Acta Oceanol. Sin., 2019, Vol. 38, No. 10, P. 100–110
https://doi.org/10.1007/s13131-019-1486-3
http://www.hyxb.org.cn
E-mail: hyxbe@263.net

Comparative analysis of CPUE standardization of Chinese Pacific saury (*Cololabis saira*) fishery based on GLM and GAM

Chuanxiang Hua[1, 2], Qingcheng Zhu[1, 2*], Yongchuang Shi[1], Yu Liu[1, 2]

[1] College of Marine Sciences, Shanghai Ocean University, Shanghai 201306, China

[2] National Engineering Research Center for Pelagic Fishery, Shanghai 201306, China

Received 2 November 2018; accepted 22 January 2019

Abstract

Pacific saury is an important high-seas fishery resource in the Northwest Pacific Ocean for the Chinese Mainland. Reliable and accurate catch per unit effort (CPUE) plays a significant rule in Pacific saury stock assessment. Many statistical models have been used in the previous CPUE standardization research. Here, we compare the performance of Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) using CPUE data collected from Chinese saury fishery in the Northwest Pacific Ocean from 2003 to 2017 (excluding data from Chinese Taipei), and evaluate the influence of spatial, temporal, environmental variables and vessel length on CPUE. Optimal GLM/GAM models were selected using the Bayesian information criterion (BIC). Explained deviance and 5-fold bootstrap cross-validation results were used to compare the performance of the two model types. Fitted GLMs accounted for 21.57% of the total model-explained deviance, while GAMs accounted for 38.95%. Predictive performance metrics and 5-fold cross-validation results showed that the best GAM performed better than the best GLM. Therefore, we recommend GAM as the preferred model for standardizing CPUE of Pacific saury in the Northwest Pacific Ocean.

Key words: *Cololabis saira*, CPUE standardization, generalized linear model, generalized additive model

Citation: Hua Chuanxiang, Zhu Qingcheng, Shi Yongchuang, Liu Yu. 2019. Comparative analysis of CPUE standardization of Chinese Pacific saury (*Cololabis saira*) fishery based on GLM and GAM. Acta Oceanologica Sinica, 38(10): 100–110, doi: 10.1007/s13131-019-1486-3

1 Introduction

Pacific saury (*Cololabis saira*) is a highly migratory fish, widely distributed in the high seas of the Northwest Pacific Ocean (Lin, 2003; Sun et al., 2003). The species is harvested primarily by Japan, Russia, South Korea, Taiwan Province, and Chinese Mainland. China began Pacific saury fishing in the high seas in 2003 and it has become one of the most important fisheries for China since then. Although various studies regarding the fishing gear and methods (Xu et al., 2005; Yu et al., 2006; Xia, 2008), distribution of fishing grounds (Huang et al., 2005; Hua et al., 2010), and basic biology (Zhang et al., 2013) have been conducted, there is no study focused on catch per unit effort (CPUE) standardization for Chinese Pacific saury fishery in the Northwest Pacific Ocean.

CPUE is commonly used as an important relative index of fish abundance and is one of the most important dataset used in fisheries stock assessment (Nishida and Chen, 2004; Chen et al., 2008). Although abundance index should, ideally, be derived from fishery-independent surveys, it is often based on fishery-dependent data, because fishery-independent data are often costly and difficult to collect (Ward et al., 2013). Nominal CPUE is the uncorrected value obtained directly from commercial fishing data (Maunder and Start, 2003). CPUE is assumed to have a linear relationship with abundance in the assessment; however, CPUE derived from fishery-dependent data may not be the case, because it is often influenced greatly by various factors, such as temporal factors (e.g., *Year*, *Month*), spatial factors (e.g., *Longitude*, *Latitude*), environmental factors (e.g., sea surface temperature (*SST*), sea surface height (*SSH*), sea surface temperature gradient (*SSTG*), and fishing capacity (e.g., *Vessellength*) (Harley et al., 2001; Erisman et al., 2011). CPUE can be misleading if these confounding factors are not taken into account.

In order to have more reliable and representative CPUE data for stock assessment, nominal CPUE values need to be standardized using statistical models, which is a process aiming to remove the impacts of the confounding factors (Maunder and Punt, 2004). For providing better management and conservation recommendations, high quality data and continued evaluation of statistical model performance should be highly valued (Martínez-Rincón et al., 2012). In recent decades, many efforts have been made to solve the problems associated with CPUE standardization. Generalized linear models (GLM) and generalized additive models (GAM) are commonly used to standardize CPUE due to the availability of well-tested and user-friendly software to perform calculations (Venables and Dichmont, 2004). Indeed, GLMs are the most common method for standardizing CPUE; they differ from ordinary linear models by allowing fitting of categorical variables and they allow to incorporate non-normal distributions of the response variable. GAMs, on the contrary, are extensions of GLMs which have smooth functions. They are often used to deal with nonlinear relationships between response and explanatory variables (Wood, 2006). However, when standardizing CPUE data, GLMs and GAMs always have their own disadvantages in error structure assumptions, dealing with interaction terms and zero data (Yu et al., 2013). In order to select the better

Foundation item: The National Sci-Tech Support Plan "Fishing Technology and New Resources in Oceanic Fisheries" under contract No. 2013BAD13B05.

*Corresponding author, E-mail: qczhu@shou.edu.cn

CPUE standardization model, comparative research should be conducted between GLMs and GAMs. Further, to date and to the best of our knowledge, there is no study about CPUE standardization for Pacific saury.

In this study, we used GLM and GAM based on Chinese fishery data (2003–2017) to conduct a comparative study on CPUE standardization of the Chinese Pacific saury fishery in the Northwest Pacific Ocean. Firstly, we selected the optimal GLM/GAM model using the Bayesian information criterion (BIC). We then used the explained Deviance and results from 5-fold bootstrap cross-validations to compare the predictive accuracy of the two models (Rodríguez-Marín et al., 2003; Ortiz and Arocha, 2004). The goal of this study was to identify the best method to the standardization of Pacific saury CPUE data and improve the quality of future stock assessment for Pacific saury.

2 Materials and methods

2.1 *Study area*

As shown in Fig. 1, the study area is outside the Exclusive Economic Zone (EEZ) of Japan and Russia. This area is at the junction of the Kuroshio warm current and the Oyashio cold current, which together provide an adequate foundation for marine life, and the area is in fact one of the high-yield sea areas of the world (Watanabe et al., 2006).

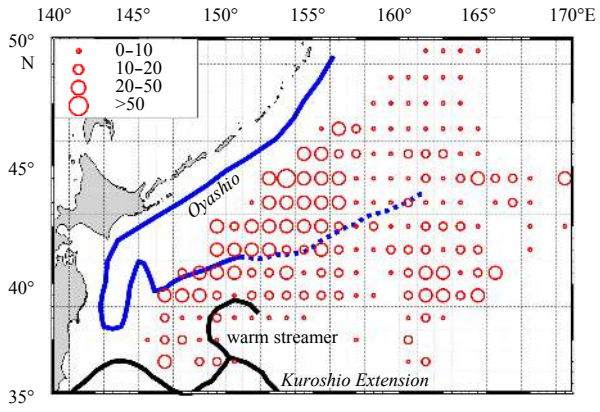

Fig. 1. Main fishing area of Pacific saury in China. The blue solid lines represent the Oyashio cold current, the blue dashed lines the Oyashio extension current, the black lines Kuroshio warm current, and the red circles CPUE, in units of tons per vessel per day.

2.2 *Data sources*

2.2.1 *Fishery data*

Fishery data were obtained from the Technical Group for Saury Fishery, Distant-water Fishery Society of China. These data included the date (with a time resolution of days), longitude, latitude, yield, vessel name, and vessel length, among others. Nominal CPUE was defined as the fishing yield of a vessel per day, in units of tons per vessel per day.

