Modeling of Chinook Salmon Runs on the North Umpqua River

Scott Jordan

PI: Thomas Dietterich
Collaborators: Rebecca Flitcroft, Gordon Grant

Abstract

To model the behavior of the spring run Chinook salmon, we expanded the current dataset of salmon counts at the Winchester Dam, Winchester, Oregon, to include historical data from 1991 through 1997. These counts were used to identify Chinook salmon behavior that resulted in defining preferable water conditions during their upstream migration in the Umpqua River. Also, we predicted the timing of Chinook salmon with a 95% confidence interval of 16 days. Additionally, we explored some of the challenges in predicting the presence or absence of Chinook salmon on a given day.

Introduction

Salmon play a vital role for the state of Oregon as they are important for economical, and ecological aspects. For example, salmon are a valuable resource and contributor to a healthy and balanced food web. Humans and other apex predators depend on salmon for food, as well as many other species that eat their eggs. One of their most important roles is bringing nutrients from the ocean back up the river to fertilize the river ecosystem (Cederholm, et al. 1999). In addition Salmon play an essential role in the Oregon economy. In 2008, 631,000 people went fishing in Oregon and spent $264.6 million on fishing trips (Dean Runyan Associates 2009). In 2013 4.66 million pounds of Salmon (Chinook and Coho) were caught commercially (The Research Group, LLC 2014)..

Unfortunately, the numbers of wild salmon returning to spawn every year are only about 6%-7% of what they used to be (Gresh, Lichatowich, and Schoonmaker 2000). Thus it is of interest to understand the factors that contribute to lower spawning to target these issue for restoration. One important factor is to understand salmon migration, so in our study we focus on Salmon migration on the North Umpqua River. The North Umpqua River has runs of Chinook, Coho, Steelhead, and other fish species, but we focused on primarily Spring Run Chinook Salmon. The Spring Run on the North Umpqua River is from March 1st through July 31. From 1992-2014 between 1,761 and 5,754 wild adult Chinook salmon have past Winchester Dam each year for the spring run.

Expanding the Dataset

Winchester Dam has a long historic dataset of fish counts going back to 1945. However, only since November 1998 has this data been stored electronically and accessible to researchers. As of October 24th 1991 a 24 hour camera has been used to monitor the fish ladder to count the number of fish. Prior to this counts were estimated by randomly sampling an 8 hour periods at least 5 days a week.

Archived Fish Counts

The first task was to expand the current count dataset to include data, taken from a camera, from the start of Oct 24, 1991. The data was archived in undocumented text files with no description. To identify what the data represented, we used a copy of the program used to hand enter the data. The program was created in 1980s and only ran on Microsoft DOS. We used a version that was last update in 1992.

We successfully used this program to identify salmon counts in the data files. These data were stored in files that represented 8-hour periods during the day. However some of the dates were absent, and some of the recordings had errors. Errors included having negative counts or excessively large total fish counts. To correct these errors we examined previous days and the following days to identify the actual number of fish for that period. All changes were logged and coded including change type, for auditing purposes. The results were compared to the official numbers posted on Oregon Departments of Fish and Wildlife's (ODFW) website. The count difference were off by only 1 fish each year except for one year where our total was short 46 Chinook salmon.

Other Data Sources

- Daily mean discharge - USGS sites at Winchester Dam and Elkton, OR
- Daily max water temperature - ODFW taken from the Winchester Dam
- Daily Precipitation - NOAA certified data set for Winchester, OR
- SWE (Snow Water Equivalent) - NWCC, SNOTEL Dataset at Diamond Lake, OR

Migration Patterns

To find trends for Spring Chinook we looked at the conditions the fish experience during their time in the river. Specifically we look at the daily max water temperature and daily mean discharge at Winchester Dam to see the impact the different conditions have on Spring Chinook.

