A comparative study on CPUE standardization of bigeye tuna in the Indian Ocean using multi-scale fisheries data and environment data

Tianjiao Zhang[1,2], Liming Song[1*], Hongchun Yuan[2], Ebango Ngando Narcisse[1]

1 .College of Marine Sciences, Shanghai Ocean University, Shanghai 201306, China

2. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China

Abstract

Environmental and spatial variability could impact the relative abundance of highly migratory species. It becomes especially problematic when the variability affect the standardization of CPUE (catch-per-unit-effort) used to assess the status of fish stocks. This paper presents CPUE standardization and model comparison procedures for bigeye tuna (*Thunnus obesus*) in the Indian Ocean based on multi-scale fisheries data and environment data from 2008 to 2015. We used the fisheries datasets from two sources for comparison: (1) the statistical longline datasets published by IOTC Secretariat with monthly catch-and-effort of the 5^o or 1^o grid; and (2) the survey datasets from the Chinese longline fishery with set by set catch-and-effort data. We calculated multiple marine environmental factors for CPUE standardization models. Beside those frequently used factors, such as sea surface temperature (SST), concentration of sea surface chlorophyll a (Chla), sea surface wind speed (WS), we also calculated factors that could possibly affect the fish distribution habitat but were rarely used in previous CPUE standardization study, such as the vertical ocean temperature and salinity factors based on 15 profiles of the ARGO buoys, the nearest distance between CPUE positions and the SST fronts (DF), and the eddy kinetic energy (EKE) derived from geostrophic velocities. We applied cluster analysis methods to identify suitable environmental locations for the target species. The cluster group parameter was then included as a categorical factor in models. We used generalized linear model (GLM) with lognormal constant analyses for CPUE standardization. We generally built three types of models based on the fishery dataset sources and the inclusion of potential environmental factors. Basically, for the whole region, using IOTC 5^o datasets could capture the underlying bigeye tuna CPUE trends; While

for the R1 region, Chinese set by set fishery dataset at higher resolution improved the model fit for subarea standardization. The inclusion of some environmental variables aided the CPUE standardization process as well. For the whole region models, habitat clusters, WS, Chla, vertical temperature at depth of 0m, 100m, 150m, and 500m, vertical salinity at depth of 5m, 200m and 500m have showed great significance in the related best model; For the R1 region, habitat clusters, EKE, vertical temperature at depth of 50m, 150m, 500m and vertical salinity at depth of 0m, 200m and 500m have contributed to some level to the related model fit. In conclusion, IOTC 5^{o} fishery dataset was in the appropriate scale for the whole region CPUE standardization, and IOTC 1^{o} or set by set fishery dataset in finer scale was suitable for estimating the subarea indices. The meaningful explanatory environmental factors in our models could be served as recommendations for further practices for CPUE standardization within multi-scale regions.

1 Introduction

Fishery-dependent time series of catch per unit effort (CPUE) is often used for estimating indices of fish abundance and therefore is an integral part of the stock assessment process (Forrestal. et al, 2018). Nominal CPUE values are often not proportional to the abundance of the stock as the variations due to changes in the spatial extent of fish population, shifts in fish movement patterns, as well as habitat environmental changes over time (Bigelow et al., 1999). Bigeye tuna (*Thunnus obesus*) is a target species of the tropical longline fishery in the Indian Ocean and the joint CPUE standardization for bigeye tuna has been implemented for years and deepened our understanding of movements, habitat utilization and stock structure of this species in the Indian Ocean (Hoyle et al., 2016). However, the best practices for incorporating environmental variables within appropriate spatial scale in CPUE standardization have not been defined, which adds uncertainty in choosing standardization methods aimed at minimizing CPUE bias.

In this study, we used the fisheries datasets from two sources for comparison: (1) the statistical longline datasets published by IOTC Secretariat with monthly catch-and-effort of

the 5º or 1º grid; and (2) the survey datasets from the Chinese longline fishery with set by set catch-and-effort data. We calculated multiple marine environmental factors for CPUE standardization models. Beside the frequently used factors, such as sea surface temperature (SST), sea surface height (SSH), concentration of sea surface chlorophyll *a* (Chla), we also calculated factors that could possibly affect the fish habitat distribution but were rarely used in previous CPUE standardization study, such as the vertical ocean temperature and salinity factors based on 15 profiles of the ARGO buoys, the nearest distance between CPUE positions and the SST fronts, and the eddy kinetic energy (EKE) derived from geostrophic velocities. We also made clustering analysis to identify the spatial effect on species composition in each cluster group. We used generalized linear model (GLM) with lognormal constant analysis for the CPUE standardization. Totally 15 comparison models were built based on the multi-scale fisheries data and the environmental factors. The goal of this work is to determine the meaningful explanatory environmental variables and the appropriate fishery datasets scales for CPUE standardization.

2 Material and Methods

2.1 Data sources

The study area is defined as 40°S–25°N, 20°E–150°E in the Indian Ocean. The regional structures in the joint analysis for bigeye tuna CPUE standardization (Hoyle et al. 2016) was adopted in this study (Figure 1).

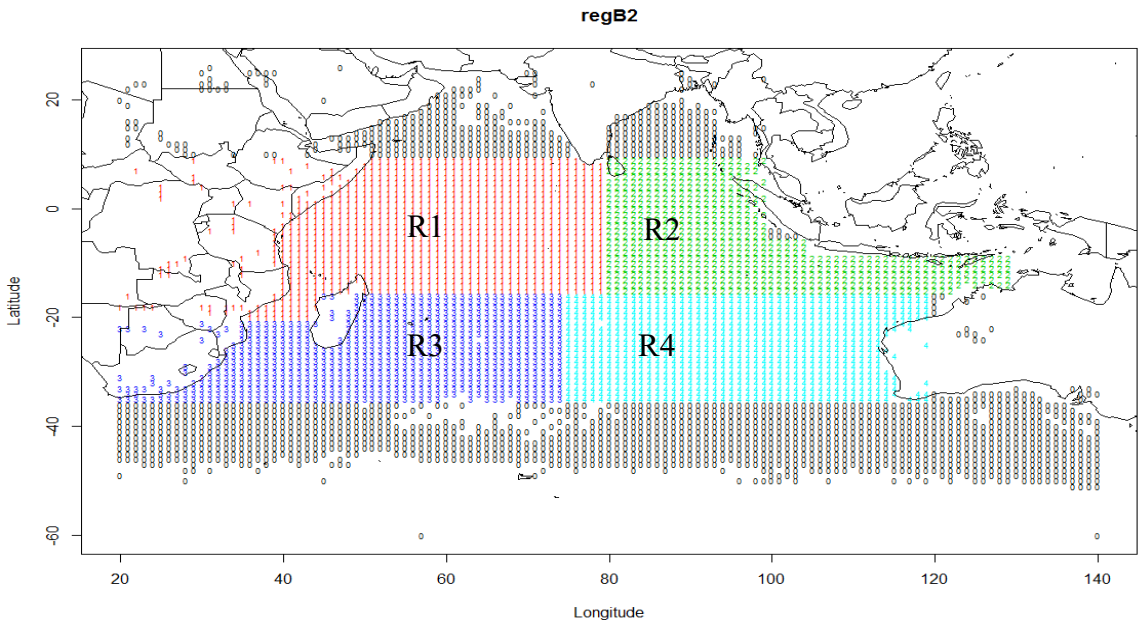


Figure 1. Map of the regional structures used to estimate bigeye tuna CPUE indices in the Indian Ocean.

The bigeye tuna data were collected from two sources: (1) the statistical longline datasets published by IOTC Secretariat. The fields of the datasets include year, month, location (5° or 1° of latitude and longitude), number of hooks, and catch in number and weight of bigeye within the whole study area; (2) Chinese longline fishery with set by set catch-and-effort data. Observations of each set included year, month, date, number of hooks, longitude and latitude, bigeye tuna catch in number, and most of the observations were within R1 region. As the Chinese fishery datasets were only available on discrete months from 2008 to 2015, we collected the IOTC datasets with all fields in the same year period for comparison.

Multiple marine environmental factors were calculated for CPUE standardization models. Beside those frequently used factors to reflect sea surface environment, such as sea surface temperature (SST), concentration of sea surface chlorophyll *a* (Chla), surface wind speed (WS), we also calculated factors that could possibly affect the fish distribution habitat but were rarely used in previous CPUE standardization study, such as the vertical ocean temperature and salinity factors based on 15 profiles of the ARGO buoys, the nearest distance between CPUE positions and the SST fronts, the eddy kinetic energy (EKE) derived from geostrophic velocities. We also considered the geographical area of each 1° or 5° grid as a

factor to minimize the spatial area inconformity between equatorial and middle and high latitudes.

The data source, time range, spatial resolution and data preprocessing for each marine environmental factor are shown in Table 1. VIF methods were used to remove the correlations between the environmental metrics.

Table 1 Marine environment data

Environmental factors	Data source	Duration of data collection	Resolution	Data preprocessing
Sea surface temperature （℃）	NOAA Ocean Watch (https://oceanwatch.pifsc.noaa.gov/)	2008—2015	0.5°	SST, Wind speed and Chla dataset were downloaded in .csv format based on the extent of study area in the related years and were matched with fishery dataset records based on latitude and longitude using R script.
Wind speed (m/s)		2008—2015	0.5°	
Concentration of Chlorophyll *a* (milligrams/m^3)		2008—2015	0.5°	
Vertical Temperature （℃）	Asia-Pacific Data-Research Center (http://apdrc.soest.hawaii.edu/data/data.php/)	2008—2015	1°	The raster datasets of 15 levels depth for both vertical temperature and vertical salinity were downloaded based on the extent of study area in the related years. Each factor was matched with the fishery dataset records based on latitude and longitude using R script.
Vertical Salinity		2008—2015	1°	
EKE (**cm^2/s^2**)	Copernicus Marine Environmental Monitoring	2008—2015	1°	The eastward velocity (u) and northward velocity (v) raster files were downloaded and EKE were calculated with the Raster Calculator function in Spatial Analyst

	Service website (http://marine.copernicus.eu)			extension in ArcGIS
Df (km)	Marine Environmental Monitoring Service website (http://marine.copernicus.eu)	2008—2015	4°	The image of MODIS (SST) were used in the Cayula-Corbillon single-image algorithm in the MGET (Marine Geospatial Ecology Tools) based on the ArcGIS to detect the SST fronts. Histogram analysis were made to detect the formal portion of the edge and calculate the nearest distance between CPUE position and the SST fronts based on ArcGIS "Field Calculator" function.
Grid area (km^2)	Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG)(https://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html)	2015	1°	High resolution continental land masses and ocean islands boundaries .shp files were downloaded. Each 1° grid area was calculated based on Projected Coordinate System: World_Goode_Homolosine_Ocean in ArcGIS.

