NPFC-2019-SSC PS05-WP01 (Rev. 1)

**Standardization of CPUE data of Pacific saury (*Cololabis saira*) caught by the Chinese stick-held dip net fishery**

**Summary**

China’s Pacific saury fishery began in 2003. Most of the Pacific saury catch is harvested by the stick-held dip net fishery. In this paper, we standardized catch per unit fishing effort (CPUE) using Generalized linear model (GLM) and generalized additive model (GAM). Four groups of independent variables were considered in the standardization: spatial variables (latitude and longitude), temporal variables (year and month), vessel length and environmental variables (SST, SSTG and SSH). Log-CPUE was treated as the dependent variable and its error was assumed to follow normal distribution in each model. The model selections of GLM and GAM were based on the BIC. From the results, Higher Spearman’s correlation and lower mean squared error were observed by GAM. Besides, the standardized CPUE trend of GAM model is similar with that of nominal CPUE. Therefore, we prefer to choose the best GAM model to estimated standardized CPUE of Pacific saury.

**1. Introduction**

At the beginning of the 20th century, the first stick-held net fishing vessel (changed from squid jigging vessel) from China went to the high seas for fishing Pacific saury in the northwest Pacific Ocean (NWP). Now, about 50 PS vessels from China operate in the NWP, after developing for more than ten years. The main fishing area of China is shown in Figure 1. The catches of Pacific saury are higher in region 146 - 157 °E and 39 - 45 °N (Fig.2).

**2. Data sources**

Full-commercial fishery data were from 2013-2018, which were derived from Technical Group for Pacific saury Fishery, Distant-water Fishery Society of China. Sea surface temperature (SST) data were derived from National Oceanic and Atmospheric Administration (NOAA; [ftp.nodc.noaa.gov](ftp://ftp.nodc.noaa.gov)). The spatial-temporal resolution of the SST data is daily at 0.1°×0.1° grid. Sea surface height (SSH) data were derived from Archiving Validation and Interpolation of Satellite Oceanographic Data (AVISO; [www.aviso.altimetry.fr](http://www.aviso.altimetry.fr)). The spatial-temporal resolution of the data is SSH daily at 0.25°×0.25° grid. Sea surface temperature gradients (SSTG) were calculated by Gradient Magnitude (GM) method (Ortiz, 2004; Howell, 2006). The formula is:



where , , and are SST values of 4 consecutive grids respectively, *i* and *j* is the numbering of row and column, is the longitudinal distance (km) between (*j*-1)th and (*j*+1)th columns, is the latitudinal distance (km) between (*i*-1)th and (i+1)th rows, is SSTG value of the current grid (°C/km).

This study extracted the corresponding oceanographic data from the nearest grid to the grid where the fishery data existed at the same date.

In this study, nominal CPUE were defined as catch per day per vessel, unit: ton/day/v.

**3. Factors that may affect the CPUE of PS fishery**

The 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, 2003); therefore, temporal variables (year and month), spatial variables (longitude and latitude) were included in the analysis. The formation of the Pacific saury fishing grounds is tightly associated with the marine environment (Zhu, 2006). Thus, the *SST*, *SSTG*, *SSH* were included in the analysis. In addition, the vessel length may affect the quantity of the catch, which was included in this study.

**4. Statistical model and model selection strategy**

Both generalized linear model (GLM) and generalized additive model (GAM) were used to standardize the CPUEs.

The full GLM model was:

*log(CPUE) =Year + Month + Longitude\_c + Latitude\_c + Sst + Sstg* *+ Ssh +Vessellength\_c + interaction+ε*

The full GAM model was:

*log(CPUE)*= *Year*+ *Month*+ *longitude\_c* + *latitude\_c + s(sst) + s(sstg) + s(ssh) +* *s(vessel length) +interaction+ε*

*interaction* is an interaction term representing the interactive effect of spatial and temporal factors for the Pacific saury. Full model interaction includes all the possible combination of year, month, longitude\_c, latitude\_c.

The optimal model was selected using the Bayesian information criterion (BIC).

The way to calculate the standardization CPUE is the yearly mean of fitted CPUE from the best model. The formula is,

where, is CPUE indices in *i*th year, is the observation number in *i*th year, is the *k*th fitted CPUE data in *i*th year.

**5. Results and** **Discussions**

In this study we used two models to standardize the CPUEs. Variance Inflation Factor (VIF) and Spearman correlation coefficient among explanatory variables were calculated (Table1). The Maximum VIF<5, indicates there is no serious multi-collinearity (Tien, 2011). Residuals from both approaches showed an approximately normal distribution around 0, which indicated that the model assumptions were satisfied. The results were shown in Fig. 3 and Fig. 4.

We used same explanatory variables in GLM and GAM analysis (Table 2). The best GLM model selected based on BIC values is shown in Table 3. The summary of fitting a GLM for the optimal model is shown in Table 4. All explanatory variables are highly significant (*p*<0.01). The best GAM model selected based on BIC values is shown in Table 6. The summary of fitting a GAM for the best model is shown in Table 7. All explanatory variables are highly significant (*p*<0.01) except for longitude\_c.

Table 9 and figure 8 shows the annual changes of nominal CPUE and standardized CPUE by GAM and GLM models. There are few differences between fitted CPUEs data by GLM and GAM, which may be related to the assumption of relationships between CPUEs and explanatory variables.

Comparing the results of cross validation tests in GLM and GAM analyses (Table 5 and 8), higher Spearman’s correlation and lower mean squared error (MSE) between observed and predicted of test data were observed by GAM, so we prefer to choose the best GAM model to estimate standardized CPUE.

We standardized CPUE in accordance with the standardization protocol (NPFC - 2017 - TWG PSSA - Report Annex D). The checklist is shown in Appendix 1.

**References**

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

ORTIZ M, AROCHA F. 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[J]. Fisheries Research, 2004, 70(1): 275-294.

Tian Y J. Variations in the abundance of Pacific saury (*Cololabis saira*)from the Northwestern Pacific in relation to oceanic—climate changes[J]. Fisheries Research, 2003, 60:439-454.

Tien Bui, D., Lofman, O., Revhaug, I., Dick, O. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Nat. Hazards ,2011,59 (3), 1413-1444.

ZHU Guo-ping, ZHU Qing-cheng. Preliminary study on relationship between Cololabis saira fishing ground and temperature factor in the Northern Pacific Ocean[J]. Marine Sciences, 2006, 30(7):91-96.

