Standardised CPUE indices of abundance for horse mackerel off South Africa based on Desert Diamond midwater trawling

Dawit Yemane[1], Deon Durlhotz[1], Tracey Fairweather[1], and Larvika Singh[1]

2024-12-04 06:16:51

Contents

1 - *Department of Forestry, Fisheries (DFFE) & the Environment*

1 Abstract

This report provides preliminary results for standardized CPUE indices of abundance generated for the horse mackerel stock targeted by the mid-water trawler Desert Diamond. Given the spatio-temporal nature of the data, the standardized index of abundance was generated based on a model that takes advantage of this information to learn about the long-term trend in the abundance of modelled stock. For the purpose of this work, a spatio-temporal Generalized Linear Mixed Effect Model (GLMM) was applied to assess the potential effects of the data used and the spatial domain of prediction grid. Multiple fits were carried out where different portions of the data were included in the mode, and the index of abundance was also computed on two prediction grids with different spatial extents. The results suggest that the effect of the spatial extent of the data and the prediction grid is minima, with the predicted index of abundance varying minimally in response to the different data-related settings.

2 Background

Standardized indices of abundance from fisheries-dependent sources are one of the inputs into the assessment of the horse mackerel stocks. Previously a non-spatial index of abundance from GLM was used to inform the model on the abundance of the horse mackerel stock. Recently, given that the data is spatial in nature, and recent observations of a drop in the catch of horse mackerel by the single mid-water trawler, Desert Diamond, there has been a request to explore the utility of the spatio-temporal data collected to develop standardized index of abundance (CPUE). In this report, results from the exploratory analysis of the spatio-temporal catch data within a framework of spatio-temporal Generalizd Linear Mixed Effect Models implemented in the R package *sdmTMB* (Anderson *et al.*, 2024).

3 Methods

A map of the study region, the south coast of South Africa, is shown in Figure 1.

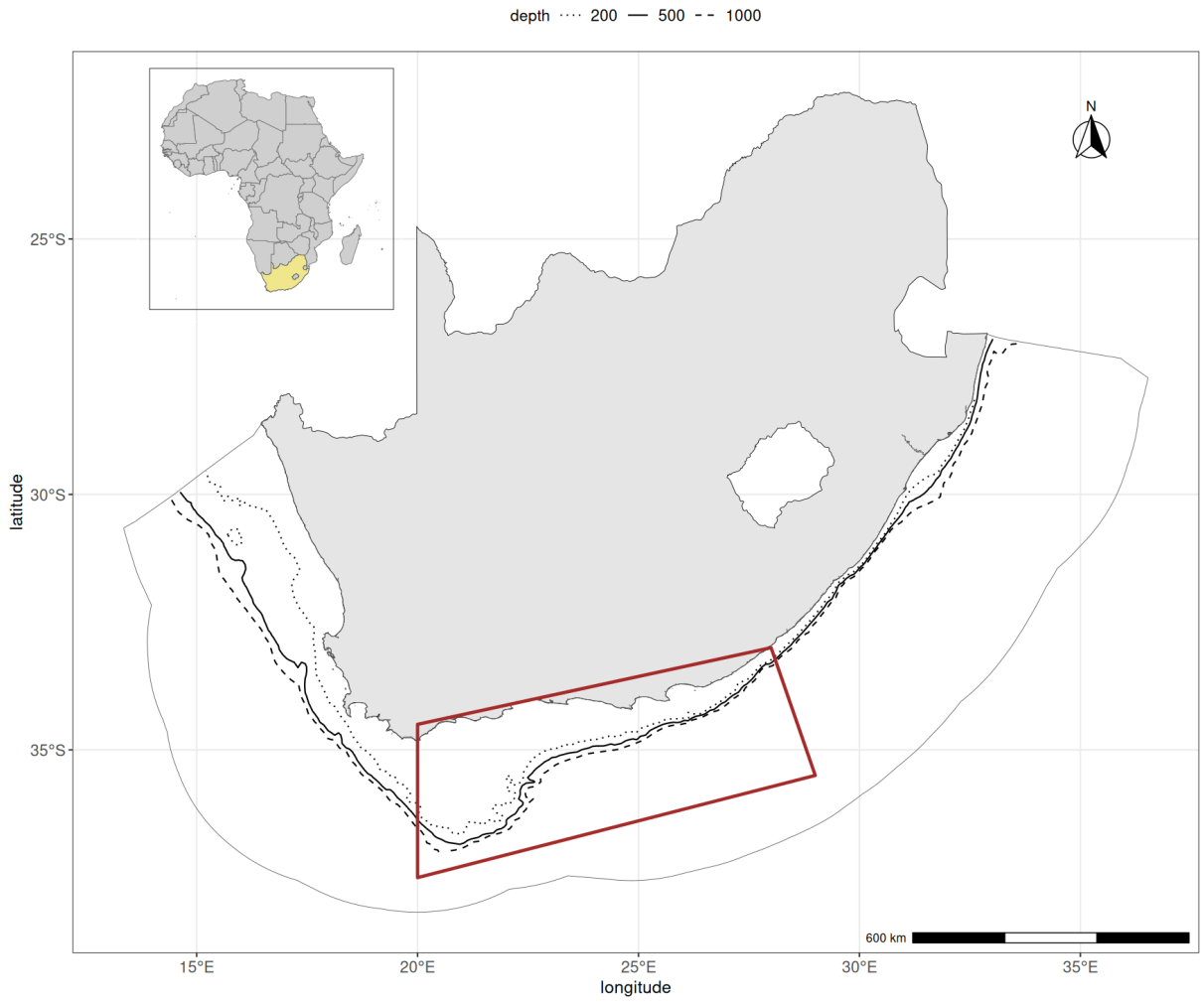

Figure 1: Map of the South Africa showing the horse mackerel study region. African continent as inset.

Prior to the modelling fitting process, multiple data filtration/exclusions were undertaken so as to retain only reliable sets of catch records: checked for missing values, unreasonable values for numbers of hours of fishing (e.g more than five hours of fishing) and excluding fishing activity from demersal trawl (so restricting to mid-water trawl).

In addition to that noted above the catch data used in this report were taken through a followup data filtration step:

- 2^{nd} step filtering:
  - exclude records with negative effort
  - retain only records with minutes fished between 10 and 1000
  - retain only records where fishing took place between 100m and 1000m

Results from initial exploration of the horse mackerel catch over the entire time series, to summarise the overall spatial pattern in the catch location, are shown in Figure 2 while the corresponding year specific catch location and magnitude are shown in Figure 3.

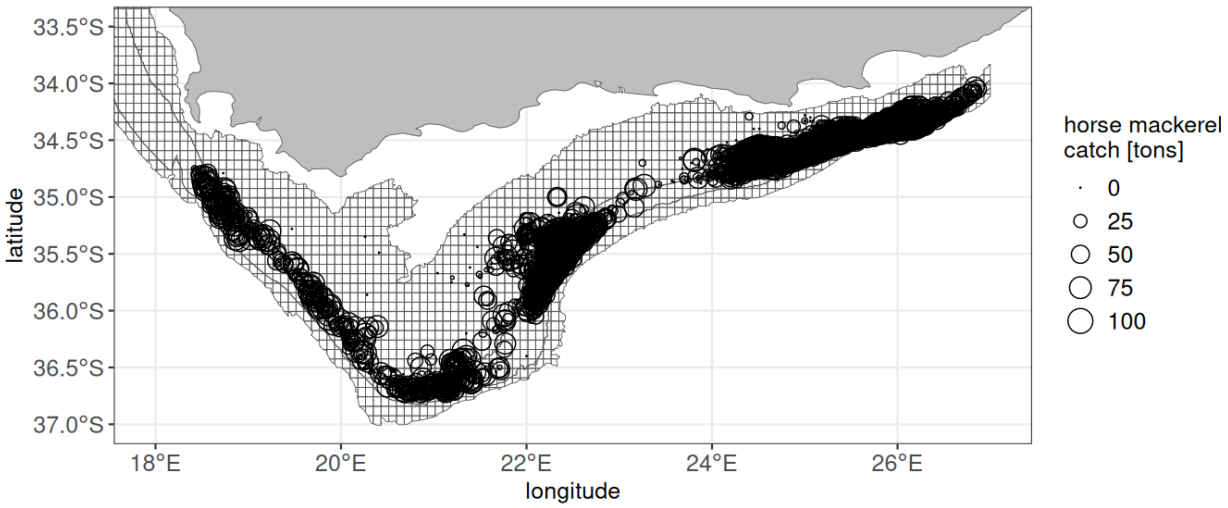

Figure 2: Pattern in horse mackerel catch level for the entire time series.

