2021 SEAK Pink Salmon Harvest Forecast

Sara Miller, Rich Brenner, Jim Murphy

November 12, 2020

Contents

1 Objective

To forecast the Southeast Alaska (SEAK) pink salmon commercial harvest in 2021.

2 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. This approach is based on a multiple regression model with juvenile pink salmon catch-per-unit-effort (CPUE) and temperature data from the Southeast Alaska Coastal Monitoring Survey (SECM; Murphy et al. 2020). The final model used for the forecast was:

$$E(y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where y is log-transformed pink salmon harvest in SEAK, β_1 is the coefficient for CPUE using the pooled-species vessel calibration coefficient, β_2 is the coefficient for the environmental covariate water temperature, and ϵ represents the normally distributed error term. The CPUE data are the average log-transformed catches standardized to an effort of a 20 minute trawl set and calibrated to the fishing power of the NOAA

Ship *John N. Cobb.* For each year, the standardized catch is either taken from June or July, whichever month had the highest average catches. Water temperature data is the average (May through July) temperature in the upper 20 m at the eight SECM stations in Icy Strait.

Leave-one-out cross validation (hindcast) and model performance metrics were used to evaluate the forecast accuracy of models. Based on the Akaike Information Criterion corrected for small sample sizes (AICc values; Burnham and Anderson 2004), the mean absolute percentage error (MASE metric), and significant coefficients in the model, the preferred model (i.e., the additive model with CPUE and temperature; model m2) predicted that the SEAK pink salmon harvest in 2021 will be in the average range with a point estimate of 28.5 million fish (80% prediction interval: 19.4 to 41.7 million fish).

3 Forecast Models

3.1 Vessel calibration coefficient

For the 2021 SEAK pink salmon forecast, there was a discussion as to which vessel calibration coefficient to use going forward (see *calibration_coefficient_discussion* document). Using the four potential vessel calibration coefficients (mixspecies, mixpooled, species-specific, pooled), the corresponding index of juvenile abundance was slightly different (i.e., CPUE; standardized pink salmon catch based on a 20 minute trawl set by year). Performance metrics such as AICc, MASE, median absolute percentage error (MEAPE), and mean absolute scaled error (MASE; Hyndman and Kohler 2006) were used to evaluate forecast accuracy of the four alternative vessel calibration coefficients using the same model (the additive model with CPUE and temperature; model m2). Based on performance metrics, the pooled-species vessel calibration coefficient was preferred. Therefore, the index of juvenile abundance was based on the pooled-species vessel calibration coefficient in the following analysis.

3.2 Analysis

3.2.1 Hierarchical models

Three hierarchical models were investigated. The full model was model m3:

$$E(y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where X_1 is the average CPUE for catches in either the June or July survey, whichever month had the highest average catches in a given year, and was based on the pooled-species vessel calibration coefficient, and X_2 is the average temperature in Icy Strait in May, June, and July at the eight SECM stations sampled in Icy Strait (Icy Strait and Upper Chatham transects; 'ISTI'), and β_3 is the interaction term between CPUE and the temperature index. The CPUE data are log-transformed in the model. If temperature is altering how CPUE is related to abundance it makes sense to restrict the temperature data to the CPUE months in the forecast model (June and July). The month of May is included as there are important migratory dynamics prior to the time juveniles are sampled in Icy Strait. In the past, the 'ISTI' variable was the average temperature in the upper 20 m from May to August at the eight SECM stations in Icy Strait. For simplicity, although the definition of the variable has changed, the variable is still called 'ISTI.'

Model m1 only contained the CPUE variable, model m2 contained the CPUE variable and a temperature variable ('ISTI'), and model m3 contained the CPUE variable, a temperature variable ('ISIT'), and an interaction term. The regression coefficients CPUE and temperature ('ISTI') were significant in the first two models (m1, m2). The interaction term was not significant in the full model (model m3; Table 1). Therefore, only the first two models will be considered further.

