in fitted effects. The bathymetry effect (Panels a and b in Figure 6) is clearly the most different of all effects given that FD and FI datasets sampled different bathymetric strata. However, both display a maximum around 50 m deep. Spatial effect differences are driven by the extent and resolution at which FD and FI datasetss sampled the study area (Panels a and b in Figure 7). The FD spatial effect shows smaller high and low density spots than the FI spatial effect due to the smaller estimated range in the FD model (mean FD range = 23 km, mean FI range = 45 km), which may be driven by the higher sampling intensity of the FD data. Despite the difference in range, hot and cold spot locations are consistent across models, suggesting that both datasets observe the same spatial pattern. Finally, temporal trends (Panels a and b in Figure 8) are quite similar both displaying an overall decreasing trend from 2013 to 2018 despite a peak in 2017.

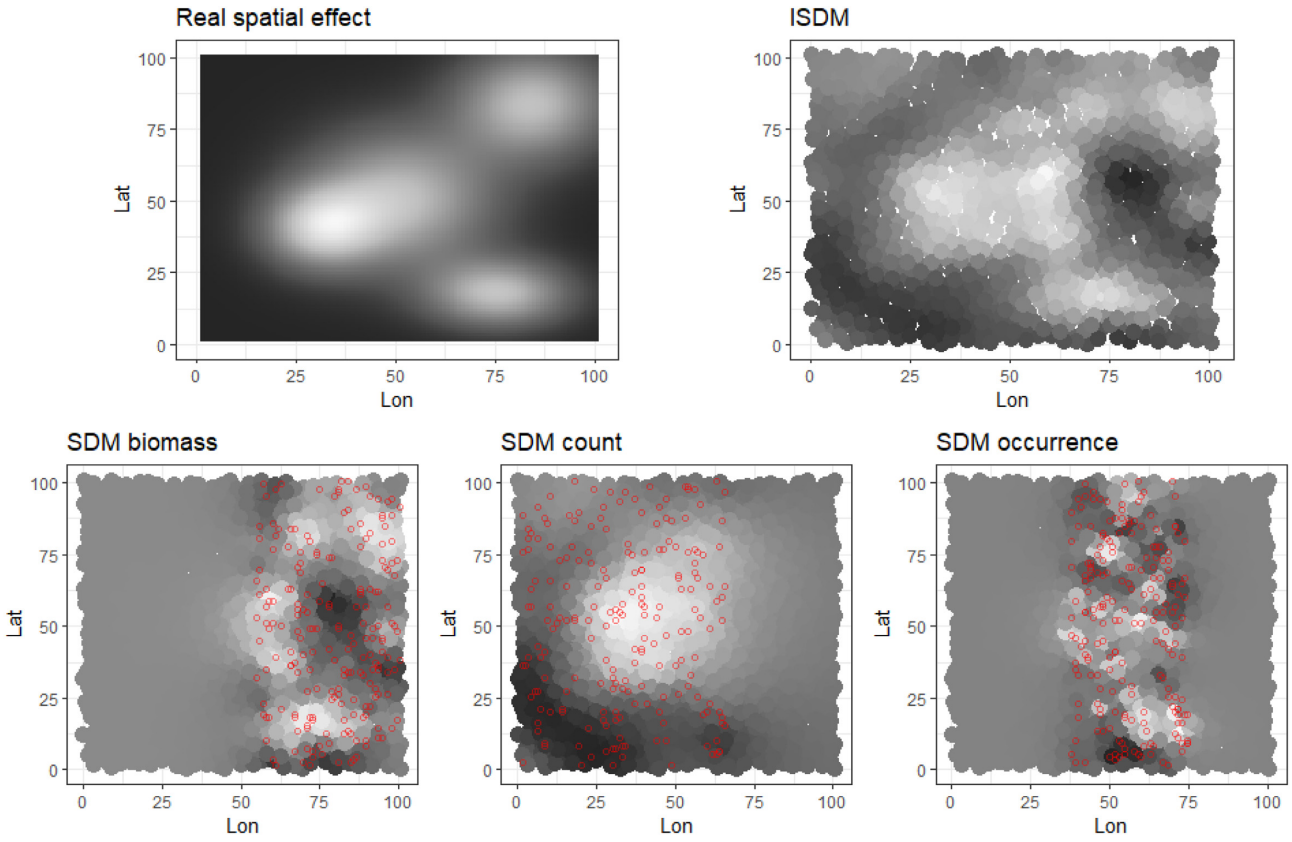

Figure 5. Simulated spatial effect (row 1, left panel) and fitted spatial effects using separate SDMs for each dataset (row 2), or combining all three datasets through ISDMs (row 1, right panel).

