SEAK Pink Salmon 2023 Forecast Process

Sara Miller

October 11, 2022

Objective

To forecast the Southeast Alaska (SEAK) pink salmon commercial harvest in 2023.

Executive Summary

Forecasts were developed using an approach originally described in Wertheimer et al. (2006), and modified in Orsi et al. (2016) and Murphy et al. (2019). We used a similar approach to Murphy et al. (2019), but assumed a log-normal error (Miller et al. 2022). This approach is based on a multiple regression model with juvenile pink salmon catch-per-unit-effort (CPUE; a proxy for abundance) and temperature data from the Southeast Alaska Coastal Monitoring Survey (SECM; Piston et al. 2021) or from satellite sea surface temperature (SST) data (Huang et al. 2017). See the document satellite_SST_process–September_2022 for details about the temperature variables. Based on prior discussions, the index of juvenile abundance (i.e., CPUE) was based on the pooled-species vessel calibration coefficient.

The model performance metric one-step ahead mean absolute percent error (MAPE) for the last five years (years 2018 through 2022) and for the last ten years (years 2013 through 2022) was used to evaluate the forecast accuracy of the 18 individual models. Based upon this performance metrics, model m11 (a model that included CPUE and a May temperature index based on northern Southeast Alaska satellite SST data; Table 1 and Table 2) was the best performing model and a forecast using this model would be in the weak range with a point estimate of 18.8 million fish (80% prediction interval: 12.3 to 28.9 million fish).

Analysis

Individual, multiple linear regression models

Biophysical variables based on data from Southeast Alaska were used to forecast the harvest of adult pink salmon in Southeast Alaska, one year in advance, using individual, multiple linear regression models (models m1–m18). The simplest regression model (model m1) consisted of only the predictor variable juvenile pink salmon CPUE (X_1), while the other 17 regression models consisted of the predictor variable juvenile pink salmon CPUE and a temperature index (X_2),

$$E(Y_i) = \hat{\alpha}_i + \hat{\beta}_{1_i} X_1 + \hat{\beta}_{2_i} X_2. \quad (1)$$

The temperature index was either the SECM survey Icy Strait temperature Index (ISTI; Murphy et al. 2019) or one of the 16 satellite-derived SST data (Huang et al. 2017). Although the simplest model only contained CPUE, including temperature data with CPUE is likely a more accurate measure of juvenile abundance if temperature affects the proportion of juveniles that migrate through Icy Strait in a given year (Murphy et

al. 2019). The response variable (Y; Southeast Alaska adult pink salmon harvest in millions) and CPUE data were natural log transformed in the model, but temperature data were not. The forecast ($\hat{Y}_i$), and 80% prediction intervals (based on output from program R; R Core Team 2021) from the 18 regression models were exponentiated and bias-corrected (Miller 1984),

$$\hat{F}_i = \exp(\hat{Y}_i + \frac{{\sigma_i}^2}{2}), \tag{2}$$

where $\hat{F}_i$ is the preseason forecast (for each model i) in millions of fish, and σ_i is the variance (for each model i).

Table 1: Annual adult pink salmon harvest data from Southeast Alaska (millions of fish), 1998-2022.

Year	Harvest
1998	42.45
1999	77.82
2000	20.25
2001	67.02
2002	45.32
2003	52.47
2004	45.31
2005	59.12
2006	11.61
2007	44.80
2008	15.91
2009	38.02
2010	24.14
2011	58.88
2012	21.28
2013	94.72
2014	37.17
2015	35.09
2016	18.37
2017	34.73
2018	8.07
2019	21.14
2020	8.06
2021	48.50
2022	18.04

Table 2: Juvenile pink salmon CPUE data collected from the SECM project and May satellite sea surface temperature data (°C) from the northern Southeast Alaska region in the juvenile years 1997–2022.