Table 1: Parameter estimates for the three potential models.

model	term	estimate	std.error	statistic	p.value
m1	(Intercept)	2.2891818	0.208	11.019	0.000
m1	CPUE	0.4379654	0.071	6.157	0.000
m2	(Intercept)	7.0773197	0.945	7.486	0.000
m2	CPUE	0.5065420	0.050	10.176	0.000
m2	ISTI	-0.5463568	0.107	-5.121	0.000
m3	(Intercept)	3.8317812	2.321	1.651	0.115
m3	CPUE	1.7886278	0.844	2.119	0.047
m3	ISTI	-0.1925848	0.254	-0.757	0.458
m3	CPUE:ISTI	-0.1387265	0.091	-1.522	0.145

3.2.2 Preferred model (model m2)

The model summary results using the performance metrics AICc, MAPE, MEAPE, and MASE (Hyndman and Kohler 2006) are shown in Table 2. For all of these metrics, the smallest value is the preferred model. The difference (Δ_i) between a given model and the model with the lowest AICc value, and the metric MASE were the primary statistics for choosing appropriate models in this analysis. Models with AICc$\Delta_i \leq 2$ have substantial support, those in which $4 \leq$ AICc$\Delta_i \leq 7$ have considerably less support, and models with AICcΔ_i > 10 have essentially no support (Burnham and Anderson 2004). The AICc in Table 2 is the AICc value and not the AICcΔ_i. The two performance metrics (i.e., AICc, MASE) suggest that model m2 is the preferred model;the differece (AICcΔ_i) in the best model (model m2; lowest AICc value; AICc$\Delta_i = 0$) and the other model (m1) is 16 (i.e., AICc$\Delta_i = 16$). Therefore, model m2 (based on CPUE and average temperature in May through July ('ISIT')) was used to forecast the 2021 pink salmon harvest (Figure 1).

Table 2: Summary of model outputs and forecast error measures.The AICc is the AICc value. Therefore, the differece in the best model (model m2) and the other model (m1) is 16.

model	terms	AdjR2	AICc	MAPE	MEAPE	MASE
m1	CPUE	0.627	30	0.116	0.105	0.399
m2	CPUE+ISTI	0.830	14	0.074	0.060	0.249

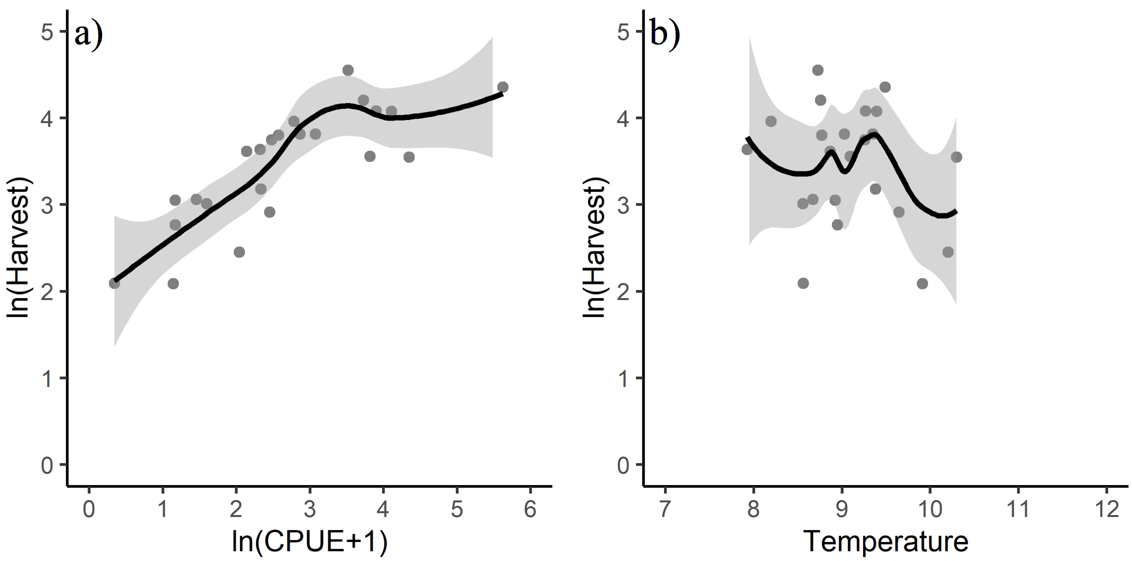

Figure 1: Relationship between a) CPUE and harvest and b) average temperature in May through July ('ISTI') and harvest. The line is a smoothing function applied to the relationship with a 95% confidence interval.