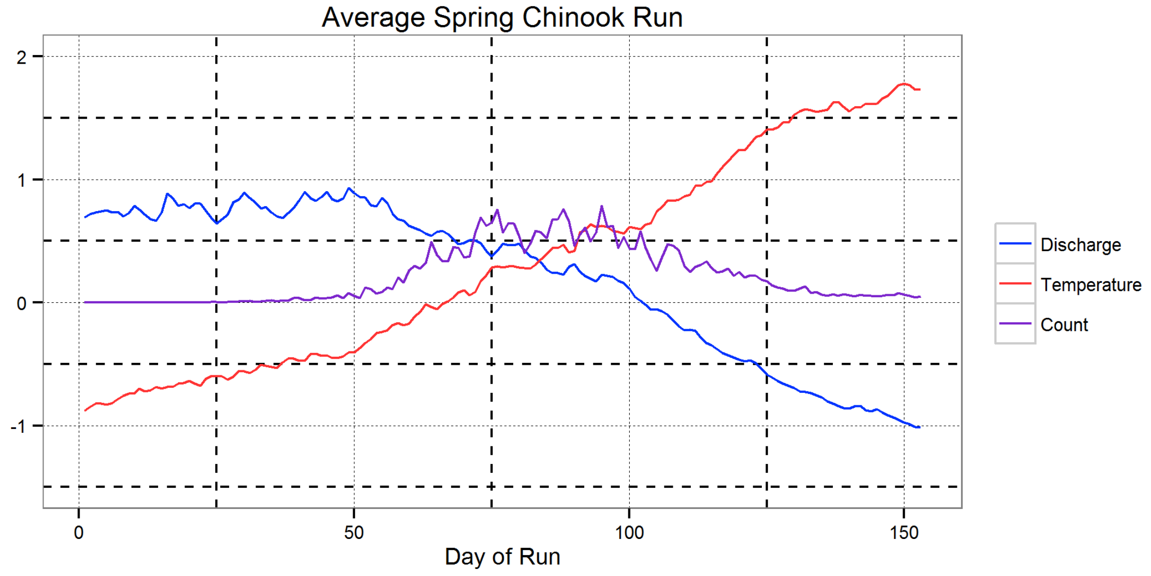


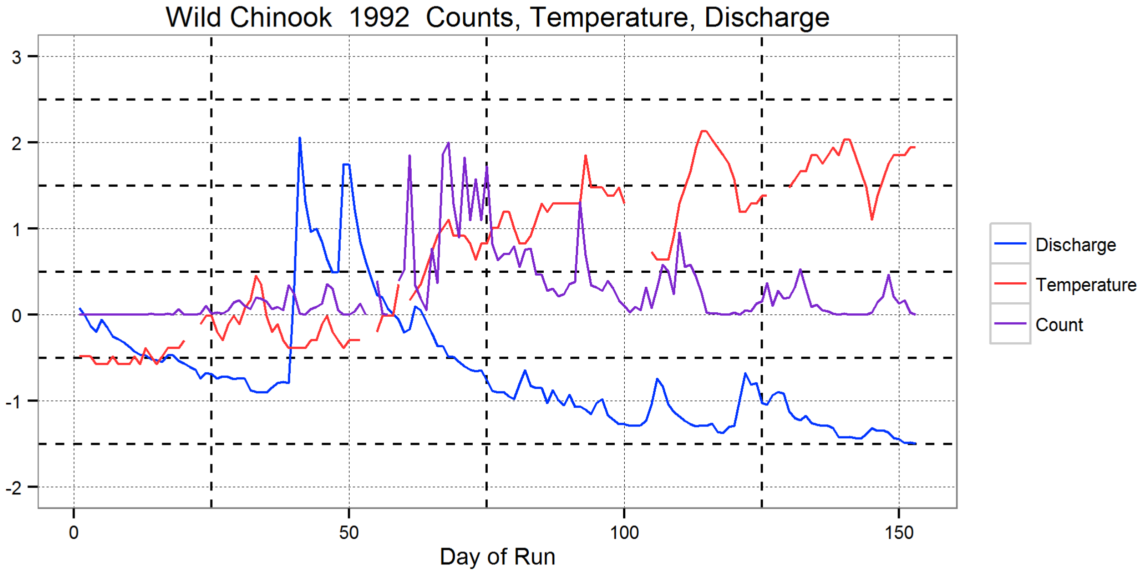
Figure 1: The average values of temperature, discharge, and number of Wild Chinook Salmon. The x-axis represents the days of the run from March 1 through July 31. Temperature and discharge were fit to a Gaussian Normal distribution. Counts are scale to fit onto the graph.