2.2 Modeling methods

2.2.1 Cluster analysis

Data were aggregated by location and then clustered on species composition in the catch, using the Ward hclust method.

For the IOTC datasets with 5° grids, clustering was carried out for the whole region; For

the IOTC datasets with 1° grids and the Chinese longline fisheries datasets, clustering was carried out for R1 region only.

2.2.2 Selecting the number of groups

Hclust method was used firstly to examine the hierarchical trees and subjectively estimate the number of distinct branches. Then kmeans deviances were then plotted with number of groups k ranging from 1 to 10. The optimal group number was the lowest value of k after which the rate of decline of deviance became slower and smoother.

2.2.3 Selecting vertical environmental factors

We used the variance inflation factor (VIF) (Rose, 1995) to quantify the severity of multicollinearity between vertical environmental variables within each model. We used the 'fmsb' package in R to calculate the VIF values (Nakazawa, 2012). As an initial step, a linear model was created which related the variables to a dummy variable. Based on the linear model, the VIF for each variable was calculated and the variable with the highest value removed. The VIFs were recalculated for the new variable set and again the variable with highest value was removed. This iterative process was repeated until all variables had a VIF < 10.

2.2.4 CPUE standardization

CPUE standardization methods generally followed the approaches used by Hoyle and Okamoto (2011) with some modifications. The operational data were standardized using generalized linear models in R.

Lognormal constant analyses were carried out using generalized linear models that assumed a lognormal distribution. In this approach the response variable log ($CPUE$+0.01) was used, and a Normal distribution assumed.

$\ln(CPUEs+k) \sim covarites + \epsilon$

We generally built three types of models based on the fishery dataset sources.

(1) Model with response variable of CPUE estimated based on the IOTC 5° grids, with model abbreviation: Model 5d series.

The covariates in the models were in three forms:

Model 5d-1, only the related time and spatial location covariates within the whole region

were included, without cluster and without environmental metrics;

Model 5d-2, the related time, spatial location and cluster covariates were included, and without environmental metrics within the whole region;

Model 5d-3, the related time, spatial location covariates and the environmental factors based on VIF analysis were included within the whole region;

Model 5d-4, the same covariates as in Model 5d-3 were used, but the response variables of CPUE were scaled by the geographical area of the IOTC 5° grids.

Model 5d-5, Model 5d-6, and Model 5d-7, the same covariates as in Model 5d-1, 2, 3 were used, but the study area was narrowed to R1.

Model 5d-8, the same covariates as in Model 5d-7 were used for R1.The same amount of catch records (295) as in Chinese fishery dataset were randomly sampled in R1 for comparison.

(2) Model with response variable of CPUE estimated based on the IOTC 1° grids, with model abbreviation: Model 1d series.

The covariates in Model 1d-1, Model 1d-2, and Model 1d-3 were also in three forms as Model 5d-1, Model 5d-2 and Model 5d-3.

Model 1d-4 used the same covariates as in Model 1d-3.The same amount of catch records (295) as in Chinese fishery dataset were randomly sampled in R1 for comparison.

(3) Model with response variable of CPUE estimated based on the Chinese set by set fishery dataset, with model abbreviation: Model Ch series.

The covariates in the models were also in three forms as Model 5d series but the study area was narrowed to R1.

The details for each model were described in Table 2.

Table 2 Description, time period and model function for 15 models

Model	Description	Time Period	Model function
Model 5d-1	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset, and the related time and spatial	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year)+ as.factor(MonthStart)+as.factor(DegreesL

	location covariates within the whole region.		atitude)+as.factor(DegreesLongtitude)
Model 5d-2	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset, and the related time, spatial location and the cluster covariates within the whole region.	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year) +as.factor(MonthStart)+as.factor(DegreesLatitude)+as.factor(DegreesLongtitude)+as.factor(cluster)
Model 5d-3	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF within the whole region.	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year) +as.factor(MonthStart)+as.factor(DegreesLatitude)+as.factor(DegreesLongtitude)+as.factor(cluster)+tem.0m+tem.100m+tem.150m+tem.500m+sal.5m+sal.75m+sal.200m+sal.500m+sea_surface_temperature +wind_speed+chla+EKE+DF+Grid area
Model 5d-4	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset scaled by the grid area, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF within the whole region.	2008-2015	glm(log(bet_cpue+0.01/grid area)~as.factor(Year) +as.factor(MonthStart)+as.factor(DegreesLatitude)+as.factor(DegreesLongtitude)+as.factor(cluster)+tem.0m+tem.100m+tem.150m+tem.500m+sal.5m+sal.75m+sal.200m+sal.500m+sea_surface_temperature +wind_speed+chla+EKE+DF+Grid area
Model 5d-5	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset, and the related time and spatial location covariates within the R1.	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year)+ as.factor(MonthStart)+as.factor(DegreesLatitude)+as.factor(DegreesLongtitude)
Model 5d-6	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year)

	dataset, and the related time, spatial location and the cluster covariates within the R1.		+as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)
Model 5d-7	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF within the R1.	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year) +as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)+tem.0m+tem.100m+tem.150m+tem.500m+sal.5m+sal.75m+sal.200m+sal.500m+sea_surface_temperature +wind_speed+chla+EKE+DF+Grid area
Model5d-8	generalized linear model (GLM) with CPUE estimated based on IOTC 5° fisheries dataset, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF within the R1, with same amount of data records (295) as in Model Ch series randomly sampled in R1.	2008-2015	glm(log(bet_cpue+0.01)~as.factor(Year) +as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)+tem.0m+tem.100m+tem.150m+tem.500m+sal.5m+sal.75m+sal.200m+sal.500m+sea_surface_temperature +wind_speed+chla+EKE+DF+Grid area
Model 1d-1	generalized linear model (GLM) with CPUE estimated based on IOTC 1° fisheries dataset, and the related time and spatial location covariates.	2014	glm(log(bet_cpue+0.01)~as.factor(Year)+ as.factor(MonthStart)+as.factor(DegreesLatitude)+as.factor(DegreesLongtitude)
Model 1d-2	generalized linear model (GLM) with CPUE estimated based on IOTC 1° fisheries dataset, and the related time, spatial location and the cluster covariates.	2014	glm(log(bet_cpue+0.01)~as.factor(Year) +as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)
Model	generalized linear model (GLM) with	2014	glm(log(bet_cpue+0.01)~as.factor(Y

1d-3	CPUE estimated based on IOTC 1° fisheries dataset, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF.		ear) +as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)+tem.0m+tem.50m+tem. 100m+tem.200m+sal.0m+sal.50m+sal.10 0m+sal.150m+sal.500m+sea_surface_tem perature+wind_speed+chla+EKE+DF+Gr id area
Model 1d-4	generalized linear model (GLM) with CPUE estimated based on IOTC 1° fisheries dataset, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF, with same amount of data records (295) as in Model Ch series randomly sampled in R1.	2014	glm(log(bet_cpue+0.01)~as.factor(Y ear) +as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)+tem.0m+tem.50m+tem. 100m+tem.200m+sal.0m+sal.50m+sal.10 0m+sal.150m+sal.500m+sea_surface_tem perature+wind_speed+chla+EKE+DF+Gr id area
Model Ch-1	generalized linear model (GLM) with CPUE estimated based on Chinese set by set fisheries dataset, and the related time and spatial location covariates.	2008,2 009,2012,20 13,2014,201 5 (with scatted months)	glm(log(bet_cpue+0.01)~as.factor(Y ear)+ as.factor(MonthStart)+as.factor(DegreesL atitude)+as.factor(DegreesLongtitude)
Model Ch-2	generalized linear model (GLM) with CPUE estimated based on Chinese set by set fisheries dataset, and the related time, spatial location and the cluster covariates.	2008,2 009,2012,20 13,2014,201 5 (with scatted	glm(log(bet_cpue+0.01)~as.factor(Y ear) +as.factor(MonthStart)+as.factor(Degrees Latitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)

		months)	
Model Ch-3	generalized linear model (GLM) with CPUE estimated based on Chinese set by set fisheries dataset, and the related time, spatial location, the cluster covariates and the environmental factors selected based on VIF.	2008,2009,2012,2013,2014,2015 (with scatted months)	glm(log(bet_cpue+0.01)~as.factor(Year) +as.factor(MonthStart)+as.factor(DegreesLatitude)+as.factor(DegreesLongtitude)+ as.factor(cluster)+tem.0m+tem.50m+tem.300m+sal.0m+sal.75m+sal.150m+sea_surface_temperature+wind_speed+chla+EKE+DF+Grid area

2.2.5 Model evaluation and comparison

Firstly, an Akaike Information Criterion was computed for each model to assess model relative fit: the lower the AIC, the better the model (Akaike, 1974).

Secondly, we examined several deviance- based quantities (null, residual and explained) as a proxy of the model reliability to standardize CPUE indices. A high explained deviance can indicate a good fit, whereas a high null deviance and a high residual deviance can indicate a bad one.

Finally, we plotted Q-Q plot of residuals to evaluate the absolute goodness- of-fit of the models, with x-axis represent the theoretical values of the quantiles of the standard normal distribution and y-axis shows the empirical values. In a perfect case, the Q-Q plot should show dots following a straight 45° line.

3 Result

3.1 Cluster analysis

The aim of the cluster analysis was to identify locations fit for each species. The hclust trip and kmeans set methods separated the 5° locations into 5 clusters in the whole region; the 1 degree locations into 4 clusters in the R1 region; and the set by set locations into 4 clusters in the R1 region (Fig 2).

Species compositions were plotted for each cluster (Fig 3). The spatial distributions of clusters in the whole region were shown in Fig 4.