3.3 Model Diagnostics

Model diagnostics for model m2 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). These observations may have significant impact on model fitting and may need to be excluded.

3.3.1 Cook's distance

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

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

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., 23), was used as a benchmark for identifying the subset of influential observations (Ren et al. 2016). Therefore, a Cook's distance cut-off of 0.20 was used; observations with a Cook's distance greater than 0.20 were investigated further (Figure 2a).

3.3.2 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., 23); $p/n = 3/23 = 0.13$ for this study). Therefore, a leverage cut-off of 0.26 was used; observations with a leverage value greater than 0.26 were investigated further (Figure 2b).

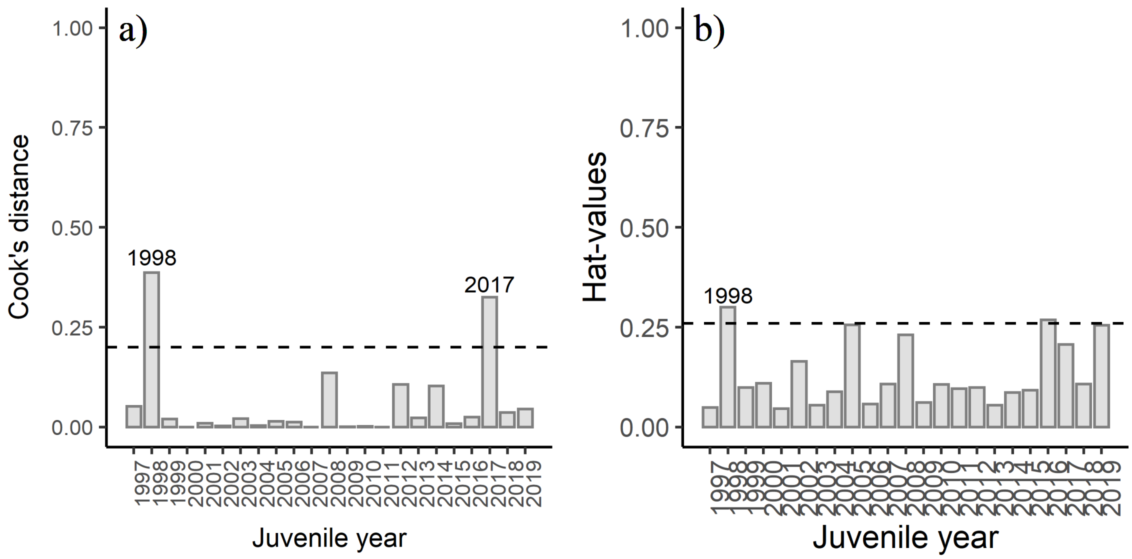

Figure 2: Diagnostics plots of influential observations including a) Cook's Distance (with a cut-off value of 0.20), and b) leverage values (with a cut-off value of 0.26) from model m2.

3.3.3 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. The CPUE term was significant in the lack-of-fit curvature test (P<0.05), suggesting some lack of fit for this term (Figure 3a). Diagnostics indicated that two of the data points were above the cut-off value for the Cook's distance (Figure 2a). Two observations had a high leverage value (Figure 2b), but none of the observations affected model fitting. Based on the Bonferroni outlier test, none of the data points had a studentized residual with a significant Bonferroni P-value suggesting that none of the data points impacted the model fitting; although observations 1, 16, 18 and 21 and were the most extreme (juvenile years 1997, 2012, 2014, and 2017 corresponding to years 1998, 2013, 2015, and 2018) based on standardized residuals (Figure 4a; Table 3). Based on the lightly curved fitted lines in the residual versus fitted plot (Figure 4b), the fitted plot shows some lack of fit of the model.