2.2.2 *Environment data*

SST was obtained from the National Oceanic and Atmospheric Administration (NOAA, ftp.nodc.noaa.gov). The spatial-temporal resolution of *SST* data is 0.1°×0.1° grid per day. *SSH* was obtained from Archiving Validation and Interpolation of Satellite Oceanographic data (AVISO, www.aviso.altimetry.fr). The spatial-temporal resolution of the data is *SSH* daily at 0.25°×0.25° grid.

SSTG was estimated from Gradient Magnitude (GM) (Ortiz and Arocha, 2004; Howell and Kobayashi, 2006), and expressed as

$$SSTG_{i,j}=\sqrt{\left(\frac{SST_{i+1,j}-SST_{i-1,j}}{\Delta x}\right)^2+\left(\frac{SST_{i,j+1}-SST_{i,j-1}}{\Delta y}\right)^2}, \quad (1)$$

where $SST_{i+1,j}$, $SST_{i-1,j}$, $SST_{i,j+1}$ and $SST_{i,j-1}$ were *SST* values of four adjacent grids respectively, i and j are row and column number, respectively, Δx is the longitudinal distance between (j–1)th and (j+1)th columns (km), Δy is the latitudinal distance between (i–1)th and (i+1)th rows (km), $SSTG_{i,j}$ is *SSTG* value of the current grid (°C/km).

In order to match fishery data and environmental data, the present study used the environmental data of the nearest grid corresponding to the grid where the fishery data existed on the same date.

2.3 *Methods for analysis*

2.3.1 *Selection of variables*

GLM and GAM both require response variables and explanatory variables that are independent from each other. Since there were no zero-catch data, the natural logarithm of CPUE (i.e., ln(CPUE)) was used as the response variable (Campbell, 2004).

The selection of explanatory variables takes into account the following facts: (1) Pacific saury is a highly migratory fish, and the distribution of its fishing grounds shows large variation during the fishing period (June–November) each year (Tian et al., 2003; Shen et al., 2004); therefore, the explanatory variables include temporal variables (*Year* and *Month*), spatial variables (*Longitude* and *Latitude*), and temporal-spatial interaction terms (*Year*×*Longitude*, *Year*×*Latitude*, *Month*×*Longitude*, and *Month*×*Latitude*); (2) the formation of Pacific saury fishing grounds is tightly associated with the marine environment (Zhu et al., 2006a, b; Zou and Zhu, 2006; Yan et al., 2012; Zhang et al., 2015). Thus, explanatory variables include *SST*, *SSTG* and *SSH*; (3) in a real fishing process, vessel performance may affect fishing efficiency; thus, explanatory variables, such as *Vessellength*, were taken into account in this study.

Mutual independence of explanatory variables was checked by the variance inflation factor (VIF) and Spearman's correlation coefficient (Table 1). In this table, the data under the dotted line are Spearman's correlation coefficient among explanatory variables, data above the line are the corresponding *P* values. The maximum VIF<10, indicated there was no serious multi-collinearity (Tien et al., 2011; Menard, 1995).

2.3.2 *GLM*

GLMs are the most common models for standardizing CPUE data. The key assumption of a GLM is that the relationship between some function of the expected value of the response variable and the explanatory variables, is linear:

$$g(\mu_i)=X_i^{\mathrm{T}}\beta, \quad (2)$$

where g is the differentiable and monotonic link function, μ_i=$E(Y_i)$, X_i is the explanatory variable for the ith response variable, β is a vector of the parameters, and Y_i is the ith random variable.

Table 1. Variance inflation factor (VIF) and Spearman's correlation coefficient among explanatory variables

Coefficient/*P* value	VIF	*Year*	*Month*	*Longitude*	*Latitude*	*SST*	*SSTG*	*SSH*	*Vessellength*
Year	1.72		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Month	3.17	–0.107 4		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Longitude	4.42	0.1537	–0.854 0		<0.001	<0.001	<0.001	<0.001	0.394 6
Latitude	3.91	–0.090 0	–0.581 2	0.602 8		<0.001	<0.001	<0.001	0.225 3
SST	1.32	0.191 7	0.301 4	–0.308 4	–0.315 9		<0.001	<0.001	<0.001
SSTG	1.38	–0.083 1	0.453 4	–0.513 8	–0.422 4	0.279 9		<0.001	0.255 4
SSH	3.39	0.268 0	0.416 3	–0.359 6	–0.819 5	0.461 6	0.347 1		<0.001
Vessellength	3.05	0.298 3	0.046 7	–0.009 1	–0.012 9	0.114 7	0.012 1	0.103 1	

GLM assumes a normal error distribution, and the full GLM is expressed as

$$\ln(CPUE) = Year + Month + Longitude + Latitude + SST + SSTG + SSH + Vessellength + Interaction + \varepsilon, \quad (3)$$

where *CPUE* is the fishing catch of a vessel per day, and *Interactions* is an interaction term representing the interactive effect of spatial and temporal factors for Pacific saury. Full model includes all the possible combination of *Year*, *Month*, *Longitude*, and *Latitude* as interaction terms; ε is the residual, which is assumed to have a normal distribution; *Year* is a categorical variable of 15 years (2003–2017). *Month* is a categorical variable including the eight calendar months from May to December. We attempted two cases (categorical and splined variable) for *Longitude* and *Latitude*, which divided at intervals of 1°. We also investigated two cases (categorical or splined variable) for each explanatory variable of environment. *Vessellength* is a categorical or continuous variable of 60–75 m vessels, which will affect the catchability (Table 2).

2.3.3 *GAM*

GAMs are extensions of the generalized linear models, which can be used to describe nonlinear relationships between response variables and explanatory variables (Tseng et al., 2013), as shown below:

$$g(\mu_i) = \alpha + \sum_{i=1} f_i(X_i) + \varepsilon_i, \quad (4)$$

where f_i is a smoothing function. The full GAM in this study is expressed as

$$\ln(CPUE) = Year + Month + s(Longitude) + s(Latitude) + s(SST) + s(SSTG) + s(SSH) + s(Vessellength) + s(Interactions) + \varepsilon, \quad (5)$$

where $s(\ )$ denotes the smoother functions. The explanatory variables used in GAM are the same as GLM (Table 2). Temporal-spatial interaction terms, including *Year*×*Longitude*, *Year*×*Latitude*, *Month*×*Longitude*, *Month*×*Latitude*, and all possible combinations were considered in this study.

2.3.4 *Model evaluation*

Explanatory variables were added to the GLM/GAM in turn and GLM models/GAM models with different number of explanatory variables were obtained (Shono, 2005). We used the Bayesian information criterion (BIC) to select the best model with minimum BIC in each of the GLM and GAM analyses (Quinn and Keough, 2002; Watanabe et al., 2006). The BIC was calculated as follows:

$$BIC = m\ln(n) + n\ln(RSS/n), \quad (6)$$

where m is the number of parameters in the model, n is the number of observed values (data points), and *RSS* is the sum of squared residuals.