For the whole region, cluster 4 and 5 consisted mostly of bigeye tuna; cluster 3 was for high proportion of yellowfin tuna; cluster 1 and 2 was mainly for albacore tuna. The proportions of swordfish catches in all clusters were obviously lower than other species. The map of the whole region showed that cluster 4 and 5 with high caches of bigeye tuna mainly concentrated tropical waters. Cluster 3 with high proportion of yellowfin tuna mainly concentrated western subtropical and temperate waters. Albacore tuna catches occupied most of the subtropical and temperate waters, while the lowest proportion of swordfish scatted on the whole region.

For the R1 region, four cluster groups generated from IOTC 1° dataset all showed a high proportion of bigeye tuna. Cluster 2 consisted of more yellowfin tuna. The proportions of albacore tuna and swordfish catches in all clusters were obviously lower than other species. Cluster 1 with the highest proportion of bigeye tuna didn't show a congregation on the map, while cluster 2 with a mix of bigeye tuna and yellowfin tuna concentrated in western tropical waters.

Clusters based on the Chinese set by set fishery dataset differentiate species compositions more clearly. Cluster 4 consisted mostly of bigeye tuna and cluster 3 consisted mostly of yellowfin tuna. Cluster 1 was mixed by high percentage of yellowfin tuna, few bigeye tuna and swordfish, while Cluster 2 was mixed by high percentage of both bigeye tuna and yellowfin tuna. The map showed that cluster 4 and 2 with relative high catch of bigeye tuna concentrated in tropical and western subtropical waters, while cluster 3 with high yellowfin tuna concentrated in the southeastern part of region R1.

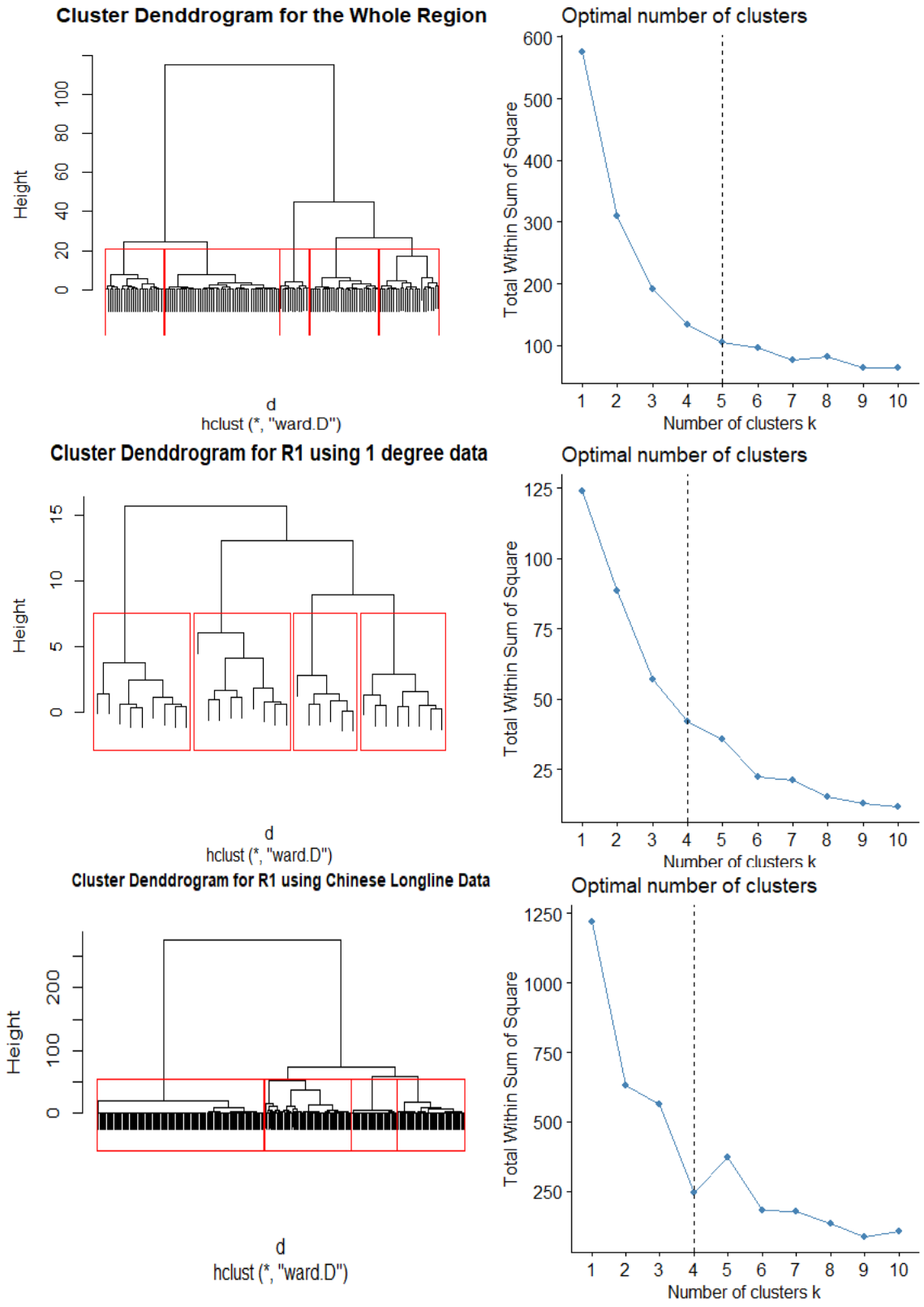


Fig 2: Cluster analysis for the whole region with 5° dataset (top), R1 region with 1° dataset (middle) and Chinese set by set datasets (bottom); The hierarchical Ward clustering analysis to estimate the number of distinct classes of species composition (left); The total sums of squares within-group from kmeans analyses with a range of numbers of clusters (right).

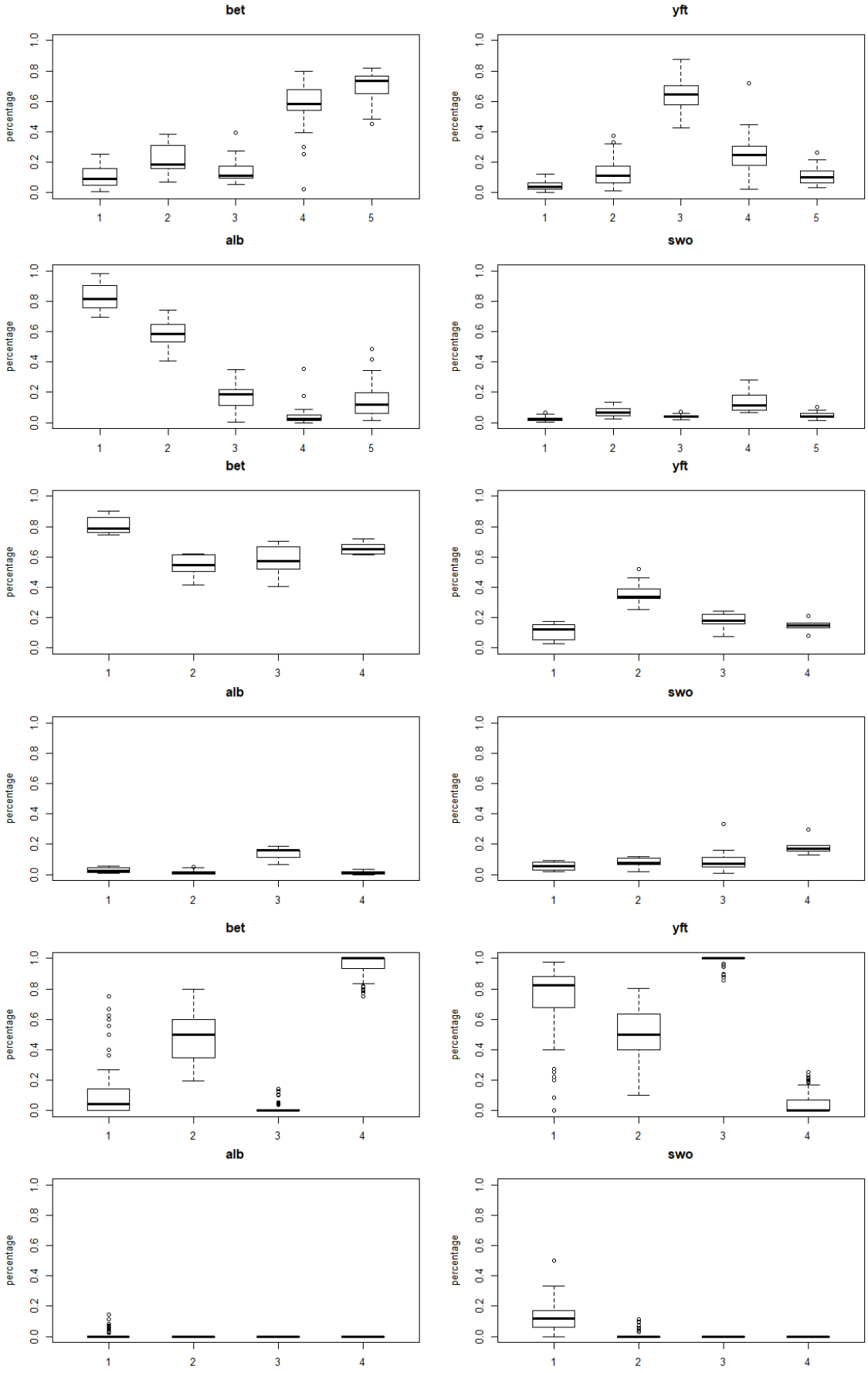

bet
yft
alb
swo
bet
yft
alb
swo
bet
yft
alb
swo
percentage

Fig 3: Boxplots showing species composition by cluster in the whole region based on IOTC 5 °fishery dataset (top four figures); species composition by cluster in the R1 region based on IOTC 1 °fishery dataset (middle four figures); species composition by cluster in the R1 region based on Chinese set by set dataset (bottom four figures);