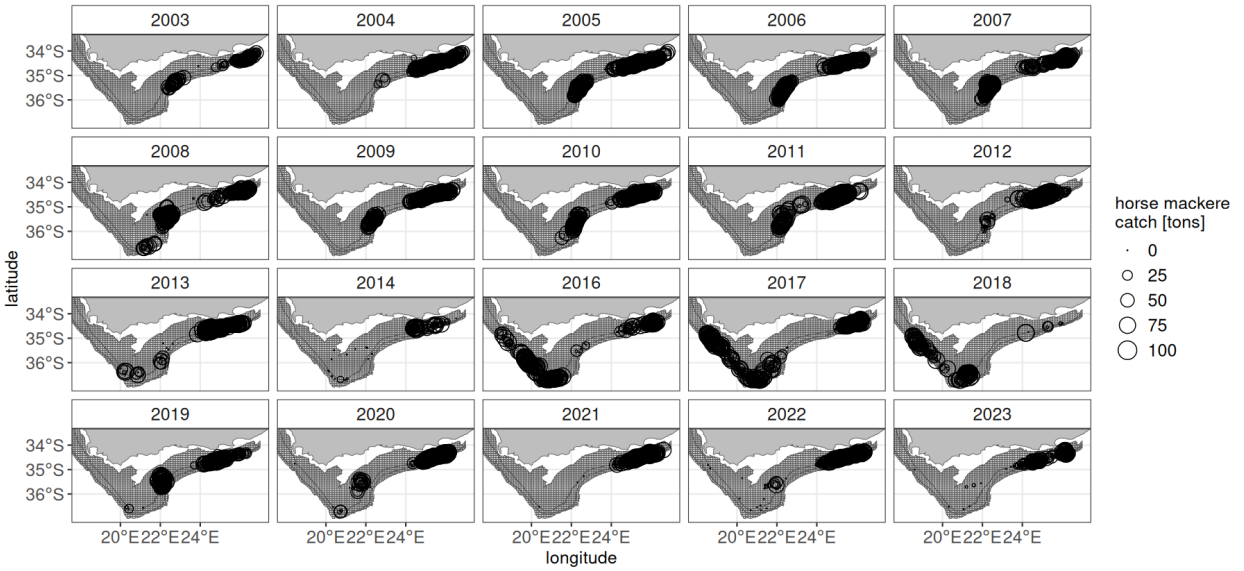

Figure 3: Spatio-temporal pattern in horse mackerel catch.

3.0.1 CPUE standardization

Given the spatio-temporal nature of the data, any potential method to be used to generate a standardized index of abundance needs to account for the effect of location of the catches and how these catch locations vary over time. In addition, the relative influence of additional covariates, e.g month, fishing depth, and other relevant variables needs to be included in the model. Although there are multiple modelling frameworks that can be used, for the purpose of this report a spatio-temporal GLMM implemented in the R package *sdmTMB* (Anderson *et al.*, 2024) was used. Although **sdmTMB**, as the name suggests species distribution model implemented in TMB, was initially intended to be used for modelling species distribution, it has since been extended to be used in a range context including generation of index of abundance both from fisheries dependent and independent sources (Anderson *et al.*, 2024). For the purpose of this study horse mackerel data from the commercial mid-water trawl fishery (*Desert Diamond*) on the south coast of South Africa were used. To properly model the spatial process the coordinates (longitude and latitude), these were projected to UTM zone 35, the zone in which most observations fall (spatial process are modeled with respect to distance). The standard GLMM with covariate effects and spatial and spatio-temporal component takes the form:

$$\mathbb{E}[y_{s,t}] = \mu_{s,t}$$
$$\mu_{s,t} = f^{-1}(\mathbf{X_{s,t}}B + \omega_s + \epsilon_{s,t})$$

where $\mathbb{E}[y_{s,t}]$ is the expected value of the observation, in this case catch at location s and time t; $\mathbf{X_{s,t}}$ is the design matrix for the main effect (e.g. covariates, fixed effect of year); β is the vector of coefficients for the main effects; $\mu_{s,t}$ the mean of the observation, here catch, at location s, and t; f^{-1} is the link function linking the mean response to the predictors (it allows to model response in multiple space: *logit*, *log*, *inverse*,and *identity*).

$$\omega_s \sim \mathbf{MVN}(\mathbf{0}, \mathbf{\Sigma}_\omega)$$

$$\epsilon_{s,t} \sim \mathbf{MVN}(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$$

where ω_s is the spatial random effect, and Σ_ω is the covariance of the spatial random field; $\epsilon_{s,t}$ is the spatio-temporal random field and $\Sigma_{\epsilon_{s,t}}$ is the covariance of the spatio-temporal random field.

Model	formula
$Model_1$	$catch \sim Year + s(depth) + month + offset(log(Hrs))$
$Model_2$	$catch \sim Year + s(depth) + offset(log(Hrs))$
$Model_3$	$catch \sim Year + offset(log(Hrs))$

The proportion of zero values over the entire time series ranged between 7% to 53% per year. Although not specifically applicable to the horse mackerel mid-water fishery, given that it is not zero dominated, when one is dealing with observations that have substantial amount of zeros, the most commonly followed approaches are either to use hurdle model, where two sub-models are fitted to the data (modelling probability of encounter/occurrence and modelling positive observations) and combine them, or to use Tweedie distribution. In the current release of *sdmTMB*, multiple types of hurdle models are implemented including delta-lognormal and delta-gamma.

All the analysis, visualisation and report generation were conducted in R (R Core Team, 2024). Multiple R packages were utilised for data processing, visualization, analysis and summary of results including (Allaire *et al.*, 2024; Anderson *et al.*, 2024; Letaw, 2015; Maechler *et al.*, 2023; Pebesma, 2024; Raiche and Magis, 2022; Robinson *et al.*, 2024; Spinu *et al.*, 2023; Wickham *et al.*, 2023, 2024; Wickham and Henry, 2023; Wood, 2023; Xie, 2024).

4 Results

4.1 Model selection

Multiple models were considered for the purpose of this report including: The type of spatio-temporal random fields; distribution family; the covariates sets considered. Of the two different types of spatio-temporal random fields (*iid* vs *ar1*) only few of the models with *ar1* for the spatio-temporal random fields converged. Similarly one of the hurdle model, with year only effect, converged. As can be seen in Table 1 the full model was the best in terms of *AIC*.