For model diagnostics, percent explained deviance was calculated in addition to *q*-*q* plot and residual plots. The most common method to evaluate the performance of different models is *k*-fold cross-validation (Arlot and Celisse, 2010). When conducting *k*-fold cross-validation, *k* sets of subsamples of roughly equal size are produced from the original samples. One set of subsamples is saved as the validation data for testing the model, and the rest sets are used as training data. Then, the cross-validation

Table 2. Summary of explanatory variables used for GLM and GAM analysis

Variables	Cases	Categorical or continuous	Detail	Note
Year	*Year*	15 categories	15 years from 2003 to 2017	
Month	*Month*	8 categories	8 months from May to December	
Longitude	*Longitude* *Longitude_c*	23 categories	*Lon*<144°, 144°≤*Lon*<145°, 145°≤*Lon*<146°,..., *Lon*>165°	at intervals of 1°
Latitude	*Latitude* *Latitude_c*	13 categories	*Lat*<38°, 38°≤*Lat*<39°, 39°≤*Lat*<40°, ..., *Lat*>48°	at intervals of 1°
Sea surface temperature	*SST* *SST_c*	spline 12 categories	*SST*<10°C, 10°C≤*SST*<11°C, 11°C≤*SST*<12°C,..., 19°C≤*SST*≤20°C, *SST*>20°C	at intervals of 1°C
Sea surface temperature gradient	*SSTG*	continues (spline)		
Sea surface height	*SSH*	continues (spline)		
Vessel length	*Vessellength* *Vessellength_c*	continues (spline) 9 categories	*Vessellength*<64 m, 64 m≤*Vessellength*<76 m,..., 76 m≤*Vessellength*	at intervals of 2 m

Note: *Lon* is *Longitude*; *Lat* is *Latitude*.

is repeated k times, each of the k sets used exactly once as the validation data. The estimation can be obtained as the average of k results from folds. The 5-fold cross-validation procedure has shown good performance in model selection (Kohavi, 2001). Thus, here we used this procedure to evaluate the performance of each model. We conducted a 5-fold cross validation test for the final model selection between the best models derived from GLM and GAM and the process in cross validation are repeated for 1 000 times. In this test, Spearman's correlation between the predicted and observed CPUEs, and mean of squared errors between two CPUEs were calculated to evaluate prediction performance.

2.3.5 *Calculation of nominal CPUE and standardized CPUE*

The yearly nominal CPUE and standardized CPUE can be calculated by the following formula:

$$CPUE_i = \frac{1}{n_i} \times \sum_{k=1}^{n_i} CPUE_k. \quad (7)$$

Nominal CPUE values are calculated as the means of original CPUE, standardized CPUE values are calculated as the means of fitted CPUE from the best model, where $CPUE_i$ (nominal/standardized CPUE) is the CPUE index in the ith year, n_i is the observation number in the ith year, and $CPUE_k$ is the kth CPUE data (fitted/original CPUE) in the ith year.

The bootstrapped 95% confidence intervals of Standardized CPUE of the optimal GLM and GAM were calculated. All statistical analyses were conducted using Matlab2016b.

3 Results

3.1 *Statistical distribution test of response variable (ln(CPUE))*

The scatter points of ln(CPUE) in the normal q-q plot appeared aligned on a straight line (Fig. 2a). Frequency distribution (Fig. 2b) indicated that ln(CPUE) showed approximate normal distribution and was acceptable for using as a response variable in the GLM and GAM.

3.2 *GLM analysis*

The best GLM model selected by BIC is shown in Table 3. A summary of the best fitting model is shown in Table 4. All explanatory variables, including the interaction terms, were statistically significant (P<0.05). In the 5-fold cross validation test, high correlation was observed for the best GLM model (Table 5). The q-q plot and the residual plots for the explanatory variables for evaluating the distribution assumption are shown in Fig. 3. Residuals showed an approximately normal distribution around 0, which indicated that the model assumptions were satisfied.

3.3 *GAM analysis*

The best GAM model selected by BIC is shown in Table 6. A summary of fitting to the best model is shown in Table 7. All explanatory variables, including the interaction terms, were statistically significant. In the 5-fold cross validation test, a high correlation was observed for the best GAM model (Table 8). The q-q plot and the residual plots for the explanatory variables for evaluating the distribution assumption are shown in Fig. 4. Residuals were plotted around 0 and displayed a normal distribution despite a few observed biases.

Figure 5 displays the residuals and explanatory variables fitted by the best GAM. The mean residuals of the GAM model over the years ranged from –1.15 to 0.94. The year 2011 showed the largest residual values for the GAM model and the year 2008 showed the largest negative residual values. The mean residuals of the GAM model over the months ranged from –0.98 to 0.32. September showed the largest positive residual values for the GAM model and December had the largest negative residual values compared with that from other month. For the spatial explanatory variables, *Longitude* and *Latitude* showed relatively smaller residuals, which were close to zero. As environmental explanatory variables, *SST*, *SSTG* and *SSH* showed similar distributions of residuals, which were also close to zero. The mean residuals of the GAM model over *Vessellength* ranged from –0.21 to 0.28.

Effects of temporal, spatial, environmental, and fisheries operational variables on Pacific saury CPUE are shown in Fig. 6. CPUE gradually declined in 2004–2008 and reached its lowest

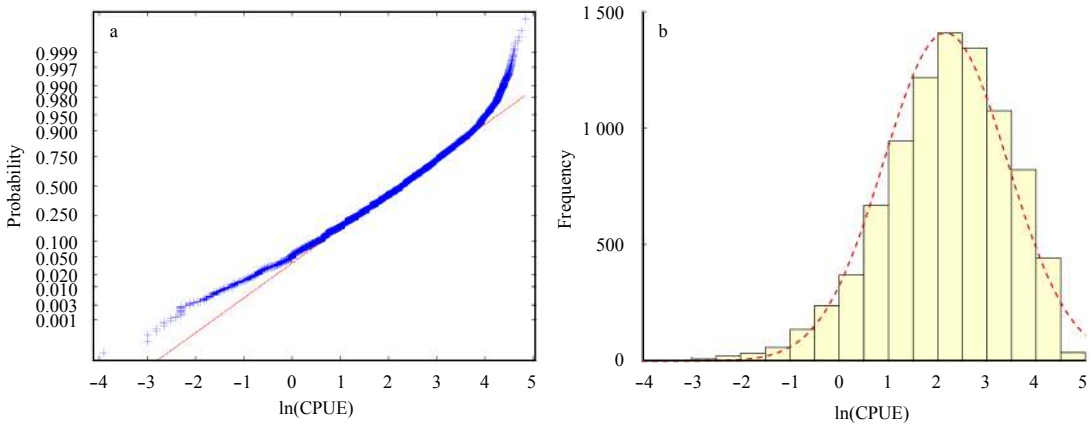

Fig. 2. ln(CPUE) distribution of the Chinese Pacific saury fishery during the period 2003–2017 and its distribution tests. a. Normal q–q plot and b. frequency distribution.

Table 3. Best GLM selected based BIC values

Best model in GLM analysis	R^2	BIC	Explained deviance/%
ln(CPUE)~*Intercept+Year+Month+Lon+Lat+SST_c+SSTG+SSH+Vessellength_c+ Year: Month+Month: Lon+Month: Lat+Lon: Lat+*ε	0.621 7	47 506.8	21.57

Table 4. Anova test for the best GLM model

Parameter name	Df	Deviance	Resid.Df	Resid.Dev	*F*	Pr(>*F*)
Intercept			15 544	24 440		
Factor(*Year*)	13	1 672.5	15 531	22 768	128.662 0	<0.001
Factor(*Month*)	7	4 492.1	15 524	18 276	641.781 2	<0.001
Factor(*Lon*)	23	491.3	15 501	17 784	21.360 3	<0.001
Factor(*Lat*)	12	361.5	15 489	17 423	30.128 4	<0.001
Factor(*SST_c*)	11	115.7	15 478	17 307	10.515 8	<0.001
SSTG	1	17.3	15 477	17 290	17.301 2	3.207×10^{-5}
SSH	1	6.8	15 476	17 283	6.806 3	0.009 093
Factor(*Vessellength_c*)	5	170.5	15 471	17 113	34.097 0	<0.001
Factor(*Year*):factor(*Month*)	48	861.4	15 423	16 251	17.946 6	<0.001
Factor(*Month*): factor(*Lon*)	73	461.5	15 350	15 790	6.322 5	<0.001
Factor(*Month*):factor(*Lat*)	37	185.5	15 313	15 604	5.014 0	<0.001
Factor(*Lon*): factor(*Lat*)	117	409.4	15 196	15 195	3.499 0	<0.001