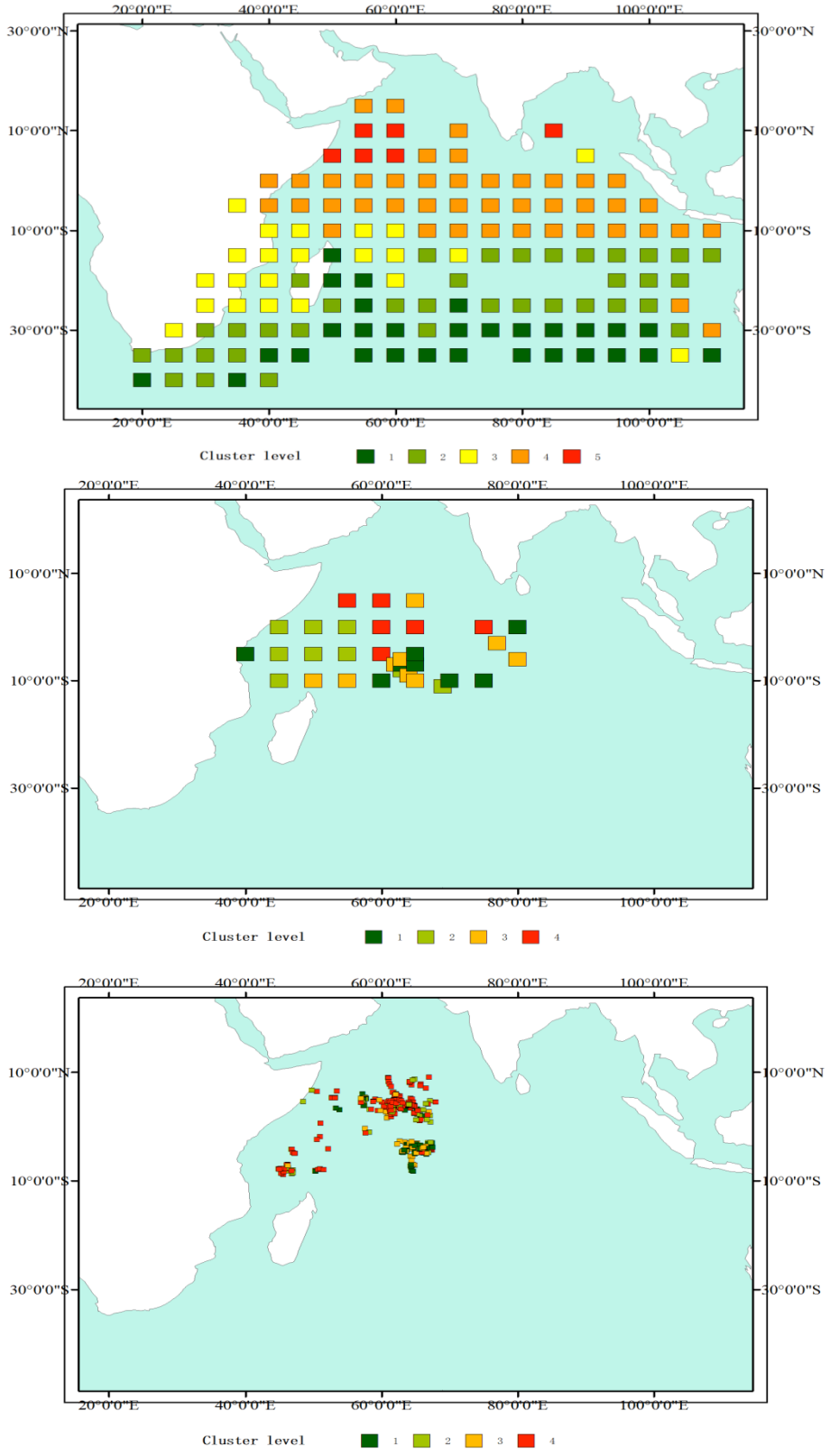


Fig 4: Maps of the spatial distributions of clusters in the whole region and R1 region

3.2 Environmental factor correlations

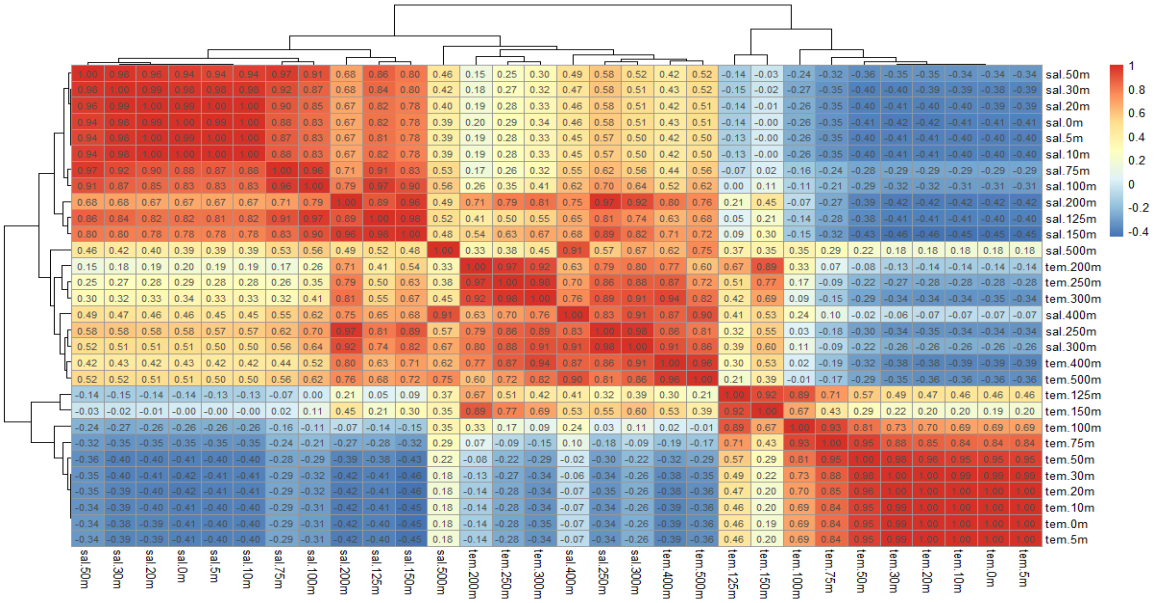
The correlations among the vertical environmental variables within each model series were shown in Fig 5.

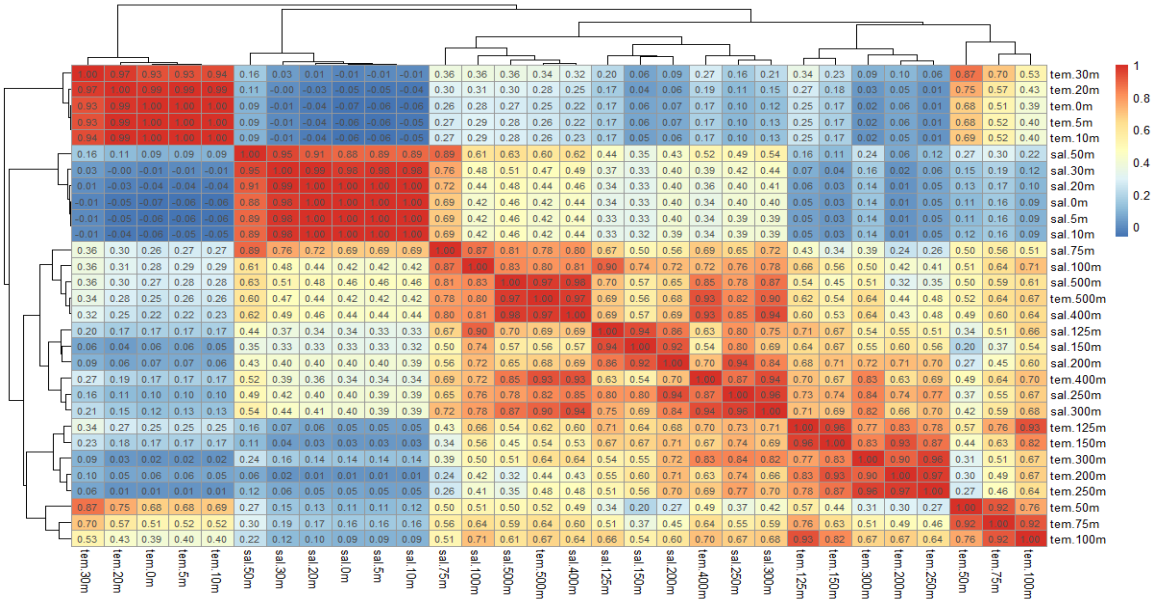
In the Model 5d series, high correlations were observed within the salinity factors of the upper 150m; within the tem factors of the upper 125m; within the salinity and the temperature at the depth of 200~500m;

In the Model 1d series, there were even higher correlations (spearman' rho >0.9) within the salinity of the upper 50m, the temperature of the upper 50m and 125~300m. High correlations also existed in the salinity and the temperature at the depth of 400~500m.

In the Model Ch series, there were also high correlations (spearman' rho >0.9) among the salinity of the upper 50m and the temperature of the upper 50m. The correlations among the salinity and the temperature at the depth of 300~500m were even higher than those in Model 1d series.