Results show that the average values for the Chinook upstream run migration period are highest when temperature rises. It can also be seen that once discharge drops the number of fish migrating per day greatly increases. Results for individual years show similar trends. Such as in 1992 the discharge level is very low for that time of the year and temperature is very warm, but the salmon start coming up based on discharge and temperature rather than time of year. Interestingly, in 1992, when discharge rapidly increased the fish halt their migration and wait for the discharge to decrease. Once discharge decreases again they resume running and come in large numbers.

Likewise in 2011 the discharge stays high for a long time keeping the water temperature cool. Once the high level of discharge dissipated and temperature raised the salmon began to migrate. Then after another storm the salmon continue with the run. Although despite what looks like favorable run conditions on days 99 and 105 we see a large drop in the number of Chinook salmon passing the dam that day. This may be because Chinook salmon prefers to travel in groups as a defense strategy from predators.

We have come up with three general theories of Spring Chinook Salmon behavior:

1. They don't like to travel at high levels of discharge
2. They wait for temperature to rise and discharge to drop before starting the main portion of the run
3. They like to travel in groups



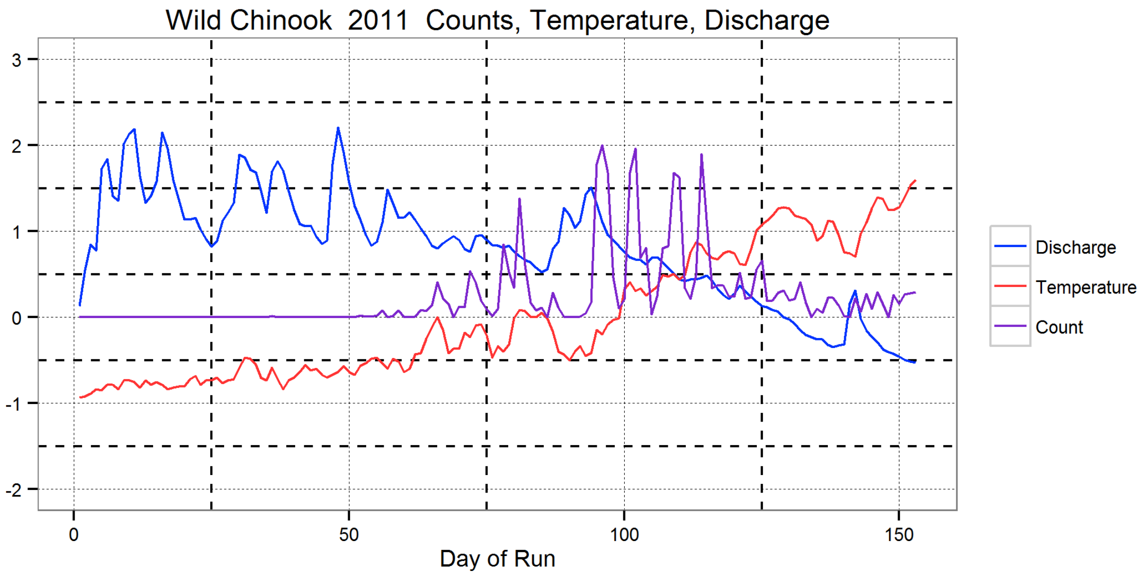


Figure 2: Above are graphs representing the Spring Run of Wild Chinook for 1992 (Top) and 2011 (Bottom).

Optimal Run Conditions

Since our results indicate that fish are responding to temperature and discharge rates our next step is to define the optimal conditions that trigger salmon on the Umpqua to migrate. Looking at the running and non-running days plotted (Figure 3) in temperature and discharge there is a triangular region that encompasses the main portion of the run. The region that represents the running period should be one that maximizes the correct classification of days and incorporates as many of the running days as possible.

To find the region we used two methods, the first being a brute force approach and the second being linear programming. The brute force method involved checking millions of triangles at different locations for each vertex. This method is exponential in computation time and memory as the number of vertices increase. The benefit to this method is that it can be done in parallel on a GPU. We were able to compute roughly 70 million vertex combinations in around one second. Using Linear Programming we wrote the optimization problem to be minimizing the number of misclassifications. Unfortunately this turned out to be slow because in order to classify each point as in the triangle or not, every point had to be checked to each side of the triangle. Then when we tried the algorithm on datasets of one year, the region that was found did not well represent the running conditions for that year. So we ended up using the brute force method to find the triangles.