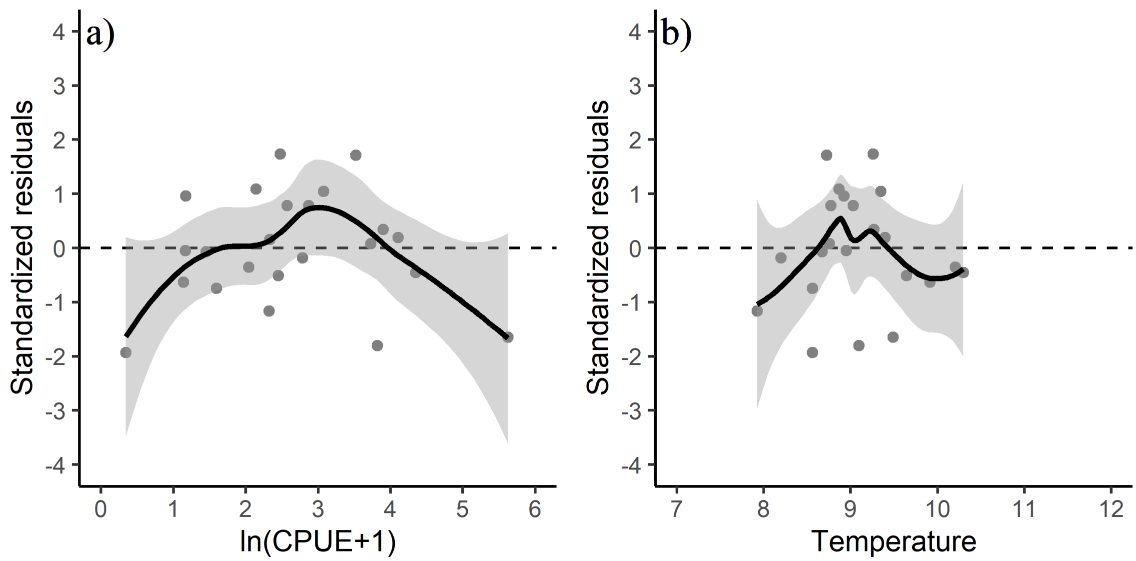

Figure 3: Standardized residuals versus predicted plots for a) CPUE and b) temperature.