Juvenile year	CPUE	Temperature
1997	2.48	7.35
1998	5.62	7.65
1999	1.60	6.70
2000	3.73	7.23
2001	2.87	6.66
2002	2.78	6.39
2003	3.08	7.57
2004	3.90	7.89
2005	2.04	8.42
2006	2.58	6.98
2007	1.17	6.90
2008	2.32	6.64
2009	2.33	7.32
2010	4.11	7.76
2011	1.51	7.25
2012	3.52	6.95
2013	2.14	6.59
2014	3.80	8.15
2015	2.45	8.92
2016	4.35	8.92
2017	0.35	7.75
2018	1.17	7.53
2019	1.14	8.42
2020	2.15	8.26
2021	0.88	7.29
2022	1.45	7.62

Performance metric: One-step ahead MAPE

The model summary results using the performance metric one-step ahead MAPE are shown in Table 3; the smallest value is the preferred model. The performance metric one-step ahead MAPE was calculated as follows.

1. Estimate the regression parameters at time t-1 from data up to time t-1.
2. Make a prediction of $\hat{Y}_t$ at time t based on the predictor variables at time t and the estimate of the regression parameters at time t-1 (i.e., the fitted regression equation).
3. Calculate the MAPE based on the prediction of $\hat{Y}_t$ at time t and the observed value of Y_t at time t,

$$\text{MAPE} = \left|\frac{\exp(Y_t) - \exp(\hat{Y}_t + \frac{{\sigma_t}^2}{2})}{\exp(Y_t)}\right|. \tag{3}$$

4. For each individual model, average the MAPEs calculated from the forecasts,

$$\frac{1}{n}\sum_{t=1}^{n}\left|\frac{\exp(Y_t) - \exp(\hat{Y}_t + \frac{{\sigma_t}^2}{2})}{\exp(Y_t)}\right|, \tag{4}$$

where n is the number of forecasts in the average (5 forecasts for the 5-year MAPE and 10 forecasts for the 10-year MAPE). For example, to calculate the five year one-step-ahead MAPE for model m1 for the 2022 forecast, use data up through year 2016 (e.g., data up through year 2016 is t -1 and the forecast is for t, or year 2017). Then, calculate a MAPE based on the 2017 forecast and the observed pink salmon harvest in 2017 using equation 3. Next, use data up through year 2017 (e.g., data up through year 2017 is t -1 and the forecast is for year 2018; t) and calculate a MAPE based on the 2018 forecast and the observed pink salmon harvest in 2018 using equation 3. Repeat this process for each subsequent year through year 2020 to forecast 2021. Finally, average the five MAPEs to calculate a five year one-step-ahead MAPE for model m1. For the 10 year one-step-ahead MAPE for model m1, the process would be repeated, but the first forecast year would be 2012.

Table 3: Summary of the adjusted R squared value and the 5-year and 10-year one-step ahead MAPEs for the 18 regression models.

Model	5-year MAPE	10-year MAPE	AdjR2
m1	0.58	0.63	0.60
m2	0.39	0.36	0.81
m3	0.31	0.25	0.79
m4	0.44	0.39	0.74
m5	0.33	0.29	0.79
m6	0.37	0.33	0.77
m7	0.33	0.25	0.78
m8	0.45	0.39	0.73
m9	0.35	0.29	0.76
m10	0.39	0.34	0.75
m11	0.30	0.25	0.78
m12	0.40	0.34	0.74
m13	0.31	0.28	0.78
m14	0.34	0.29	0.76
m15	0.34	0.29	0.76
m16	0.42	0.37	0.73
m17	0.34	0.31	0.77
m18	0.37	0.32	0.75

Results

Based upon the 5-year and 10-year one-step ahead MAPE, the best performing model was model m11.

Model Diagnostics

Model diagnostics for model m11 included residual plots, the curvature test, and influential observation diagnostics using Cook's distance (Cook 1977), the Bonferroni outlier test, and leverage plots. Model diagnostics were used to identify observations that were potential outliers, had high leverage, or were influential (Zhang 2016).

Table 4: Detailed output for model m11. Juvenile years 1998, 1999, 2005, 2012, and 2019, and 2020 (years 1999, 2000, 2006, 2013, 2020, and 2021) show the largest standardized residual (Std. residuals). Fitted values are bias-corrected.