Table 5. Five-fold cross validation of the best GLM

Case	Correlation	Mean square error
1	0.493 9	1.179 7
2	0.477 8	1.227 3
3	0.466 7	1.310 8
4	0.485 2	1.247 3
5	0.495 6	1.215 5

point in 2008. CPUE increased steadily from 2008 to 2014, and then it decreased slightly in 2014 (Fig. 6a). During the main fishing season (May–December), CPUE showed a gradual increase with the *Month* (Fig. 6b). The spatial factor *Longitude* had a great influence on the CPUE in the range of 145°–155°E (Fig. 6c). The effect of *Latitude* on CPUE decreased gradually from the South to the North (Fig. 6d). As for environmental factors, *SST* had different effects on CPUE within different temperature ranges. When the temperature was between 10°C and 18°C, the effect of *SST* on CPUE was relatively stable. Between 18°C and 23°C, the effect showed a decreasing trend first, which then reversed (Fig. 6e). The effect on CPUE is a gradual increase with increasing *SSTG* (Fig. 6f). The effect of *SSH* also showed some fluctuations (Fig. 6g). *Vessellength* was one of the most important factors affecting CPUE; as Figure. 6h shows, *Vessellength* had a positive correlation with CPUE.

3.4 *Yearly trend of standardized CPUE*

We used the best GLM and GAM models to estimate the yearly trend of standardized CPUE values with a 95% confidence interval (Fig. 7a). The estimates of standardized and nominal CPUE values from 2003 to 2017 are shown in Table 9. Besides 2004–2005 and 2008–2009, standardized CPUE values by the GLM model were significantly lower than the corresponding nominal CPUE values. As for the GAM model, the standardized CPUE values were significantly lower than the corresponding nominal CPUE values. However, there was little difference between CPUEs values standardized by GLM and GAM; this may be related to the assumption of relationships between CPUEs and explanatory variables. So linear relations maybe weak between CPUEs with explanatory variables, deviance explained is 21.57% (Table 3), however, more nonlinearity relations maybe available for them, deviance explained is 38.95% (Table 6). CPUE de-

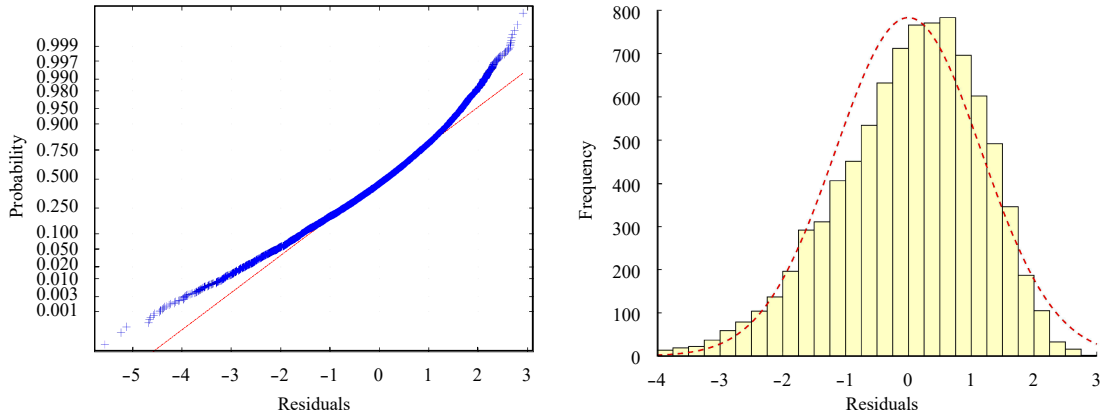

Fig. 3. Normal distribution checks, *q*-*q* plot, and histogram of residuals for the best GLM.

Table 6. Best GAM selected based on BIC values

Best model in GAM analysis	R^2	BIC	Explained deviance/%
ln(CPUE)~*Intercept*+*Year*+*Month*+*s*(*Lon*)+*s*(*Lat*)+*s*(*SST*)+*s*(*SSTG*)+*s*(*SSH*)+*s*(*Vessellength*)+*s*(*Year: Month*)+*s*(*Year: Lon*)+*s*(*Month: Lon*)+*s*(*Month: Lat*)+*s*(*Lon: Lat*)+*ε*	0.572 4	45 254.56	38.95%

Table 7. Anova test for the best GAM model

	Parametric terms		
	Df	*F*	*P*-value
Factor(*Year*)	13	4.666	4.18×10^{-8}
Factor(*Month*)	7	3.287	0.00171
factor(*Lon*)	23	2.763	1.19×10^{-5}
Factor(*Lat*)	12	2.189	0.009 88
Factor(*Year*): factor(*Month*)	61	5.353	<0.001
Factor(*Year*): factor(*Lon*)	166	3.819	<0.001
Factor(*Month*): factor(*Lon*)	92	3.712	<0.001
factor(*Month*): factor(*Lat*)	51	3.993	<0.001
Factor(*Lon*): factor(*Lat*)	152	3.634	<0.001
	Approximate significance of smooth terms		
	Edf	Ref.df	*F*	*P*-value
s(*SST*)	7.850	8.638	11.921	<0.001
s(*SSTG*)	5.538	6.743	4.509	6.75×10^{-5}
s(*SSH*)	1.824	2.394	3.602	0.0211
s(*Vessellength*)	8.838	8.974	103.935	<0.001

Table 8. Five-fold cross validation of the best GAM

Case	Correlation	Mean square error
1	0.617 2	0.975 2
2	0.631 3	0.968 2
3	0.606 3	1.012 9
4	0.605 7	1.027 0
5	0.602 7	0.985 7

Note: Spearman's correlation coefficient is shown.

creased from May to August (Fig. 7b), whereas it increased and then decreased from September to December, with the highest CPUE value occurring in October.

4 Discussion

4.1 *Effect of temporal-spatial factors on CPUE*

The GLM and GAM analyses indicated that temporal (i.e., *Year*, *Month*) and spatial factors (i.e., *Longitude* and *Latitude*), all had a significant effect on CPUE values (p<0.05). GAM analysis suggested that the two most important factors, which accounted for the largest percentage of the CPUE value, were *Longitude* and *Year*. Fluctuations of resources, as well as changes in marine environmental conditions, climate, and fishing effort over time, eventually led to yearly and seasonal fluctuations in CPUE. During 2005–2008, the annual mean CPUE value showed obvious decreases and dropped to the lowest level over the past 12 years in 2008 (Fig. 7a). In 2009, CPUE showed a significant increase, but then gradually decreased. Changes in fishing effort may be the main cause of yearly fluctuations and declines in CPUE value. Further, CPUE differed significantly among months (p<0.01), decreasing from May to August (Fig. 7b), while increasing and then decreasing from September to December, peaking in October, a result that was consistent with those reported by Wu et al. (2015) and Yan (2012). The seasonal migration of saury and the production time of the saury fishery are the main reasons for the seasonal fluctuation in CPUE value; thus, the time variable month showed an impact on CPUE value.