The triangle found (Figure 3) to fit all the data does a good job of representing the running conditions. It classifies running days and non-running days with 84.09% accuracy. It includes 90.1% of the running days in the triangle, but also includes 27.9% of non-running days. While the triangle does a good job of capturing the running days it does miss most of the days at the beginning of the run. Also it includes almost all non-running days after the run has started.

A lot of the missed running days come from the wide variation in run conditions from year to year.

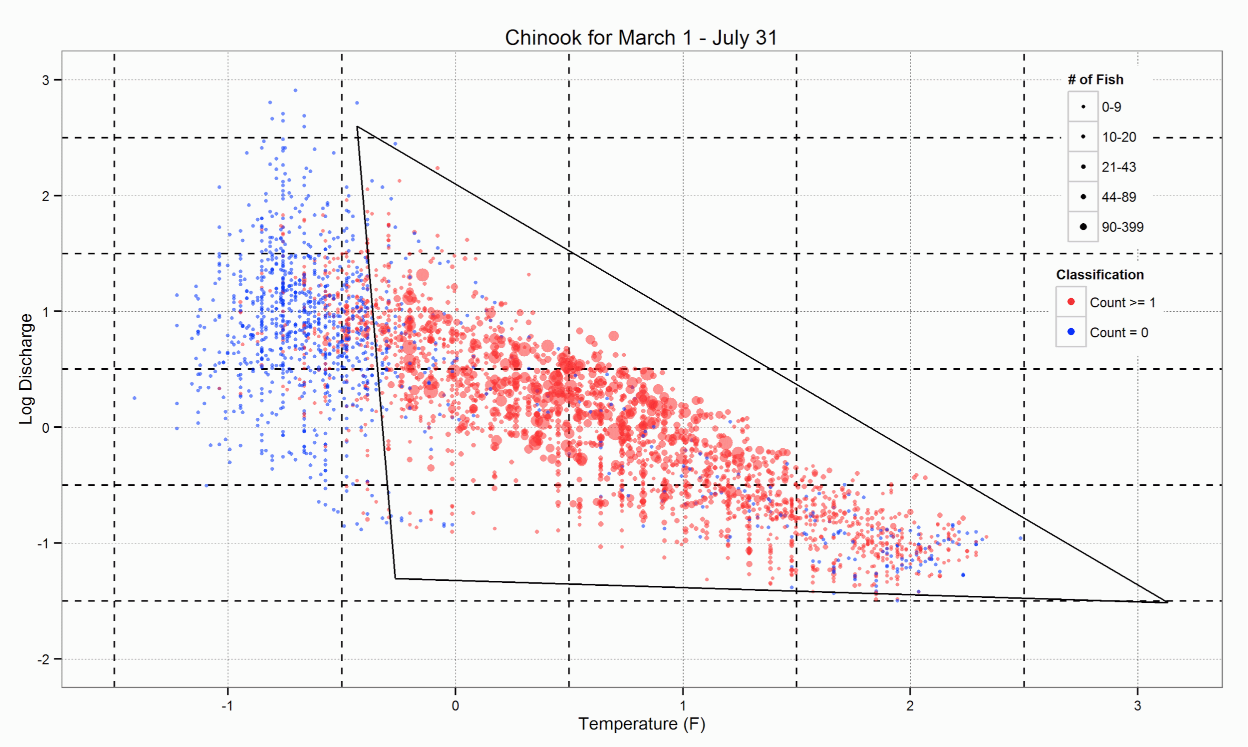


Figure 3: This is a plot of temperature and discharge space for Spring Chinook. Running days are marks in red and not running days are in blue. The more fish that passed the dam that day the larger the circle. Time is flowing in general from top left to bottom right as temperature increases and discharge lessens. The black triangle represents the region found to represent the running conditions.

To improve on the ability to capture the run conditions we look at fitting triangles to individual years and then combine them. The region found for 1992 does a good job of separating the run days from non-run days in the beginning and end of the run. Still the region misses the non-running days occurring during the middle of the run. In 2011 the region found, again does well of separating the running and non-running days. When all the regions are combined (figure 5), there is a nice overlapped region where most intersect. When plotting the data points on top it can be seen that most of the running days are contained within the overlapped region. So we can say that we have found a region that captures the main portion of the run. That region spans from about 55.1˚F to 73.2˚F and roughly from 1,850 cu ft/sec to 7,855 cu ft/sec. The variation in run conditions from year can also be seen and the result of the variation is “cloudy” areas near the edge of the region that have many running and non-running days.