Year	Harvest	Residuals	Hat values	Cooks distance	Std. residuals	Fitted values
1998	42.45	0.28	0.04	0.01	0.91	33.71
1999	77.82	-0.43	0.29	0.37	-1.65	126.17
2000	20.25	-0.32	0.11	0.05	-1.07	29.17
2001	67.02	0.12	0.09	0.00	0.39	62.62
2002	45.32	-0.11	0.11	0.01	-0.37	53.11
2003	52.47	-0.03	0.15	0.00	-0.10	56.70
2004	45.31	0.16	0.05	0.00	0.52	40.59
2005	59.12	0.18	0.09	0.01	0.59	52.02
2006	11.61	-0.39	0.12	0.08	-1.32	17.94
2007	44.80	0.14	0.06	0.00	0.46	40.89
2008	15.91	-0.28	0.11	0.04	-0.95	22.10
2009	38.02	-0.04	0.10	0.00	-0.14	41.62
2010	24.14	-0.23	0.04	0.01	-0.74	31.80
2011	58.88	0.02	0.11	0.00	0.07	60.52
2012	21.28	0.00	0.07	0.00	-0.01	22.43
2013	94.72	0.45	0.10	0.08	1.50	63.67
2014	37.17	0.00	0.11	0.00	0.00	39.07
2015	35.09	-0.19	0.11	0.02	-0.66	44.75
2016	18.37	0.08	0.21	0.01	0.31	17.72
2017	34.73	-0.15	0.26	0.04	-0.56	42.44
2018	8.07	-0.24	0.19	0.06	-0.86	10.79
2019	21.14	0.26	0.09	0.03	0.87	17.14
2020	8.06	-0.34	0.19	0.11	-1.19	11.84
2021	48.50	0.93	0.10	0.36	3.14	20.04
2022	18.04	0.14	0.12	0.01	0.47	16.48

Cook's distance

Cook's distance is a measure of influence, or the product of both leverage and outlier. Cook's distance,

$$D_i = \frac{e^2_{PSi}}{k+1} * \frac{h_i}{1-h_i}, \tag{5}$$

where e^2_{PSi} is the standardized Pearson residuals, h_i are the hat values (measure of leverage), and k is the number of predictor variables in the model, is a measure of overall influence of the i_{th} data point on all n fitted values (Fox and Weisburg 2019). A large value of Cook's distance indicates that the data point is an influential observation. Cook's distance values greater than $4/(n-k-1)$, where n is the number of observations (i.e., 25), was used as a benchmark for identifying the subset of influential observations (Ren et al. 2016). Therefore, a Cook's distance cut-off of 0.18 was used; observations with a Cook's distance greater than 0.18 may be influential observations (Figure 1a).

Leverage

An observation that is distant from the average covariate pattern is considered to have high leverage or hat-value. If an individual observation has a leverage value h_i greater than 2 or 3 times p/n (Ren et al. 2016), it may be a concern (where p is the number of parameters in the model including the intercept (i.e., 3), and

n is the number of observations in the model (i.e., 25); $p/n = 3/25 = 0.12$ for this study). Therefore, a leverage cut-off of 0.24 was used; observations with a leverage value greater than 0.24 may affect the model properties (e.g., summary statistics, standard errors, predicted values) (Figure 1b).

