ICES Journal of Marine Science, 2024, Vol. 81, Issue 5, 984–995
https://doi.org/10.1093/icesjms/fsae056
Received: 29 November 2023; **revised:** 2 April 2024; **accepted:** 4 April 2024
Advance access publication date: 2 May 2024
Original Article

Accounting for spatio-temporal and sampling dependence in survey and CPUE biomass indices: simulation and Bayesian modeling framework

Alba Fuster-Alonso [1,2,*], David Conesa [2], Marta Cousido-Rocha [3], Francisco Izquierdo [3], Iosu Paradinas [4], Santiago Cerviño [3], Maria Grazia Pennino [5]

[1]Instituto de Ciencias del Mar (ICM-CSIC), Renewable Marine Resources Department, Barcelona 08003, Spain
[2]Universidad de Valencia (UV), Statistics and Operations Research Department (VaBar), Valencia 46100, Spain
[3]Instituto Español de Oceanografía (IEO-CSIC). C.O. de Vigo, Vigo 36390, Spain
[4]AZTI, Txatxarramendi Ugartea z/g, Sukarrieta, Bizkaia 48395, Spain
[5]Instituto Español de Oceanografía (IEO-CSIC). C.O. de Madrid, Madrid 28002, Spain

*Corresponding author: Instituto de Ciencias del Mar (ICM-CSIC), Renewable Marine Resources Department, Pg. Marítim de la Barceloneta, 37, Ciutat Vella, 08003 Barcelona, Spain. E-mail: afuster@icm.csic.es

Abstract

Estimating changes in the biomass of a fish stock is crucial for successful management. However, fishery assessment may be affected by the quality of the inputs used in stock assessment models. Survey biomass indices derived from fishery-independent and catch per unit effort (CPUE) biomass indices derived from fishery-dependent data are key inputs for model calibration. These indices have biases that could compromise the accuracy of the stock assessment models results. Therefore, there are plenty proposed methods to standardize survey or CPUE biomass data. From simpler models like generalized linear models (GLMs) to more complex models that take into account spatio-temporal correlation, like geostatistical models, and sampling dependence, like marked point processes. But many of them do not consider the underlying spatio-temporal or sampling dependence of these data. Hence, the goal of the study is to present a spatio-temporal simulation and Bayesian modeling framework to assess the impact of applying models that do not consider spatio-temporal and sampling dependence. Results indicate that geostatistical models and marked point processes achieve the lowest measures of error. Hence, to capture the underlying spatio-temporal process of the survey and CPUE biomass indices and data sampling preferentiality, it is essential to apply models that consider the spatio-temporal and sampling dependence.

Keywords: survey biomass indices; CPUE biomass indices; simulation; statistical modeling; preferential sampling and spatio-temporal effects

Introduction

Fisheries assessment implies understanding the past and predicting the present and future status of a fish stock (Neis et al. 1999, Xu et al. 2020). For this reason, a large number of stock assessment models have been developed to evaluate the state of a exploited fish stock (ICES 2012, Maunder and Punt 2013, Methot Jr and Wetzel 2013, Peterman 1990). Most stock assessment models require a common input to calibrate the estimated population biomass trends, which is a temporal series of survey or catch per unit effort (CPUE) biomass indices. These indices of survey and CPUE biomass play a key role in the stock assessment, as their estimated trends are assumed to be representative of the stock biomass.

These inputs mainly depend on two sources of information, fishery-dependent data (fisheries activities) and fishery-independent data (oceanographic surveys) (Maunder et al. 2020). In an oceanographic survey the information is frequently obtained as relative biomass (survey biomass), while in commercial fisheries the information is derived as CPUE. The main difference between the two sources of information is that in fisheries CPUE are sampling-dependent (preferential sampling), because fishers know or suppose to know the areas with the highest biomass of target species and systematically use the same fishing grounds (Diggle et al. 2010, Pennino et al. 2019). In contrast, many oceanographic surveys, such as swept area methods (Gunderson 1993), have a design based on randomness behind, and consequently, the survey biomass is independent of sampling. This distinction between fishery-dependent and fishery-independent data is very important because it has a direct influence on the statistical modeling (Pennino et al. 2016).

It is important to mention that, survey and CPUE biomass data have potential biases (different fishing efforts, selectivity of the gear, year variations, environmental variability, etc.) that may affect the accuracy of the stock assessment model results (Tagliarolo et al. 2021). For this reason, there are many proposals to standardize these data (Maunder and Punt 2004, Stock et al. 2020, Zhou et al. 2019). From simpler models such as generalized linear models (GLMs) or generalized additive models (GAMs) (Hazin et al. 2007), to more complex models such as geostatistical models (Cao et al. 2017, Hoyle et al. 2024, Kai 2019, Shelton et al. 2014, Tremblay-Boyer et al. 2017, Xu et al. 2019). Hence, the modeling of the temporal series of survey and CPUE biomass indices may be quite a challenge, since there are no structured protocols on how should we model these data.

In recent years, there has been an increased focus on research examining the effects of incorporating spatial correlation into the modeling of survey and CPUE biomass data (Hoyle et al. 2024, Maunder and Punt 2004). However, it remains a common practice to use models that do not account for the underlying spatio-temporal process, especially the sampling dependence of CPUE biomass data in the model (Pennino et al. 2019) Consequently, by using these indices as inputs into stock assessment models, potential biases can affect the model's results (Maunder et al. 2020).

Therefore, the goal here is to conduct a simulation study based on an underlying spatio-temporal structure to assess the impact of applying models that do not consider spatio-temporal correlation and sampling dependence. For this purpose, (1) we simulate a spatio-temporal scenario of biomass of a fish stock; (2) we generated 30 replicates of each simulated biomass sampling using two different approaches (i.e., random sampling as in oceanographic surveys and preferential sampling as in fisheries activities), and through sampling we obtain the survey and CPUE biomass data for each one; (3) we apply different regression models to obtain the standarized series of survey and CPUE biomass indices for each sample; (4) finally, we validate the standardized series with respect to the median of the simulated biomass series (for each year and replica) by calculating different measures such as the RMSE (root mean square error), MAPE (mean absolute percentage error), Spearman correlation, standard deviationm IRQ (interquartile range), median symmetric accuracy (ζ), and symmetric signed percentage bias (SSPB) (Morley et al. 2018) (see supplementary material 1), and also representing some of the mentioned measures through a Taylor diagram.

Our hypothesis is that ignoring the underlying spatio-temporal process and the sampling dependence in the modeling approach can worsen the accuracy of the assessment. Whereas the application of preferential and/or spatio-temporal models, such as geostatistical or marked point process, may reduce the uncertainty of survey and CPUE biomass indices, explaining part of the variability that otherwise would not be accounted for.

Material and methods

In what follows, we present our proposal for analyzing whether the survey and CPUE biomass trends estimated by the models follow the behavior of the simulated biomass. The process begins with the simulation of the 'real' biomass that would describe a potential fish stock. From this biomass, a sampling process is carried out imitating the data collection performed in practice, providing the survey and CPUE biomass data. Once the simulation process is established, the next step is to standardize the survey and CPUE biomass data using different regression models.

All the necessary code to reproduce the results is available in a repository on GitHub.: GitHub url. The repository contains the simulation protocol and the code for fitting geostatistical and marked point processes to standardize survey and CPUE biomass data. The code for the modeling could be used in a real-world case study, and the simulation protocol can be employed with different parameters to conduct other simulation studies.

Simulation process

Here, we describe the data simulation process in detail, first describing the generation of the biomass and then obtaining the survey and CPUE biomass data by sampling from the simulated biomass.

Spatio-temporal biomass scenario

The biomass of a given fish stock is usually influenced by different external factors (i.e., environmental, antropogenic, etc.), as well as spatially structured biological processes (i.e., predation, competition, etc.). While some approaches incorporate external factors in modeling the biomass of the species, in some cases, spatially structured biological processes are not considered since this requires to model spatial autocorrelation. This implies that the modeling part of the intrinsic spatial variability of the data is left unexplained. To solve this issue, the inclusion of a spatial component in these models is essential (Izquierdo et al. 2022, 2021, Paradinas et al. 2017, 2022a, Pennino et al. 2022, 2019).