**Tables**:

Table 1 Variance Inflation Factor (VIF) and Spearman correlation coefficient among explanatory variables

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| coefficient/p value | VIF | year | month | Longitude | Latitude | SST | SSTG | SSH | vessellength |
| year | 1.16 |  | <0.001 | <0.001 | 0.8468 | <0.001 | <0.001 | <0.001 | <0.001 |
| month | 3.34 | -0.0988 |  | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Longitude | 4.80 | 0.2087 | -0.8245 |  | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Latitude | 4.57 | 0.0014 | -0.5530 | 0.6586 |  | <0.001 | <0.001 | <0.001 | <0.001 |
| SST | 1.46 | 0.1269 | 0.3799 | -0.3298 | -0.4342 |  | <0.001 | <0.001 | 0.0022 |
| SSTG | 1.35 | -0.1205 | 0.3907 | -0.4888 | -0.4074 | 0.2183 |  | <0.001 | <0.001 |
| SSH | 3.54 | 0.1593 | 0.3902 | -0.3860 | -0.7938 | 0.5196 | 0.3007 |  | 0.3749 |
| vessellength | 1.02 | 0.1260 | -0.0497 | 0.0693 | 0.0402 | 0.0224 | -0.0289 | -0.0065 |  |

1. Spearman correlation coefficient are under the slope line; *p* values are above the slope line.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Cases | Categorical or continuous | Details | Note |
| Year | *Year* | 6 categories | 6 years from 2013 to2018 |  |
| Month | *Month* | 8 categories | 8 months from May to December |  |
| Longitude | *Longitude\_c* | 23 categories | Longitude<144° ; 144°≦Longitude＜145°，145°≦Longitude＜146，…, Longitude>165° | at intervals of 1° |
| Latitude | *Latitude\_c* | 13 categories | Latitude<38° ; 38°≦Latitude＜39°，39°≦Latitude＜40，…, Latitude >48° | at intervals of 1° |
| Sea surface temperature | *Sst*  *Sst\_c* | spline  12 categories | Sst<10℃;10℃≦Sst＜11℃，11℃≦Sst＜12℃，…, 19℃≦Sst≤20℃; Sst>20℃ | at intervals of 1℃ |
| Sea surface temperature gradient | *Sstg* | continues（spline） |  |  |
| Sea surface height | *Ssh* | continues（spline） |  |  |
| Vessel length | *Vessellength*  *Vessellength\_c* | continues（spline）  9 categories | Vessellength＜64m，64m≦Vessellength＜76m，…, 76m≦Vessellength | at intervals of 2m |

Table 3 Best GLM selected based on AIC and BIC values

|  |  |  |  |
| --- | --- | --- | --- |
| Best model in GLM analysis | *R*2 | BIC | deviance explained% |
| Ln(CPUE)~*Intercept*+*Year*+*Month*+*Longitude\_c*+*Latitude\_c*+*Sst*+*Sstg*+*Ssh*+*Vessellength\_c*+*Year:Month*+ *Year: Latitude\_c+ε* | 0.3440 | 55250 | 33.94 |

Table 4 Anova test for best GLM model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Df | Deviance | Resid. Df | Resid. Dev | *F* | Pr(>*F*) |  |
| NULL |  |  | 18721 | 29789 |  |  |  |
| factor(year) | 5 | 1726.1 | 18716 | 28063 | 328.41 | < 2.2E-16 | \*\*\* |
| factor(month) | 7 | 5732.5 | 18709 | 22331 | 779.07 | < 2.2E-16 | \*\*\* |
| factor(longitude\_c) | 23 | 641.7 | 18686 | 21689 | 26.54 | < 2.2E-16 | \*\*\* |
| factor(latitude\_c) | 12 | 381.3 | 18674 | 21308 | 30.23 | < 2.2E-16 | \*\*\* |
| Sst | 1 | 30.5 | 18673 | 21277 | 29.01 | 7.27E-08 | \*\*\* |
| Sstg | 1 | 19.1 | 18672 | 21258 | 18.20 | 2.00E-05 | \*\*\* |
| Ssh | 1 | 20.6 | 18671 | 21237 | 19.63 | 9.47E-06 | \*\*\* |
| factor(vessellength\_c) | 5 | 168.3 | 18666 | 21069 | 32.03 | < 2.2E-16 | \*\*\* |
| factor(year):factor(month) | 28 | 910.0 | 18638 | 20159 | 30.92 | < 2.2E-16 | \*\*\* |
| factor(year):factor(latitude\_c) | 49 | 618.7 | 18589 | 19540 | 12.01 | < 2.2E-16 | \*\*\* |

Significant code: \*\*\* 0.001, \*\*0.01, \*0.05

Table 5. The Five-fold cross validation for the best GLM.

|  |  |  |
| --- | --- | --- |
| case | cor\_GLM\_test | MSE\_GLM\_test |
| 1 | 0.5874 | 1.0740 |
| 2 | 0.5851 | 1.0448 |
| 3 | 0.5828 | 1.0812 |
| 4 | 0.5951 | 1.0319 |
| 5 | 0.5716 | 1.0804 |

The spearman’s correlation coefficient is showed in the table.

Table 6 Best GAM selected based on AIC and BIC values

|  |  |  |  |
| --- | --- | --- | --- |
| Best model in GAM analysis | *R*2 | BIC | deviance explained% |
| Ln(CPUE)~*Intercept*+*Year*+*Month*+*Longitude*\_*c*+*Latitude*\_*c*+s(*Sst)*+s(*Sstg)*+s(*Ssh)*+s(*Vessellength)*+*Year:Month*+*Year:Latitude\_c* +*ε* | 0.3712 | 54508 | 37.63 |

Table 7. Anova test for best GAM model

Parametric Terms:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | df | *F* | P-value |  |
| factor(year) | 5 | 5.84 | 2.15E-05 | \*\*\* |
| factor(month) | 7 | 7.37 | 7.28E-09 | \*\*\* |
| factor(longitude\_c) | 23 | 11.61 | < 2.2E-16 | \*\*\* |
| factor(latitude\_c) | 12 | 2.00 | 0.020177 | \* |
| factor(year):factor(month) | 31 | 16.89 | < 2.2E-16 | \*\*\* |
| factor(year):factor(latitude\_c) | 54 | 11.30 | < 2.2E-16 | \*\*\* |

Approximate significance of smooth terms:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Edf | Ref.df | *F* | p-value |  |
| s(sst) | 7.76 | 8.49 | 4.49 | 2.22E-05 | \*\*\* |
| s(sstg) | 4.14 | 5.20 | 5.98 | 1.24E-05 | \*\*\* |
| s(ssh) | 7.72 | 8.53 | 3.65 | 0.000286 | \*\*\* |
| s(vessellength) | 8.89 | 8.99 | 116.43 | < 2.2E-16 | \*\*\* |

Significant code: \*\*\* 0.001, \*\*0.01, \*0.05

Table 8. The cross validation for the best GAM.

|  |  |  |
| --- | --- | --- |
| case | cor\_GAM\_test | MSE\_GAM\_test |
| 1 | 0.5978 | 1.0233 |
| 2 | 0.6218 | 1.0335 |
| 3 | 0.6172 | 1.0021 |
| 4 | 0.6081 | 0.9995 |
| 5 | 0.6091 | 0.9976 |

The spearman’s correlation coefficient is showed in the table.