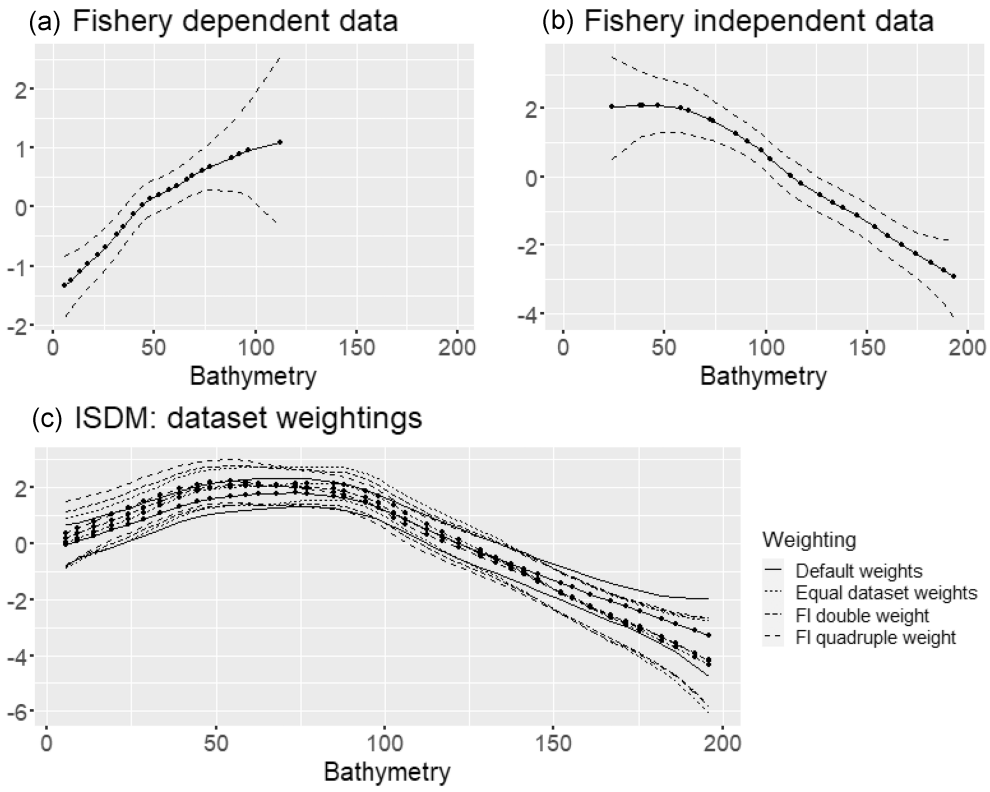

Figure 6. Visualization of the fitted bathymetric effects by the different models. Panels a and b represent modelled FD and FI SDM bathymetric effects, respectively. Panel c shows ISDM bathymetric effects using different weights for the FD and FI data: default equal weight per data point, equal weight per dataset, double weight assigned to the FI dataset, and quadruple weight assigned to the FI dataset.