For this reason, our underlying model for the biomass includes a spatio-temporal structure, a bathymetry effect and a temporal trend as key effects. All fish populations exhibit a spatial and temporal structured somehow (Hoyle et al. 2024). Therefore, in this work we simulate the biomass scenario assuming a spatio-temporal effect. Additionally, the biomass values differ between years and across the bathymetry (depth of fishing) levels, hence both effects are considered in the simulation process:

$$\begin{aligned} Y(s,t) &\sim \text{Gamma}(\mu(s,t),\phi), \\ \log(\mu(s,t)) &= \beta_0 + f(t) + f(\text{X(s)}) + U(s,t), \end{aligned} \tag{1}$$

where, the response $Y(s, t)$ represents the biomass at time t in the location s following a Gamma distribution with parameters $\mu(s, t)$ (mean) and ϕ (dispersion); the mean $\mu(s, t)$ is linked to the predictor by the logarithm link function; β_0 is the intercept; $f(t)$ stands for the temporal trend in the year t, and $f(\text{X})$ is a deterministic function for the bathymetry. Lastly, $U(s, t)$ refers to the spatio-temporal structure.

Once we have determined which is the model structure for the biomass simulation, we now describe how to deal with each one of the terms included in the predictor in (1). The spatio-temporal structure is simulated as a Gaussian Markov Random Field (GMRFs) correlated with an autoregressive $AR(1)$ model with parameter of autocorrelation ρ_{sp} (Krainski et al. 2018). For the bathymetry $f(\text{X}(s))$, we have selected a range between 0 and 800 meters and also a nonlinear effect that indicates that the highest biomass values are found at intermediate depths. Lastly, for the temporal trend $f(t)$, changes in the average biomass values over time are included by simulating a vector of values from an autoregressive model of order 1 with parameter of autocorrelation ρ_t. Once the terms of the predictor have been obtained, the biomass (response variable) has to be constructed by simulating from the corresponding Gamma distribution.

Sampling process

Once we have the "real" biomass of a fish stock (our absolute biomass from now on), our next step is to perform a sampling process imitating the real data collection. Although there are other differences between scientific survey data and fishery de-

pendent data (Nielsen 2015), our focus here is to reproduce the different behavior of survey and CPUE biomass data that affects directly in the statistical modeling.

A critical factor that plays a key role in our modeling and differs between survey and CPUE biomass data is the dependence on sampling. CPUE biomass data lack independence in observations, as the sampling is preferential. In contrast, in survey biomass data, the observations are independent of each other.

Therefore, in order to mimic these two data collection procedures and reproduce their sampling behavior, we need two different sampling processes: random sampling (mimicking fishery-independent data) and preferential sampling (that mimics fishery-dependent data), the latter being a process in which the locations with highest values of biomass have a large probability of being sampled.

In particular, for the fishery-dependent data, we have based our preference sampling as in Diggle et al. (2010), where the process determining the locations (fishing hauls) and the response variable (biomass) are dependent from each other. But, there could be other factors influencing fishing locations base on the fisher behavior, such as vessel density, revenue, cost, risk, targeting or tradition (see Girardin et al. 2017). This sampling process is repeated a total of 30 times for each sampling scenario. The consideration of multiple replicates increases the robustness of conclusions.

Catch data are usually presented aggregated in space as lattice or areal data (for instance when it comes from logbooks or other similar sources). However, here we reproduce the sampling that provides the geolocated catches, as the underlying nature of this type of data is continuous. Therefore, the sampling procedures provide the exact locations and the corresponding absolute biomass values. From this, the corresponding catch values are derived.

Here, catch data is obtained by multiplying the biomass by a constant q, which is called the catchability coefficient (Arreguín-Sánchez 1996). However, it is important to highlight that there could be other sources of bias that might affect the catch, such as depth distribution, gear selectivity, target species, effort allocation, catchability, survey fish detection, unreported catch, size composition, or environmental factors. Nielsen (2015).

In the random sampling, the fishing time is constant for each haul, and so, the catch can be understood as relative biomass (survey biomass), that is

$$\text{Survey biomass} = \text{Absolute biomass} \times q\,. \tag{2}$$

In contrast, in the context of preferential sampling the effort is not constant (i.e. each fisher decides the time that the fishing gear remains active), which may influence the catchability coefficient. In this case, catch is understood as CPUE and it is expressed as:

$$\text{CPUE} = \frac{\text{Absolute Biomass} \times q}{\text{effort}}\,, \tag{3}$$

where effort is simulated from a normal distribution and varies in each catch.

Statistical modeling framework

We now describe the common models that are used to standardize the survey and CPUE biomass. In particular, we mention here simple approaches such as GLMs or GAMs, but also more complex formulations such as geostatistical models or marked point process. All of them will be particularized in the context of our proposal model in (1).

(1) **Generalized** linear model
GLMs are one of the most popular models used to standardize survey and CPUE biomass indices (Hardin et al. 2007) by relating the mean of the survey and CPUE biomass data (via a convenient link) to the linear predictor of the covariates of interest. This can be particularized according to our proposal model in (1) in the following GLM that includes both a bathymetry and a temporal effect:

$$\begin{aligned} Z(s,t) &\sim \text{Gamma}(\mu(s,t),\phi), \\ \log(\mu(s,t)) &= \beta_0 + \beta_1 X(s) + \beta_2 X^2(s) + \alpha(t), \end{aligned} \tag{4}$$

where $Z(s, t)$ represents the response variable (survey and CPUE biomass data) in year t and location s and follows a Gamma distribution with ϕ and $\mu(t)$ the dispersion and mean parameters at time t, respectively; $\mu(s, t)$ is linked to the linear predictor by the log link function; β_0 refers to the intercept; β_1 and β_2 correspond to the coefficients of the polynomial associated with the bathymetry effect (X); and $\alpha(t)$ is the fixed factor related to the categorical year.

(2) **Generalized** additive model
GAMs go beyond GLMs by allowing us to include in the model the assumption of nonparametric relationships between the variable of interest and the predictor terms (Hastie 2017). Indeed, GAMs allow us to include the spatial variability as a bivariate tensor spline on the latitude and longitude covariates. The resulting GAM that particularizes our proposal in (1) stands as:

$$\begin{aligned} Z(s,t) &\sim \text{Gamma}(\mu(s,t),\phi), \\ \log(\mu(s,t)) &= \beta_0 + f(X(s)) \\ &\quad + f(x(t),y(t))(s,t) + \alpha(t), \end{aligned} \tag{5}$$

where $f(X)$ corresponds now to a smooth function of the bathymetry; $f(x, y)(t)$ represents a bivariate smoothing function for coordinates x and y at year t; and the remaining terms are those in (4).

(3) **Geostatistical** model
A better way in order to explicitly include in the model the existing spatial dependence is to work with geostatistical models (Cressie 1989). In this case, the model that particularizes our proposal in (1) is:

$$\begin{aligned} Z(s,t) &\sim \text{Gamma}(\mu(s,t),\phi), \\ \log(\mu(s,t)) &= \beta_0 + f(X(s)) + f(t) + U(s,t), \end{aligned} \tag{6}$$

where $f(t)$ and $U(s, t)$ represent respectively a temporal trend and a spatial effect correlated over time, and the remaining terms are those in (4) and (5). In particular, $U(s, t) = W(s, t) + \rho_1 W(s, t - 1)$, so the spatial effect $W(s, t)$ changes over subsequent time events through a first order autoregressive model (AR1). The spatial effect is modeled as a Gaussian Markovian Random Field (GMRF), $W(s, t) \sim \text{GMRF}(0, Q^{-1}(\sigma_w, \tau))$, being Q the sparse matrix with variance σ_w and range τ.