Table 9 Nominal and standardized CPUE from 2013 to 2018.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Nominal CPUE | SD of Nominal CPUE | Standardized CPUE by GLM | SD by GLM | 95% CI by GLM | | Standardized CPUE by GAM | SD by GAM | 95% CI by GAM | |
| 2013 | 20.80 | 19.17 | 13.79 | 4.84 | [13.44 | 14.18] | 14.01 | 5.53 | [13.58 | 14.44] |
| 2014 | 22.11 | 20.62 | 15.86 | 8.27 | [15.49 | 16.23] | 16.27 | 9.33 | [15.87 | 16.66] |
| 2015 | 23.48 | 21.21 | 17.57 | 10.55 | [16.95 | 18.18] | 17.78 | 11.13 | [17.16 | 18.36] |
| 2016 | 15.02 | 18.87 | 9.06 | 6.31 | [8.80 | 9.30] | 9.36 | 7.06 | [9.07 | 9.65] |
| 2017 | 12.12 | 12.82 | 8.39 | 4.22 | [8.22 | 8.55] | 8.57 | 4.72 | [8.38 | 8.76] |
| 2018 | 23.13 | 24.48 | 15.58 | 10.92 | [15.15 | 16.11] | 15.96 | 11.87 | [15.52 | 16.39] |