Table 1: Comparison of model performance, based on AIC, models with different fixed effect structure. Using all the data ('all'), and data east of 20E ('east_20E').

data_source	family	formulas	anisotropy	spatiotemp	AIC
all	tweedie	0 + as.factor(year) + fmonth + s(depth_scaled, k = 3)	FALSE	iid	64,761.57
	tweedie	0 + as.factor(year) + s(depth_scaled, k = 3)	FALSE	iid	64,900.95
	tweedie	0 + as.factor(year)	FALSE	iid	64,905.22
	delta-gamma	0 + as.factor(year)	FALSE	iid	65,935.38
	tweedie	fmonth	TRUE	ar1	64,751.73
	tweedie	fmonth	FALSE	ar1	64,762.12
east_20E	tweedie	0 + as.factor(year) + fmonth + s(depth_scaled, k = 3)	FALSE	iid	62,233.27
	tweedie	0 + as.factor(year) + s(depth_scaled, k = 3)	FALSE	iid	62,358.70
	tweedie	0 + as.factor(year)	FALSE	iid	62,363.31
	delta-gamma	0 + as.factor(year)	FALSE	iid	63,341.15
	tweedie	fmonth	FALSE	ar1	62,240.12

Visual summary of estimated range from the different models are shown in Figure 4. The range was relatively comparable among the different model, and when data east of 18^oE and data east of 20^oE is used. Although based on data from narrower portion of the distribution of the stock it appears to suggest that horse mackerel appear to be chracterize by patchy distribution.

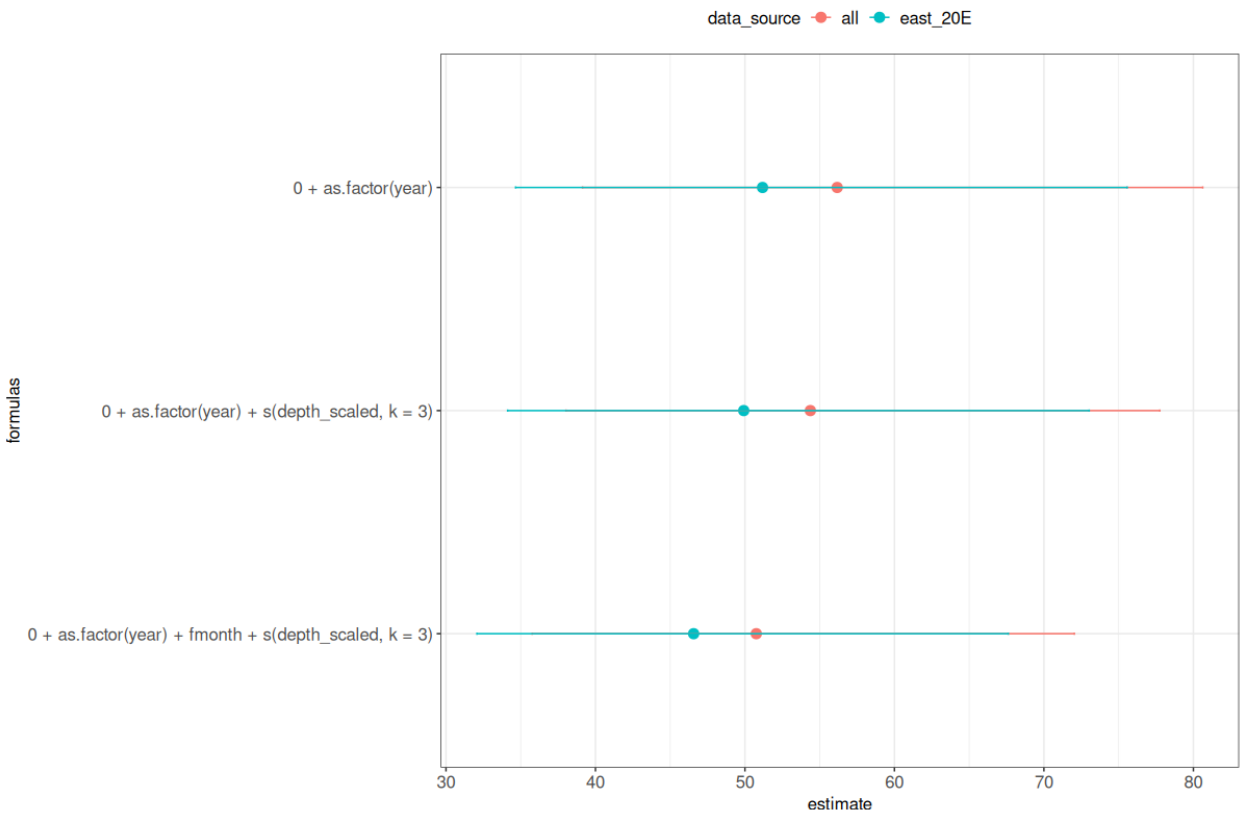

Figure 4: Summary of the estimated range from the different models (Models with different fixed effect structure).

4.2 Model diagnostics

Standard model diagnostics have been checked. To save space, only those corresponding to the model that used the entire data set are presented below. The quantile-quantile plot of residuals, randomized quantile residuals (if the data is consistent with the model residuals should be distributed $N(0, 1)$), for the model that uses all the data is shown in Figure 5 and the corresponding spatio-temporal pattern is shown in Figure 6. Partial effects of bottom depth, normalized (depth_scaled), and month are shown in Figure 7.

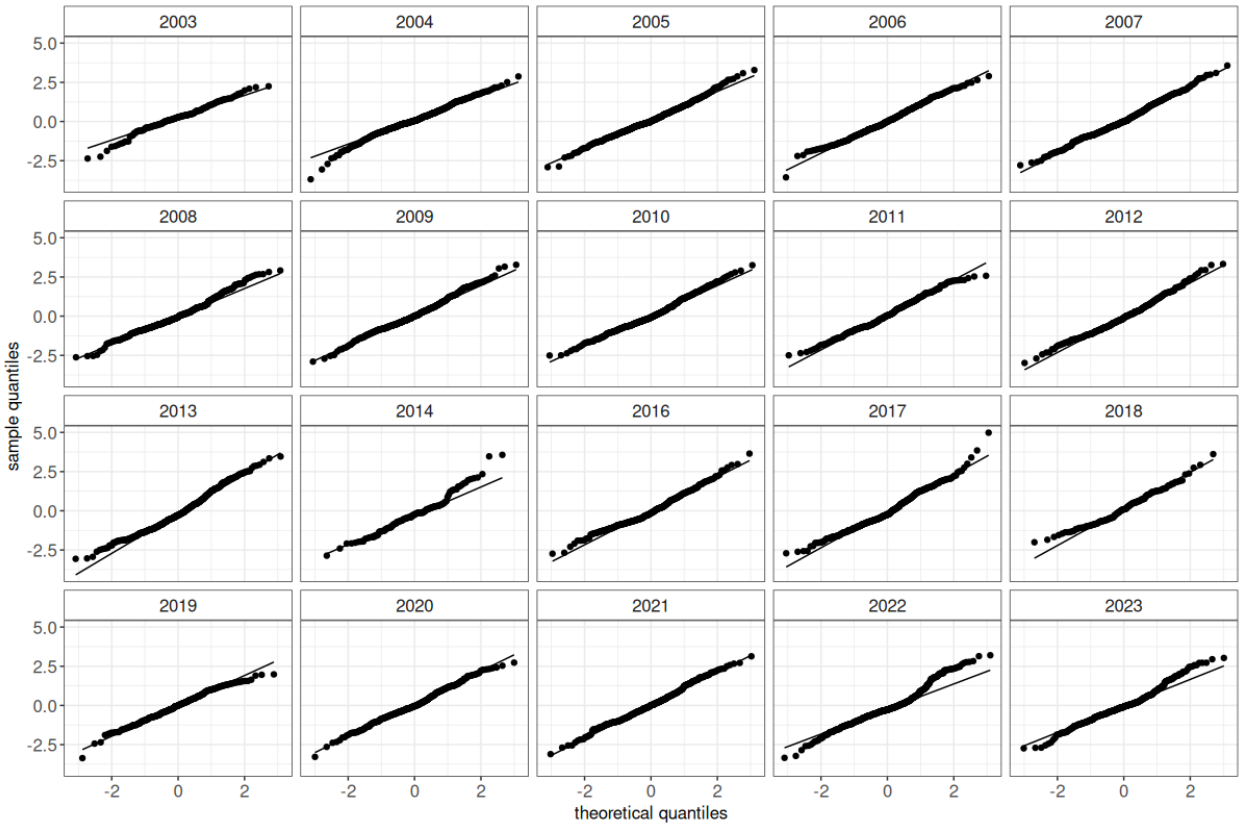

Figure 5: Quantile-quantile plots residuals for the model that uses the entire data sets (east of 18E). Residuals for the model with a all fixed effects.