(4) **Marked** point process
So far, all proposed models assume that fishery locations are independent of each other. However, this assumption does not always hold, like in fisheries were

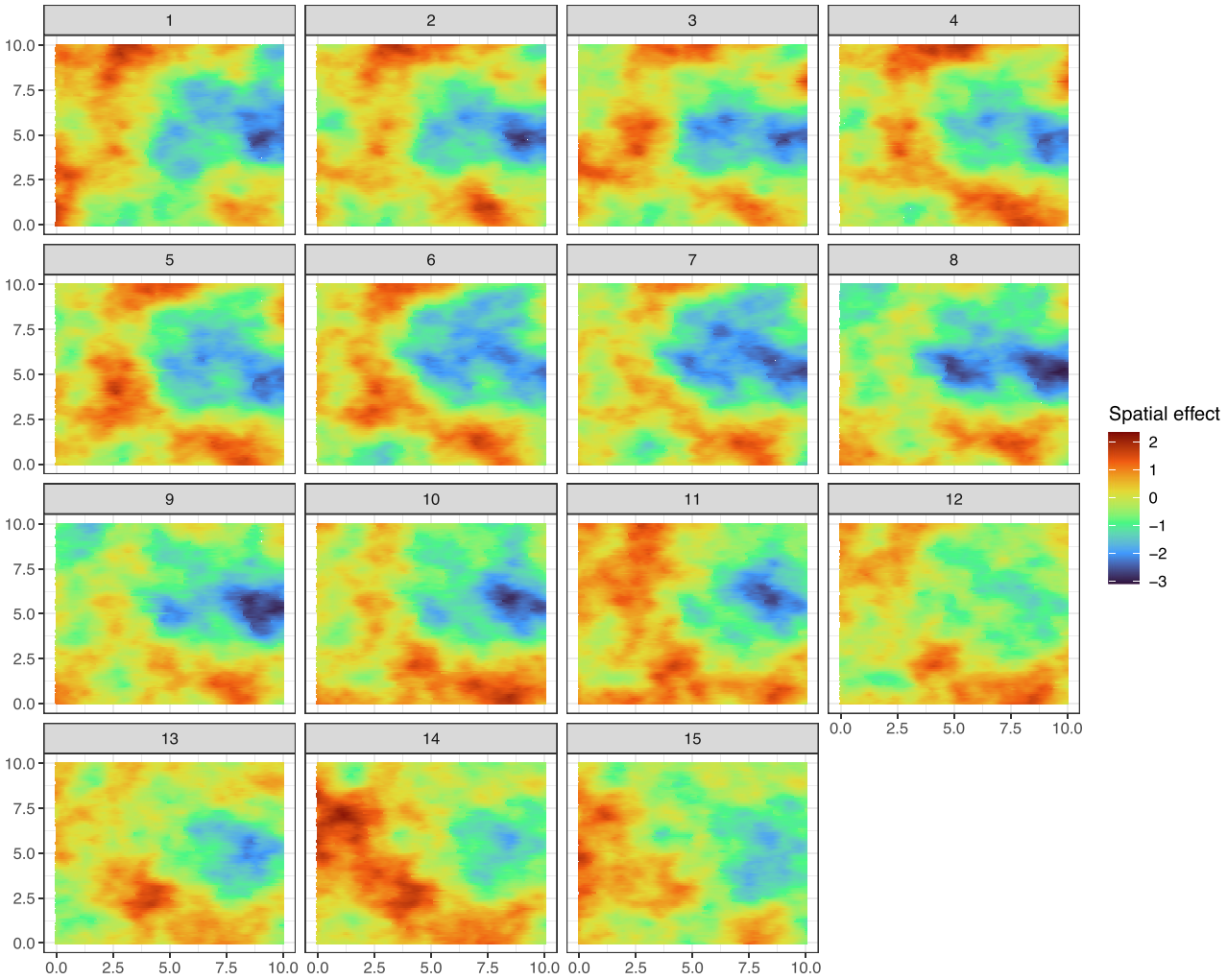

Figure 1. Maps with the simulated spatio-temporal effect in the hypothetical study field and with a time window of 15 years. Red values indicate positive relationship between points over space, while blue values indicate negative relationship between points over space.

locations clearly depend on the biomass (fishers go to fish where fishes are, or at least where they think there are). For this reason, we describe a model that allows to take into account this preferential sampling (Conn et al. 2017, Diggle et al. 2010, Shirota and Gelfand 2022, Watson 2021) in the context of the biomass and CPUE (Alglave et al. 2022, 2023, Pennino et al. 2019, Rufener et al. 2021). This model is named as a Marked point process and it is formulated as follows:

$$\begin{aligned} Z(s,t) &\sim \text{Gamma}(\mu(s,t),\phi), \\ \log(\mu(s,t)) &= \beta_{0\text{CPUE}} + \beta_{1CPUE}X(s) + f(t) + U(s,t), \\ PP(s,t) &\sim \text{LGCP}(\lambda(s,t)), \\ \log(\lambda(s,t)) &= \beta_{0\text{PP}} + \beta_{1PP}X(s) + \alpha_{pp}U(s,t), \end{aligned} \tag{7}$$

where PP stands for a point pattern process PP(s, t) modeled as a Log-Gaussian Cox Process, which is based on a Poisson process of varying intensity $\lambda(s, t)$ and the remaining terms are those in (4), (5), and (6). This type of process tries to model the average intensity with which events occur, and so the logarithmic intensity $\lambda(s, t)$ of the Cox process is linked with a Gaussian predictor. The joint model also includes a similar geostatistical part as in (6). It is worth noting that the predictors of both parts of the joint model have the same components, except for the temporal trend $f(t)$, and more importantly, that the correlated spatial effect of both likelihoods is related by a scaling parameter α_{pp}, which is the one that express the degree of preferentiality.

In this work, inference and prediction of the model parameters for all these statistical models is performed from a Bayesian perspective. With respect to the software used here, `R-INLA` (Lindgren et al. 2011, Rue et al. 2009), `inlabru` (Bachl et al. 2019) and `R2BayesX` (Belitz et al. 2022, Umlauf et al. 2015) packages have been employed. Vaguely informative prior distributions of `R-INLA`, `inlabru`, and `R2BayesX` (most of them the default ones) have been used for all the parameters and hyperparameters. There are various approaches for inference and prediction in spatio-temporal modeling, allowing the inclusion of spatial and temporal autocorrelation. INLA is one such method, but it's not the only option. Other methods, like VAST (Thorson 2019), sdmTMB (Anderson et al. 2022) (both frequentist approaches), and glmmfield (a Bayesian approach) (Anderson and Ward 2019), are also available.

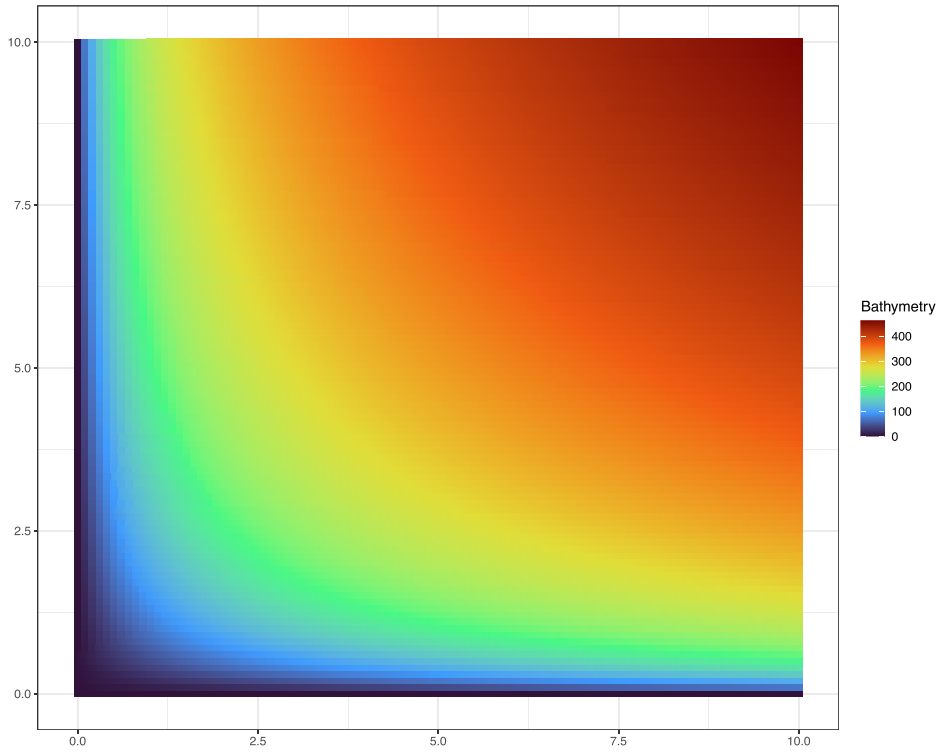

Figure 2. Map with the covariate bathymetry in the study field, which indicates the depth of the seabed and remains constant over time.

Finally, in order to compare the performance of all regression models respect with to the "real" biomass, we use different measures: RMSE and the MAPE as error measures; the Spearman correlation coefficient for analyzing the trends; 3) IQR and the SD as uncertainty; and 4) the Taylor diagram to represent the performance of the different models considering the RMSE, the correlation and also the SD. We specifically utilize the estimated values from each regression model for every year and compare them with the respective median of the simulated biomass for each year.

Results

We now present the results obtained when first simulating a specific spatio-temporal biomass, then reproducing the survey and CPUE biomass data, and finally comparing the behavior of the previously presented modelings for survey and CPUE biomass indices in terms of fitting. Besides, all the necessary code to reproduce the results is available in a repository on GitHub.: GitHub url.