The results of this study indicated that, in the range from 38.5° to 44°N and from 145° to 155°E, overall, CPUE values increased with increasing *Longitude* and *Latitude* but fluctuated, peaking at a marine site near 44°N, 155°E; further, this fluctuation may be associated with the migratory route of Pacific saury. Before August, Pacific saury migrates from the South to the North for feeding, and they start to reverse the route from near 46°N in September for feeding and over-winter migration (Tian et al., 2003; Shen

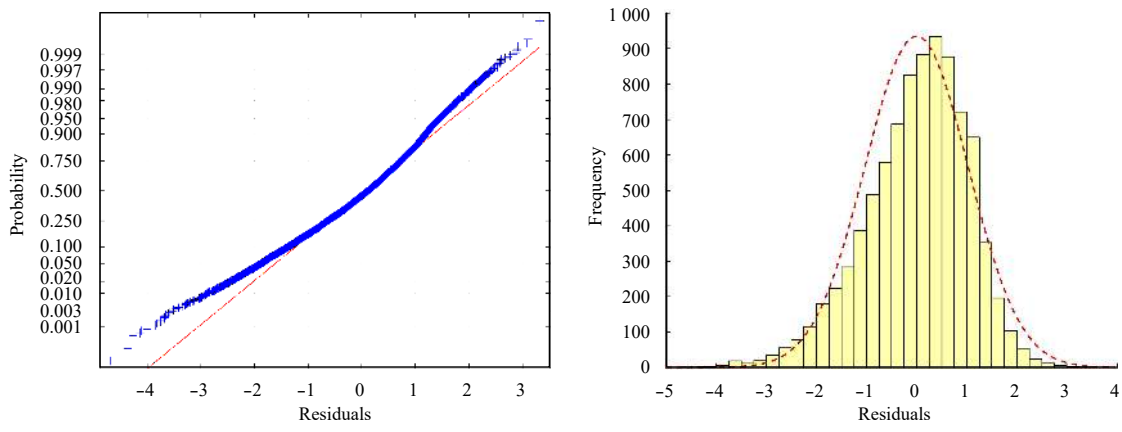

Fig. 4. Normal distribution checks, *q*-*q* plot, and histogram of residuals for the best GAM.

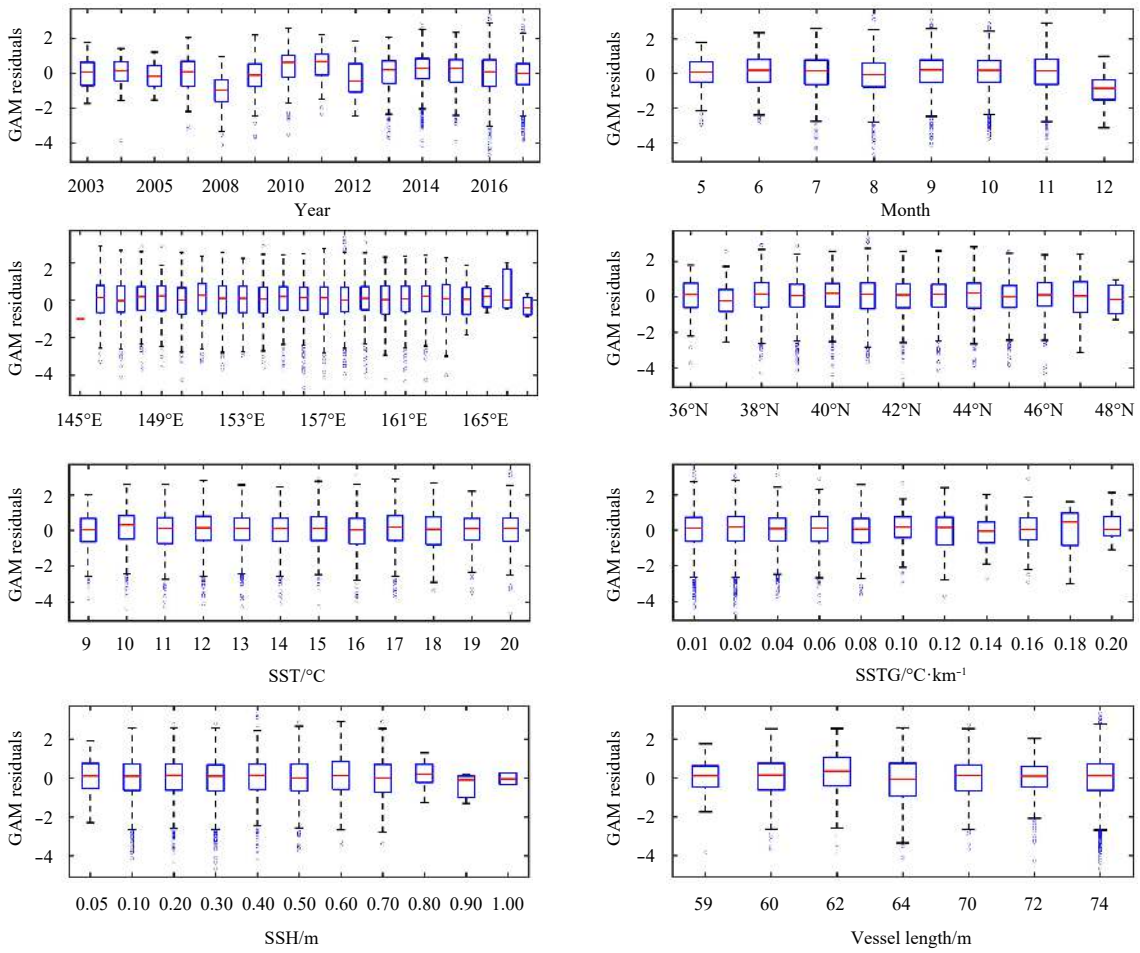

Fig. 5. Boxplot of residuals and explanatory variables fitted by best GAM.

et al., 2004). During its northward migration, Pacific saury is widely distributed, and the fishing grounds are relatively dispersed; whereas, during its southward migration, plankton growth is promoted by the increasing strength of the Oyashio Current and the richness in nutrients of seawater, which causes the school of Pacific saury to gradually concentrate near the coastal waters (He et al., 1999), thereby increasing the fishery catch continually until the end of the fishing season in early December.

4.2 *Effect of environmental factors on CPUE*

The distribution of Pacific saury is tightly associated with marine environmental factors (Takasuka et al., 2014; He et al., 1999; Wang et al., 2012), as reported in many studies based on environmental factor data obtained via remote sensing (Stephens and MacCall, 2004). The *SST* data in this study were obtained via on-board monitoring. Our results indicated that CPUE was significantly correlated with *SST* (p<0.01), and slowly increased with increasing *SST* within a temperature range from 11 to 15°C, which was consistent with the findings reported by Zhu et al. (2006b). CPUE was significantly correlated with *SSH* (p<0.01) and *SSTG* (p<0.01). For the *SSH*, CPUE increased with increasing *SSH* within a range from 0.2 to 0.6 m and have some fluctuation between a range from 0.6 to 1.2 m. CPUE increased with increasing *SSTG* within a range from 0.05 to 0.2°C/km. A similar result was obtained by Tian (2004) while studying Pacific saury fishery data in Japan. It can be seen that environmental factors have a significant impact on CPUE and should be taken into account in future studies.

4.3 *Effect of vessel performance on CPUE*

Pacific saury fishing is conducted using a stick-held net under light induction (Yu et al., 2006), where fish-gathering lamps are used to induce a school of fish into a fishing net on one side of the vessel, and the net is then hauled to catch the fish (Yang et al., 2005). It is generally accepted that the *Vessellength* significantly affects Pacific saury fishing. Indeed, the results from the GLM and GAM models indicated that the *Vessellength* significantly affected CPUE (p<0.01); the reason for this may be that vessel length affected the working space in Pacific saury fishing vessels, the size of the fishing gear, the convenience of the fishing process, and the capacity for processing and freezing caught fish, all of which, together, resulted in a large impact on CPUE of Pacific saury fishing vessels.