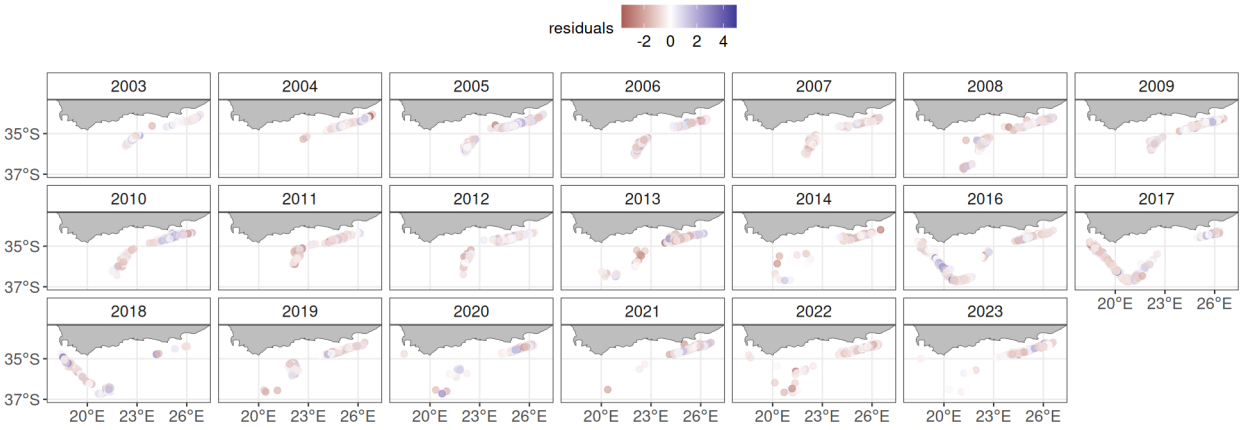

Figure 6: Spatio-temporal pattern in residuals from the model that uses all the data. Residuals for the model with all fixed effects.

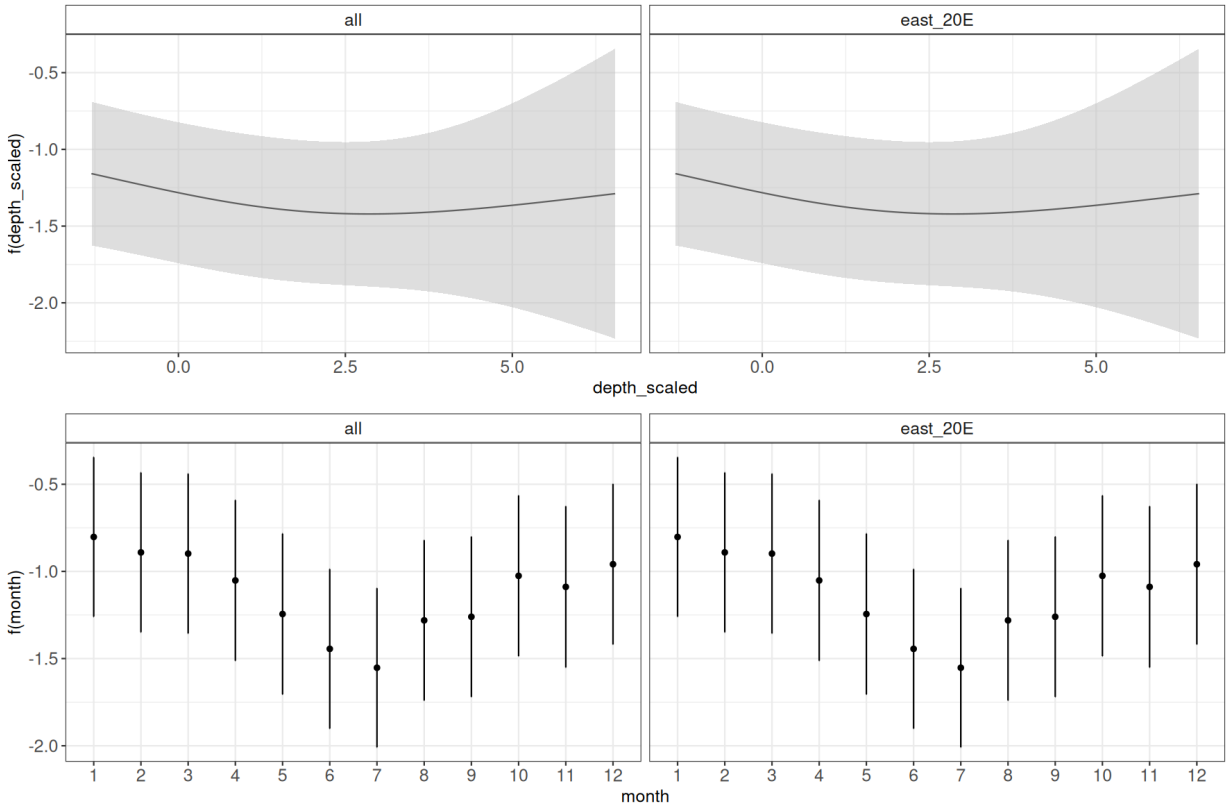

Figure 7: Partial effects of the bottom depth and month from the best model. The models that use all the data are used here.

The predicted density for the best model that uses all the data (18^oE to 27^oE), and fixed effects (year,depth, and month) are shown in Figure 8 and Figure 9.