**Figures**:

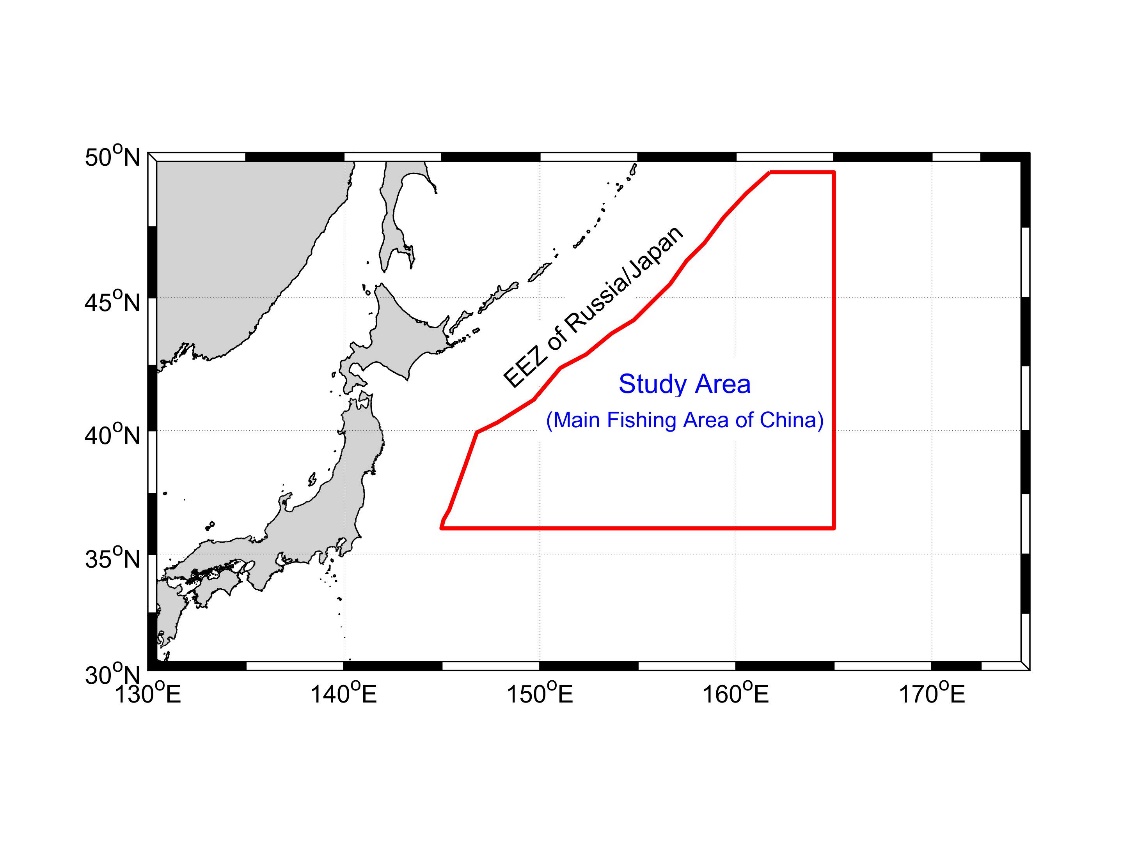


Fig. 1 The main fishing area of China

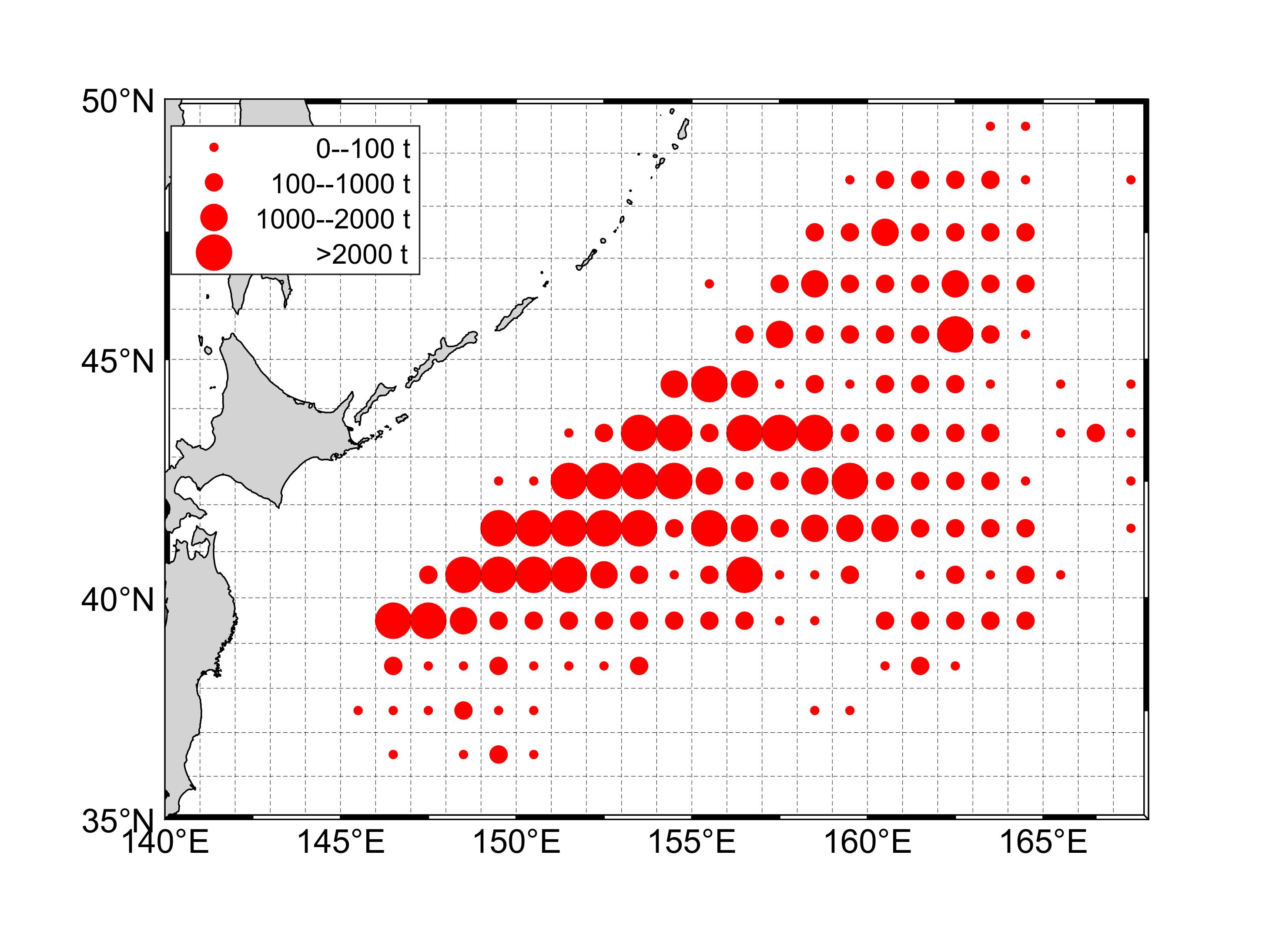


Fig. 2 Distribution of catch (ton) for