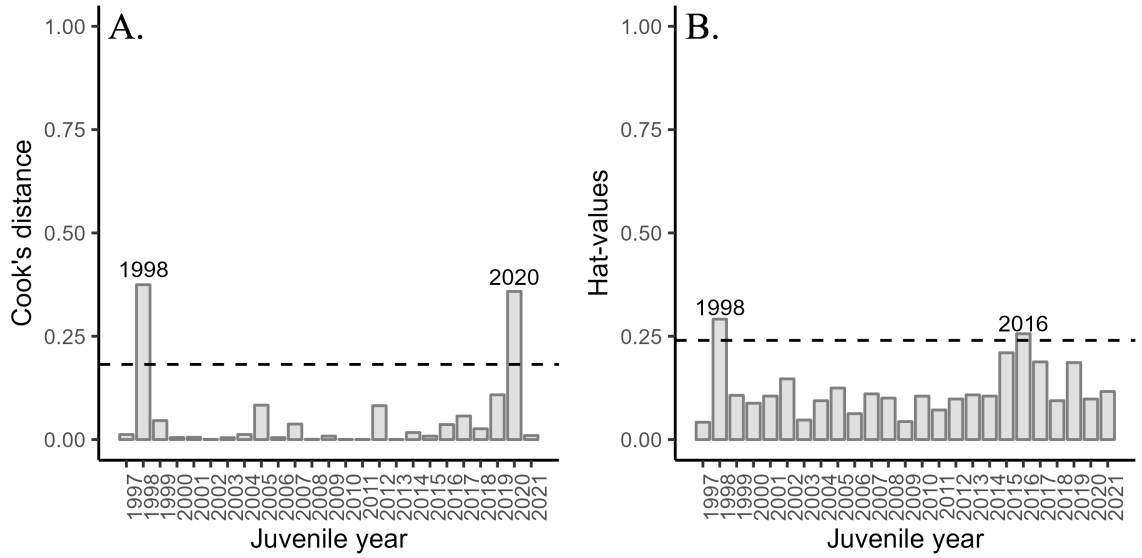

Figure 1: Diagnostics plots of influential observations including A. Cook's distance (with a cut-off value of 0.18), and B. leverage values (with a cut-off value of 0.24) from model m11.

Influential datapoints

To determine if a variable has a relationship with residuals, a lack-of fit curvature test was performed. In this test, terms that are non-significant suggest a properly specified model. No terms were significant in the lack-of-fit curvature test ($P<0.05$) (Figure 2a; Figure 2b). Diagnostics indicated that two of the data points were above the cut-off value for the Cook's distance (Figure 1a; 1998 and 2020). Two observations had high leverage values (Figure 1b; 1998 and 2016). Based on the Bonferroni outlier test, one of the data points had a studentized residual with a significant Bonferroni P-value suggesting that one of the data points impacted the model fitting (observation 24; juvenile year 2020); although observations 2, 3, 9, 16, 23, and 24 and were the most extreme (juvenile years 1998, 1999, 2005, 2012, 2019, and 2020) based on standardized residuals (Figure 3a; Table 4). Based on the lightly curved fitted lines in the residual versus fitted plot (Figure 3b), the fitted plot shows some lack of fit of the model.

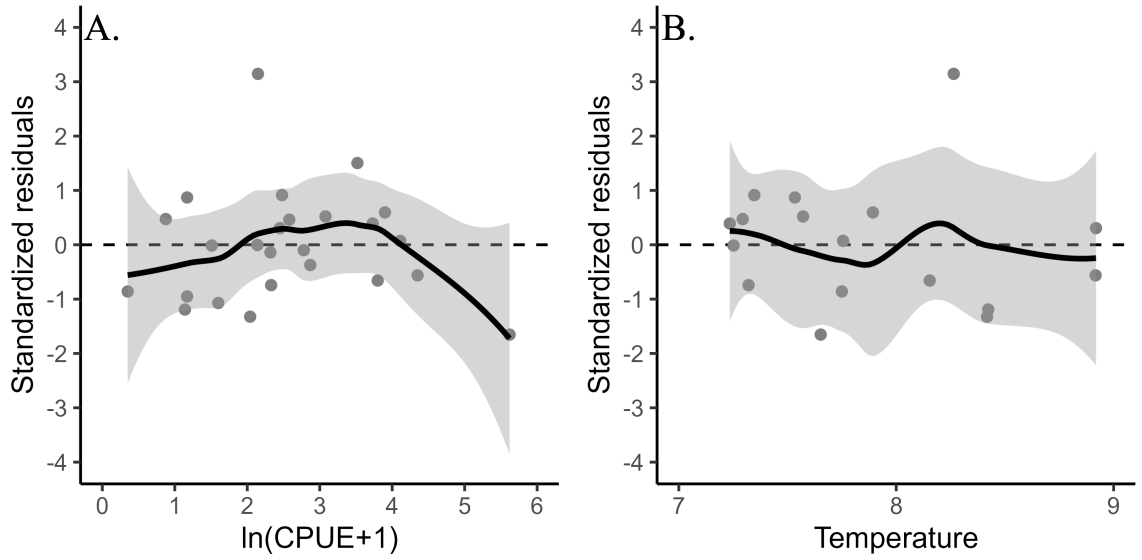

Figure 2: Standardized residuals versus predicted plots for A. CPUE and B. temperature.