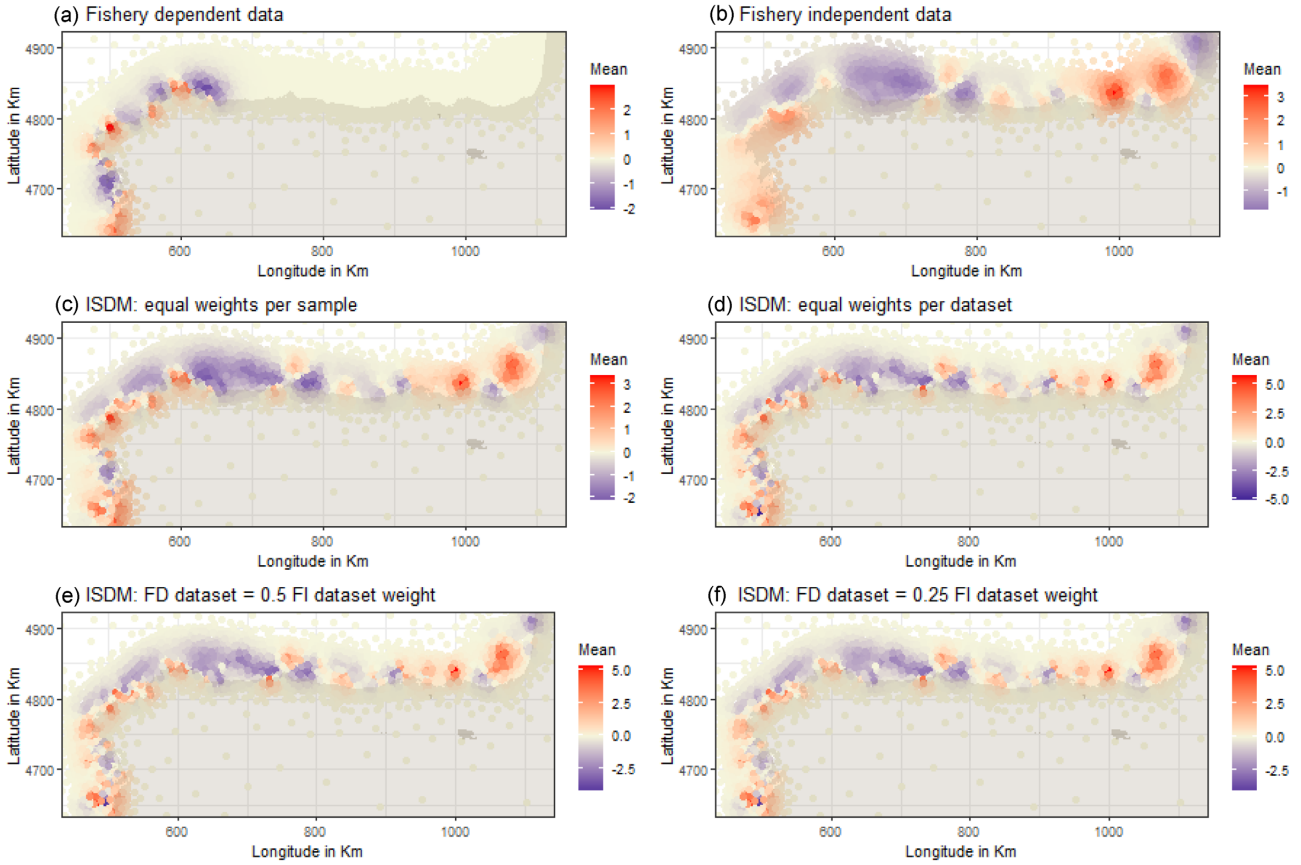

Figure 7. Visualization of the fitted spatial effects by the different models. Panels a and b represent FD and FI SDM spatial effects, respectively. Panels c to f show ISDM spatial effects using different weights: default equal weight per data point, equal weight per dataset, double weight assigned to the FI dataset, and quadruple weight assigned to the FI dataset.

ISDMs produced shared bathymetric, spatial, and temporal trend effects. Different weightings of the likelihood produced very consistent results suggesting that both FD and FI datasets sample the same fraction of the target population despite using different gears (i.e. proportional catchability between gears). The joint bathymetric effects integrated both FD and FI data to produce a much better characterization of common sole's bathymetric niche, with a maximum between 40 and 100 m deep (Panels c to f in Figure 6). Joint spatial effects produced hot spot sizes similar to that of the FD model (Panels c to f in Figure 7). Estimated spatial range parameter estimates vary between 27 and 18 km, which is coherent with the higher resolution of the data obtained by integrating both datasets. While different weightings estimated similar spatial effects, there were some minor localized differences (e.g. south-westernmost part of the study area). Lastly, ISDM fitted temporal trends display narrower credibility intervals than those obtained by the FI and FD SDMs. Results were very similar across weightings (Panels c to f in Figure 8), but as with the spatial effect, there were some minor differences (e.g. transitions between 2013–2014 and 2015–2016). Ensembled temporal trend results (Panels g to j in Figure 8) displayed a slightly smaller range in mean estimates, and credibility intervals were a little bigger but overall results showed remarkably similar patterns to those fitted using ISDMs.

We predicted the distribution of common sole in the western part of the study area for year 2016 (*n* Figure 9). Maps display the consequences of the aforementioned bathymetric and spatial effects in the ISDM, FD, and FI models. FD and FI model predictions were constraint by their limited bathymetric and spatial coverage, while ISDMs produced more coherent maps. We highlighted a few areas in Figure 9 to visualize the differences between the ISDM, FD, and FI models: area A, shows a coastal hot-spot in the FI model, driven by the hot-spot inferred in the spatial effect and the fitted positive bathymetric effect in shallow waters. In the contrary, the FD model predicts low abundance given that there were no observations in the area, and its bathymetric effect expects low abundances in shallow waters. The ISDM, however, produces more sensible predictions with high abundances constrained to a narrow band corresponding to approximately the 50–100 m deep range, i.e. the optimum bathymetric range fitted by the ISDM; in area B, FI data only sampled deeper waters where low specimen counts were observed and therefore predicted relatively low abundances. In the contrary, FD data collected moderately high numbers of specimens in the shallower waters, and consequently, the spatial effect inferred a small hot-spot that produced high mean predictions at its deeper part driven by the incomplete bathymetric effect fitted by this model. The better informed effects of the ISDM predicted a more sensibly defined hot-spot in the area; in area C, both FI and FD models inferred positive spatial effects (Panel a in Figure 5). FD model predictions in Figure 9 display a hot-spot in the deeper waters driven by its incomplete bathymetric effect. The ISDM, once again, predicts a more sensibly shaped hot-spot in the area.