The selection procedures based on VIF considerably reduced the number of vertical variables. The removal of correlated variables resulted in Model 5d series containing 8 variables, Model 1d series containing 9 variables, and Model Ch series containing only 6 variables. The variables remained in each model series could basically represent the vertical environmental condition at the depth above 500m and the correlations among the variables have been reduced substantially (Fig 6).


sal.50m
sal.30m
sal.20m
sal.0m
sal.5m
sal.10m
sal.75m
sal.100m
sal.200m
sal.125m
sal.150m
sal.500m
tem.200m
tem.250m
tem.300m
sal.400m
sal.250m
sal.300m
tem.400m
tem.500m
tem.125m
tem.150m
tem.100m
tem.75m
tem.50m
tem.30m
tem.20m
tem.10m
tem.0m
tem.5m
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4


tem.30m
tem.20m
tem.0m
tem.5m
tem.10m
sal.50m
sal.30m
sal.20m
sal.0m
sal.5m
sal.10m
tem.75m
tem.100m
tem.500m
sal.400m
sal.125m
sal.150m
tem.200m
tem.400m
sal.250m
sal.300m
tem.125m
tem.150m
tem.300m
tem.200m
tem.250m
tem.50m
sal.75m
sal.100m
sal.500m
tem.100m
1
0.8
0.6
0.4
0.2
0

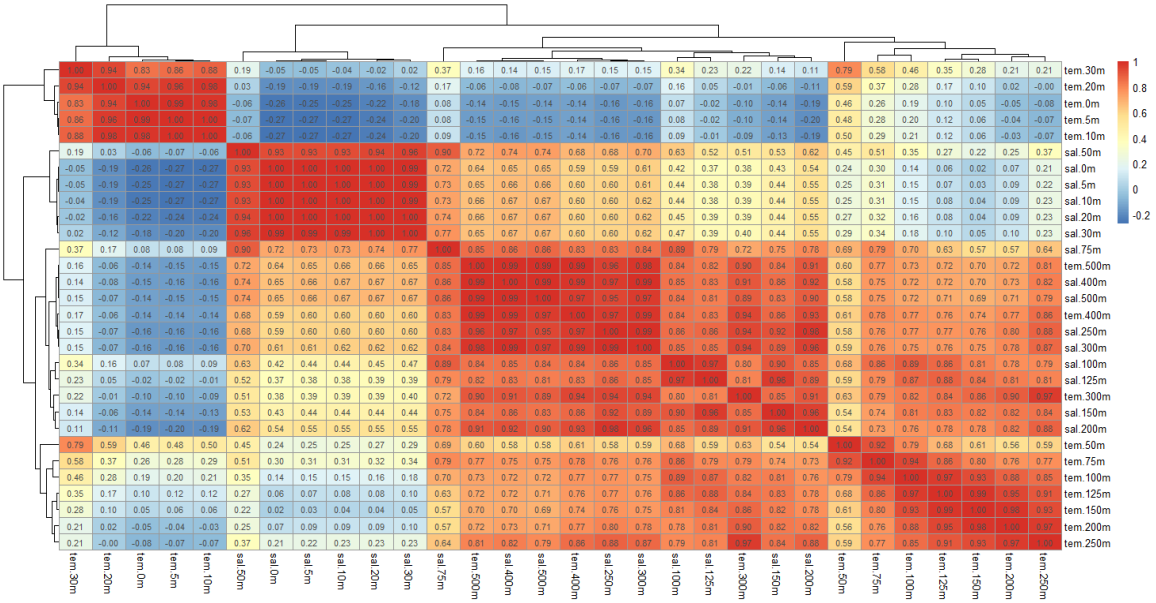
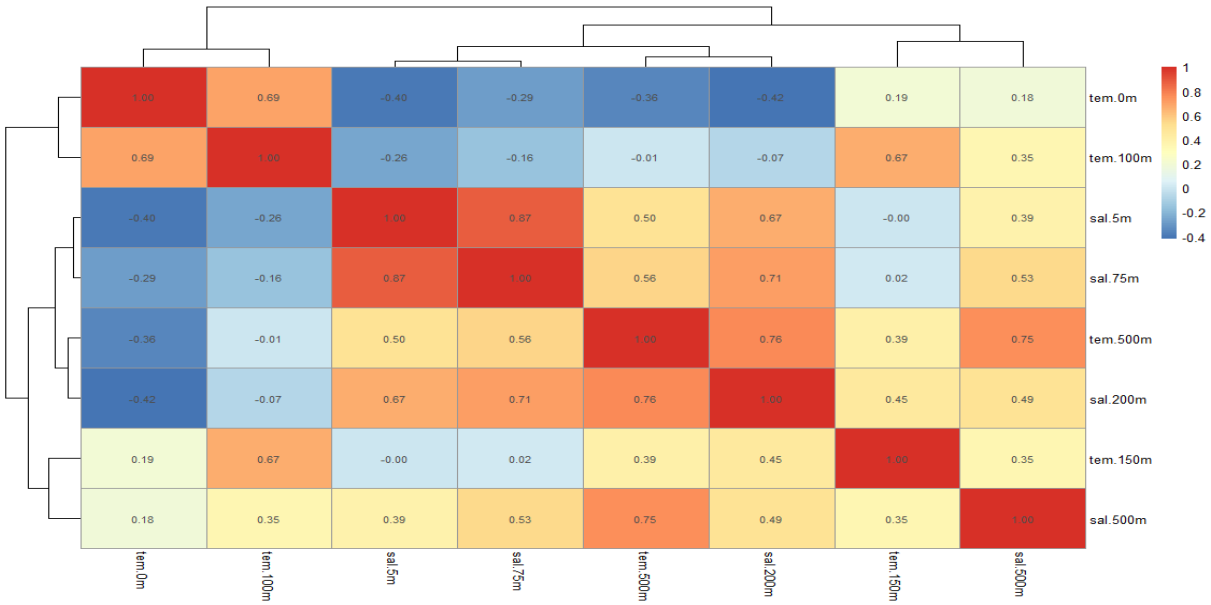
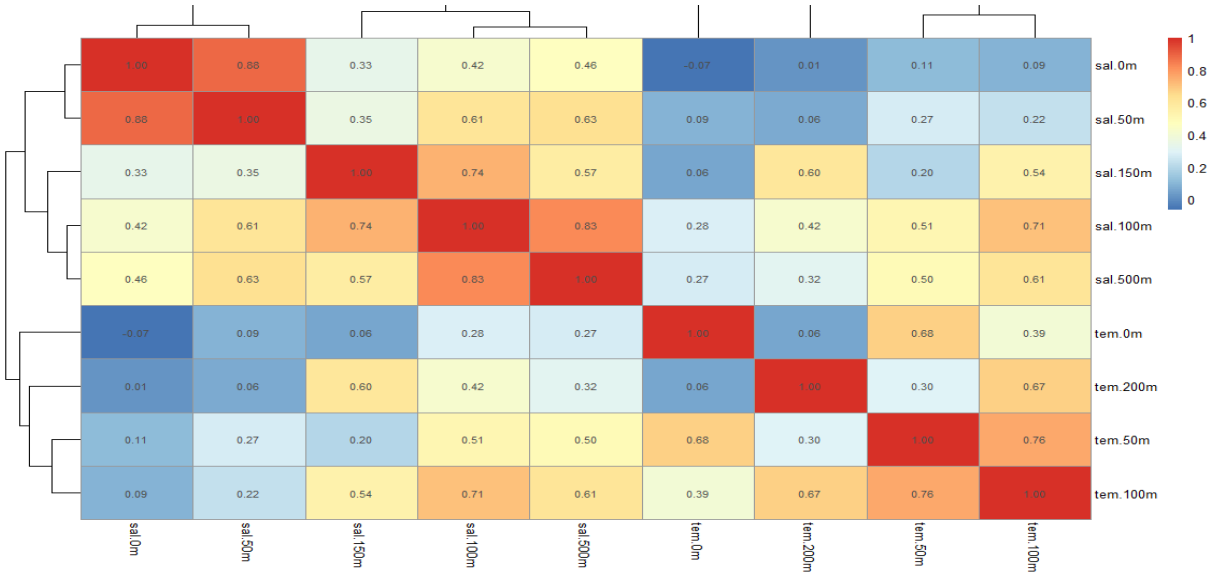

tem.30m
tem.20m
tem.0m
tem.5m
tem.10m
sal.50m
sal.0m
sal.5m
sal.10m
sal.20m
sal.30m
sal.75m
tem.500m
sal.400m
sal.500m
tem.400m
tem.250m
sal.300m
sal.100m
sal.125m
tem.300m
sal.150m
sal.200m
tem.50m
tem.75m
tem.100m
tem.125m
tem.150m
tem.200m
tem.250m
1
0.8
0.6
0.4
0.2
0
-0.2

Fig 5. Pearson Correlation between vertical environmental variables within each model series (Model5d series, top; Model1d series, middle; ModelCh series, bottom)


1.00 0.69 -0.40 -0.29 -0.36 -0.42 0.19 0.18 tem.0m
0.69 1.00 -0.26 -0.16 -0.01 -0.07 0.67 0.35 tem.100m
-0.40 -0.26 1.00 0.87 0.50 0.67 -0.00 0.39 sal.5m
-0.29 -0.16 0.87 1.00 0.56 0.71 0.02 0.53 sal.75m
-0.36 -0.01 0.50 0.56 1.00 0.76 0.39 0.75 tem.500m
-0.42 -0.07 0.67 0.71 0.76 1.00 0.45 0.49 sal.200m
0.19 0.67 -0.00 0.02 0.39 0.45 1.00 0.35 tem.150m
0.18 0.35 0.39 0.53 0.75 0.49 0.35 1.00 sal.500m
tem.0m tem.100m sal.5m sal.75m tem.500m sal.200m tem.150m sal.500m
1 0.8 0.6 0.4 0.2 0 -0.2 -0.4


1.00 0.88 0.33 0.42 0.46 -0.07 0.01 0.11 0.09 sal.0m
0.88 1.00 0.35 0.61 0.63 0.09 0.06 0.27 0.22 sal.50m
0.33 0.35 1.00 0.74 0.57 0.06 0.60 0.20 0.54 sal.150m
0.42 0.61 0.74 1.00 0.83 0.28 0.42 0.51 0.71 sal.100m
0.46 0.63 0.57 0.83 1.00 0.27 0.32 0.50 0.61 sal.500m
-0.07 0.09 0.06 0.28 0.27 1.00 0.06 0.68 0.39 tem.0m
0.01 0.06 0.60 0.42 0.32 0.06 1.00 0.30 0.67 tem.200m
0.11 0.27 0.20 0.51 0.50 0.68 0.30 1.00 0.76 tem.50m
0.09 0.22 0.54 0.71 0.61 0.39 0.67 0.76 1.00 tem.100m
sal.0m sal.50m sal.150m sal.100m sal.500m tem.0m tem.200m tem.50m tem.100m
1 0.8 0.6 0.4 0.2 0