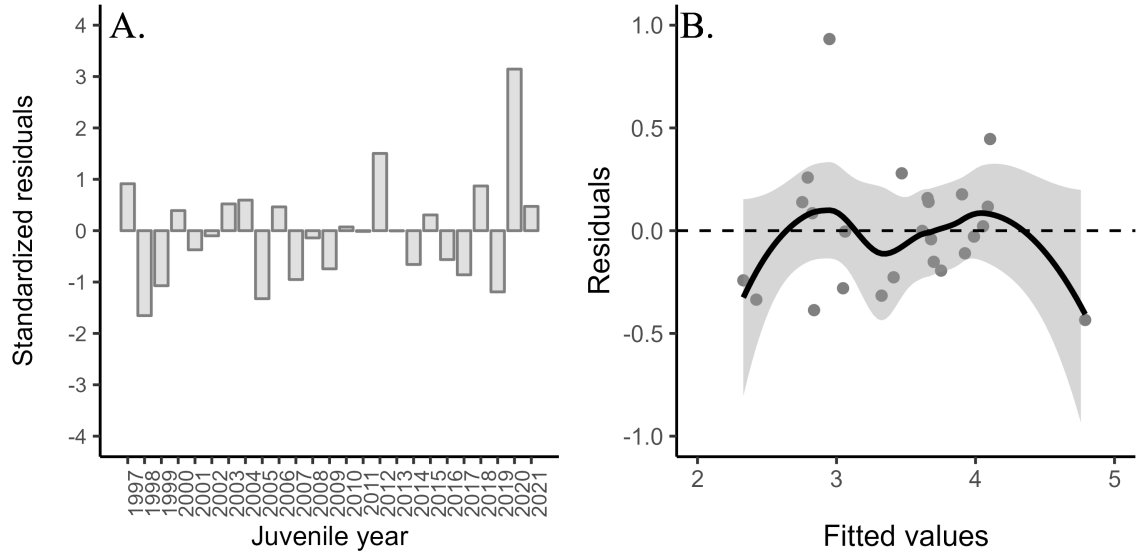

Figure 3: A. Standardized residuals versus juvenile year and B. residuals versus fitted values for model m11. Positive residuals indicate that the observed harvest was larger than predicted by the model.

The best performing model, based on the performance metric one-step ahead MAPE, was model m11 (i.e., the model containing CPUE and May NSEAK SST). The adjusted R^2 value for model m11 was 0.78 (Table 3) indicating overall a good model fit. Based upon a model that includes juvenile pink salmon CPUE and May NSEAK SST (model m11), the 2023 SEAK pink salmon harvest would be in the weak range with a point estimate of 18.8 million fish (80% prediction interval: 12.3 to 28.9 million fish).