Obtaining the simulated field of study

As above mentioned, the spatio-temporal simulation of the biomass of a fish stock is done by considering three terms: (1) a spatio-temporal effect, (2) a bathymetry effect and (3) a temporal trend. Hence, a series of parameters must be set for all the terms in Equation (1).

Our hypothetical field of study is a regular grid (10×10) over a 15-year time window. We have chosen a resolution that provides a reasonable computation time framework, as all variables are simulated at the same resolution. Moreover, the established parameters simulation have been chosen to simulate effects that are relevant for the model, especially the range of the spatial effect. This allows us to assess the effects of not considering an underlying spatial effect.

We start simulating the correlated spatial effect setting the values of the range ($r = 5.7$), variance ($\sigma = 1$) and temporal correlation ($\rho_{sp} = 0.9$). Figure 1 represents this simulated spatio-temporal effect, in which we see an autoregressive behavior over time. Then, we construct a bathymetry effect (constant over time) in a range from 0 to 800 meters by means of the following formula $100 * \log(xy + 1)$, where x and y are the coordinates, reflecting that the closer to the axis the lower the bathymetry (Fig. 2). As the relationship between bathymetry and biomass is usually nonlinear, we set up a polynomial of degree two to achieve a quadratic relationship ($\beta_1 = -2.5, \beta_2 = -1.5$) between them. With respect to the temporal trend, we simulate a vector of values from an autorregresive model of order 1, where $\rho_t = 0.9$.

Then, all the terms of the predictor in Equation (1) are summed and used as the mean of a Gamma distribution in order to simulate the biomass behavior of a stock. In addition, we have the changes in the median of the biomass over time, which has an increasing trend in the first years and a decreasing trend in the last years (Fig. 3). In Fig. 4, we illustrate an example of a simulated biomass, where it is possible to observe the quadratic (nonlinear) relationship with the bathymetry and the autoregressive behavior of the spatial effect.

Once, we have the biomass simulated, we reproduce the two sampling scenarios; random sampling as in an oceanographic survey and preferential sampling as in fisheries. Both scenarios are replicated 30 times each. In the first case, we simply randomly selected 50 locations for each year and each sampling scenario (see section 2 of the supplementary material). In the

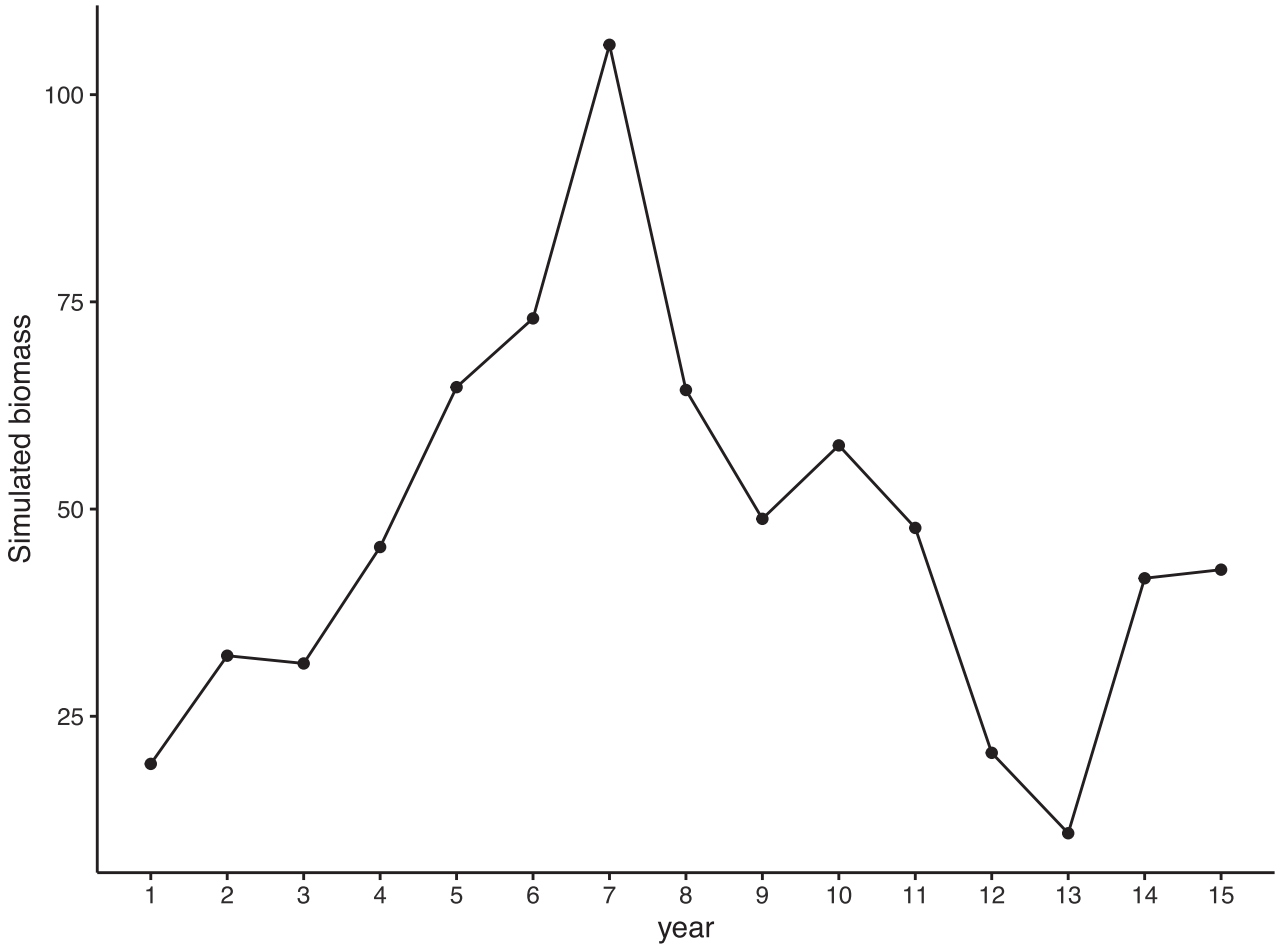

Figure 3. Temporal trend of simulated biomass over 15 years. The x-axis represents each of the years considered in the simulation, and the y-axis represents the median of the simulated biomass for each year. The dots represent the median of the entire simulated study field for each year.

latter one, we selected 50 locations for each year and sampling scenario according to a vector of probabilities, obtained from the simulated biomass by transforming those simulated values to a range from 0 to 1 and then use them as the probability vector. As it can be appreciated in Figures from 30 to 59 of the section 2.1 of supplementary material, higher probabilities are associated to locations with large biomass values. Finally, with the select biomass values, we reproduce the survey and CPUE biomass data as mentioned in Equations (3) and (2), setting a catchability coefficient q of 0.3 in the case of random sampling and 0.6 for the preferential sampling (see section 2.2 of the supplementary material). We assume a nonvariable catchability coefficient to simplify the framework, albeit q may change over space and time.

Results of the analysis of the proposed models

With the survey and CPUE biomass data obtained from the simulated biomass, we now present the results of the inference and predictive processes for the different models mentioned in Section 2.2;.

Regarding survey biomass data, Fig. 5a shows the different temporal trends obtained from the survey biomass indices compared to the "real" simulated biomass series. The geostatistical model seems to better fit the temporal trend with respect to the other models, albeit it tends to overestimate the biomass value in some years. It is also worth noting that GLMs and GAMs have a large variability in the different replicates compared to the geostatistical model, specially when the biomass increases.

Figure 6a presents the analysis conducted to compare simulated biomass against biomass estimated using various models. In the first graph of Fig. 6a, we can observe the estimated values compared to the observed ones, highlighting that the GLM and GAM exhibit a significantly higher deviation compared to the geostatistical model. Furthermore, the estimated values in the geostatistical model appear to better fit the simulated biomass in most points along the line (first graph of Fig. 6a). With respect to the correlation coefficients, the geostatistical model shows a higher median than the GLM or GAM, as well as lower RMSE (second and third graph of Fig. 6a). Additionally, in the Taylor diagram, we can observe that the geostatistical model bears the closest resemblance to the simulated data (fourth graph of Fig. 6a). This last Taylor diagram represents the median of all replicates. In section 3.2. of the supplementary materials, we have Taylor diagrams for each model and replicate.