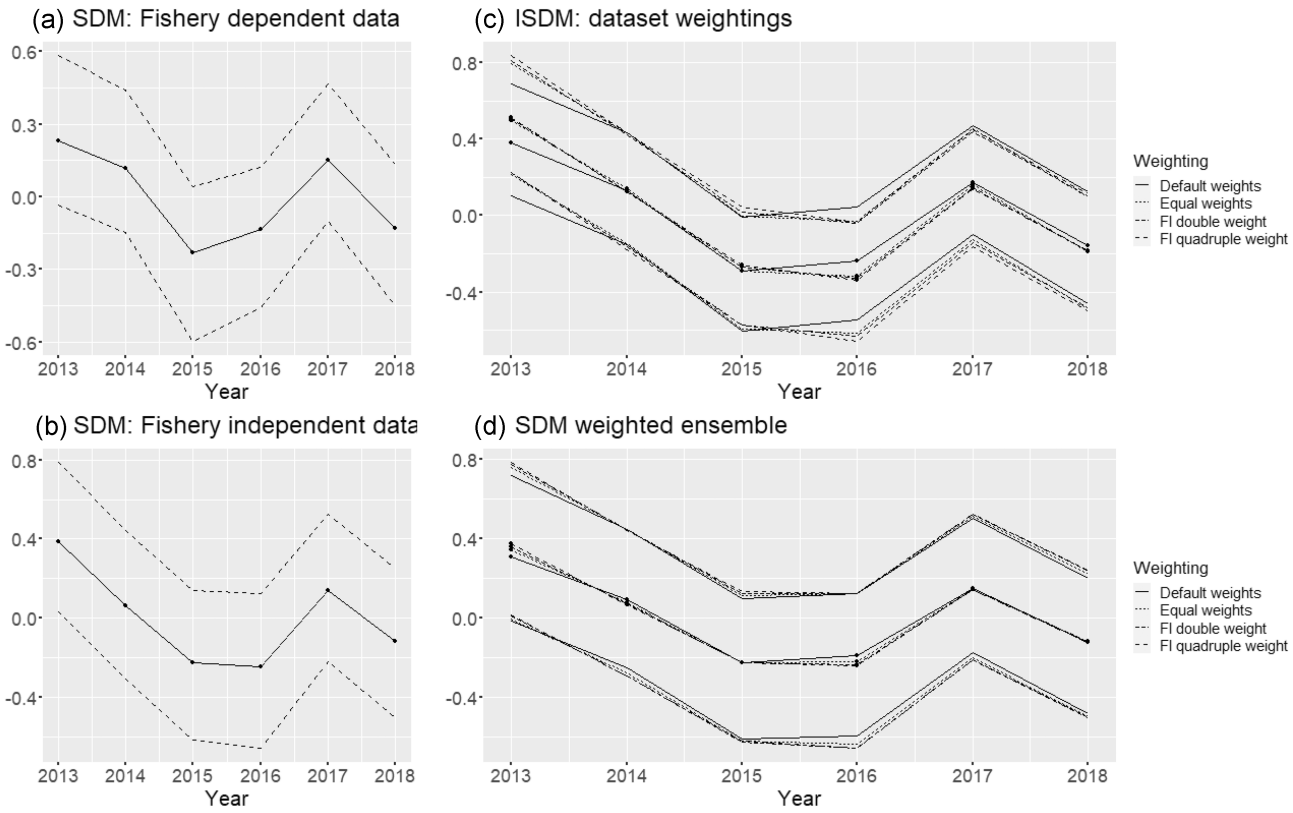

Figure 8. Visualization of the fitted temporal trend effects, mean and 95% *CI*, by the different models. Panels A and B represent fishery dependent and fishery independent SDM fits, respectively. Panel C shows ISDM model fit using different weights as described in each panel title, while panel d shows SDM ensemble model results using the same weights as for the ISDMs.