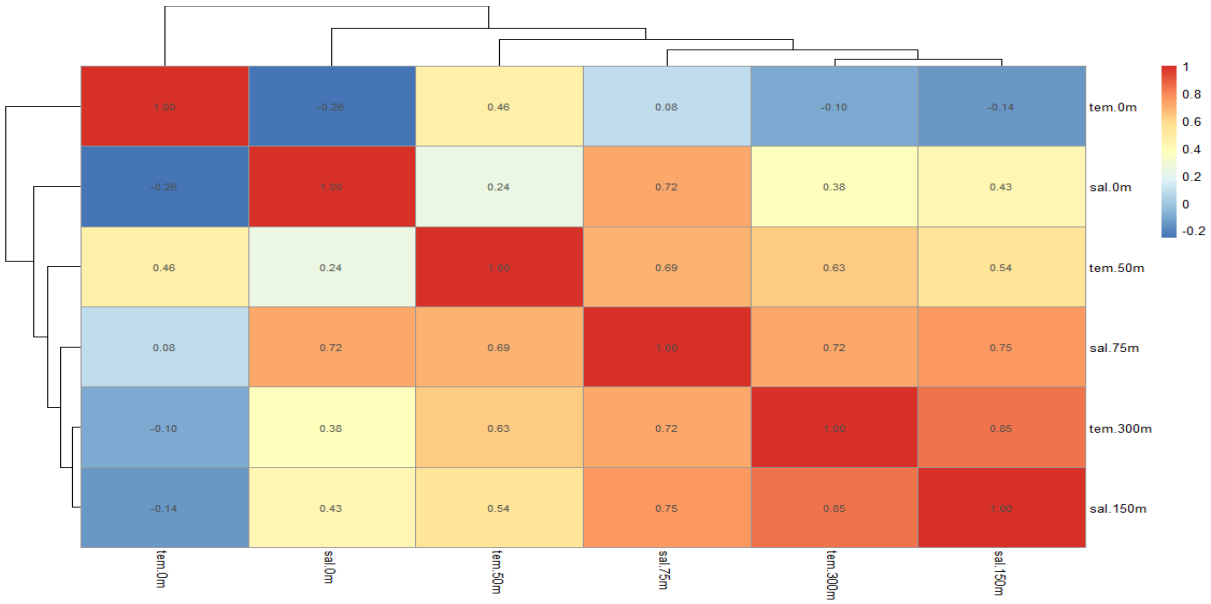

1.00 -0.26 0.46 0.08 -0.10 -0.14 tem.0m
-0.26 1.00 0.24 0.72 0.38 0.43 sal.0m
0.46 0.24 1.00 0.69 0.63 0.54 tem.50m
0.08 0.72 0.69 1.00 0.72 0.75 sal.75m
-0.10 0.38 0.63 0.72 1.00 0.85 tem.300m
-0.14 0.43 0.54 0.75 0.85 1.00 sal.150m
tem.0m sal.0m tem.50m sal.75m tem.300m sal.150m
1 0.8 0.6 0.4 0.2 0 -0.2

Fig 6. Pearson Correlation between vertical environmental variables within each model series after VIF analysis (Model5d series, top; Model1d series, middle; ModelCh series, bottom)

3.3 Evaluation and comparison of models

The indices: AIC, explained deviance, null deviance and residual deviance for each model were shown in Table 3. Among the model series for the whole region from Model 5d-1 to 5d-4, Model 5d-3 showed the lowest AIC and the highest explained deviance. Among the model series for R1, the highest explained deviance was computed for the Model Ch series, which indicated a better model fit compared to Model 1d series that, despite a low AIC, showed very low explained deviance. The explained deviances varied between 68.82% for the Model Ch-1 and 77.55% for the Model Ch-3. In addition, the model fit of Model 5d-8 and Model 1d-4 have been obviously improved with random sampling the same amount of catch records with Chinese set by set datasets. Therefore, the best model for the whole region was Model 5d-3; while for R1 the best models included: Model 5d-8, Model 1d-4, and Model Ch-3.

The normal Q-Q plots for the best models were also shown in Fig. 7, and they indicated that the residuals approximated to be normal distributions.

Table 3. Indices used for the comparison of the 15 models

Model	AIC	Explained deviance	Null deviance	Residual deviance
Model(5d-1)	20150	39.12%	13099	7975
Model(5d-2)	19722	40.18%	12856	7691
Model(5d-3)	14658	40.31%	9385	5602
Model(5d-4)	14673	37.53%	8998	5621
Model(5d-5)	12034	25.03%	5876	4405
Model(5d-6)	11787	26.23%	5784	4267
Model(5d-7)	9051	29.08%	4557	3232
Model(5d-8)	672	35.54%	287	185
Model(1d-1)	450	22.69%	119	92
Model(1d-2)	409	16.67%	96	80

Model(1d-3)	357	29.27%	82	58
Model(1d-4)	100	71.79%	39	11
Model(Ch-1)	1075	68.82%	1456	454
Model(Ch-2)	1015	75.00%	1456	364
Model(Ch-3)	832	77.55%	1207	271

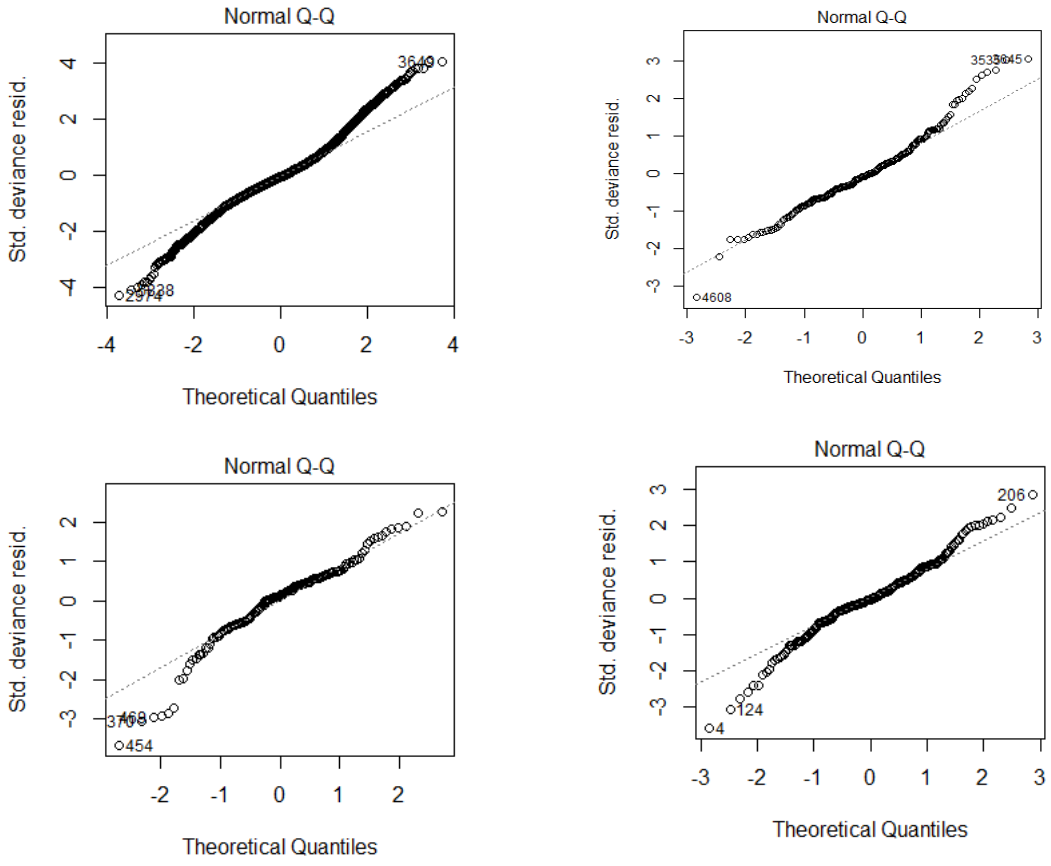


Fig. 7 The normal Q-Q plots for Model 5d-3(left-top), Model 1d-3(right-top) and Model Ch-3(bottom)

3.4 CPUE standardization

For the whole region, we estimated the CPUE indices based on Model 5d-3 (Fig. 8). The long term tropical indices in the whole region showed a relative low level before the end of 2011 and then sharply increased in months of 2012 and 2013. Then declining trend subsequently resumed and continued until the end of 2015. The CPUE was estimated to be at its lowest level from 2014 to 2015. The estimated indices based on Model 5d-4 with the response variable scaled by geographical area showed almost the same year-month trends as Model 5d-3. The scatter plot of the relation between the CPUE indices before and after scaled

by geographical area of each 5° grid was almost on a straight 45° line (Fig. 9). The spearman correlation coefficient was above 0.95, which indicated that the standardized CPUE didn't get effect from the spatial area inconformity between equatorial and middle and high latitudes.

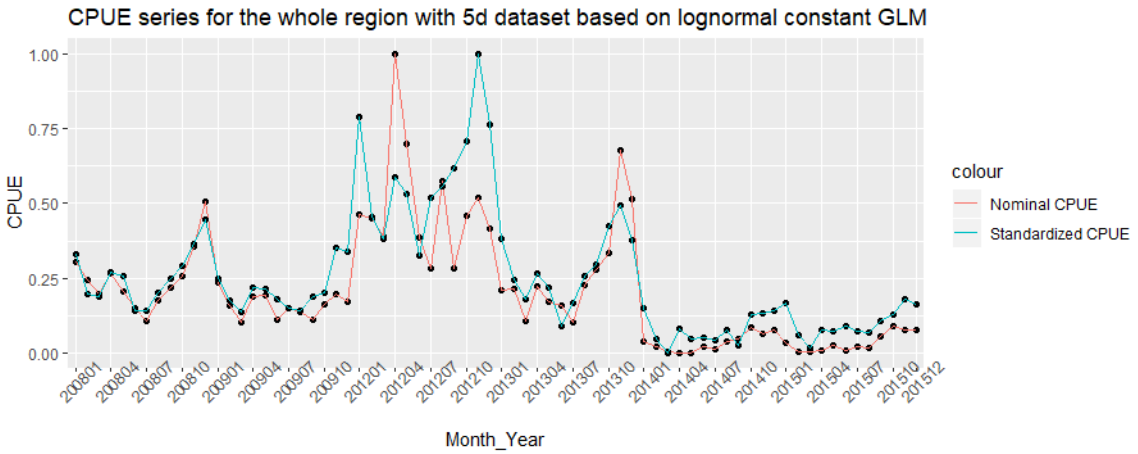


Fig. 8 Nominal CPUE and Standardized CPUE indices estimated based on Model 5d-3

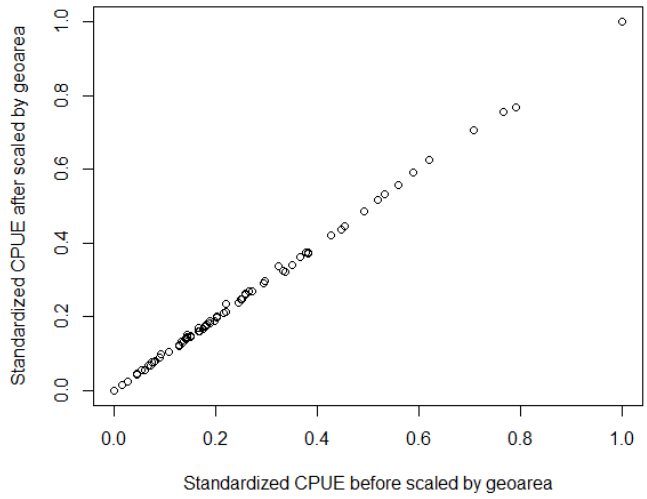


Fig. 9 Scatter plot of the relation between the Standardized CPUE indices before and after scaled by geographical area of each 5° grid