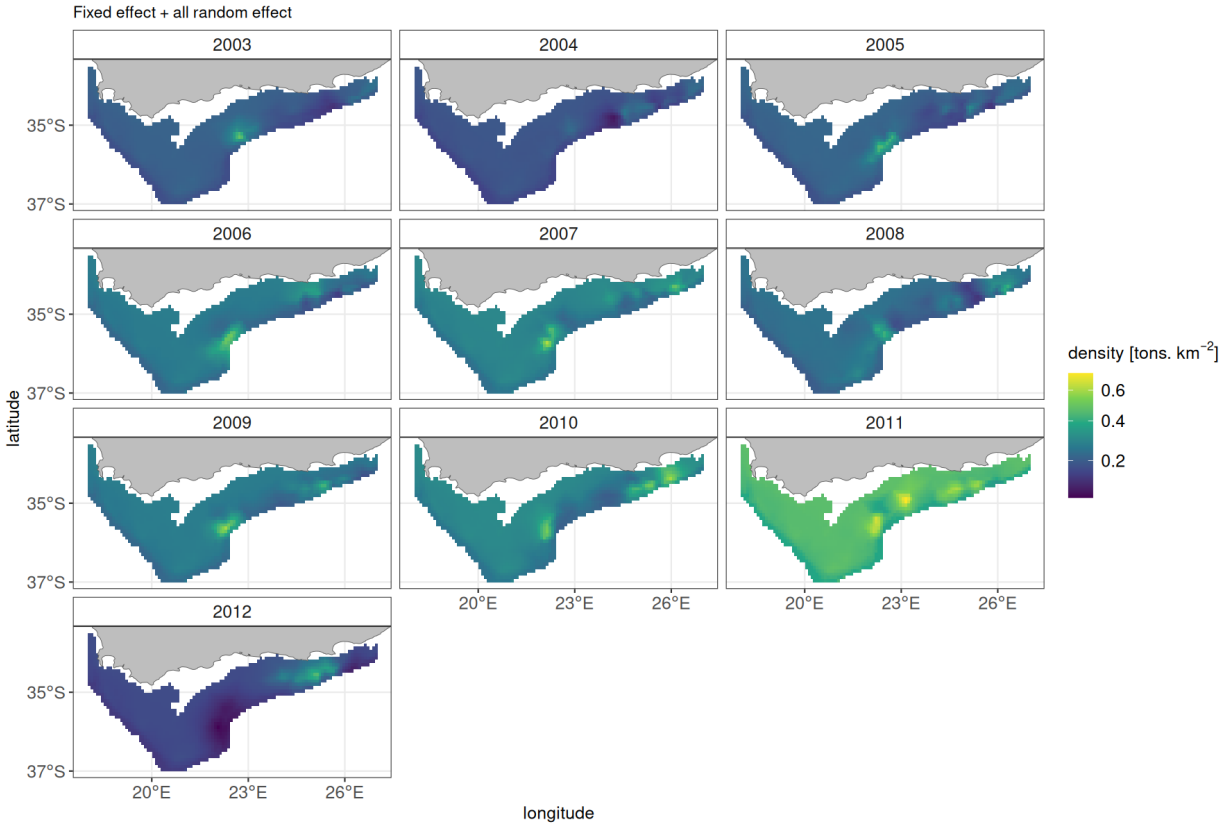

Figure 8: Predictions from the best model, including all fixed and random effects, that uses all data (18E to 27E) for model fitting and prediction. for the years 2003 - 2012

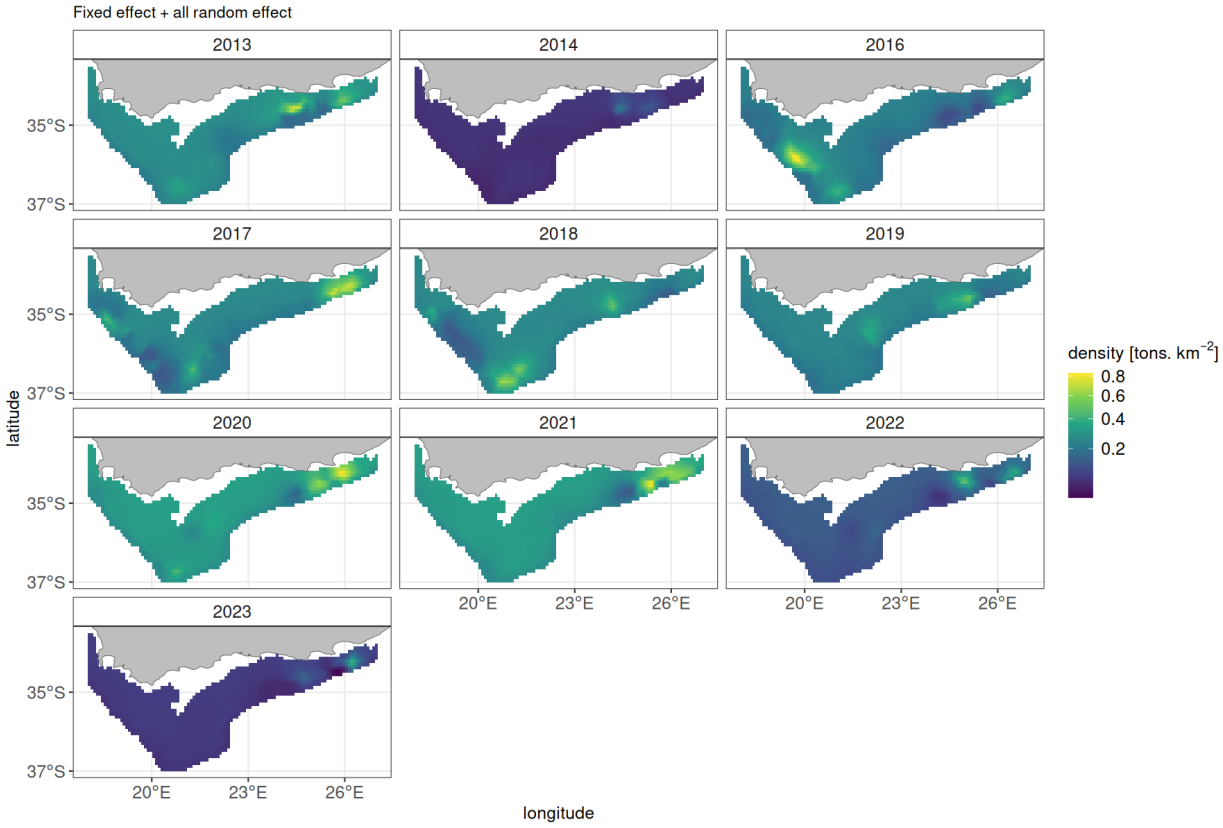

Figure 9: Predictions from the best model, including all fixed and random effects, that uses all data (18E to 27E) for model fitting and prediction. for the years 2013 - 2023.

The prediction error for the model that is based on all the data (18^oE) is shown in Figure 10 and Figure 11. Large part of the prediction grid with limited observation, catch data, is characterized by relatively higher standard error.

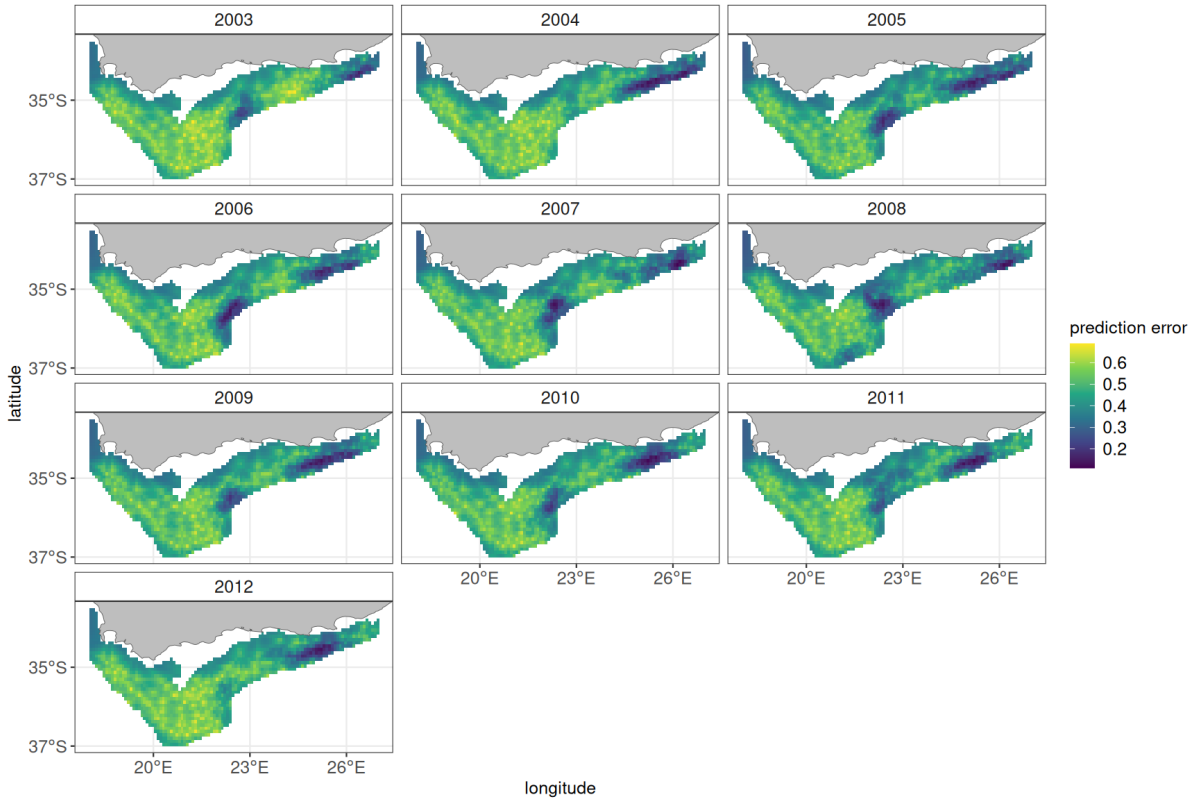

Figure 10: Prediction error from the best model, including all fixed and random effects, that uses all data (18E to 27E) for model fitting and prediction. for the years 2003 - 2012