Discussion

Integrating different fishery datasets is an attractive approach to improve species distribution models. Different fishing gears have different species-specific gear efficiencies and different sampling schemes may collect different types of data (e.g. biomass, abundance, occurrence), which complicates the integration of different fish distribution data sources. Different Species Distribution Model studies have proposed different approaches to integrate datasets (Fletcher *et al.*, 2019), but only ISDMs allow a formal integration of different sampling processes, as well as types and scales of data.

ISDMs use joint-likelihood techniques to fit shared linear predictors by integrating observations from different types of data. If gear efficiencies are proportional across gears, we can accommodate the difference by scaling the linear predictors. The scaling approach will depend on the link function applied to the linear predictors. When using identity links, linear predictors should incorporate an extra scaling parameter to do the scaling. In contrast, the use of logarithmic and logit links imply multiplicative effects, and therefore intercepts perform as scalars. It is important to note that both intercepts and scaling parameters are fixed values and therefore require proportional gear efficiencies across gears. Non-proportional gear efficiencies would require modelling fish length class abundances and including a smooth term over length class for each fishing gear to capture the length-specific gear selectivity (Munro and Somerton, 2001; Fryer *et al.*, 2003; Miller, 2013; Gonzalez *et al.*, 2021).

We tested the use of ISDMs through a simulation study and a case study. The simulation study combined a Poisson, a gamma, and a Bernoulli likelihood to integrate different sampling processes and showed a better niche characterization of ISDMs than that obtained through individual SDMs. The case study integrated FD and FI data that used different fishing gears and sampled different habitats of the same common sole population in the northern Iberian Atlantic waters. As compared to individual SDMs, ISDMs clearly improved the characterization of the bathymetric niche, and estimated coherent joint temporal trend and spatial effects. Therefore, ISDMs also produced better prediction maps than individual SDMs.

Testing different data source weights may be a good data sensitivity analysis for ISDMs. By checking whether results change with different data weightings ,one can assess whether the model is well specified for the different data sources. If results are affected by data weighting, one would probably downweight the less reliable data source(s) or simply remove it from the model. In our case study, the FD dataset was larger than the FI, which could produce FD dominated results and *a priori* FD data are thought to be less reliable than FI data. Therefore, we tested different levels of downweighitng FD data, but results showed very consistent results across the different weightings. Such consistency may reflect that both datasets sample the same population, i.e. proportional gear efficiencies between the FI trawler and the FD trammel net, and that models are well specified for the data. If gear efficiencies were not proportional and /or one predictor was not well specified, we could have expected larger differences in fitted relationships across the different weightings (Francis, 2011; Thorson *et al.*, 2017; Wang and Maunder, 2017).