For R1 region, Nominal CPUEs estimated from different fisheries datasets were presented in Fig. 10 and the standardized CPUEs estimated based on three best models: Model 5d-8, Model 1d-4, and Model Ch-3 for R1 were shown in Fig 11. Nominal CPUEs estimated based on different fisheries datasets showed totally different trends over the same period from Dec, 2013 to Apr, 2014 for R1 and the differences were mirrored in standardized CPUEs.

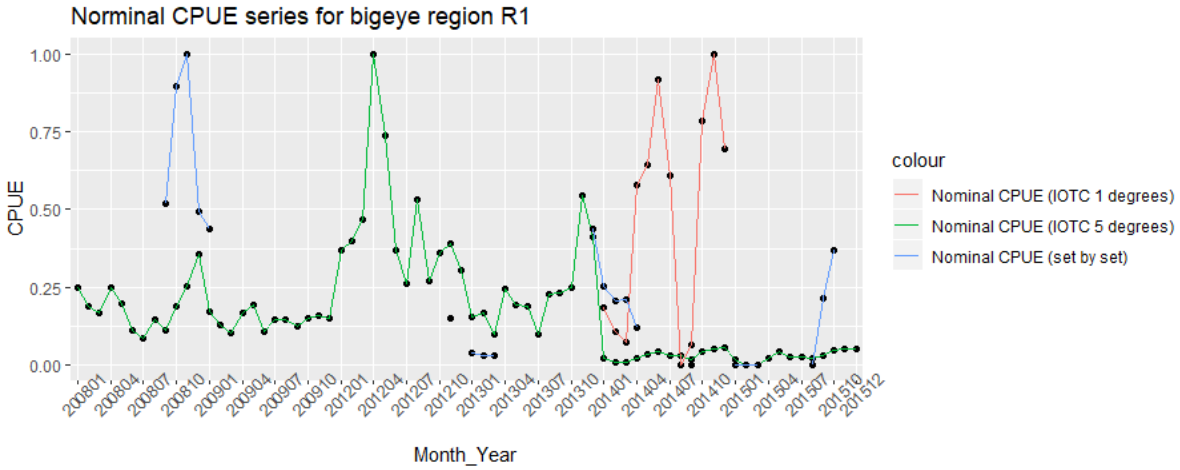


Fig. 10 Nominal CPUEs estimated from three different fisheries datasets

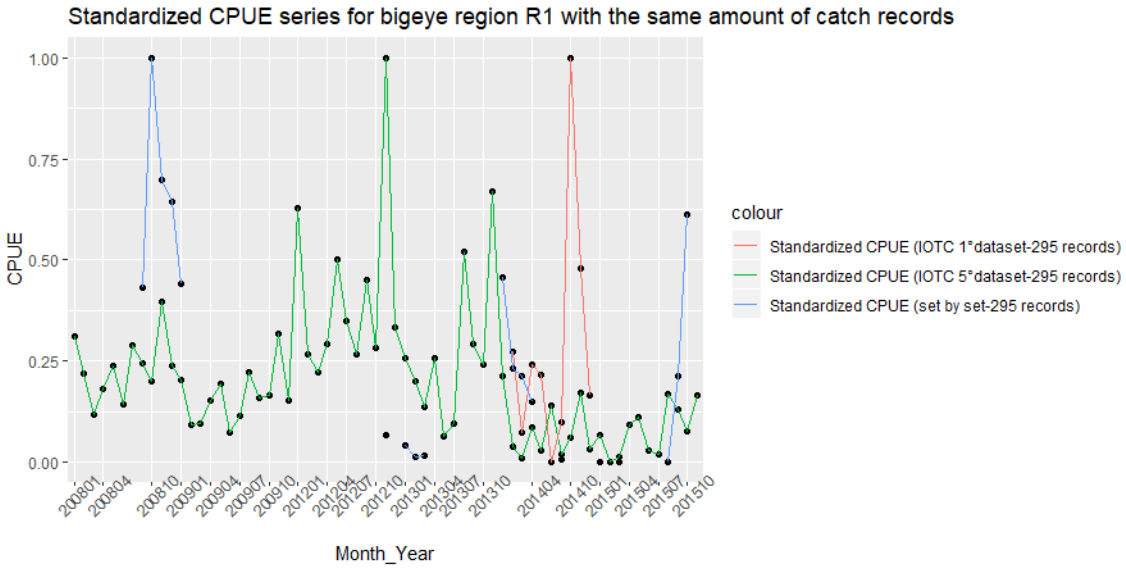


Fig.11 Standardized CPUEs estimated based on three best models with the same amount of catch records: Model 5d-8, Model 1d-4 and Model Ch-3

For bigeye tuna in region 1, IOTC 1° dataset were only available for 2014 (Fig 12). The Standardized CPUE dropped at the first three months and then increased until June and then declined again until August. The indices arrived at the peak on November and then declined again.

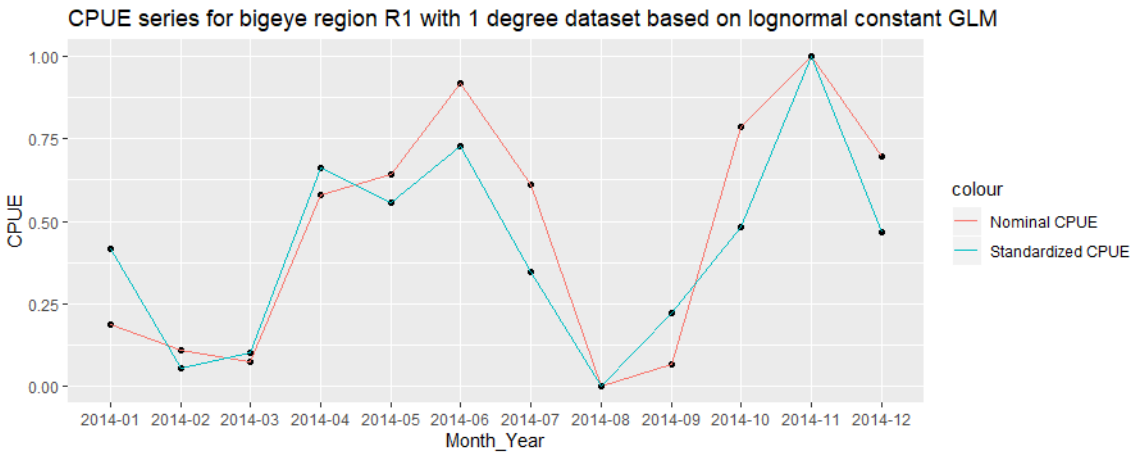


Fig. 12 The Nominal CPUE and Standardized CPUE estimated based on 1° datasets within Model 1d-3 for R1.

Chinese fishery dataset for bigeye in R1 were relatively sparse for the months from 2008 to 2015. For the last four months in 2008, standardized CPUEs were at a high level, with the scaled values above 0.5; but then the values declined fast during 2009, 2012 and 2013. The indices peaked at the end of 2013 but then declined again until August, 2015 and then increased again.

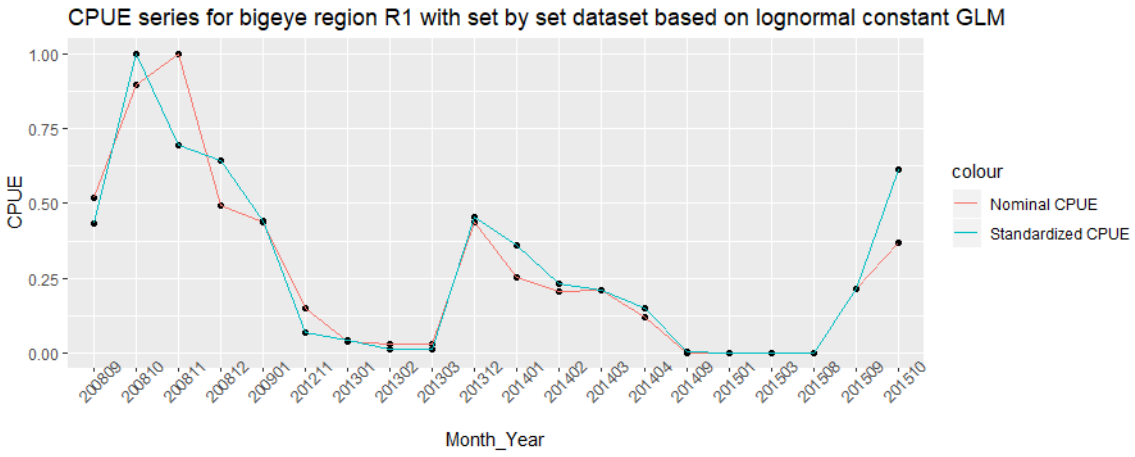


Fig. 14 The Nominal CPUE and Standardized CPUE estimated based on Chinese set by set dataset within Model Ch-3 for R1.