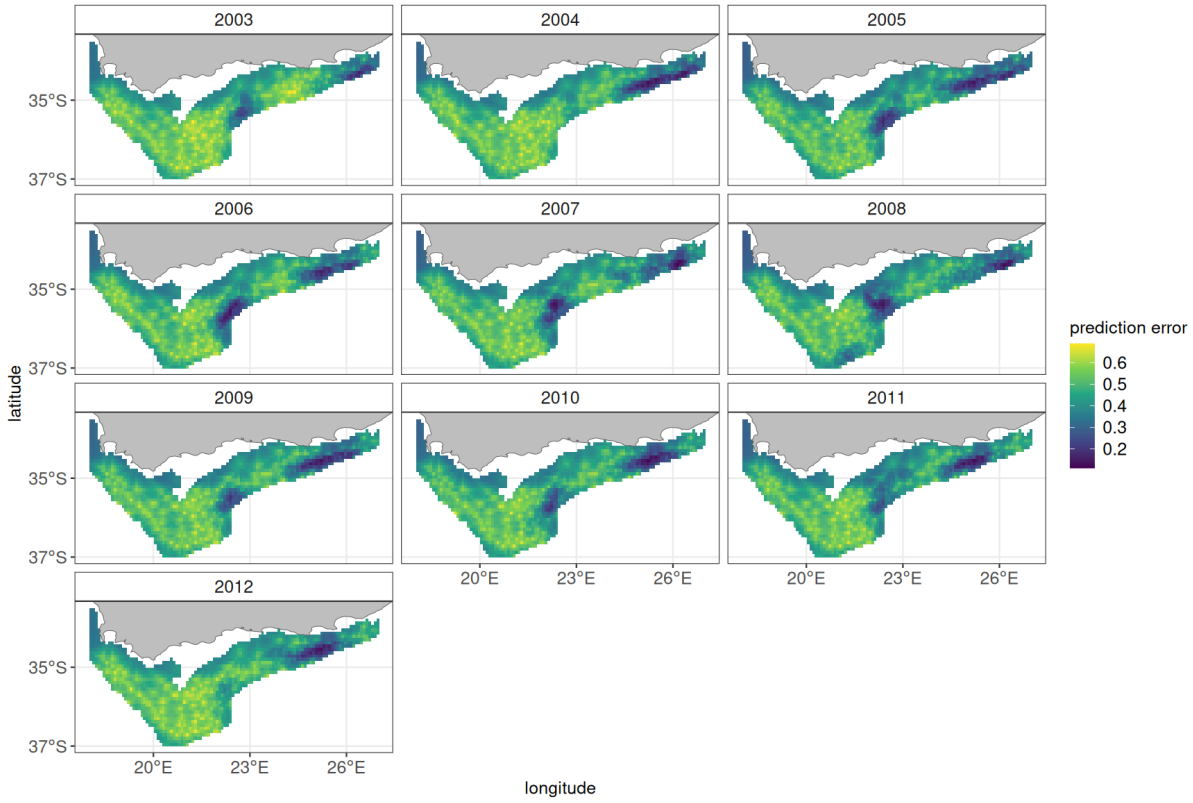

Figure 11: Prediction error from the best model, including all fixed and random effects, that uses all data (18E to 27E) for model fitting and prediction. for the years 2013 - 2023.

In addition for the best model that uses all the data spatial random effect are shown in Figure 12. The spatio-temporal random effect are shown in Figure 13 for 2003 - 2012 and Figure 14 for years 2013 - 2023.

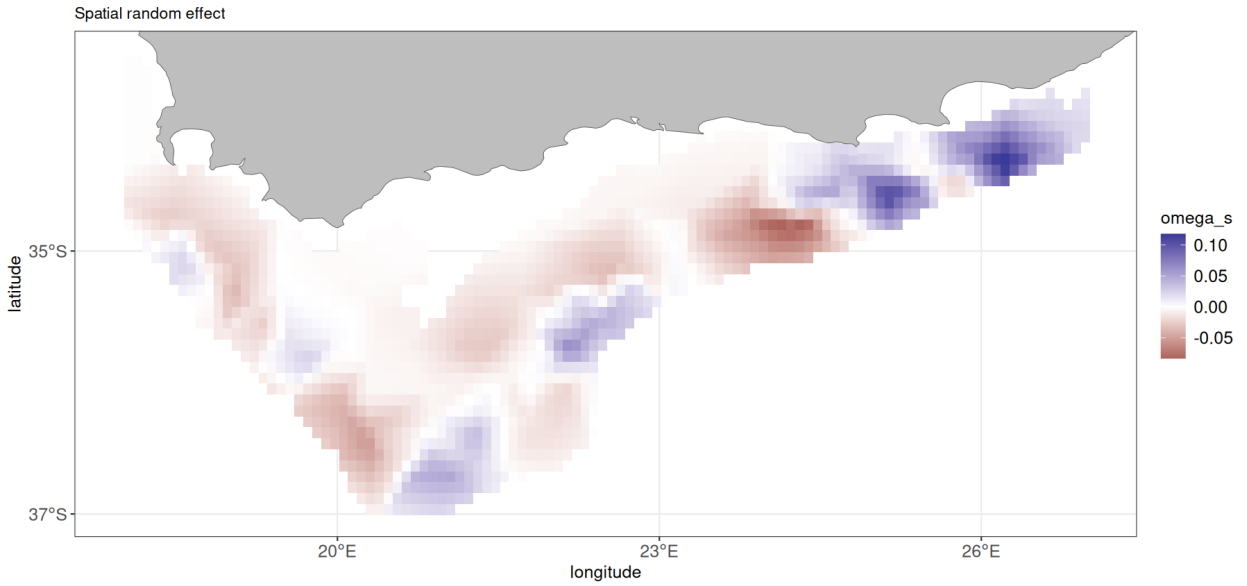

Figure 12: Spatial random effects from the model that uses all the data. The spatial random effect is expected to account for time invariant effects (both biotic and abiotic) that are not taken into account by the current fixed effect structure.