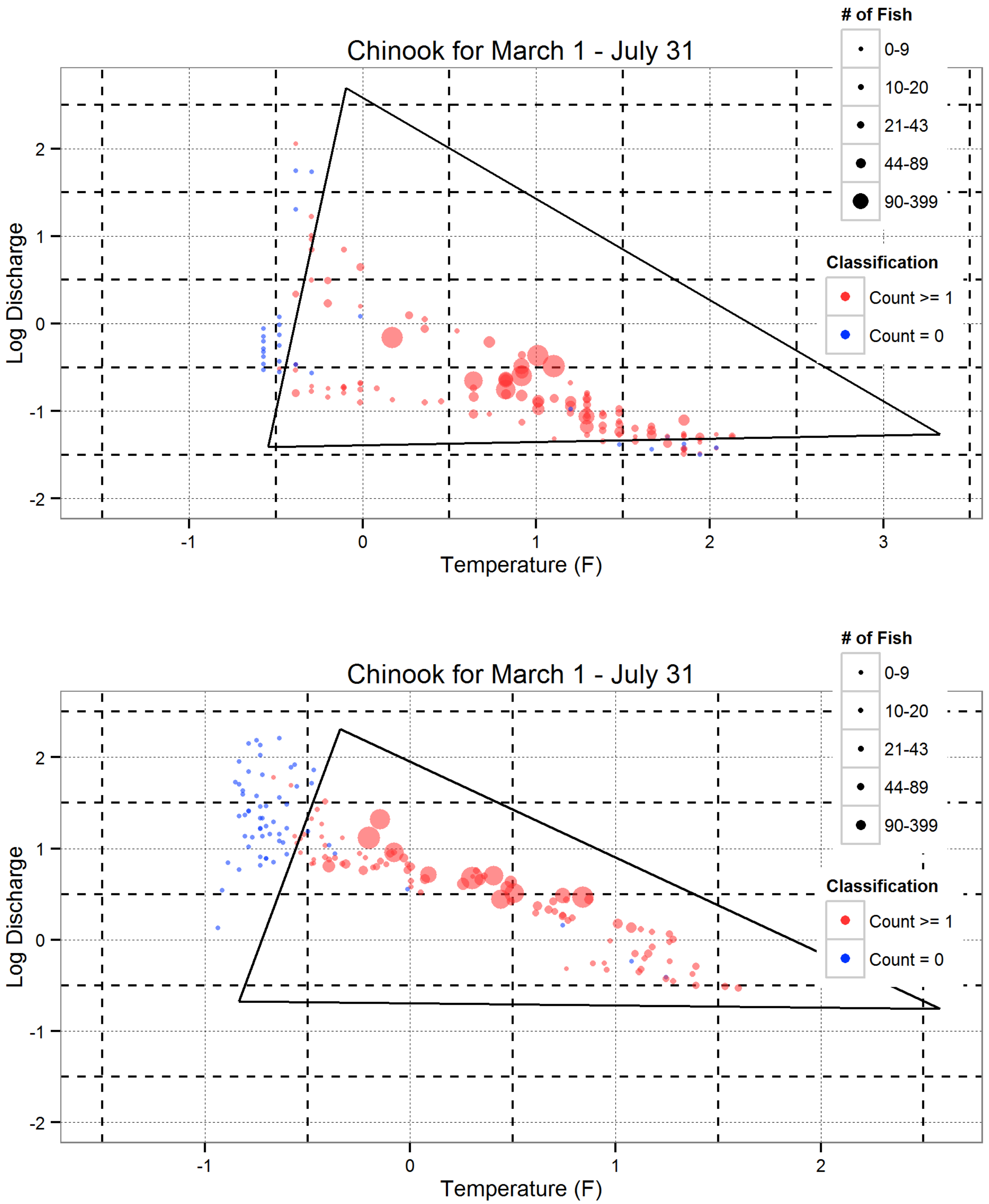


Figure 4: Triangles found to represent the running conditions for 1992 (Top) and 2011 (Bottom).