In particular, among all the survey biomass indices series estimated with the different models, the ones that obtained the lowest RMSE, MAPE and ζ were the series coming from the

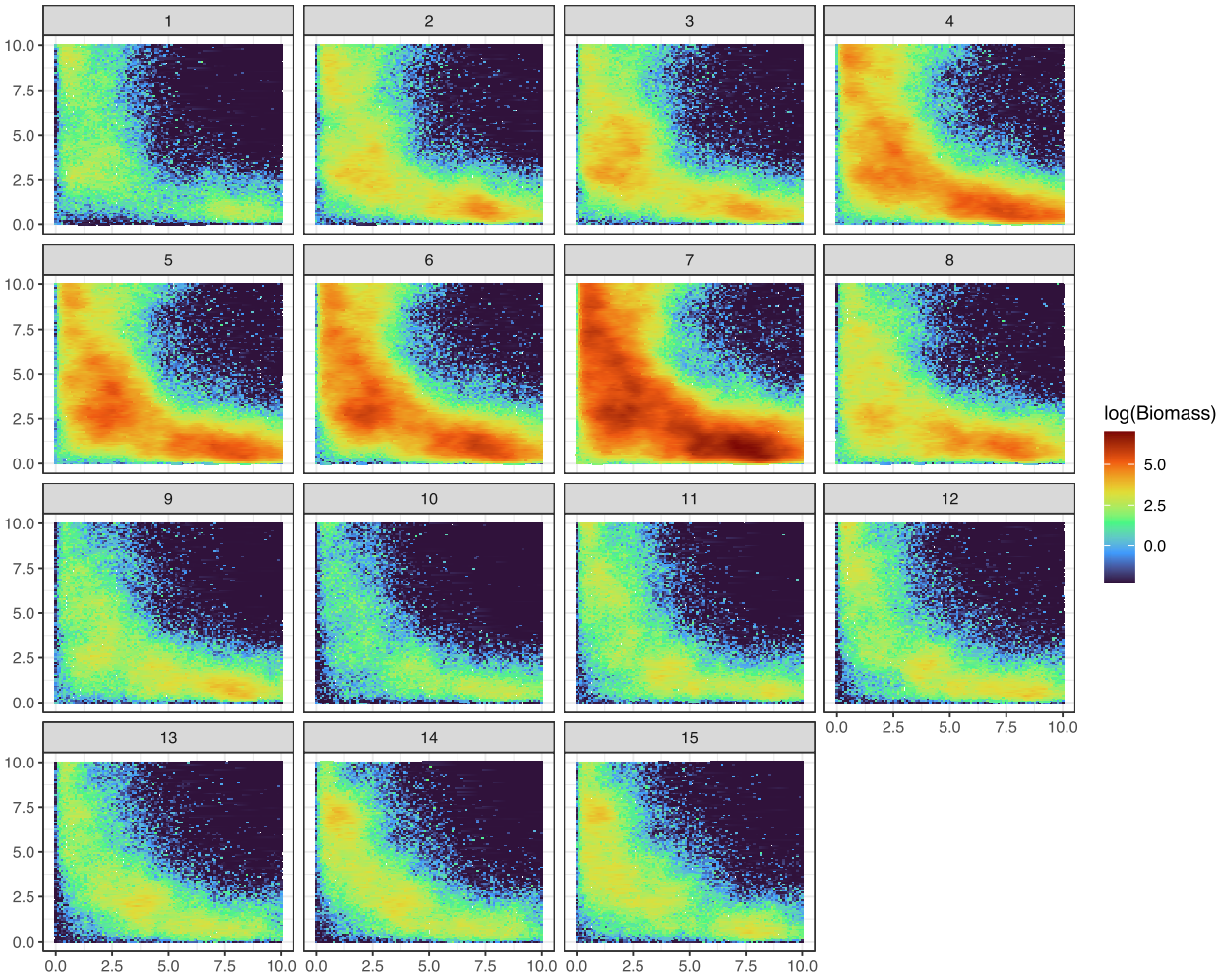

Figure 4. Maps with the simulated biomass scenario of a fish stock in the hypothetical study field and with a time window of 15 years. Red values indicate high biomass values, while blue values indicate low biomass values.

geostatistical model (Table 1). After the geostatistical model, in general the models with the lowest RMSE and MAPE were the GAMs and, finally the GLMs, whose results were very similar (Table 1). Same results where observed for the correlation, where the geostatistical model obtained a median coefficient of 0.96 (Table 1). Moreover, for the SSPB measure, we can find the results for each replica in Fig. 184of the supplementary material, where GLMs provide values far from 0. GAMs seem to provide a good estimator in some replicas but with large uncertainty, and the geostatistical method is not very close to 0 but appears more consistent. However, all SSPB measures indicate that the models are overestimating the biomass. Table 1 also presents the median values for both the SD and the IQR. Considering that the SD and IQR for the simulated biomass are 7.84 and 4.67, respectively, all models seem to produce results close to these numbers. For a detailed breakdown of the results for each replication, please refer to section 3.2. in the supplementary materials.

Considering the results for the survey biomass indices, we could also refer to the whole spatio-temporal maps for the simulated biomass and observe the resemblance with the prediction map obtained with the geostatistical model. Indeed, in section 3.1. of the supplementary material the median of the posterior predictive distribution obtained in space-time with the geostatistical model appears to have captured the spatio-temporal distribution of the simulated stock biomass.

With respect to the CPUE biomass indices, Fig. 5b highlights that most models are overestimating for some years compared to the simulated stock biomass. However, the geostatistical model reduces the overestimation by considering spatio-temporal dependence. Moreover, the marked point process avoids overestimation by considering both spatio-temporal and sampling dependence, making it the model that best represents the simulated spatio-temporal stock biomass.

Table 1 highlights that the marked point process stands out as the model with the lowest RMSE, MAPE, and ζ among all the CPUE biomass indices series predicted using the proposed models. Likewise, in Figure 184 of the supplementary material, we observe in the SSPB measure that all models tend to overestimate the biomass, except for the marked point process, which is the one closest to 0. However, in this sampling scenario, the correlation coefficient for all models is notably high, indicating that there are no relevant differences between them in this regard. Moreover, in Fig. 6b, we can clearly observe how the marked point process outperforms the others in terms of matching the simulated biomass, as evidenced in both the Taylor diagram and the real versus estimated plot.