4.4 *Comparison between GLM and GAM*

GLMs and GAMs are commonly used to standardize CPUE values; however, both show advantages and limitations. GLMs assume that there is a linear relationship between response variables and the explanatory variables. However, nonlinear relationships are common between fish densities and environmental factors (Walsh and Kleiber, 2001; Denis et al., 2002). For example, the three most important predictors in the main effects models were *SST*, *Latitude*, and *Longitude*, all of which may have

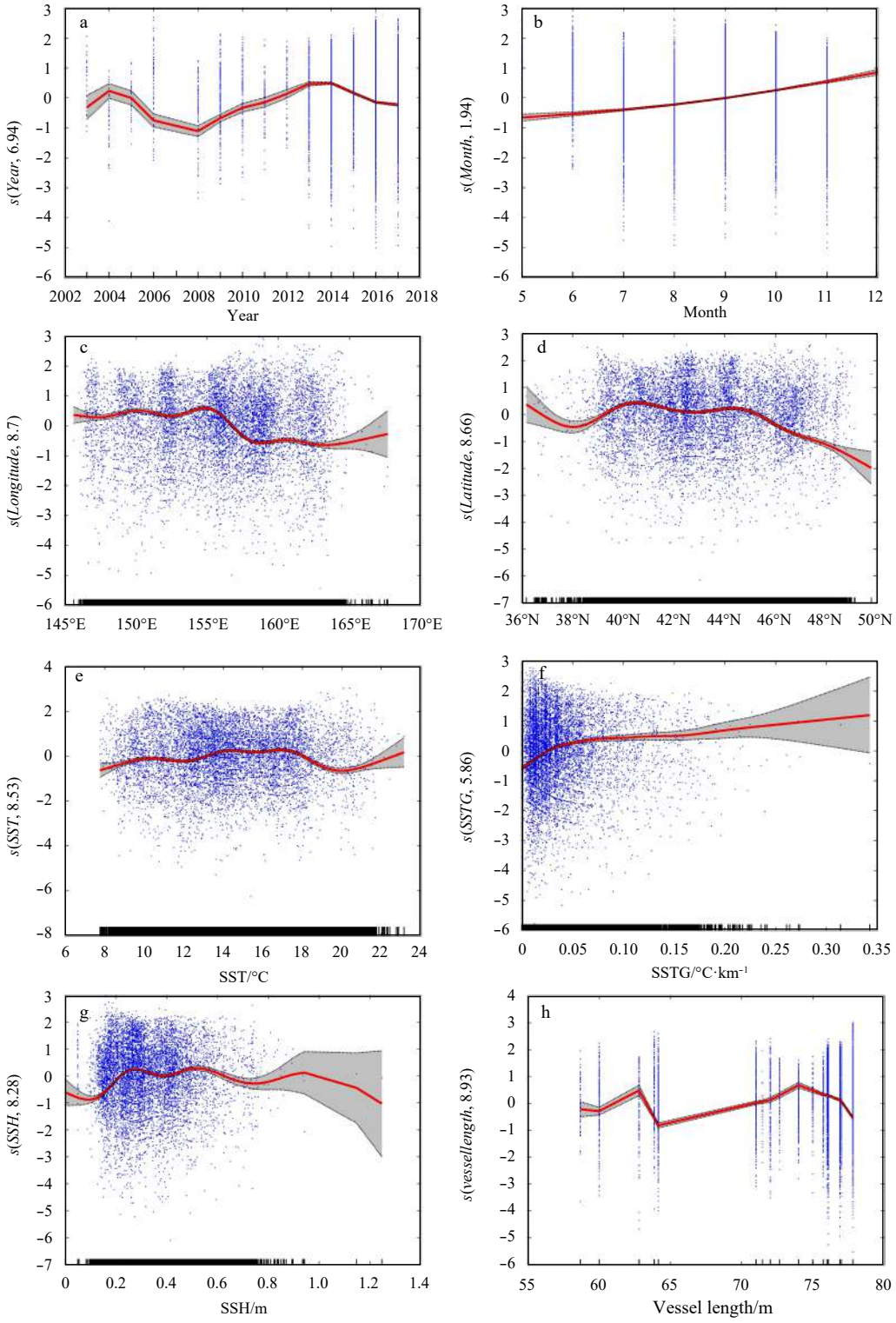

Fig. 6. Effects of temporal, spatial, environmental, and fisheries operational variables on Pacific saury CPUE selected for the best GAM model. Gray lines indicate 95% confidence intervals.

non-linear relationships with Pacific saury CPUE. Contrarily, GAMs are extensions of GLMs in which the explanatory variables have been replaced with smooth functions to deal with nonlinear relationships between the response variable and explanatory variables. However, the smoothening functions of GAM models cannot infer predictions outside of the range of the training data that were used to build the model effectively (Frescino et al., 2001). Any value in the test sets out of the training data range would be assigned to the closest maximum or minimum values of the training data set. Nonetheless, both GLM and GAM showed good performance for estimating relative abundance index in this study. Our experience confirms that different models

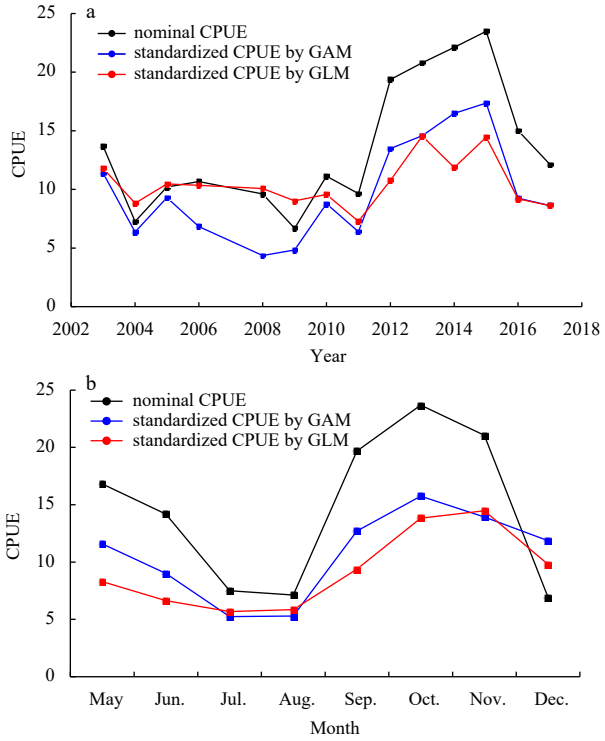

Fig. 7. Standardized CPUE (ton per vessel per day) for the best GLM and GAM in the Northwest Pacific Ocean.

are suitable for different situations and data sources.

In our study, GLM analysis indicated that the *Year*, *Month*, *Longitude*, *Latitude*, *SST*, *SSTG*, *SSH*, *Vessellength* and the interaction terms *Year*×*Month*, *Month*×*Longitude*, and *Month*×*Latitude*, *Longitude*×*Latitude* were all significant variables with highly significant effects on CPUE (p<0.05), whereas GAM analysis indicated that interaction terms include *Year*×*Month*, *Year*×*Longitude*, *Month*×*Longitude*, *Month*×*Latitude*, *Longitude*×*Latitude*. The GAM model effectively explained a higher average percentage of Deviance than the GLM model (Tables 3 and 6). We can see that linear relations may be weak between CPUE values and an explanatory variable (deviance explained was only 21.57% in this case); however, more nonlinear relations were available for CPUE values, whereby deviance explained reached 38.95%. Comparing the results of cross-validation tests in GLM and GAM analyses (Tables 5 and 8), higher Spearman's correlation and lower mean squared error (MSE) between observed and predicted values of test data were observed by GAM. Therefore, GAM is likely to be more suitable than GLM for CPUE standardization of Pacific saury fishing in the Northwest Pacific Ocean.