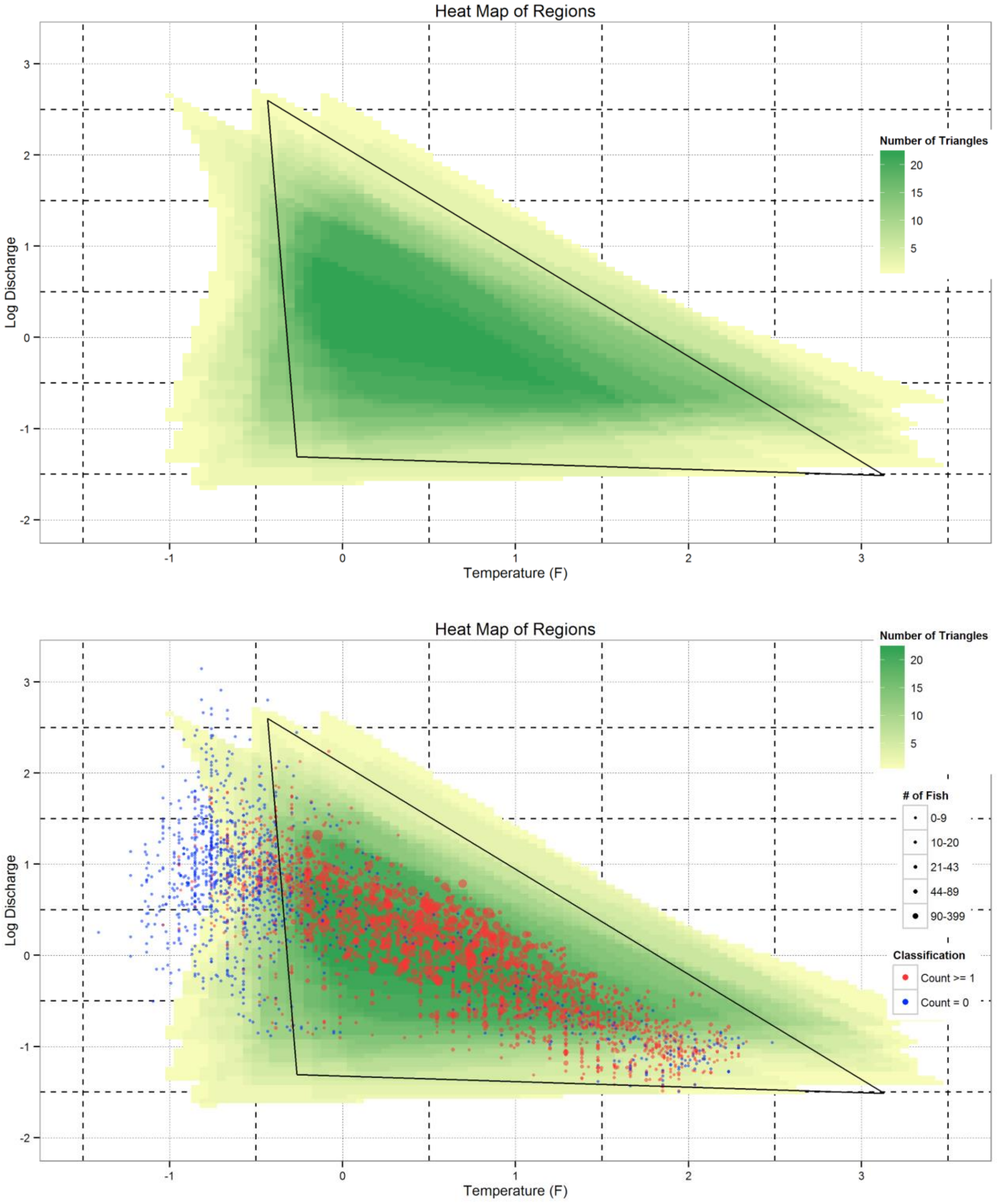


Figure 5: Heat Map of regions found for run conditions of individual years. The greener the area the more regions that overlap. The black triangle is the region found for all the years.