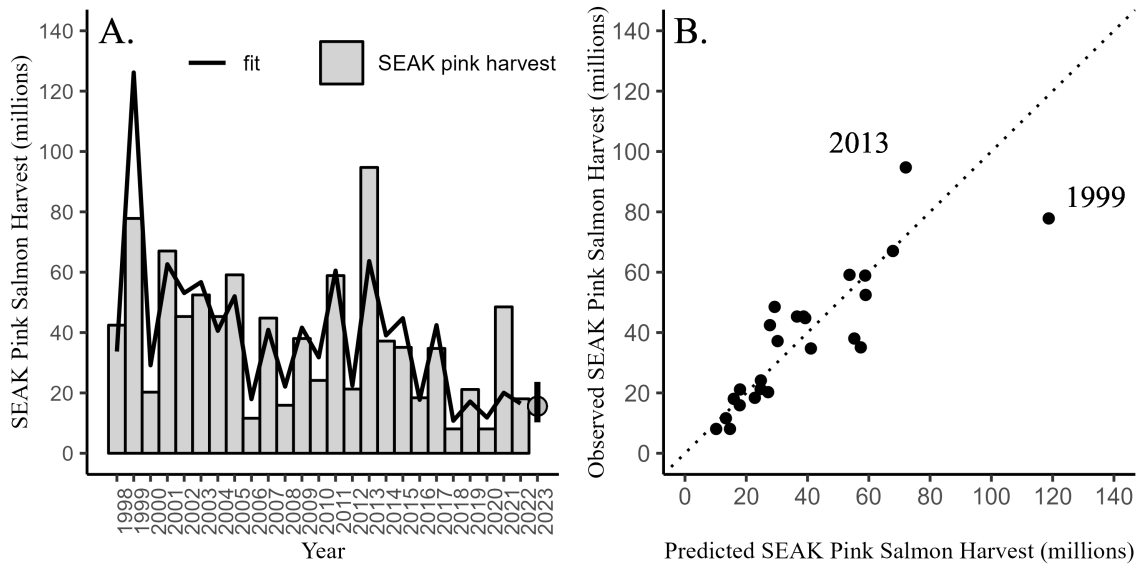

Figure 4: A. SEAK pink salmon harvest (millions) by year with the model fit (line) based upon the best performing model (model m11). The predicted 2023 forecast is symbolized as a grey circle with an 80% prediction interval (12.3 to 28.9 million fish). B. SEAK pink salmon harvest (millions) against the fitted values from model m11 by year. The dotted line is a one to one reference line.

References

Cook, R. D. 1977. Detection of influential observations in linear regression. Technometrics 19: 15-18.

Fox, J. and S. Weisburg. 2019. An R Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage Publications, Inc.

Huang, B., P. W. Thorne, V. F. Banzon,, T. Boyer, G. Chepurin, J. H. Lawrimore, M. J. Menne, T. M. Smith, R. S. Vose, and H. M. Zhang. 2017. Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. Journal of Climate 30:8179–8205.

Miller, D. M. 1984. Reducing transformation bias in curve fitting. The American Statistician 38: 124-126.

Miller, S. E., J. M. Murphy, S. C. Heinl, A. W. Piston, E. A. Fergusson, R. E. Brenner, W. W. Strasburger, and J. H.Moss. 2022. Southeast Alaska pink salmon forecasting models. Alaska Department of Fish and Game, Fishery Manuscript No. 22-03, Anchorage.

Murphy, J. M., E. A. Fergusson, A. Piston, A. Gray, and E. Farley. 2019. Southeast Alaska pink salmon growth and harvest forecast models. North Pacific Anadromous Fish Commission Technical Report No. 15: 75-81.

NOAA Coral Reef Watch (NOAA_DHW_monthly dataset). 2022, updated daily. NOAA Coral Reef Watch Version 3.1 Monthly 5km SST and SST Anomaly, NOAA Global Coral Bleaching Monitoring Time Series Data, May 1997-June 2021. College Park, Maryland, USA: NOAA/NESDIS/STAR Coral Reef Watch program. Data set accessed 2022-09-12 at https://coastwatch.pfeg.noaa.gov/erddap/griddap/NOAA_DHW_monthly.html.

NOAA Coral Reef Watch (NOAA_DHW dataset). 2022, updated daily. NOAA Coral Reef Watch Daily Near-real-Time Global 5km SST and SST Anomaly, NOAA Global Coral Bleaching Monitoring Time Series Data, July 2021 to July 2022. College Park, Maryland, USA: NOAA/NESDIS/STAR Coral Reef Watch program. Data set accessed 2022-09-12 at https://coastwatch.pfeg.noaa.gov/erddap/griddap/NOAA_DHW.html.