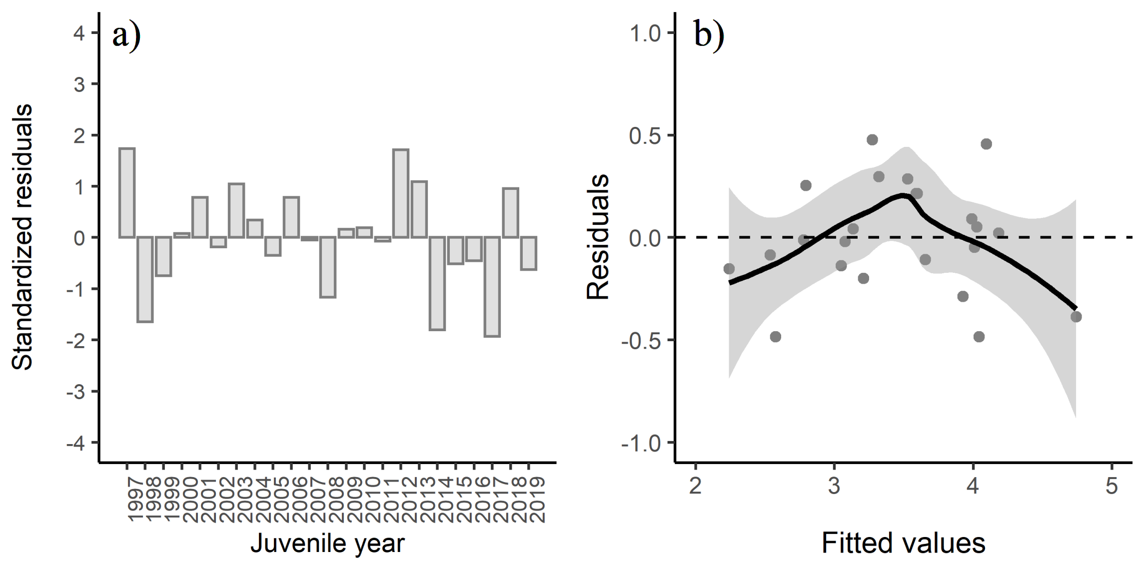

Figure 4: a) Standardized residuals versus juvenile year and b) residuals versus fitted values for model m2. Positive residuals indicate that the observed harvest was larger than predicted by the model.

Table 3: Detailed output for model m2. Juvenile years 1997, 2012, 2014, and 2017 (years 1998, 2013, 2015, and 2018) show the largest standardized residual. The variable SEAKCatch is commercial harvest of adult fish in millions, the variable CPUE is the average log-transformed catches standardized to an effort of a 20 minute trawl set, and the variable ISTI is the average (May through July) temperature in the upper 20 m at the eight SECM stations in Icy Strait. Year refers to the forecast year. Fitted values are log-transformed.

year	juvenile_year	SEAKCatch	CPUE	ISTI	resid	hat_values	Cooks_distance	std_resid	fitted
1998	1997	42.5	2.478	9.259	0.476	0.049	0.052	1.737	3.273
1999	1998	77.8	5.622	9.489	-0.387	0.300	0.387	-1.645	4.741
2000	1999	20.3	1.598	8.560	-0.199	0.099	0.021	-0.747	3.210
2001	2000	67.0	3.730	8.756	0.022	0.110	0.000	0.082	4.183
2002	2001	45.3	2.869	9.028	0.215	0.046	0.010	0.784	3.598
2003	2002	52.5	2.785	8.198	-0.048	0.165	0.002	-0.188	4.009
2004	2003	45.3	3.078	9.349	0.285	0.054	0.021	1.043	3.528
2005	2004	59.1	3.899	9.269	0.091	0.089	0.004	0.338	3.988
2006	2005	11.6	2.040	10.203	-0.085	0.256	0.014	-0.352	2.536
2007	2006	44.8	2.573	8.771	0.214	0.058	0.013	0.785	3.588
2008	2007	15.9	1.168	8.951	-0.012	0.108	0.000	-0.046	2.778
2009	2008	38.0	2.323	7.925	-0.287	0.231	0.136	-1.164	3.925
2010	2009	24.0	2.333	9.378	0.043	0.061	0.001	0.158	3.135
2011	2010	58.9	4.108	9.395	0.051	0.107	0.001	0.191	4.025
2012	2011	21.3	1.455	8.669	-0.019	0.096	0.000	-0.071	3.078
2013	2012	94.7	3.522	8.725	0.457	0.099	0.107	1.712	4.094
2014	2013	37.2	2.143	8.865	0.297	0.055	0.023	1.088	3.319
2015	2014	35.1	3.817	9.093	-0.485	0.086	0.102	-1.804	4.043
2016	2015	18.4	2.453	9.645	-0.138	0.092	0.009	-0.514	3.050
2017	2016	34.7	4.351	10.297	-0.109	0.268	0.025	-0.452	3.655
2018	2017	8.1	0.346	8.560	-0.484	0.207	0.325	-1.933	2.576
2019	2018	21.1	1.172	8.925	0.255	0.107	0.037	0.959	2.795
2020	2019	8.1	1.142	9.911	-0.153	0.255	0.045	-0.630	2.241

3.4 Results

The best regression model based on the AICc value, the MASE metric, and significant coefficients in the model was model m2 (i.e., the model containing CPUE, and a May through July temperature variable). The adjusted R^2 value for model m2 was 0.83 (Table 2) indicating overall a good model fit. Based upon a model that includes juvenile pink salmon CPUE and May through July temperature (model m2), the 2021 SEAK pink salmon harvest is predicted to be in the average range with a point estimate (model mean) of 28.5 million fish (80% prediction interval: 19.4 to 41.7 million fish). The 80% prediction interval around the forecast were calculated using the car package (Fox and Weisberg 2019) in program R (R Core Team).