In summary, the application of statistical models such as GAMs, which perform better than GLMs with the nonlinearity of predictors and spatial autocorrelation, should be given priority for the future CPUE standardization. Furthermore, in most cases, fishery data are commercial fishery data, which do not correspond to a design. The predicted relative resource abundance index will be unreliable and remain considerable uncertainty if the survey area is not enough due to the lack of sampling locations or biased designs. In this situation, GAMs should be preferred (Yu et al., 2013). Therefore, GAMs are appropriate for standardization of Pacific saury CPUE. This study contributes to Pacific saury fishery research by recommending potential statistical approaches for standardizing CPUE, which can provide a solid support to fishery stock assessment.

5 Conclusions

Fishery management and conservation depend heavily on accurate fish stock assessments which always need reliable relative abundance index. In order to get a more reliable fishery CPUE, a well-performed CPUE standardization model is needed. In this study we evaluated the performance of statistical methods including GLMs and GAMs, using CPUE data collected from the Chinese saury fishery in the Northwest Pacific Ocean from 2003 to 2017. Further, we evaluated the impact of spatial, temporal, environmental variables and vessel length on CPUE. The significant variables were used individually in the GLM/GAM to select an optimal GLM/GAM model based on BIC. Subsequently, we used deviance explained and results from 5-fold bootstrap cross-validations to compare the performance of the two types of model. The standardized indices of abundance from the two models studied here suggested a gradual fluctuation trend in Pacific saury CPUE for the Northwest Pacific Ocean. The performance of GLM and GAM were evaluated to determine the most robust model for standardizing CPUE as an index of abundance for Pacific saury. GAM models were more suitable than GLM models for fitting the fishery data. The reason was that GAMs could fit the nonlinear relationships that exist between the response variables and the explanatory variables. These methods could be reproduced efficiently and used to examine the spatial/temporal dy-

Table 9. Nominal and standardized CPUE from 2003 to 2017

Year	Nominal CPUE	Standardized CPUE by GAM	95% CI by GAM	Standardized CPUE by GLM	95% CI by GLM
2003	13.70	11.40	[8.57, 14.23]	11.79	[11.10, 12.48]
2004	7.26	6.36	[5.18, 7.54]	8.82	[8.32, 9.33]
2005	10.20	9.29	[7.72, 10.85]	10.47	[9.88, 11.07]
2006	10.67	6.83	[6.24, 7.42]	10.35	[9.32, 11.39]
2008	9.62	4.36	[3.83, 4.89]	10.08	[8.08, 10.89]
2009	6.69	4.84	[4.52, 5.17]	9.02	[8.69, 9.35]
2010	11.13	8.77	[7.92, 9.62]	9.57	[8.77, 10.38]
2011	9.64	6.40	[5.69, 7.11]	7.28	[7.09, 7.48]
2012	19.37	13.48	[12.60, 14.36]	10.77	[10.25, 11.28]
2013	20.80	14.58	[14.22, 14.95]	14.55	[14.06, 15.04]
2014	22.11	16.48	[16.14, 16.82]	11.87	[11.68, 12.05]
2015	23.48	17.36	[16.93, 17.79]	14.45	[14.19, 14.71]
2016	15.02	9.25	[9.04, 9.45]	9.19	[9.04, 9.34]
2017	12.12	8.62	[8.47, 8.77]	8.62	[8.50, 8.75]

namics of fishing activities.

Although we have obtained some valuable research results, we believe there are still some important shortcomings: (1) The relationship between fish distribution and the environmental factors over time is not static and fishing space is important for discovering the real fish abundance variation. (2) Fish species are not randomly distributed in the ocean, but tend to cluster in certain habitats and may be rare or totally absent in others. (3) AIC is the most popular criterion for model selection when GLMs/GAMs are used to estimate CPUE data. However, AIC may overestimate the effect of the number of parameters in the case of small samples, which may cause unreliable results. In future studies, we will emphasize on model improvement in order to provide better recommendations for management and conservation.

6 Data availability statement

Conflict of interest: The authors declare that they have no conflict of interest.

Data availability: The datasets created during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Human and animal rights: All applicable international, national, and/or institutional guidelines for the care and use of animals were followed.

Acknowledgements

We thank Siquan Tian, Bai Li, Jie Cao and Luoliang Xu for their valuable comments and advice. Thanks are also given to other laboratory colleagues for field and laboratory work assistance.

References

Arlot S, Celisse A. 2010. A survey of cross-validation procedures for model selection. Statistics Surveys, 4: 40–79, doi: 10.1214/09-SS054

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(2–3): 209–227, doi: 10.1016/j.fishres.2004.08.026

Chen Xinjun, Liu Bilin, Chen Yong. 2008. A review of the development of Chinese distant-water squid jigging fisheries. Fisheries Research, 89(3): 211–221, doi: 10.1016/j.fishres.2007.10.012

Denis V, Lejeune J, Robin J P. 2002. Spatio-temporal analysis of commercial trawler data using general additive models: patterns of Loliginid squid abundance in the north-east Atlantic. ICES Journal of Marine Science, 59(3): 633–648, doi: 10.1006/jmsc.2001.1178

Erisman B E, Allen L G, Claisse J T, et al. 2011. The illusion of plenty: hyperstability masks collapses in two recreational fisheries that target fish spawning aggregations. Canadian Journal of Fisheries and Aquatic Sciences, 68(10): 1705–1716, doi: 10.1139/f2011-090

Frescino T S, Edwards T C Jr, Moisen G G. 2001. Modeling spatially explicit forest structural attributes using generalized additive models. Journal of Vegetation Science, 12(1): 15–26, doi: 10.1111/j.1654-1103.2001.tb02613.x

Harley S J, Myers R A, Dunn A. 2001. Is catch-per-unit-effort proportional to abundance?. Canadian Journal of Fisheries and Aquatic Sciences, 58(9): 1760–1772, doi: 10.1139/f01-112

He Min, Song Wenling, Chen Xingfang. 1999. Typhoon activity in the Northwest Pacific in relation to El Niño/La nina events. Journal of Tropical Meteorology (in Chinese), 15(1): 17–25

Howell E A, Kobayashi D R. 2006. El Niño effects in the Palmyra Atoll region: oceanographic changes and bigeye tuna (*Thunnus obesus*) catch rate variability. Fisheries Oceanography, 15(2): 477–489

Hua Chuanxiang, Zhu Qingcheng, Xu Wei. 2010. Fishing ground distribution of *cololabis saira* in the Northwestern Pacific. Shandong Fisheries (in Chinese), 27(10): 10–13

Huang Hongliang, Zhang Xun, Xu Baosheng, et al. 2005. Preliminary analysis on the fishing grounds of *Cololabis saira* in the North Pacific Ocean. Marine Fisheries, 27(3): 206–212

Kohavi R. 2001. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence. Stanford, CA: Morgan Kaufmann Publishers Inc

Lin Longshan. 2003. Fishery survey of Stick-held Net for *Cololabis saira* in Taiwan. Marine Fisheries (in Chinese), (4): 200–203

Martínez-Rincón R O, Ortega-García S, Vaca-Rodríguez J G. 2012. Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (*Acanthocybium solandri*) in the Mexican tuna purse-seine fishery. Ecological Modelling, 233: 20–25, doi: 10.1016/j.ecolmodel.2012.03.006

Maunder M, Punt A E. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research, 70(2–3): 141–159, doi: 10.1016/j.fishres.2004.08.002

Maunder M N, Start P J. 2003. Fitting fisheries models to standardised CPUE abundance indices. Fisheries Research, 63(2): 43–50