Predicting the Median Day of the Run

Now that we have identified that the salmon are responding to specific temperature and discharge, we want to try and predict when the run will happen. There has been some work already done on predicting the timing of spring run Chinook on the Columbia River (Keefer et al. 2008). Their work on predicting the median day of the run is done by using monthly averages of air temperature, discharge, Pacific Decadal Oscillation (PDO), and North Pacific Index (NPI), then fitting a linear model to the median day. The best model they achieved with an r^2 correlation of .49 was:

Median Day = April discharge + January PDO + January NPI

Looking to fit our data with a similar model we investigated the spread of our median days. The earliest median day occurred on May 15, 1992, latest on June 12, 2011, and the average falls around May 28/29. There is about a month of spread in when the median day occurs, so to answer the question of why is one year earlier or later than another, we looked at the earliest and latest runs. In 1992 we see the discharge is very low and temperature is warm really early in the run. In 2011 discharge remains high and temperature doesn't warm up until later in the year. So one reason as to why the discharge would remain main high and temperature low, would be from snow melting down into the river. So if there is more snow in the mountain we should see higher levels of discharge later in the year.

Next we need to examine the correlations between monthly averages and median day of the run. We looked to use water temperature, discharge, snow water equivalent (SWE), and precipitation as predictors of median day of the run. For temperature and discharge May contains the best correlation, but since most of the median days occur in May we can't use it as a predictor, so April is the next best choice. We can see that higher the temperature the earlier the median of the run will occur. Inversely the higher the discharge the later the run with happen. Fitting our theory the more snow the later the run will happen. Checking the precipitation, only March has a significant correlation to when the median day happens.

To create a model to predict the median day we use linear regression. The best single variable predictors are SWE and precipitation achieving r^2 correlations of 0.5778 and 0.4918 respectively. These by themselves are as good as or better correlated than the predictor for the Spring Salmon Run on the Columbia found by Keefer et al. However it can be improved even more by combining the variables (table 1). Our best model with a correlation of 0.7134 is:

Median Day = April SWE + April Temperature + March Precipitation

Adding discharge in to the model does slightly increase performance, but by adding an extra degree of freedom the performance gain is negligible. The model performs well when looking at the prediction and has a 95% confidence interval of ± 8.02 days.

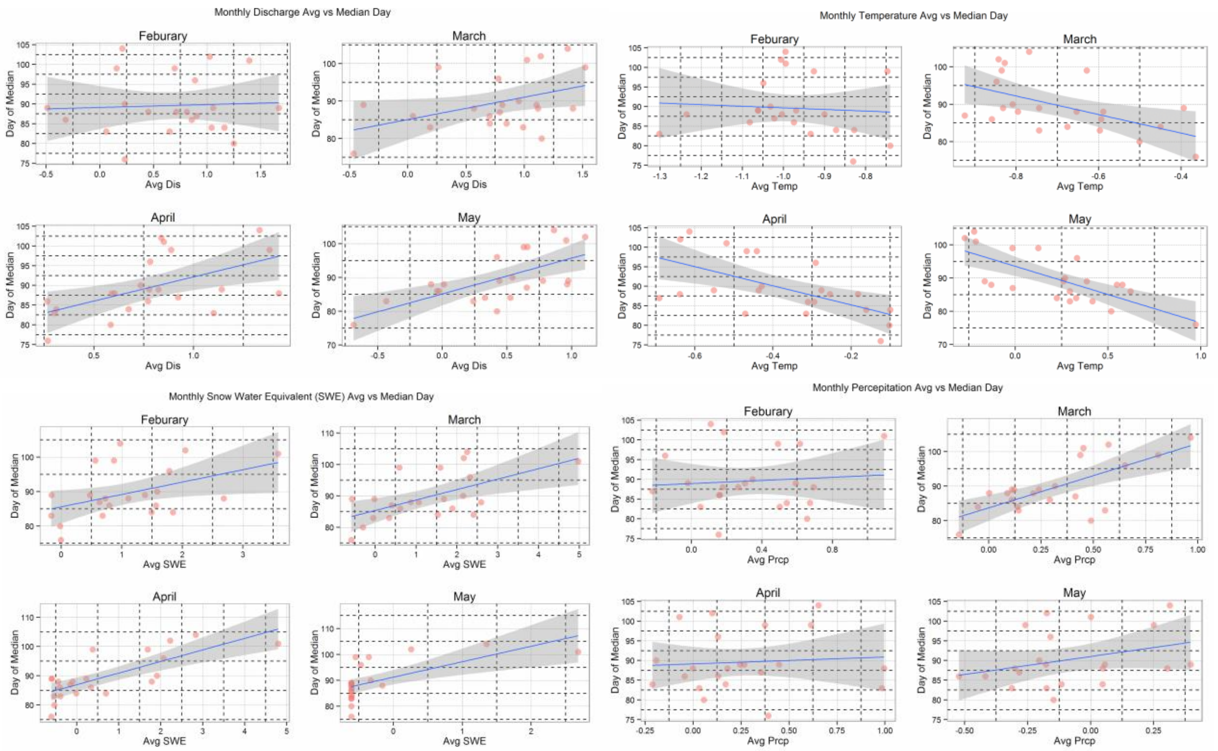


Figure 6: Correlations for monthly averages of February through May of discharge (top left), temperature (top right), snow water equivalent (bottom left), and precipitation (bottom right)