China Pacific saury fishing fleets in the Northwestern Pacific Ocean from 2013 to 2018

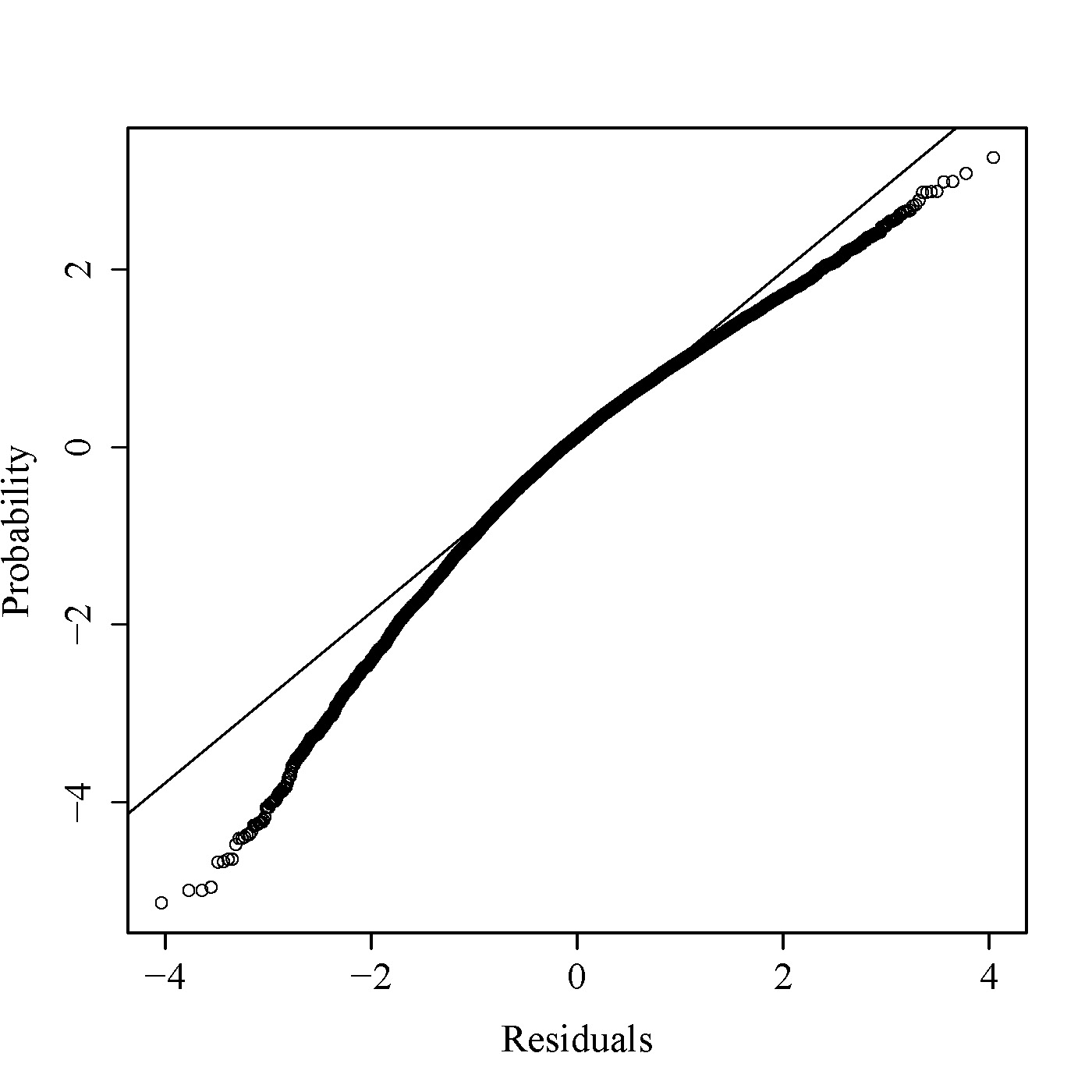
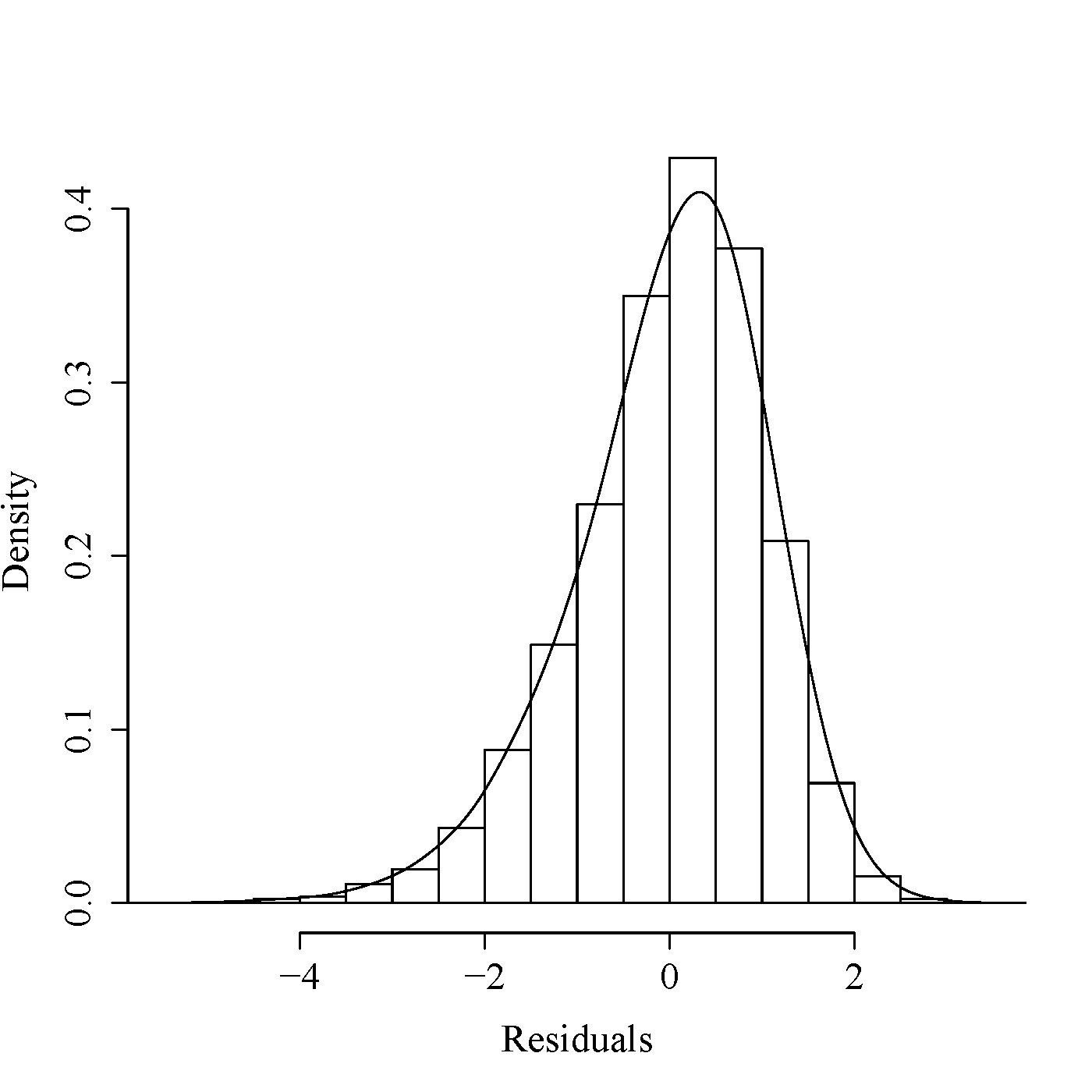
 