Menard SW. 1995. Applied Logistic Regression Analysis. Thousand Oaks, CA: SAGE

Nishida T, Chen Dinggeng. 2004. Incorporating spatial autocorrelation into the general linear model with an application to the yellowfin tuna (*Thunnus albacares*) longline CPUE data. Fisheries Research, 70(2–3): 265–274, doi: 10.1016/j.fishres.2004.08.008

Ortiz M, Arocha F. 2004. Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: billfish species in the Venezuelan tuna longline fishery. Fisheries Research, 70(2–3): 275–297, doi: 10.1016/j.fishres.2004.08.028

Quinn G P, Keough M J. 2002. Experimental Design and Data Analysis for Biologists. Cambridge: Cambridge University Press

Rodríguez-Marín E, Arrizabalaga H, Ortiz M, et al. 2003. Standardization of bluefin tuna, (*Thunnus thynnus*) catch per unit effort in the baitboat fishery of the Bay of Biscay (Eastern Atlantic). ICES Journal of Marine Science, 60(1): 1216–1231

Shen Jianhua, Han Shixin, Fan Wei, et al. 2004. Saury Resource and Fishing Grounds in the Northwest Pacific. Marine Fisheries (in Chinese), 26(1): 61–65

Shono H. 2005. Is model selection using Akaike's information criterion appropriate for catch per unit effort standardization in large samples?. Fisheries Science, 71(5): 978–986, doi: 10.1111/j.1444-2906.2005.01054.x

Stephens A, Maccall A. 2004. A multispecies approach to subsetting logbook data for purposes of estimating CPUE. Fisheries Research, 70(2–3): 299–310, doi: 10.1016/j.fishres.2004.08.009

Sun Manchang, Ye Xuchang, Zhang Jian, et al. 2003. Probe into Pacific saury fisheries in the northwest Pacific Ocean. Marine Fisheries (in Chinese), 25(3): 112–115

Takasuka A, Kuroda H, Takeshi O, et al. 2014. Occurrence and density of Pacific saury *Cololabis saira* larvae and juveniles in relation to environmental factors during the winter spawning season in the Kuroshio Current system. Fisheries Oceanography, 23(4): 304–321, doi: 10.1111/fog.12065

Tian Y J, Akamine T, Suda M. 2003. Variations in the abundance of Pacific saury (*Cololabis saira*) from the Northwestern Pacific in relation to oceanic-climate changes. Fisheries Research, 60(2–3): 439–454, doi: 10.1016/S0165-7836(02)00143-1

Tian Y J, Ueno Y, Suda M, et al. 2004. Decadal variability in the abundance of Pacific saury and its response to climatic/oceanic regime shifts in the northwestern subtropical Pacific during the last half century. Journal of Marine Systems, 52: 235–257, doi: 10.1016/j.jmarsys.2004.04.004

Tien B D, Lofman O, Revhaug I, et al. 2011. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Natural Hazards, 59(3):

1413–1444, doi: 10.1007/s11069-011-9844-2

Tseng C T, Su N J, Sun C L, et al. 2013. Spatial and temporal variability of the Pacific saury (*Cololabis saira*) distribution in the northwestern Pacific Ocean. ICES Journal of Marine Science, 70(5): 991–999, doi: 10.1093/icesjms/fss205

Venables W N, Dichmont C M. 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fisheries Research, 70(2–3): 319–337, doi: 10.1016/j.fishres.2004.08.011

Walsh W A, Kleiber P. 2001. Generalized additive model and regression tree analyses of blue shark (*Prionace glauca*) catch rates by the Hawaii-based commercial longline fishery. Fisheries Research, 53(2): 115–131, doi: 10.1016/S0165-7836(00)00306-4

Wang Zhizu, Zuo Juncheng, Chen Meixiang, et al. 2012. Relationship between El Niño and sea surface temperature variation in coastal region of Yellow Sea and East China Sea. Journal of Hohai University (Natural Sciences) (in Chinese), 40(4): 461–468

Ward H G M, Askey P J, Post J R. 2013. A mechanistic understanding of hyperstability in catch per unit effort and density-dependent catchability in a multistock recreational fishery. Canadian Journal of Fisheries and Aquatic Sciences, 70(10): 1542–1550, doi: 10.1139/cjfas-2013-0264

Watanabe K, Tanaka E, Yamada S, et al. 2006. Spatial and temporal migration modeling for stock of Pacific saury *Cololabis saira* (Brevoort), incorporating effect of sea surface temperature. Fisheries Science, 72(6): 1153–1165, doi: 10.1111/j.1444-2906.2006.01272.x

Wood S N. 2006. Generalized Additive Models: An Introduction with R. London: Chapman and Hall/CRC, 410

Wu Yue, Huang Hongliang, Liu Jian, et al. 2015. Spatiotemporal distribution pattern of saury fishing grounds and catch yield per unit effort in the Northern Pacific high sea in 2014. Fishery Modernization (in Chinese), 42(3): 61–64

Xia Hui. 2008. The illumination distribution model of the pacific saury (*Cololabis saira*) stick-held dip net fishing (in Chinese) [dissertation]. Shanghai: Shanghai Ocean University, 1-54

Xu Wei, Zhu Qingcheng, Zhang xiancun, et al. 2005. Bouke net fishing technology of Pacific saury in the Northweatern Pacific. Shangdong Fisheries (in Chinese), 22(10): 43–46

Yan Lei. 2012. The Relationship between the distribution of saury fishing ground and its environmental factors (in Chinese) [dissertation]. Shanghai: Shanghai Ocean University, 1–51

Yan Lei, Zhu Qingcheng, Zhang Yang, et al. 2012. Fishing ground distribution of saury and its correlation with SST in the Northern Pacific high sea in 2010. Journal of Shanghai Ocean University (in Chinese), 21(4): 609–615

Yang Xiulan, Wang Pengfei, Jiao Yulong, et al. 2005. Study on the culture technique in the middle stage and the growing character of *Apostichopus japonicus*. Shandong Fishery (in Chinese), 22(10): 43–46

Yu Hao, Jiao Yan, Carstensen L W. 2013. Performance comparison between spatial interpolation and GLM/GAM in estimating relative abundance indices through a simulation study. Fisheries Research, 147: 186–195, doi: 10.1016/j.fishres.2013.06.002

Yu Yuefeng, Zhang Xun, Huang Hongliang, et al. 2006. Study on attracting fish method of stick-held net for *Cololabis saira*. Journal of Zhejiang Ocean University (Natural Science) (in Chinese), 25(2): 154–156

Zhang Xiaomin, Zhu Qingcheng, Hua Chuanxiang. 2015. Fishing ground distribution of saury and its correlation with marine environment factors in the Northern Pacific high sea in 2013. Journal of Shanghai Ocean University (in Chinese), 24(5): 773–782

Zhang Yang, Zhu Qingcheng, Yan Lei, et al. 2013. Preliminary study on biological characteristics of *Cololabis Saira* in the Northwest Pacific ocean in Spring. Transactions of Oceanology and Limnology (in Chinese), (1): 53–60

Zhu Qingcheng, Hua Chuanxiang, Xu Wei, et al. 2006a. The fishing ground distribution of *Cololabis saira* and its relationship with water temperature factors in the Northwestern Pacific from July to September. Marine Fisheries (in Chinese), 28(3): 228–233

Zhu Guoping, Zhu Qingcheng, Chen Jintao, et al. 2006b. Preliminary study on relationship between *Cololabis saira* fishing ground and temperature factor in the Northern Pacific Ocean. Marine Sciences (in Chinese), 30(7): 91–96

Zou Xiaorong, Zhu Qingcheng. 2006. Preliminary analysis on the relationship between the distribution of fishing ground of pacific saury (*Cololabis saira*) and SST in northwest pacific. Journal of Zhanjiang Ocean University (in Chinese), 26(6): 26–30