The coefficients and significant codes of the environment covariates in the 'best' models were shown in Table 4. For the whole region, we could find that most of the covariates used in Model 5d-3 contributed to the model, while Year, Month, Latitude, cluster group, tem 100m, tem 500m and Sal.500m were important factors with confidence coefficient >0.999. For R1,

Year, Month and cluster group were the most significant factors in Model Ch-3. Other variables such as tem.0m, Sal.500m, Sal.0m, tem.50m, EKE, chla, windspeed, and DF also contributed to some level to model fit.

*Table 4: The Coefficients (and Signif. codes) of the environment covariates in the 'best' models. (Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)*

Covariate	Model(5d-3)	Model(5d-8)	Model(1d-4)	Model(Ch-3)
Year	-4.097e-01(***)	-7.332e-01(***)		4.533e+00(***)
Month	-2.519e-01(***)	-6.953e-01(.)		-5.903e+00(***)
Latitude	6.450e-01(***)	2.372e+00(.)	-4.228e+00(*)	3.269e+00(.)
Longitude	-1.477e+00(.)	1.116e+00(.)	-1.513e+00(.)	4.135e+00(.)
cluster	9.830e-01(***)	-6.69(.)		1.806e+00(***)
tem.0m	-9.122e-02(**)			
tem.50m				1.208e+00(**)
tem.100m	6.718e-02(***)			
tem.150m	-6.648e-02(*)	1.604e-01 (.)		
tem.500m	-4.815e-01(***)	-6.251e-01 (.)		
Sal.0m				-5.112e+00(.)
Sal.5m	3.167e-01(**)		-7.908e+00(*)	
Sal.150m				
Sal.200m	6.470e-01(*)			
Sal.500m	3.674e+00(***)		1.005e+01(.)	
windspeed	-3.123e-02(*)			
Chla	-5.184e-01(.)			
DF	-1.245e-03(**)			
EKE		-9.973e-01(.)	8.652e-01(.)	

4 Discussion

The aim of this study was to examine the differences in performance of CPUE standardization models at various spatial scales with more explanatory environmental

variables. Results of our model series were discussed as following:

1) For the whole region, using IOTC 5° datasets could capture the underlying CPUE trends, while for R1 region, using dataset with a higher resolution could improve the model fit.

Indices based on Model 5d series showed similar trends with the joint indices developed in 2016 (Hoyle, et al, 2016), which was characterized by a decline during 2013-2015 followed by a sharp increase in 2012. We have used 11294 catch records in IOTC 5° datasets for training the models in 6 years (72 months) and the explained deviance was relatively high (above 40%) compared to previous study. That means the quantity and the spatial scale of 5° monthly fishery datasets could be sufficient for training GLM models and capture the underlying CPUE trends. The 2013 IOTC CPUE workshop (Anon 2013) recommended using 5° area to account for changes in effort distribution. We scaled the response variables of CPUE by the geographical area of the IOTC 5°grids in Model 5d-4, but no significant difference occurred. This may be because GLM model was not sensitive to slight scale change of the response variable. We should explore further by using area effects as the adjusted statistical weights to allow for changing effort concentration as recommended (Campbell 2004).

For R1 region, Model Ch series had the highest explained deviance among the three model series. Even using the same amount of catch records (295) as Chinese set by set datasets for training Model Ch and 1d series, Model Ch series still performed best. The reason may lie in the hypothesis of GLM lognormal model: samples were selected randomly and distributed uniformly. The dataset at a higher resolution could represent more spatial details and the related Nominal CPUE log forms tend to better met the model requirement. However, several zero records in the Model Ch series may lead to model over-fit, which may also lead to high explained deviance. Therefore, we should explore more about the set by set fishery data for testing their reliability.

2) The addition of environmental variables obviously improved the accuracy of the estimated indices.

In each of our model series, cluster analysis was made to separate species composition based on the spatial locations of the catches. Our results indicated that clusters based on the 5°

dataset and the Chinese set by set fishery dataset differentiated the spatial distribution of species compositions more clearly, and therefore they played important roles in the related model 5d and Ch series. We infer that identifying species composition change in different spatial environment and figuring out the main spatial distribution area of target fish could improve the accuracy of CPUE standardization models.

Our study used both vertical temperature and vertical salinity in 15 levels at depths to 500m in the models and almost each vertical factor has contributed to the model fit, though not in the same model. Previous studies have mentioned the environmental values at depth influenced species' distribution (Forrestal, et al, 2018). Our results helped to explain why longline sets target bigeye tuna typically during the day at depths of 100–400 m (Abascal, et al, 2018).

The common used factor: concentration of sea surface chlorophyll *a* (Chla) and sea surface wind speed (WS) also appeared to affect the model indices. However, the factor sea surface temperature (SST), which has been verified to have indication significance for tuna distribution, did not show its importance in our models. The reason may be the high correlation between SST and the vertical temperatures, which made the models failed to identify SST. The factor DF appeared to have a negative correlation with CPUE indices, and this was in line with the conclusion of previous study: when fishing locations were close to the SST fronts, higher CPUEs were observed (Tseng, et al, 2014). We calculated EKE to reflect the influence of ocean eddy conditions to the distribution of our target fish. However, the factor only contributed to Model 1d series for R1 region, but not to the whole region, which may be the high spatiotemporal variability in the distribution of the mesoscale eddies.

3) Our exercise highlights some further work to improve the methods for CPUE standardization.

Although our methods have shown the use of environmental variables increased the model accuracy, but we didn't consider the interactions among the variables. We added all the environment covariates in the models at once and this may cause problems for interpretation of some correlative covariates. We have used the variance inflation factor (VIF) to remove

some vertical temperature and salinity factors that related to a dummy variable, but we could still find collinearity between the remained factors. In addition, more factors such as HBF and vessel effect have been confirmed to be related to the CPUE trends, but we were not able to include them because of unavailable of these data.

We suggest the following priorities for further work:

(1) Include more valuable environmental variables to improve the accuracy of the estimated indices.

(2) Explore the amount of fishery dataset that could meet the model accuracy requirement based on the spatial scale.

(3) Employ more factorial design to further explore the interaction of factors within the CPUE standardization methods.

(4) Use grid area effects as the adjusted statistical weights in the GLM models to test the effect of spatial area inconformity on the standardization process.

5 Conclusions

This study used fisheries datasets at various spatial scales and more explanatory environmental variables to derive a set of “best practices”. Overall, using IOTC 5° datasets could capture the underlying CPUE trends for the whole region, and Chinese set by set fishery dataset at higher resolution could improve the model fit for R1 region. The inclusive environmental variables such as, clusters that identified species composition based on locations, vertical temperature and vertical salinity, Chla, surface wind speed, and the nearest distance from SST fronts contributed to CUPE standardization models and improved the model fit. This exercise highlights the usefulness of finding finer spatial scale for subarea standardization and also verified the importance of factors that affected bigeye tuna distribution habitat but were rarely used in previous study. We suggest further work to explore the interaction of the more environmental variables with more spatial scales of fishery dataset in the CUPE standardization models.

Acknowledgements

The project was funded by Shanghai Sailing Program (17YF1407700), the Ministry of Agriculture of the People’s Republic of China under the Projects of Fishery Exploration in

High Seas in 2005 and 2006 (Project No. Z05-30, Z06-43), the National Natural Science Foundation of China (Project No.41776142, 71601113). We thank the general manager, Jingmin Fang, vice general manager, Fuxiong Huang, the crews of the tuna longliners, and the others of Guangyuang Fishery Group Ltd of Guangdong province for their support of this project. The authors wish to thank Yong Zhang of College of Information Technology in Shanghai Ocean University for gathering and compiling the oceanographic data. We also wish to thank Bo Song in Shanghai Maritime University for explaining the model theories and helping coding.

Reference

Forrestal FC, Schirripa M, Goodyear CP, et al. Testing robustness of CPUE standardization and inclusion of environmental variables with simulated longline catch datasets[J]. Fisheries Research, 2019, 210:1-13.

Bigelow KA , Boggs CH , Xi HE . Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery[J]. Fisheries Oceanography, 1999, 8(3):178-198.

Hoyle SD, Kim DN, Lee SI, et al. 2016. Collaborative study of tropical tuna CPUE from multiple Indian Ocean longline fleets in 2016. IOTC–2016–WPTT18–14.

Hoyle SD, Okamoto H, Yeh Y-m, et al. 2015. IOTC–CPUEWS02 2015: Report of the 2nd CPUE workshop on longline fisheries, 30 April – 2 May 2015. 126 p.

Rose EL, 1995. Multivariate analysis of categorical data: Theory. Structural Equation Modeling: A Multidisciplinary Journal 2, 274-276. doi:10.1080/10705519509540014.

Nakazawa M. fmsb: Functions for Medical Statistics Book with Some Demographic Data. R Package Version 0.3.4, 2012. Available online: http://CRAN.R-project.org/package=fmsb (accessed on 9 August 2012).

Hoyle SD, Okamoto H. 2011. Analyses of Japanese longline operational catch and effort for bigeye and yellowfin tuna in the WCPO, WCPFC-SC7-SA-IP-01.

Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723. https://doi.org/10.1109/TAC.1974.1100705

He X, Bigelow KA, Boggs CH. Cluster analysis of longline sets and fishing strategies

within the Hawaii-based fishery. Fisheries Research 1997, 31(1-2): 147-158

Abascal FJ, Peatman T, Leroy B, et al. Spatiotemporal variability in bigeye vertical distribution in the Pacific Ocean[J]. Fisheries Research, 2018, 204:371-379.

Tseng CT, Sun CL, Belkin IM, et al. Sea surface temperature fronts affect distribution of Pacific saury (Cololabis saira) in the Northwestern Pacific Ocean[J]. Deep Sea Research Part II: Topical Studies in Oceanography, 2014, 107:15-21.

Anon. 2013. Report of the IOTC CPUE Workshop, San Sebastian, Spain, 21–22 October, 2013. IOTC–2013–SC16–12[E], Indian Ocean Tuna Commission.

Campbell, R. A. 2004. CPUE standardisation and the construction of indices of stock abundance in a spatially varying fishery using general linear models. Fisheries Research 70:209-227.