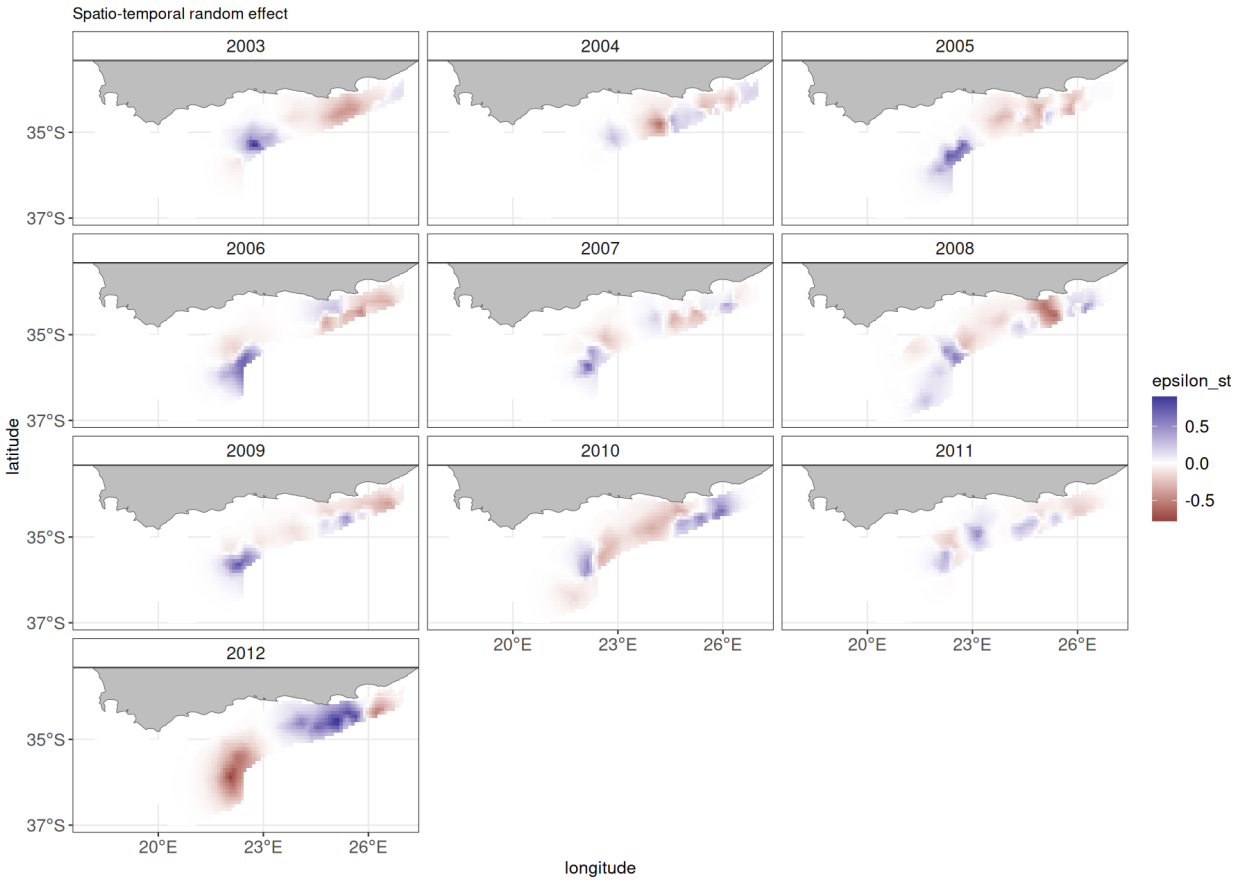

Figure 13: Spatio-temporal random effects accounting for deviation from the fixed effect prediction and spatial random effect. These represent temporally varying biotic and abiotic effects. for the years 2003 - 2012.

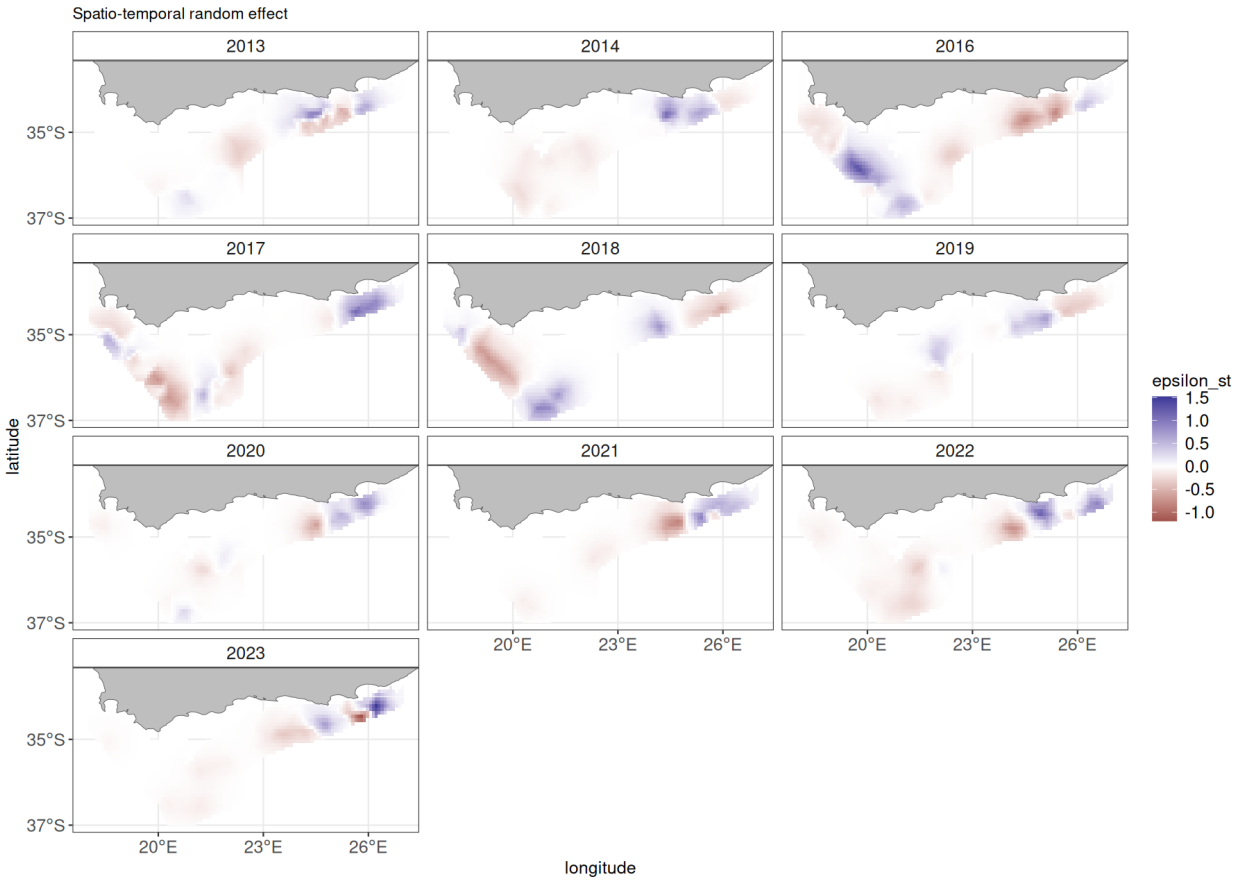

Figure 14: Spatio-temporal random effects accounting for deviation from the fixed effect prediction and spatial random effect. These represent temporally varying biotic and abiotic effects. for the years 2013 - 2023.

4.3 Limiting prediction grid

Although there were some years where catches were taken over the Agulhas bank, the majority were on the shelf-edge thus prediction over most of the prediction grid is extrapolation based on model trained on thin section of the shelf edge. Thus it is not unexpected, as shown in Figure 13 and Figure 14, to see large section of the prediction grid associated with substantial standard error (CV of up to 80%). Thus to show how well normalized index from prediction grid that is more resembling the location of most of the catches to that from using the bigger prediction grid (shown above) the additional prediction were made over smaller regions that resembles the distribution of the catch. This results are shown in Figure 15 and Figure 16 for the entire region $18^oE - 27^oE$ and east of 20^oE respectively.