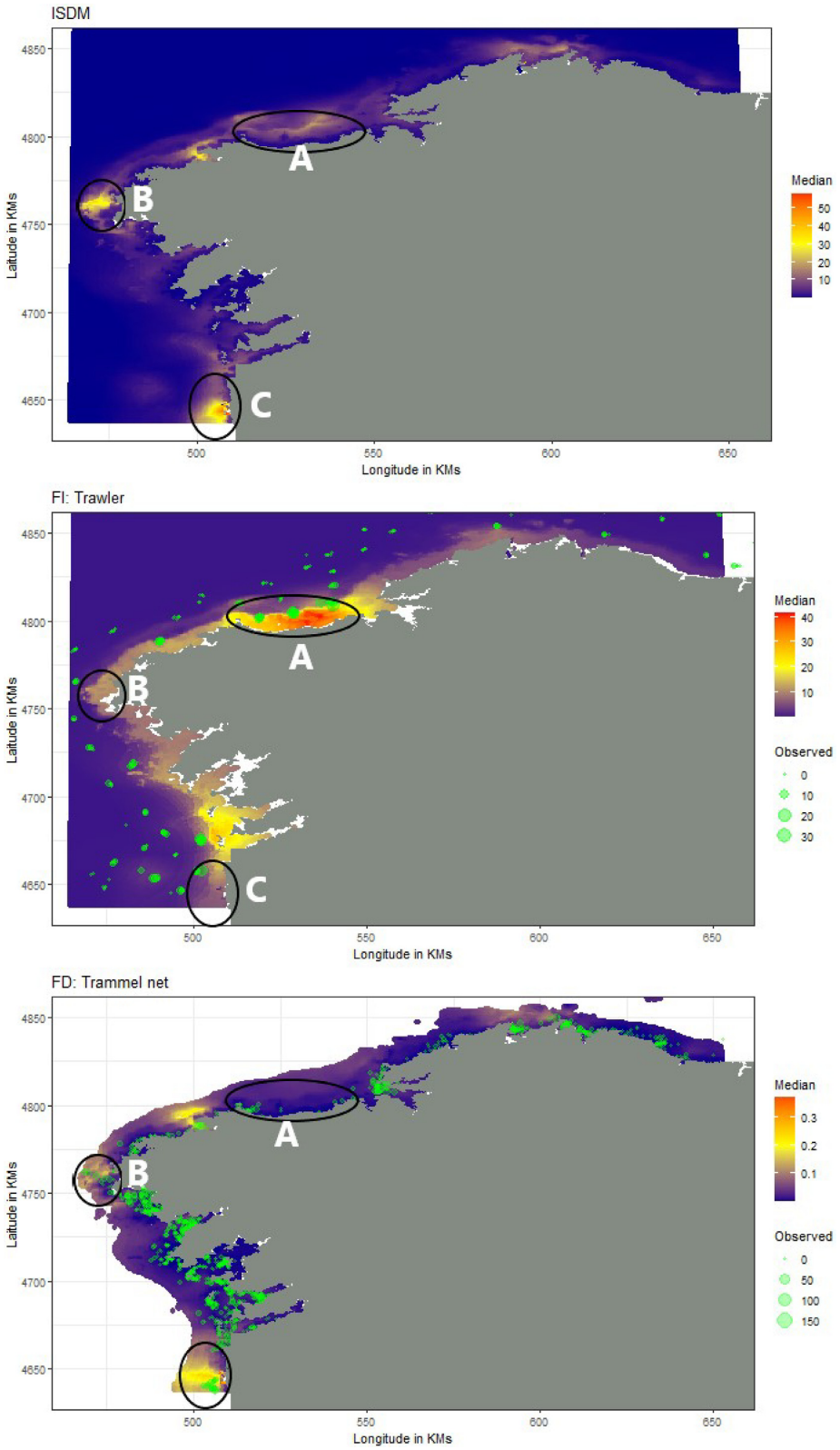

Figure 9. Prediction maps for the western study area produced using ISDM, FD, and FI models for year 2016. Green dots represent the location and abundance of the samples used in the FI and FD models. Three areas named as A, B, and C, have been highlighted to be commented in the text.

Temporal trends are particularly interesting to fisheries management (i.e. stock assessment models), thus, we paid special attention to them. Given that both sampling schemes observed common sole abundance data, and both predictors had the same link function, we compared ISDM temporal trends with SDM trends as well as ensembled temporal trends based on the FD and FI SDMs. Despite some minor differences in the range and credibility interval of yearly estimates, patterns

were remarkably similar across individual SDMs, ISDMs, and ensembled models. In particular, ensemble modelling and ISDMs produced almost identical results suggesting that ensembled modelling may be a good alternative when link functions and the types of data are the same. We stress, however, that ISDMs are more flexible and allow a formal integration of different types of data and/or link functions.

Last but not least, ISDMs provide an appropriate framework to improve estimates when a particular dataset is missing data. Missing data can be spatial (e.g. an area that could not be sampled due to bad weather), temporal (e.g. a particular year where there was no budget to perform the survey), or a covariate space gap such as hard bottom substrate in our FI data in the case study. We did not have a type of substrate map to perform our predictions, but if we had it, we could have used the inferred type of substratum effects using FD data to improve predictions over the entire study area.

This study provides an attractive approach to integrate different sources of data; however, the model proposed in this study could be expanded in several directions. First, one should expand the spatial model here presented into a spatio-temporal framework. Both the VAST (Thorson, 2019) and INLA packages for R may be specially useful in this task. Future extensions of ISDMs could also include preferential sampling approaches to account for the sampling bias of opportunistically collected data (Pennino *et al.*, 2018; Rufener *et al.*, 2021), model population size structures by partitioning the data into different fish length classes as discussed earlier.

Conclusions

ISDMs use joint likelihoods to fit shared linear predictors across different data sources. By increasing the amount of data available to inform the model, ISDMs are able to improve species niche characterization, spatial prediction, and narrow population trend credibility intervals. Similarly, ISDMs may be used to improve estimates when facing missing data (e.g. unsampled area or year due to bad weather) in the main data source (e.g. FI survey) by using accessory data source(s).

Data weighting could be used to somehow validate the merging of the different data sources. Ideally, results should be invariant to data weighting, which implies that integrated data sources sample the same process and the ISDM is well specified.

It is advisable to integrate data collected using proportional gear efficiencies (i.e. their quotient at different fish-lengths is a constant value), otherwise each dataset may sample different fractions of the underlying population and therefore observations may not be proportional (i.e. scalable through a constant factor). Under such circumstances, one would need to model fish length class abundances including a function that accommodate the length-specific gear selectivity of each fishing gear included in the ISDM (Munro and Somerton, 2001; Fryer *et al.*, 2003; Miller, 2013; Gonzalez *et al.*, 2021).