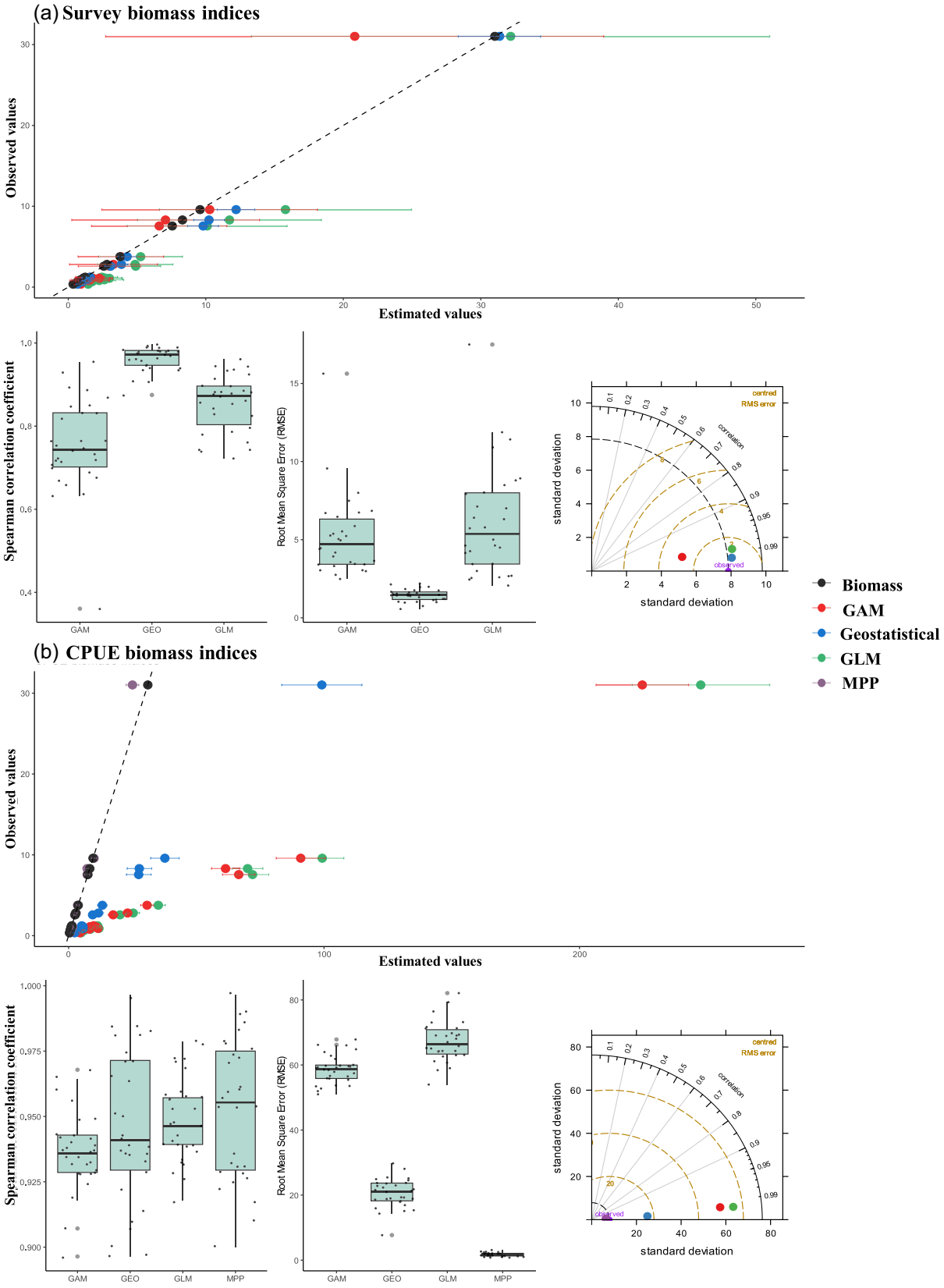

Figure 5. Comparison of the median of the simulated "real" biomass of a fish stock (named Biomass in the axis) vs. the standardized biomass series derived from the different models proposed for CPUE (a) and survey biomass (b) data and each replicate. The boxplot represented the values for each replication, and the blue dots the median of the simulated biomass.

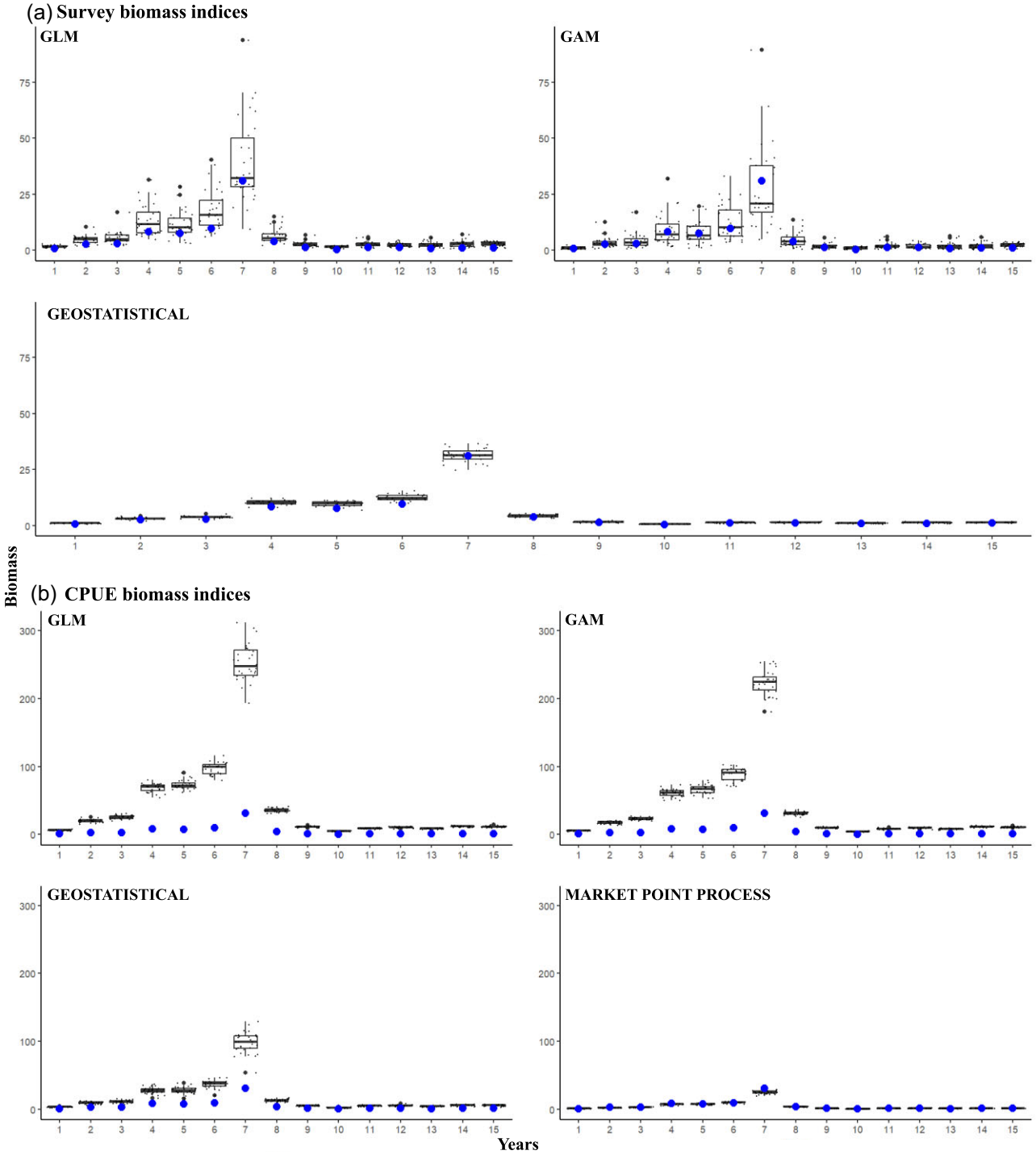

Figure 6. Estimated biomass series compared to simulated biomass. In the first graph, real biomass values (name observed values in the y-axis) are plotted against standardized values (name estimated values in the x-axis) for each of the proposed models, where the dots represent the median of the different replications and the lines the SD. Boxplots are used to depict the correlation and RMSE values for each replica. We display a Taylor diagram with respect to the medians of all replicated and based on the Pearson correlation, SD and centered RMSE error.

Finally, it is also worth noting that for the marked point process model, the median of the posterior predictive distribution of CPUE biomass indices for each year has been able to capture the trend in space and time compared to the stock biomass simulation (see section 3.1. of the supplementary material).

Discussion and conclusions

In this study, our simulated spatio-temporal scenario has allowed us to analyze the behavior of several regression models for standardizing survey and CPUE biomass data. These series of survey and CPUE biomass indices are fundamental in stock assessment models, since they determine the estimation

Table 1. Error measures for each replication of the simulation.

Model	Correlation	RMSE	MAPE	ζ	SD	IQR
			Survey biomass indices			
GLM	0.87	5.37	0.50	97.40	9.42	5.45
GAM	0.74	4.71	0.58	62.90	6.38	3.80
Geostatistic	0.97	1.45	0.23	29.82	8.07	5.77
			CPUE biomass indices			
GLM	0.95	66.40	0.89	832.85	63.50	42.10
GAM	0.94	58.80	0.88	745.30	57.50	37.10
Geostatistic	0.94	21.20	0.77	322.60	25.00	15.20
MPP	0.96	1.74	0.22	22.86	6.31	4.11

The table includes correlation Spearman coefficients (correlation), RMSE, MAPE, median symmetric accuracy (ζ), SD, and IQR for survey and CPUE biomass indices. All measures are calculated by comparing the actual values of simulated biomass with the standardized or estimated values generated by each model over time.

of the changes in the temporal trend of biomass (Cousido-Rocha et al. 2022, Santos et al. 2022). Therefore, it is important to consider a modeling that reduces the uncertainty of these series of indices.

Although GLMs and GAMs are widely used for standardizing survey and CPUE biomass indices, they have limitations, especially when considering spatio-temporal structures (Chiarini et al. 2022, Hinton and Maunder 2004, Hoyle et al. 2022, Hsu et al. 2022, Oshima et al. 2009, Pons et al. 2010). Indeed, spatially correlated variation in biomass is observed in almost all fisheries data collected from both fishery-dependent and fishery-independent sources (Martínez-Minaya et al. 2018). However, spatial variation is often ignored or not properly considered in statistical analysis. As a result, less accurate and imprecise estimates of relative indices could be derived as well as misleading interpretations of species life traits (Thorson 2015). On the contrary, the spatio-temporal index standardization can provide more precise biomass indices than design-based estimators or conventional models by explaining spatial variations (Shelton et al. 2014).