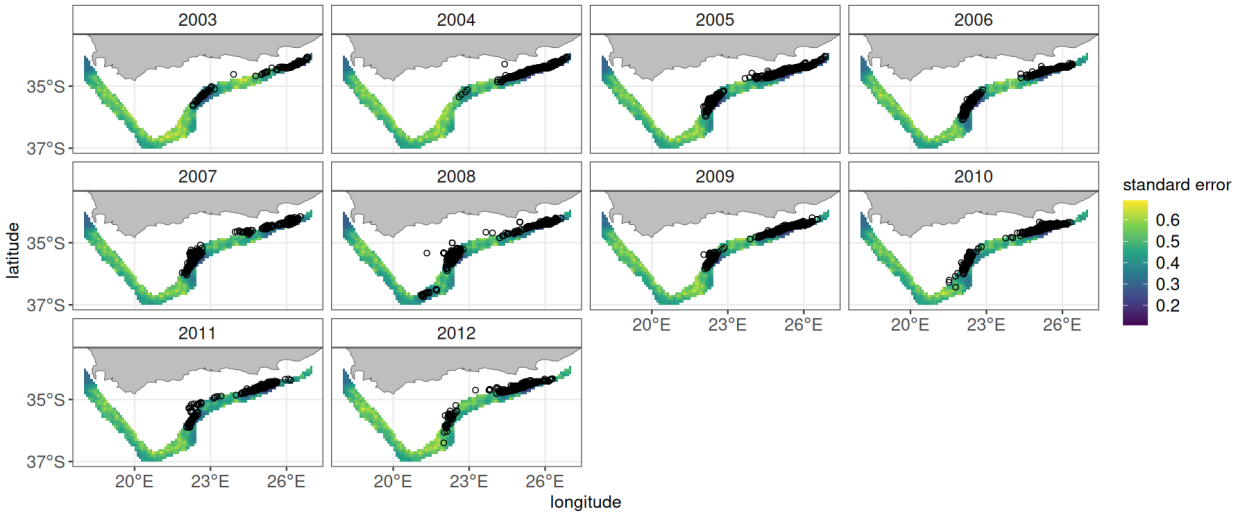

Figure 15: Prediction error from the best model, including all fixed and random effects, that uses all data (18E to 27E) for model fitting but only taking prediction grid deeper than 200m. open circle represent location of catch. for the years 2003 - 2012

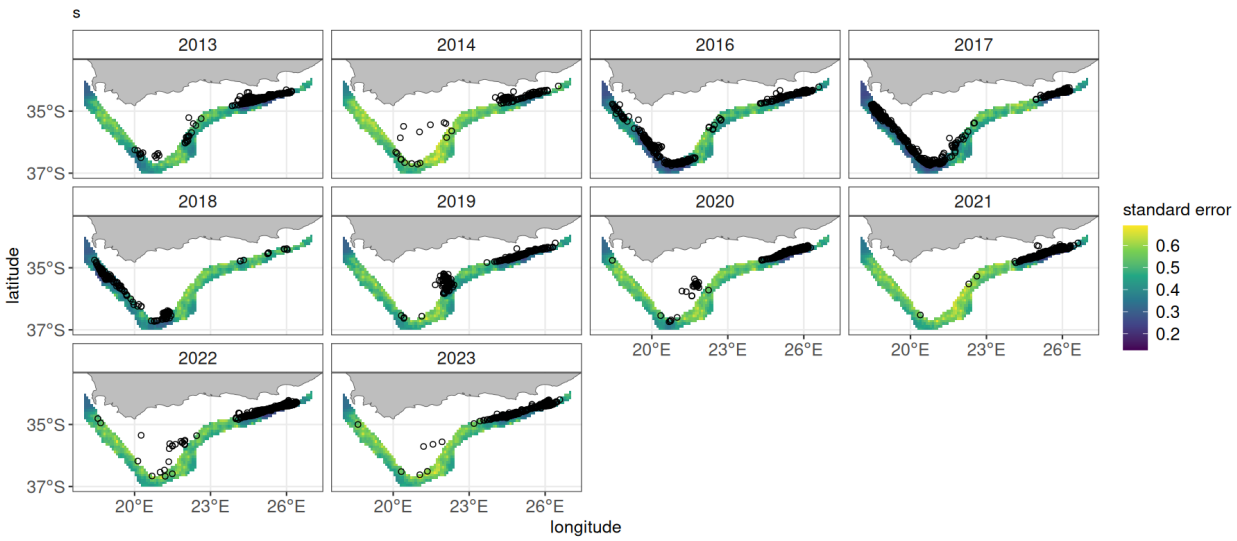

Figure 16: Prediction error from the best model, including all fixed and random effects, that uses all data (18E to 27E) for model fitting but only taking prediction grid deepter than 200m. open circle represent location of catch. for the years 2013 - 2023.

4.4 Standardized indices of abundance

Standardized indices of abundance for horse mackerel are shown in Figure 17. As it can be seen in the figure normalized index based on larger and relatively smaller prediction grid were almost identical.

Figure 17: Standardized index from the best model: the model with Tweedie distribution that uses that uses all fixed effects. Models based on bigger vs smaller prediction grid. Filled points are the standardized index from delta-lognormal GLM (Larvika et al.).

5 Discussion

The result from this work shows the value of spatio-temporal GLMM in generating standardized *CPUE* for the mid-water trawl fishery targeting horse mackerel stock of the south coast South Africa.

6 Session information

Table 2: System and session info for reproducibility

Setting	Value
version	R version 4.4.2 (2024-10-31)
os	Ubuntu 24.04.1 LTS
system	x86_64, linux-gnu
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	Africa/Johannesburg

Table 3: R packages for reproducibility

	Package	Loaded version	Date		Package	Loaded version	Date
1	broom	1.0.7	2024-09-26	13	lubridate	1.9.3	2023-09-27
2	captioner	2.2.3.9000	2024-09-11	14	mgcv	1.9-1	2023-12-21
3	cluster	2.1.6	2023-12-01	15	nFactors	2.4.1.1	2022-10-10
4	devtools	2.4.5	2022-10-11	16	nlme	3.1-165	2024-06-06
5	dplyr	1.1.4	2023-11-17	17	patchwork	1.3.0	2024-09-16
6	flextable	0.9.7	2024-10-27	18	purrr	1.0.2	2023-08-10
7	forcats	1.0.0	2023-01-29	19	readr	2.1.5	2024-01-10
8	ggplot2	3.5.1	2024-04-23	20	sf	1.0-19	2024-11-05
9	ggrepel	0.9.6	2024-09-07	21	stringr	1.5.1	2023-11-14
10	kableExtra	1.4.0	2024-01-24	22	tibble	3.2.1	2023-03-20
11	knitr	1.49	2024-11-08	23	tidyr	1.3.1	2024-01-24
12	lattice	0.22-5	2023-10-24	24	tidyverse	2.0.0	2023-02-22
				25	usethis	3.0.0	2024-07-29

date	2024-12-04
pandoc	3.2 @ /usr/lib/rstudio/resources/app/bin/quarto/bin/tools/x86_64/ (via rmarkdown)

References

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., *et al.* 2024. Rmarkdown: Dynamic documents for r. https://github.com/rstudio/rmarkdown.

Anderson, S. C., Ward, E. J., English, P. A., Barnett, L. A. K., and Thorson, J. T. 2024. sdmTMB: Spatial and spatiotemporal SPDE-based GLMMs with TMB. https://pbs-assess.github.io/sdmTMB/.

Letaw, A. 2015. Captioner: Numbers figures and creates simple captions. https://github.com/adletaw/captioner.

Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M. 2023. Cluster: "Finding groups in data": Cluster analysis extended rousseeuw et al. https://svn.r-project.org/R-packages/trunk/cluster/.

Pebesma, E. 2024. Sf: Simple features for r. https://r-spatial.github.io/sf/.

R Core Team. 2024. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raiche, G., and Magis, D. 2022. nFactors: Parallel analysis and other non graphical solutions to the cattell scree test. https://CRAN.R-project.org/package=nFactors.

Robinson, D., Hayes, A., and Couch, S. 2024. Broom: Convert statistical objects into tidy tibbles. https://broom.tidymodels.org/.

Spinu, V., Grolemund, G., and Wickham, H. 2023. Lubridate: Make dealing with dates a little easier. https://lubridate.tidyverse.org.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., *et al.* 2024. ggplot2: Create elegant data visualisations using the grammar of graphics. https://ggplot2.tidyverse.org.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. 2023. Dplyr: A grammar of data manipulation. https://dplyr.tidyverse.org.

Wickham, H., and Henry, L. 2023. Purrr: Functional programming tools. https://purrr.tidyverse.org/.

Wood, S. 2023. Mgcv: Mixed GAM computation vehicle with automatic smoothness estimation. https://CRAN.R-project.org/package=mgcv.

Xie, Y. 2024. Knitr: A general-purpose package for dynamic report generation in r. https://yihui.org/knitr/.