Orsi, J. A., E. A. Fergusson, A. C. Wertheimer, E. V. Farley, and P. R. Mundy. 2016. Forecasting pink salmon production in Southeast Alaska using ecosystem indicators in times of climate change. N. Pac. Anadr. Fish Comm. Bull. 6: 483–499. (Available at https://npafc.org)

Piston, A. W., J. Murphy, J. Moss, W. Strasburger, S. C. Heinl, E. Fergusson, S. Miller, A. Gray, and C. Waters. 2021. Operational Plan: Southeast coastal monitoring, 2021. ADF&G, Regional Operational Plan No. ROP.CF.1J.2021.02, Douglas.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.r-project.org/index.html

Ren, Y. Y., L. C. Zhou, L. Yang, P. Y. Liu, B. W. Zhao and H. X. Liu. 2016. Predicting the aquatic toxicity mode of action using logistic regression and linear discriminant analysis, SAR and QSAR in Environmental Research, DOI: 10.1080/1062936X.2016.1229691

Wertheimer A. C., J. A. Orsi, M. V. Sturdevant, and E. A. Fergusson. 2006. Forecasting pink salmon harvest in Southeast Alaska from juvenile salmon abundance and associated environmental parameters. In Proceedings of the 22nd Northeast Pacific Pink and Chum Workshop. Edited by H. Geiger (Rapporteur). Pac. Salmon Comm. Vancouver, British Columbia. pp. 65–72.

Zhang, Z. 2016. Residuals and regression diagnostics: focusing on logistic regression. Annals of Translational Medicine 4: 195.

Appendix

Table 5: Parameter estimates for the 18 individual models.

model	term	estimate	std.error	statistic	p.value
m1	(Intercept)	2.3626779	0.195	12.088	0.000
m1	CPUE	0.4226366	0.069	6.139	0.000
m2	(Intercept)	7.2871053	0.970	7.509	0.000
m2	CPUE	0.4848713	0.049	9.884	0.000
m2	ISTI20_MJJ	-0.5612768	0.110	-5.124	0.000
m3	(Intercept)	5.5996662	0.700	8.002	0.000
m3	CPUE	0.4822600	0.051	9.422	0.000
m3	Chatham_SST_May	-0.4489866	0.095	-4.722	0.000
m4	(Intercept)	6.5965044	1.197	5.510	0.000
m4	CPUE	0.4479234	0.056	7.933	0.000
m4	Chatham_SST_MJJ	-0.4384368	0.123	-3.568	0.002
m5	(Intercept)	6.3205791	0.853	7.412	0.000
m5	CPUE	0.4685021	0.051	9.252	0.000
m5	Chatham_SST_AMJ	-0.5310697	0.113	-4.706	0.000
m6	(Intercept)	6.7144293	1.058	6.348	0.000
m6	CPUE	0.4570951	0.053	8.571	0.000
m6	Chatham_SST_AMJJ	-0.5091081	0.123	-4.156	0.000
m7	(Intercept)	5.2035018	0.672	7.741	0.000
m7	CPUE	0.4944671	0.054	9.104	0.000
m7	Icy_Strait_SST_May	-0.4200420	0.097	-4.331	0.000
m8	(Intercept)	6.1471085	1.131	5.433	0.000
m8	CPUE	0.4529550	0.058	7.836	0.000
m8	Icy_Strait_SST_MJJ	-0.3843450	0.114	-3.380	0.003
m9	(Intercept)	5.8708819	0.878	6.689	0.000
m9	CPUE	0.4768410	0.055	8.688	0.000
m9	Icy_Strait_SST_AMJ	-0.4879745	0.120	-4.058	0.001
m10	(Intercept)	6.1928902	1.037	5.972	0.000
m10	CPUE	0.4630551	0.056	8.253	0.000
m10	Icy_Strait_SST_AMJJ	-0.4491176	0.120	-3.736	0.001
m11	(Intercept)	5.2720785	0.670	7.871	0.000
m11	CPUE	0.4592580	0.052	8.879	0.000
m11	NSEAK_SST_May	-0.4004154	0.090	-4.449	0.000
m12	(Intercept)	6.3874737	1.110	5.757	0.000
m12	CPUE	0.4319995	0.056	7.780	0.000
m12	NSEAK_SST_MJJ	-0.4103958	0.112	-3.664	0.001
m13	(Intercept)	6.0123915	0.838	7.179	0.000
m13	CPUE	0.4492940	0.052	8.715	0.000
m13	NSEAK_SST_AMJ	-0.4915044	0.111	-4.425	0.000
m14	(Intercept)	6.4166400	1.021	6.285	0.000
m14	CPUE	0.4412112	0.054	8.221	0.000
m14	NSEAK_SST_AMJJ	-0.4722658	0.118	-4.016	0.001
m15	(Intercept)	5.2684073	0.730	7.219	0.000
m15	CPUE	0.4567703	0.054	8.486	0.000
m15	SEAK_SST_May	-0.3696992	0.091	-4.069	0.001
m16	(Intercept)	6.1473033	1.122	5.480	0.000
m16	CPUE	0.4245370	0.057	7.456	0.000
m16	SEAK_SST_MJJ	-0.3657944	0.107	-3.410	0.003
m17	(Intercept)	5.9715141	0.883	6.762	0.000