Fig. 3 Normal distribution checks, Q-Q plot and histogram of residuals for the GLM optimal model.

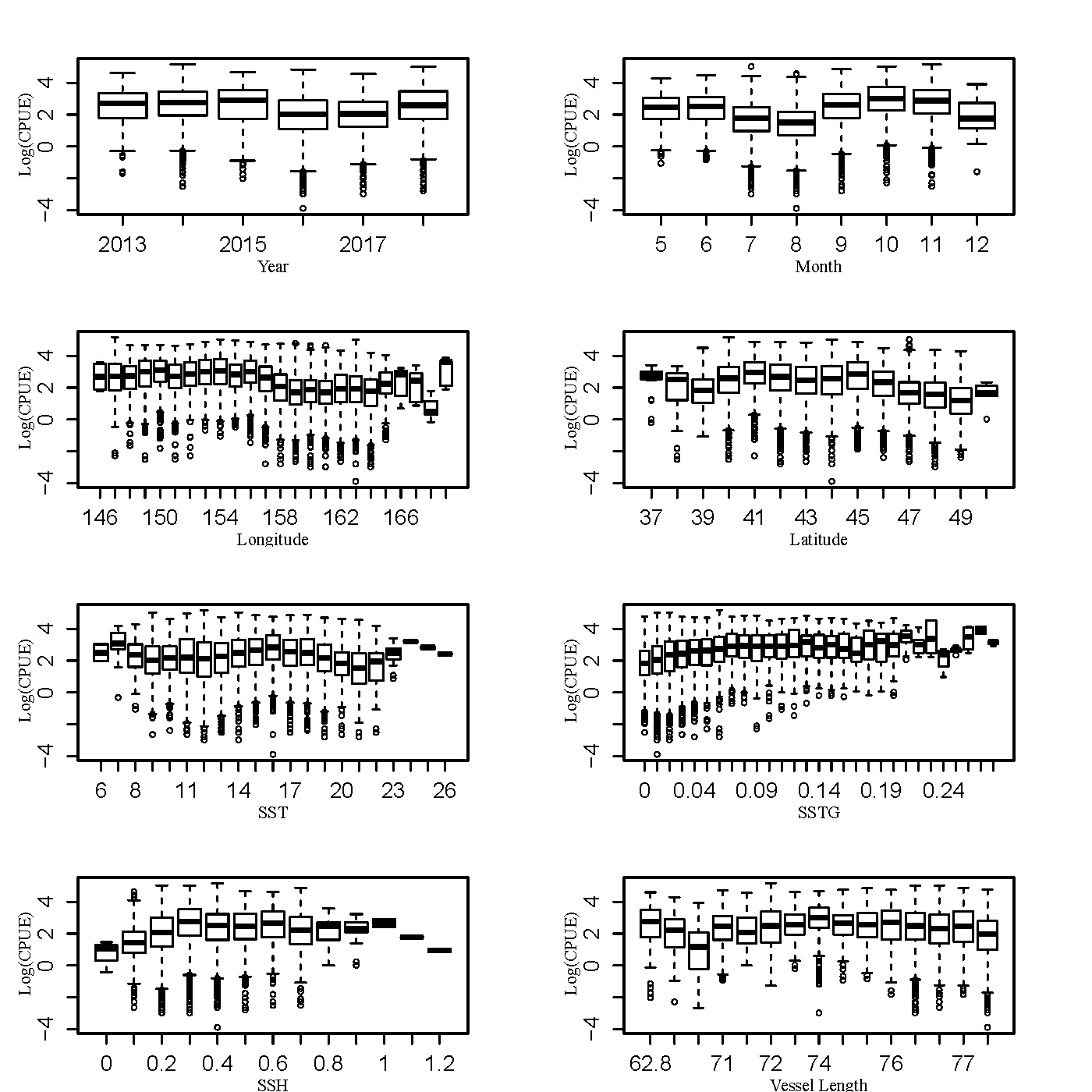


Fig. 4 Boxplots of nominal CPUE and explanatory variables

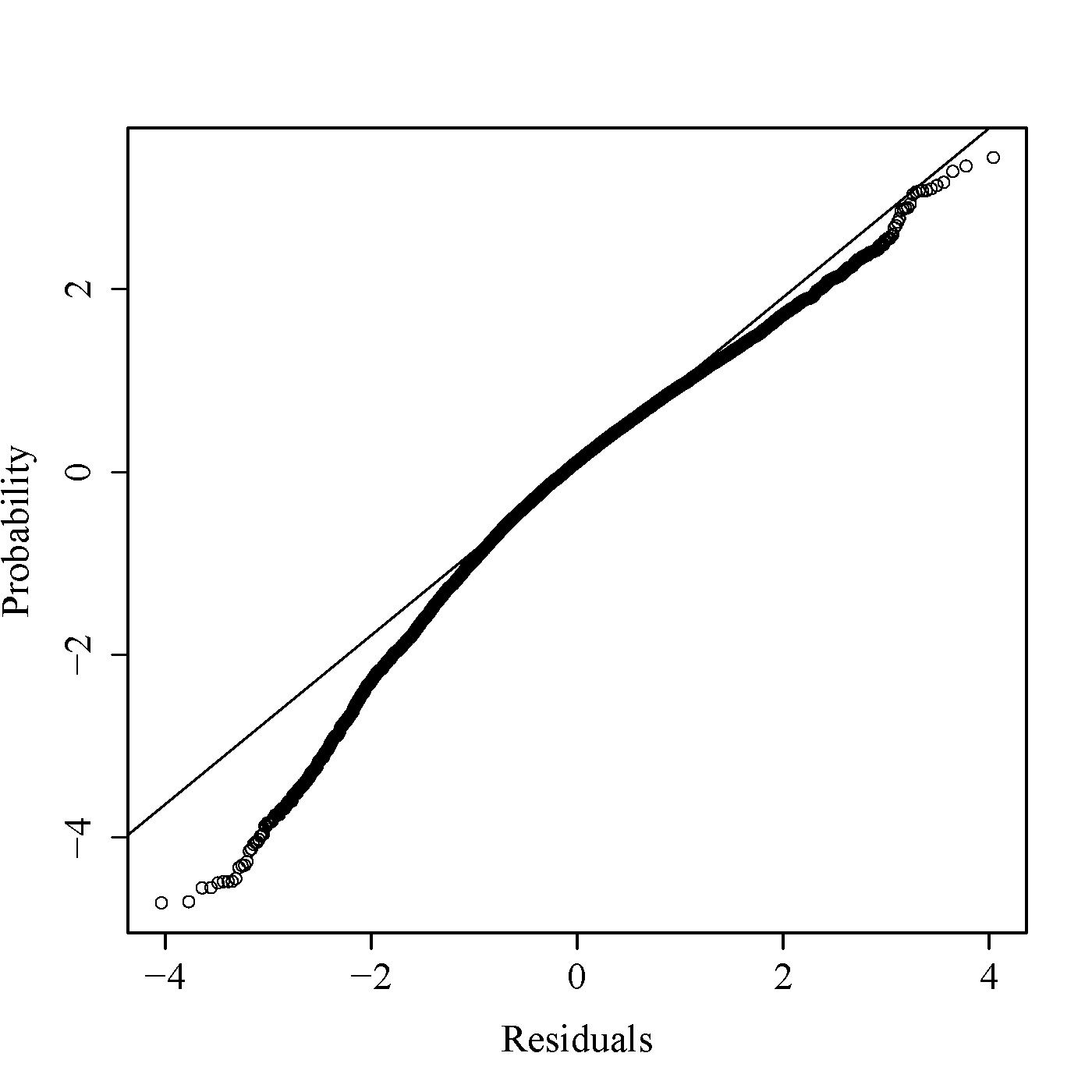
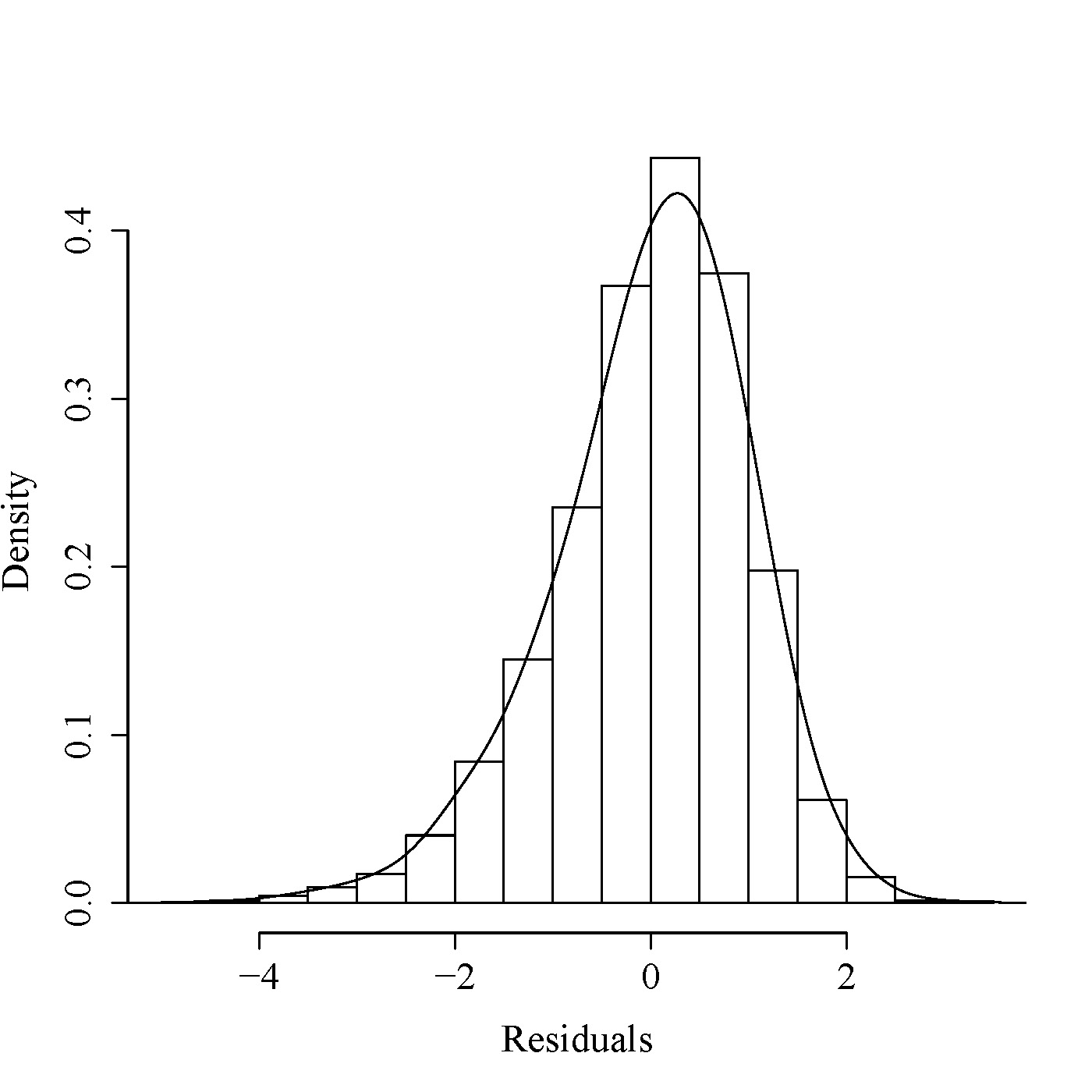
 

Fig. 5 Normal distribution checks, Q-Q plot and histogram of residuals for the GAM optimal model.