In particular, biomass in different locations are assumed to have distinct expected values based on the environment and spatial terms, and biomass at nearby sites are more similar than the ones at geographically distant locations. In contrast, conventional models assume the mean of an area is fixed and all locations within that area provide exchangeable samples of a single mean. Thus, the classical estimated of survey and CPUE biomass indices are often more sensitive to outlier observations (Shelton et al. 2014).

In line with this, models that ignored the spatio-temporal structure failed to adequately capture the behavior of the simulated stock biomass over time and lead to inaccurate series of relative indices. Note that the error measures (RMSE and MAPE) for all models was magnified in the CPUE biomass indices compared to the survey biomass indices due to the nature of the sampling, intensifying the importance of considering spatio-temporal and sample dependence. Besides, for the modeling of CPUE biomass indices, the marked point process has managed to considerably reduce the overestimation of the fish stock biomass compared to the rest of the models (Diggle et al. 2010, Pennino et al. 2019), always having in mind that the remaining models do not take into account the preferential sampling.

However, Ducharme-Barth et al. (2022) argues that if the preferentiality is not very strong, a geostatistical model can achieve satisfactory results in capturing the behavior of the biomass of the fish stock (Izquierdo et al. 2022). In our results, we observed a high correlation coefficient among all the models in this scenario of preferentiality, which highlights how the intensity of the preference and also the type of dependence can affect the results, since this preferenciality may also depend on many other covariates independent from fish distribution Alglave et al. (2022), Girardin et al. (2017).

Maunder et al. (2020) already emphasized the need for more complex models, especially for the standardization of CPUE biomass data. Indeed, Hoyle et al. (2024) argue that all fish populations, in one way or another, exhibit spatial structure, and therefore, it is important to consider it in models, although there are several ways to introduce this variability. Likewise, Paradinas et al. (2022b) discusses how considering a spatial effect can improve the prediction and smooth the effect of variables not considered in the modeling.

Furthermore, efforts to implement more complex frameworks in real-world applications have been undertaken in the last years to address various challenges related to relative indices modeling. These challenges include addressing issues like handling zero-inflation data and the modeling of preferential sampling (Alglave et al. 2023, Izquierdo et al. 2021, Rufener et al. 2021). Future research could focus on evaluating the importance of considering the spatial effect and sampling dependence according to the degree of influence that these effects have in the simulated scenario.

We envision important topics for present and future applications of spatio-temporal estimation methods. Besides scientific surveys, information registered by on-board observers is frequently used to build indices of biomass for later integration into stock assessment (Maunder et al. 2020). Despite several constrains (i.e. fishing activity is nonrandomly spatially distributed, unfished strata, differences in catchability among boats, etc.), in particular scenarios like small-scale fisheries, fishery-dependent data can provide a better spatio-temporal coverage of fishing activity to depict spatial variation in coastal areas or seasonal trends. In addition, spatio-temporal methods could be used to estimate density for different size or age-classes of many other stocks. These estimates could then be processed to generate age or size-composition data for assessment models. A model-based approach to

estimate age- or size-composition may be more statistically efficient for species with spatial segregation of size or age groups (e.g. life history stages).

In conclusion, the simulation and modeling framework here developed has been successful for the evaluation of which modeling of survey and CPUE biomass data best captures the stock biomass behavior over time. Indeed, the results obtained highlight that failure to consider the underlying spatiotemporal process and the sampling dependence lead to less accurate survey and CPUE biomass indices to inform stock assessment models, which may have a negative impact on the stock status assessment.

Acknowledgements

The authors are grateful to the Centro de Supercomputación de Galicia (Cesga) for access to its advanced computing infrastructure and services, which were necessary to carry out this research. Alba Fuster-Alonso acknowledges institutional support of the 'Severo Ochoa Center of Excellence' accreditation (CEX2019-000928-S).

Author contributions

A.F.A., D.C., and M.G.P. conceptualized the work. A.F.A. developed and implemented the model approach, performed the analysis, and generated visualizations. M.C.R, I.P. and F.I. contributed to implementing the model approach. M.G.P. and D.C. envisioned the study and supervised the work. A.F.A. drafted the manuscript. All authors reviewed the manuscript the final manuscript.

Supplementary data

Supplementary material is available at the *ICES Journal of Marine Science* online version of the manuscript.

Conflict of interest: The authors have no conflict of interest to declare.

Funding

This study is a contribution to the project IMPRESS (RTI2018-099868-B-I00), ERDF, Ministry of Science, Innovation and Universities - State Research Agency and also of the projects financed by the European Union-Next Generation EU. Componente 3. Inversión 7. CONVENIO ENTRE EL MINISTERIO DE AGRICULTURA, PESCA, Y ALIMENTACIÓN Y LA AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS M.P. -A TRAVÉS DEL INSTITUTO ESPAÑOL DE OCEANOGRAFÍA- PARA IMPULSAR LA INVESTIGACIÓN PESQUERA COMO BASE PARA LA GESTIÓN PESQUERA SOSTENIBLE. Eje4, FishClim: Conocimiento científico para la adaptación al cambio climático del sector pesquero español and Eje6, Math4Fish: Nuevas herramientas para el modelado matemático en el asesoramiento científico de pesquerías españolas. Also, we were supported by the GAIN [Agencia Gallega de Innovación] - Xunta de Galicia, GRC-MERVEX (IN607A 2022/04). David Conesa was supported by grant PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades of Spain (MCIN/AEI/10.13039/501100011033/FEDER, UE) and the European Regional Development Fund. David Conesa was also funded by grant CIAICO/2022/165 funded by Generalitat Valenciana. Alba Fuster-Alonso received funding from the Spanish project ProOceans (Ministerio de Ciencia e Innovación, Proyectos de I + D + I (RETOSPID2020-118097RB-I00)). A.F.-A. was supported by Ministerio de Ciencia e Innovación, Grant no. PRE2021-099287 from the project ProOeans (PID2020-118097RB-I00). Alba Fuster-Alonso was supported by the Barcelona municipal government through the iMARES research group at the Institute of Marine Sciences (ICM-CSIC) in Barcelona.

References

Alglave B, Rivot E, Etienne MP *et al.* Combining scientific survey and commercial catch data to map fish distribution. *ICES J Mar Sci* 2022;**79**:1133–49.

Alglave B, Vermard Y, Rivot E *et al.* Identifying mature fish aggregation areas during spawning season by combining catch declarations and scientific survey data. *Can J Fish Aquat Sci* 2023;**80**:808–24.

Anderson SC, Ward EJ. Black swans in space: modelling spatiotemporal processes with extremes. *Ecology* 2019;**100**:e02403. https://doi.org/10.1002/ecy.2403.

Anderson SC, Ward EJ, English PA *et al.* sdmtmb: an r package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *bioRxiv* 2022; **p. 2022.03. 24.485545**

Arreguín-Sánchez F, Catchability: a key parameter for fish stock assessment. *Rev Fish Biol Fish* 1996;**6**:221–42.

Bachl FE, Lindgren F, Borchers DL *et al.* `inlabru`: an `r` package for Bayesian spatial modelling from ecological survey data. *Methods Ecol Evol* 2019;**10**:760–6.

Belitz C, Brezger A, Kneib T *et al.* 2022. *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*. Version 1.1.

Cao J, Thorson JT, Richards RA *et al.* Spatiotemporal index standardization improves the stock assessment of northern shrimp in the gulf of maine. *Can J Fish Aquat Sci* 2017;**74**:1781–93.

Chiarini M, Guicciardi S, Angelini S *et al.* Accounting for environmental and fishery management factors when standardizing CPUE data from a scientific survey: A case study for nephrops norvegicus in the pomo pits area (central adriatic sea). *PloS ONE* 2022;**17**:e0270703.

Conn PB, Thorson JT, Johnson DS. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods Ecol Evol* 2017;**8**:1535–46.

Cousido-Rocha M, Pennino MG, Izquierdo F *et al.* Surplus production models: a practical review of recent approaches. *Rev Fish Biol Fish* 2022;1–18.

Cressie N. Geostatistics. *Am Stat* 1989;**43**:197–202.

Diggle PJ, Menezes R, Su Tl. Geostatistical inference under preferential sampling. *J R Stat Soc Ser C (Appl Stat)* 2010;**59**:191–232.

Ducharme-Barth ND, Grüss A, Vincent MT *et al.* Impacts of fisheries-dependent spatial sampling patterns on catch-per-unit-effort standardization: A simulation study and fishery application. *Fish Res* 2022;**246**:106169.