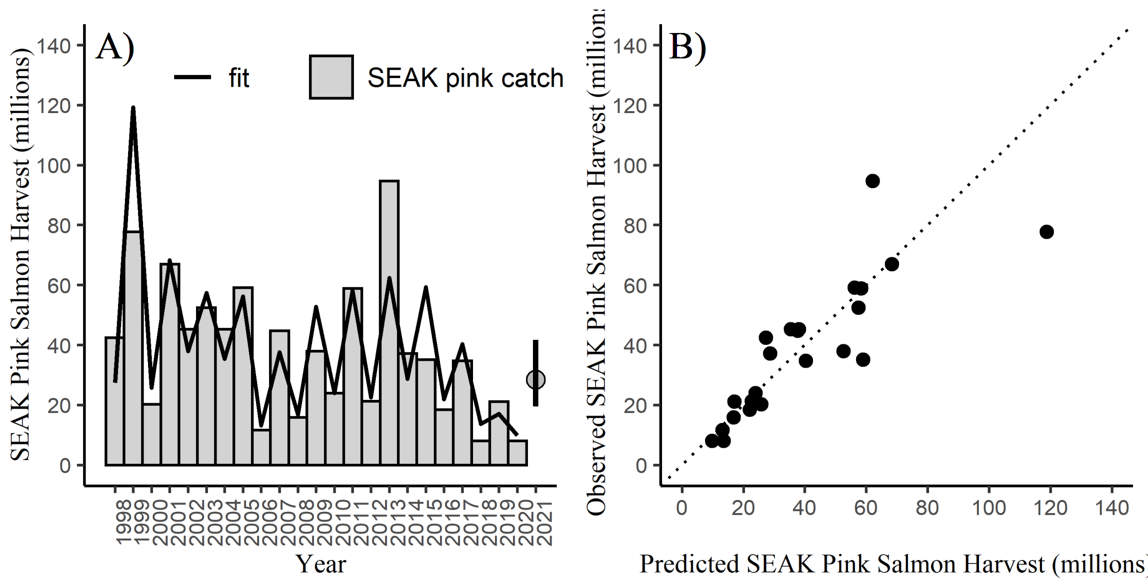

Figure 5: A) SEAK pink salmon harvest (millions) by year with the model fit (line). The predicted 2021 forecast is symbolized as a grey circle with an 80% prediction interval (19.4 to 41.7 million fish). B) SEAK pink salmon harvest (millions) against the fitted values from model m2 by year. The dotted line is a one to one line.

4 Potential variables and performance metrics

1) May 3-m water temperature (°C) adjusted to a standard date of May 23 (May 3-m Water Temperature); (potential variable in the 2013 forecast; Wertheimer et al., 2013)

2) May upper 20-m integrated average water temperature (°C) adjusted to a standard date of May 23 (May 20-m Integrated Water Temperature);

3) Sea surface temperature (SST) from satellite data (May temperature near the SECM stations);

4) Sea surface temperature (SST) from satellite data (May-July temperature near the SECM stations).

In addition to these environmental variables, it would be beneficial to add additional performance metrics to rank the models (e.g., number of years the model performed the best; number of leverage point years (run the the two competing models without the leverage years and review the performance metrics)). Also, a retrospective analysis.

5 References

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods and Research 33: 261-304.

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.

Hyndman, R. J. and A. B. Koehler. 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22: 679-688.

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.

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)

R Core Team. 2020. 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 27(9). 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.