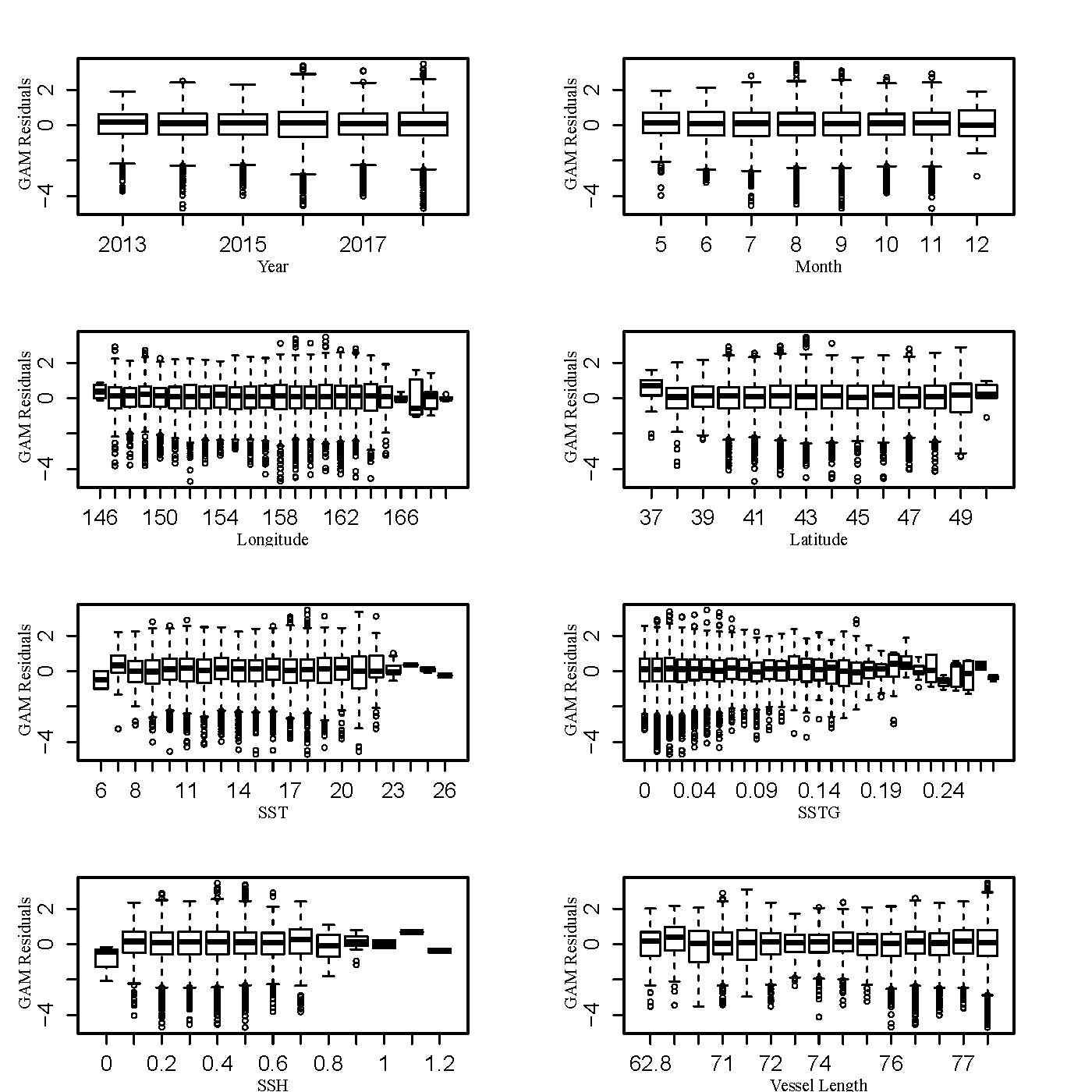


Fig. 6 Boxplots of residuals and explanatory variables fitted by best GAM

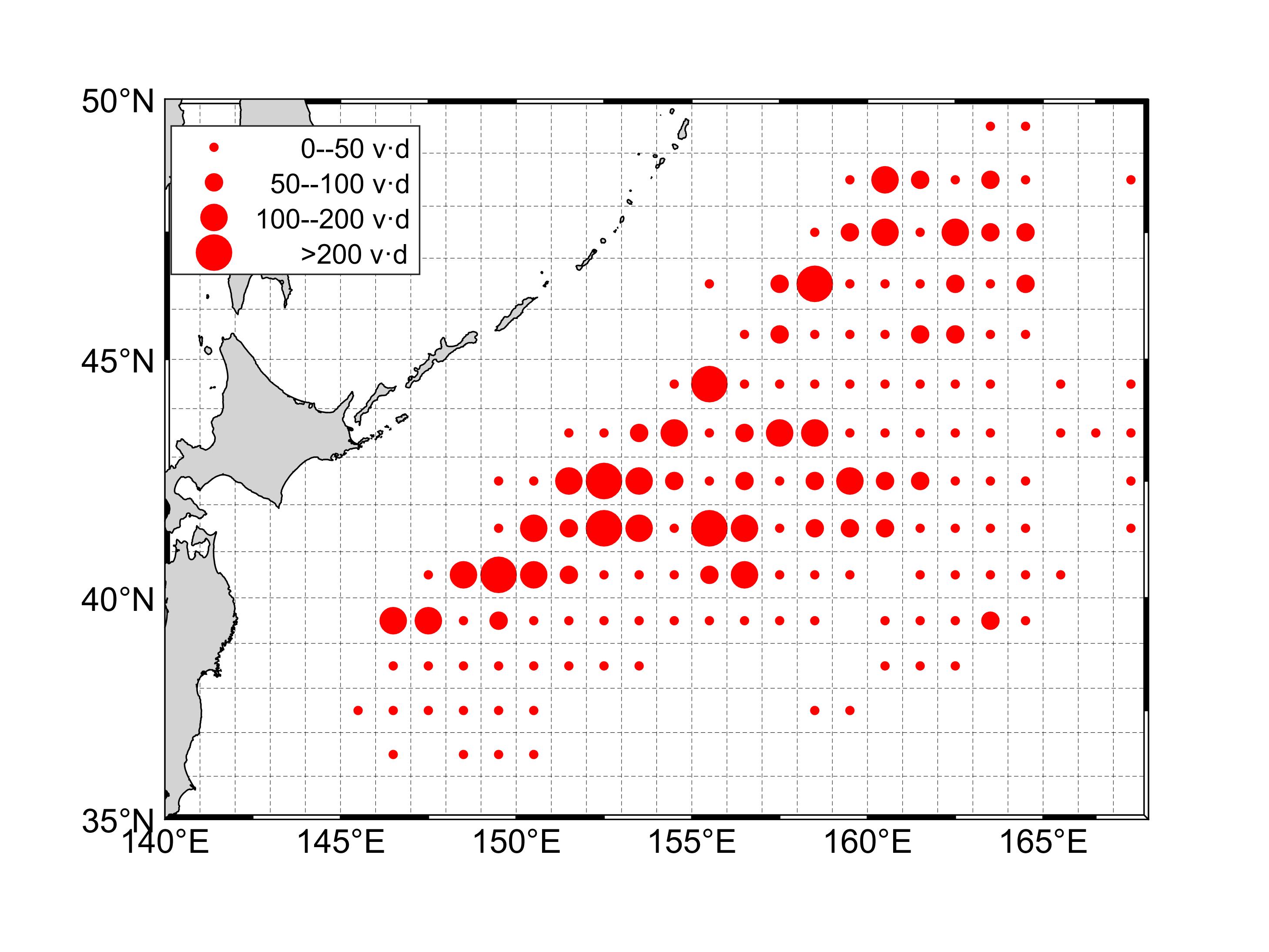


Fig. 7 Distribution of fishing effort for China Pacific saury fishing fleets in the Northwestern Pacific Ocean from 2013 to 2018

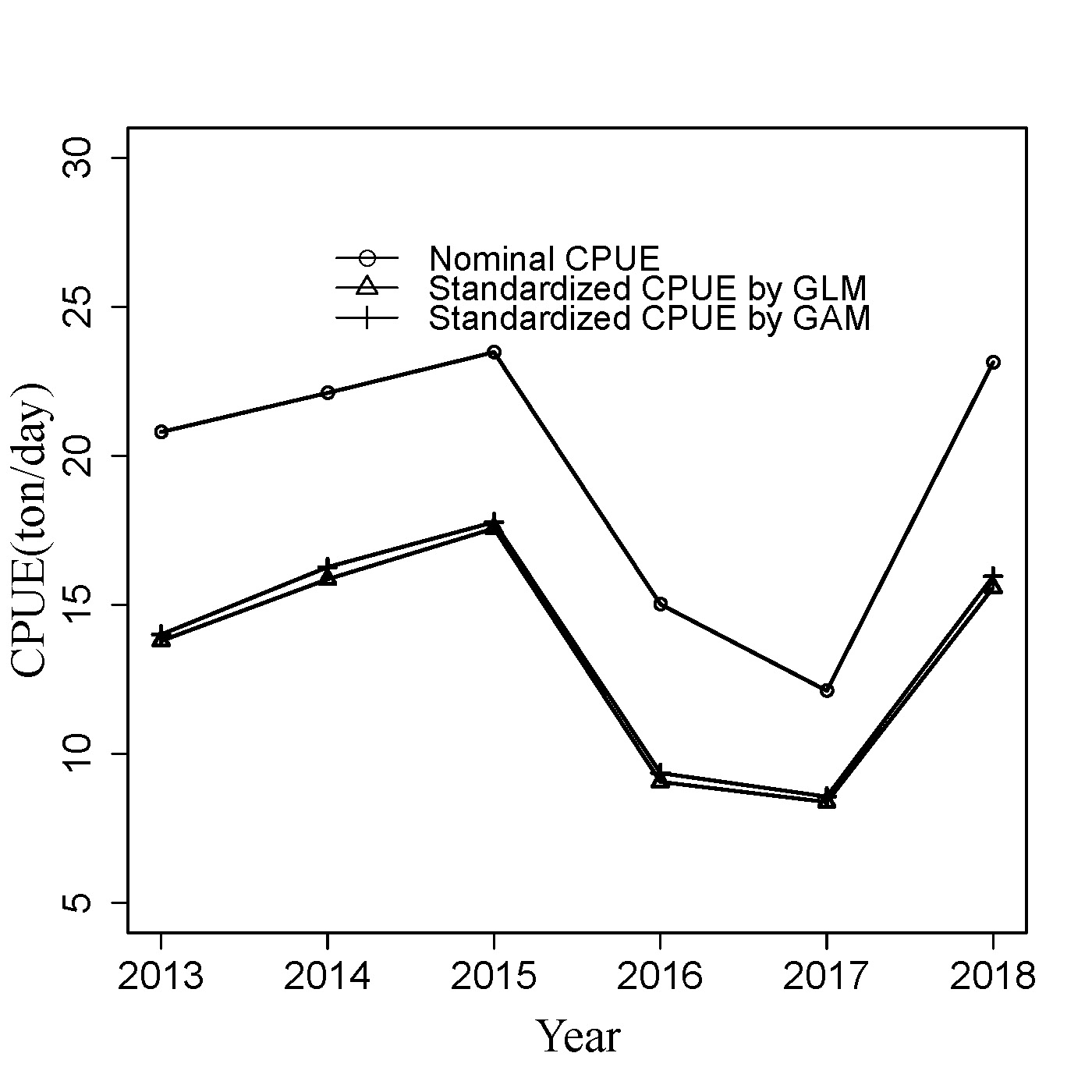


Fig.8 Annual changes in nominal, GAM and GLM estimated standardized CPUEs

Appendix1. Checklist for the CPUE standardization protocol

|  |  |  |
| --- | --- | --- |
| (1) | Conduct a thorough literature review to identify key factors (i.e., spatial, temporal, environmental, and fisheries variables) that may influence CPUE values; | Yes (*see* Factors that may affect the CPUE of PS fisheries) |
| (2) | Determine temporal and spatial scales for data grouping for CPUE standardization; | Yes (*see* table 2) |
| (3) | Plot spatio-temporal distributions of fishing efforts and catch to evaluate spatio-temporal patterns of fishing effort and catch; | Yes (*see* Fig.2 and Fig.7) |
| (4) | Calculate correlation matrix to evaluate correlations between each pair of those variables; | Yes (*see* table 1) |
| (5) | Identify potential explanatory variables based on (1)-(4) to develop full model for the CPUE standardization; | Yes |
| (6) | Make statistical assumptions on the full models and fit the data to the assumed statistical models (i.e., GLM, GAM, Delta-lognormal GLM, Neural Networks, Regression Trees, Habitat based models, and Statistical habitat based models); | Yes (GLM and GAM) |
| (7) | Select and evaluate the models using methods such as likelihood ratio, AIC, BIC or cross validation; | Yes (*see* Table3 and Table6) |
| (8) | Evaluate if distributional assumptions are satisfied and if there is a consistent spatial/temporal distribution of residuals in CPUE standardization modeling; | Yes (*see* Figs3, 4, 5) |
| (9) | Determine the optimal model to estimate yearly standardized CPUE and their associated uncertainty. | Yes (*see* Table5 and Table8) |
| (10) | Plot nominal and standardized CPUEs over time.  Overall remarks Recommendations | Yes (*see* Fig. 8) |