Acknowledgements

This study is indebted with all the on-board observers that carried out the sampling, and with the UTPB that runs the monitoring program of the artisanal fishery sector in Galician waters. The authors express their gratitude to all the people that work in the SP-NSGFS Q4 surveys. SP-NSGFS Q4 surveys were co-funded by the EU within the Spanish national program for the collection, management, and use of data in the fisheries sector and support for scientific advice regarding the Common Fisheries Policy.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Conflict of interest

The authors have no conflicts of interest to declare.

Funding

IP would like to thank the European Commission for the funding (GAP-847014). IP is grateful to the MSCA fellowship that supported his research. MGP thanks the project IMPRESS (RTI2018-099868-B-I00), ERDF, Ministry of Science, Innovation, and Universities - State Research Agency.

Author contributions

IP, SS, and JI conceived the ideas and designed the methodology. IP analysed the data and wrote the code. IP led the writing of the manuscript and SS contributed significantly. All authors contributed critically to the drafts and gave final approval for publication.

Data availability

Fishery independent data are available in the DATRAS database. Unfortunately, fisheries dependent data are not publicly available due to privacy issues, but could be accessible under specific agreement with Xunta de Galicia.

References

Alglave, B., Rivot, E., Etienne, M-P., Woillez, M., Thorson, J. T., and Vermard, Y. 2022. Combining scientific survey and commercial catch data to map fish distribution. ICES Journal of Marine Science, 79: 1133–1149.

Alonso-Fernández, A., Otero, J., Bañón, R., Campelos, J. M., Quintero, F., Ribó, J., Filgueira, F. *et al.* 2019. Inferring abundance trends of key species from a highly developed small-scale fishery off NE Atlantic. Fisheries Research, 209: 101–116.

Araújo, M. B., and New, M. 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution, 22: 42–47.

Elith, J., and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40: 677–697.

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

Francis, R. C. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences, 68: 1124–1138.

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

Fraser, H. M., Greenstreet, S. P., and Piet, G. J. 2007. Taking account of catchability in groundfish survey trawls: implications for estimating demersal fish biomass. ICES Journal of Marine Science, 64: 1800–1819.

Fryer, R., Zuur, A. F., and Graham, N. 2003. Using mixed models to combine smooth size-selection and catch-comparison curves over