model	term	estimate	std.error	statistic	p.value
m17	CPUE	0.4468392	0.053	8.422	0.000
m17	SEAK_SST_AMJ	-0.4509711	0.109	-4.146	0.000
m18	(Intercept)	6.2139777	1.050	5.916	0.000
m18	CPUE	0.4341526	0.055	7.850	0.000
m18	SEAK_SST_AMJJ	-0.4210499	0.114	-3.708	0.001

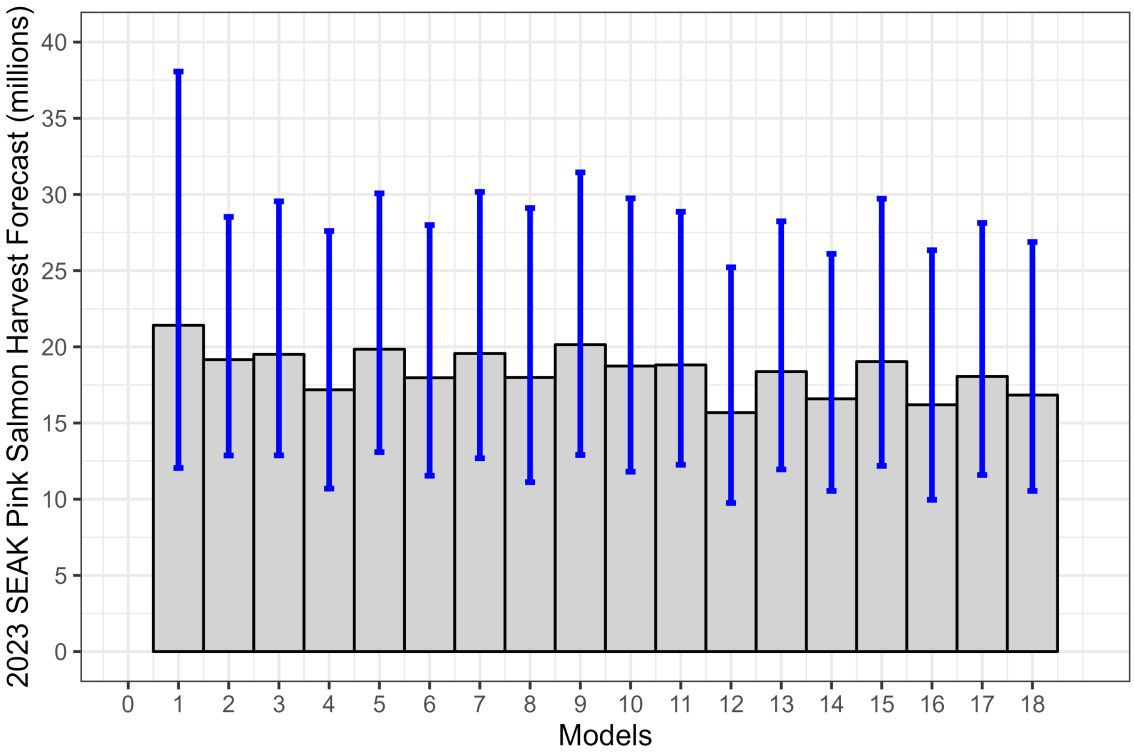

Figure 5: Bias-corrected forecasts (grey bars) for the eighteen regression models with 80% prediction intervals (blue lines). Based upon the performance metrics, the best performing model was model m11; a model that included CPUE and a May temperature index based on northern Southeast Alaska satellite SST data.