Girardin R, Hamon KG, Pinnegar J *et al.* Thirty years of fleet dynamics modelling using discrete-choice models: What have we learned? *Fish Fish* 2017;**18**:638–55.

Gunderson DR. *Surveys of fisheries resources*. New York: John Wiley and Sons, 1993.

Hardin JW, Hardin JW, Hilbe JM *et al.* *Generalized linear models and extensions*. USA: Stata press, 2007.

Hastie TJ. Generalized additive models. In: *Statistical models in S*, 1st edn. London: Routledge, 2017, 249–307.

Hazin HG, Hazin F, Travassos P *et al.* Standardization of swordfish CPUE series caught by Brazilian longliners in the Atlantic Ocean, by GLM, using the targeting strategy inferred by cluster analysis. *Collect Vol Sci Pap ICCAT* 2007;**60**:2039–47.

Hinton MG, Maunder MN. Methods for standardizing CPUE and how to select among them. *Col Vol Sci Pap ICCAT* 2004;**56**: 169–77.
Hoyle SD, Campbell RA, Ducharme-Barth ND *et al.* Catch per unit effort modelling for stock assessment: A summary of good practices. *Fish Res* 2024;**269**:106860.
Hoyle SD, Lee SI, Kim DN, CPUE standardization for southern bluefin tuna (thunnus maccoyii) in the korean tuna longline fishery, accounting for spatio-temporal variation in targeting through data exploration and clustering. *PeerJ* 2022;**10**:e13951.
Hsu J, Chang YJ, Ducharme-Barth ND. Evaluation of the influence of spatial treatments on catch-per-unit-effort standardization: A fishery application and simulation study of pacific saury in the northwestern pacific ocean. *Fish Res* 2022;**255**:106440.
ICES. Report on the classification of stock assessment methods developed by sisam. ICES Document CM 2012/ACOM/SCICOM: 01, 2012.
Izquierdo F, Menezes R, Wise L *et al.* Bayesian spatio-temporal CPUE standardization: Case study of European sardine (*Sardina pilchardus*) along the western coast of portugal. *Fish Manag Ecol* 2022;**29**:670–80.
Izquierdo F, Paradinas I, Cerviño S *et al.* Spatio-temporal assessment of the European hake (*Merluccius merluccius*) recruits in the northern Iberian Peninsula. *Front Mar Sci* 2021;**8**:1.
Kai M, Spatio-temporal changes in catch rates of pelagic sharks caught by Japanese research and training vessels in the western and central north Pacific. *Fish Res* 2019;**216**:177–95.
Krainski E, Gómez-Rubio V, Bakka H *et al. Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Boca Raton, FL: Chapman and Hall/CRC, 2018.
Lindgren F, Rue H, Lindström J. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *J R Stat Soc Ser B: Stat Methodol* 2011;**73**:423–98.
Martínez-Minaya J, Cameletti M, Conesa D *et al.* Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stoch Environ Res Risk Assess* 2018;**32**:3227–44.
Maunder MN, Punt AE. Standardizing catch and effort data: a review of recent approaches. *Fish Res* 2004;**70**:141–59.
Maunder MN, Punt AE. A review of integrated analysis in fisheries stock assessment. *Fish Res* 2013;**142**:61–74.
Maunder MN, Thorson JT, Xu H *et al.* The need for spatio-temporal modeling to determine catch-per-unit effort based indices of abundance and associated composition data for inclusion in stock assessment models. *Fish Res* 2020;**229**:105–594.
Methot RD Jr, Wetzel CR. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. *Fish Res* 2013;**142**:86–99.
Morley SK, Brito TV, Welling DT. Measures of model performance based on the log accuracy ratio. *Space Weather* 2018;**16**:69–88.
Neis B, Schneider DC, Felt L *et al.* Fisheries assessment: what can be learned from interviewing resource users?. *Can J Fish Aquat Sci* 1999;**56**:1949–63.
Nielsen JR. Methods for integrated use of fisheries research survey information in understanding marine fish population ecology and better management advice: improving methods for evaluation of research survey information under consideration of survey fish detection and catch efficiency. Netherlands: Wageningen University and Research, 2015.
Oshima K, Takeuchi Y, Miyabe N. Standardized bluefin CPUE from the Japanese longline fishery in the atlantic up to 2007. *Collect Vol Sci Pap ICCAT* 2009;**64**:594–612.
Paradinas I, Conesa D, López-Quílez A *et al.* Spatio-temporal model structures with shared components for semi-continuous species distribution modelling. *Spat Stat* 2017;**22**:434–50.
Paradinas I, Giménez J, Conesa D *et al.* Evidence for spatiotemporal shift in demersal fishery management priority areas in the western Mediterranean. *Can J Fish Aquat Sci* 2022a;**79**:1641–54.
Paradinas I, Illian J, Smout S. Understanding spatial effects in species distribution models. *Authorea Preprints* 2022b;**18**:e0285463.
Pennino MG, Conesa D, Lopez-Quilez A *et al.* Fishery-dependent and-independent data lead to consistent estimations of essential habitats. *ICES J Mar Sci* 2016;**73**:2302–10.
Pennino MG, Izquierdo F, Paradinas I *et al.* Identifying persistent biomass areas: The case study of the common sole in the northern Iberian waters. *Fish Res* 2022;**248**:106196.
Pennino MG, Paradinas I, Illian JB *et al.* Accounting for preferential sampling in species distribution models. *Ecol Evol* 2019;**9**:653–63.
Peterman RM. Statistical power analysis can improve fisheries research and management. *Can J Fish Aquat Sci* 1990;**47**:2–15.
Pons M, Domingo A, Sales G *et al.* Standardization of CPUE of loggerhead sea turtle (caretta caretta) caught by pelagic longliners in the southwestern atlantic ocean. *Aquat Living Resour* 2010;**23**:65–75.
Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)* 2009;**71**:319–92.
Rufener MC, Kristensen K, Nielsen JR *et al.* Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. *Ecol Appl* 2021;**31**:e02453.
Santos R, Crespo O, Medeiros-Leal W *et al.* Error distribution model to standardize lpue, cpue and survey-derived catch rates of target and non-target species. *Modelling* 2022;**3**:1–13.
Shelton AO, Thorson JT, Ward EJ *et al.* Spatial semiparametric models improve estimates of species abundance and distribution. *Can J Fish Aquat Sci* 2014;**71**:1655–66.
Shirota S, Gelfand AE. Preferential sampling for bivariate spatial data. *Spat Stat* 2022;**51**:100674.
Stock BC, Ward EJ, Eguchi T *et al.* Comparing predictions of fisheries bycatch using multiple spatiotemporal species distribution model frameworks. *Can J Fish Aquat Sci* 2020;**77**:146–63.
Tagliarolo M, Cope J, Blanchard F. Stock assessment on fishery-dependent data: Effect of data quality and parametrisation for a red snapper fishery. *Fish Manage Ecol* 2021;**28**:592–603.
Thorson JT. Spatio-temporal variation in fish condition is not consistently explained by density, temperature, or season for california current groundfishes. *Mar Ecol Prog Ser* 2015;**526**:101–12.
Thorson JT. Guidance for decisions using the vector autoregressive spatio-temporal (vast) package in stock, ecosystem, habitat and climate assessments. *Fish Res* 2019;**210**:143–61.
Tremblay-Boyer L, McKechnie S, Pilling G *et al.* Geostatistical analyses of operational longline CPUE data. 2017. Technical Report: Technical Report WCPFC-SC13-2017/SA-WP-03, Rarotonga, Cook Islands, 8-16 August2018.
Umlauf N, Adler D, Kneib T *et al.* Structured additive regression models: An R interface to BayesX. *J Stat Softw* 2015;**63**:1–46.
Watson J. A perceptron for detecting the preferential sampling of locations and times chosen to monitor a spatio-temporal process. *Spat Stat* 2021;**43**:100500.
Xu H, Thorson JT, Methot RD *et al.* A new semi-parametric method for autocorrelated age-and time-varying selectivity in age-structured assessment models. *Can J Fish Aquat Sci* 2019;**76**:268–85.
Xu L, Mazur M, Chen X *et al.* Improving the robustness of fisheries stock assessment models to outliers in input data. *Fish Res* 2020;**230**:105641.
Zhou S, Campbell RA, Hoyle SD. Catch per unit effort standardization using spatio-temporal models for Australia's eastern tuna and billfish fishery. *ICES J Mar Sci* 2019;**76**:1489–504.

Handling Editor: Stan Kotwicki