hauls. Canadian Journal of Fisheries and Aquatic Sciences, 60: 448–459.
Gneiting, T., and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102: 359–378.
Gonzalez, G. M., Wiff, R., Marshall, C. T., and Cornulier, T. 2021. Estimating spatio-temporal distribution of fish and gear selectivity functions from pooled scientific survey and commercial fishing data. Fisheries Research, 243: 106054.
Hogan, J. W., and Laird, N. M. 1997. Mixture models for the joint distribution of repeated measures and event times. Statistics in Medicine, 16: 239–257.
Hutchinson, G. 1957. Concluding remarks-cold spring harbor symposia on quantitative biology. reprinted in 1991: classics in theoretical biology. Bulletin of Mathematical Biology, 53: 193–213.
Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N. *et al.* 2020. Data integration for large-scale models of species distributions. Trends in Ecology & Evolution, 35: 56–67.
Knorr-Held, L., and Best, N. G. 2001. A shared component model for detecting joint and selective clustering of two diseases. Journal of the Royal Statistical Society: Series A (Statistics in Society), 164: 73–85.
Laman, E. A., Rooper, C. N., Turner, K., Rooney, S., Cooper, D. W., and Zimmermann, M. 2018. Using species distribution models to describe essential fish habitat in Alaska. Canadian Journal of Fisheries and Aquatic Sciences, 75: 1230–1255.
Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields 670 and Gaussian Markov random fields: the spde approach (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73: 423–498.
Maunder, M. N., and Punt, A. E. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries research, 70: 141–159.
Miller, T. J. 2013. A comparison of hierarchical models for relative catch efficiency based on paired-gear data for us northwest Atlantic fish stocks. Canadian Journal of Fisheries and Aquatic Sciences, 70: 1306–1316.
Moriarty, M., Sethi, S. A., Pedreschi, D., Smeltz, T. S., McGonigle, C., Harris, B. P., Wolf, N. *et al.* 2020. Combining fisheries surveys to inform marine species distribution modelling. ICES Journal of Marine Science, 77: 539–552.
Munro, P. T., and Somerton, D. A. 2001. Maximum likelihood and non-parametric methods for estimating trawl footrope selectivity. ICES Journal of Marine Science, 58: 220–229.
Paradinas, I., Conesa, D., López-Quílez, A., and Bellido, J. M. 2017. Spatio-temporal model structures with shared components for semi-continuous species distribution modelling. Spatial Statistics, 22: 434–450.
Paradinas, I., Conesa, D., López-Quílez, A., Esteban, A., López, L. M. M., Bellido, J. M., and Pennino, M. G. 2020. Assessing the spatiotemporal persistence of fish distributions: a case study on two red mullet species (*Mullus surmuletus* and *M. barbatus*) in the western Mediterranean. Marine Ecology Progress Series, 644: 173–185.
Paradinas, I., Conesa, D., Pennino, M. G., Muñoz, F., Fernández, A. M., López-Quílez, A., and Bellido, J. M. 2015. Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. Marine Ecology Progress Series, 528: 245–255.
Paradinas, I., Gimenez, J., Conesa, D., Lopez-Quilez, A., and Pennino, M. G. 2021. Evidence for spatiotemporal shift in demersal fisheries management priority areas in the western Mediterranean. Canadian Journal of Fisheries and Aquatic Sciences, 79: 1641–1654.
Pennino, M. G., Conesa, D., Lopez-Quilez, A., Munoz, F., Fernández, A., and Bellido, J. M. 2016. Fishery-dependent and-independent data lead to consistent estimations of essential habitats. ICES Journal of Marine Science, 73: 2302–2310.
Pennino, M. G., Cousido-Rocha, M., Maia, C., Rocha, A., Figueiredo, I., Alonso-Fernández, A., Silva, C., *et al.* 2022a. This is what we know: assessing the stock status of the data-poor common sole on the Iberian Coast. Estuarine, Coastal and Shelf Science, 266: 107747.
Pennino, M. G., Izquierdo, F., Paradinas, I., Cousido, M., Velasco, F., and Cerviño, S. 2022b. Identifying persistent biomass areas: the case study of the common sole in the northern Iberian Waters. Fisheries Research, 248: 106196.
Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa, D. 2018. Accounting for preferential sampling in species distribution models. Ecology and Evolution, 9: 653–663.
Punt, A. E. 2015. Some insights into data weighting in integrated stock assessments. Fisheries Research, 192: 52–65.
Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., and Krainski, E. 2009. Inla: functions which allow to perform a full bayesian analysis of structured additive models using integrated nested laplace approximation. R package version 22.12.16. https://github.com/hrue/r-inla.
Rufener, M.-C., Kristensen, K., Nielsen, J. R., and Bastardie, F. 2021. Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. Ecological Applications, 31: e02453.
Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. *et al.* 2017. Penalising model component complexity: a principled, practical approach to constructing priors. Statistical Science, 32: 1–28.
Tanner, S. E., Vasconcelos, R. P., Cabral, H. N., and Thorrold, S. R. 2012. Testing an otolith geochemistry approach to determine population structure and movements of european hake in the northeast Atlantic Ocean and Mediterranean Sea. Fisheries Research, 125: 198–205.
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: 1369–1382.
Thorson, J. T. 2019. Guidance for decisions using the vector autoregressive spatio-temporal (vast) package in stock, ecosystem, habitat and climate assessments. Fisheries Research, 210: 143–161.
Thorson, J. T., Johnson, K. F., Methot, R. D., and Taylor, I. G. 2017. Model-based estimates of effective sample size in stock assessment models using the dirichlet-multinomial distribution. Fisheries Research, 192: 84–93.
Thorson, J. T., Shelton, A. O., Ward, E. J., and Skaug, H. J. 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for west coast groundfishes. ICES Journal of Marine Science, 72: 1297–1310.
Thorson, J. T., and Ward, E. J. 2013. Accounting for space–time interactions in index standardization models. Fisheries Research, 147: 426–433.
Tolimieri, N., Wallace, J., and Haltuch, M. 2020. Spatio-temporal patterns in juvenile habitat for 13 groundfishes in the California Current ecosystem. PLoS One, 15: e0237996.
Wang, S-P., and Maunder, M. N. 2017. Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? Fisheries Research, 192: 41–51.
Watanabe, S. 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11: 3571–3594.
Zhou, S., Klaer, N. L., Daley, R. M., Zhu, Z., Fuller, M., and Smith, A. D. 2014. Modelling multiple fishing gear efficiencies and abundance for aggregated populations using fishery or survey data. ICES Journal of Marine Science, 71: 2436–2447.

Handling Editor: Shijie Zhou