Model	r^2	P-Val	Sd Error
Temp March	0.2858	0.0103	6.32
Temp April	0.3639	0.0029	5.97
Dis March	0.1831	0.0469	6.76
Dis April	0.3124	0.0069	6.21
SWE Feb	0.2004	0.0367	6.695
SWE March	0.3427	0.0042	6.07
SWE April	**0.5778**	**4.05E-05**	**4.87**
Prcp March	**0.4918**	**0.0003**	**5.34**
Temp+Dis+SWE April	0.6614	4.20E-06	4.36
Temp+SWE April	0.6540	5.24E-06	4.40
SWE April + Prcp March	**0.6842**	**2.06E-06**	**4.21**
SWE April + Temp April + Prcp March	**0.7134**	**7.68E-07**	**4.01**
SWE April + Temp April + Dis April + Prcp March	0.7139	7.52E-07	4.00

Table 1: This table depicts the correlation score, p-value from a significant t-test, and the standard deviation of the number of days the prediction is off.

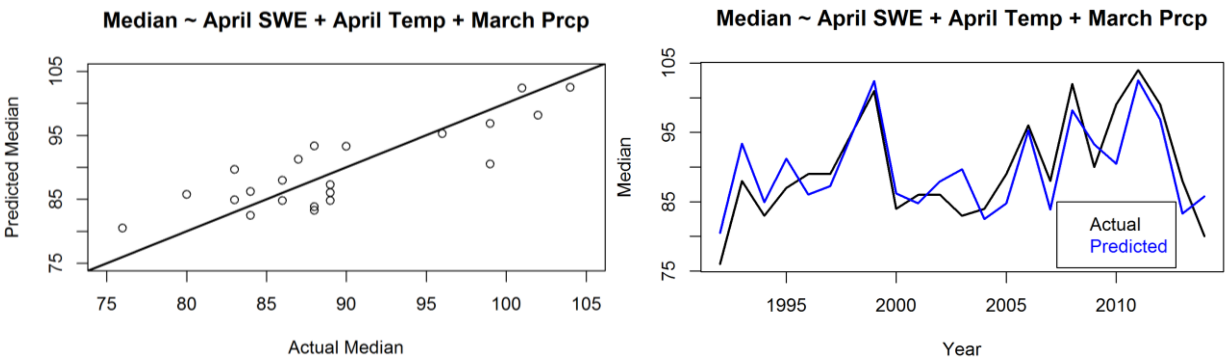


Figure 7: On the left is the correlation between predicted and actual median. On the right is the predicted median in blue and actual median in black for each year.

Predicting Running and Non-Running Days

Narrowing the focus of our research we wanted to try and predict whether or not individual days would be running days or not. To do this we used a support vector machine, a common machine learning algorithm. The goal was to have the algorithm learn what factors constituted a running day. Specifically we wanted it to learn to predict the non-running days during the middle and later parts of the run.

To train the algorithm we used cross validation leaving out one year for testing. To gain a baseline performance measure of what the algorithm could learn, we trained it using temperature and discharge. It learned to find a boundary very similar to what the triangular region found. So we should expect to miss running days early in the run and non-running days later in the run. It can be seen in figure 8 that this is just the case. The performance metrics (table 2), show that it did not handle the variation in run conditions well. It has huge confidence interval for recalling the run, which means that it was unable to find a boundary that captured most of the running days for all the years.

To improve on this model we added to our feature set, the day of the run, changes in temperature and discharge, and took a 3 day history of these values. It improved in almost every performance metric. The biggest gain is in the decrease in the confidence intervals. The 10% decrease in the confidence interval in recall run shows that it did a much better job in handling year to year variation. It was also able to learn the correctly classify a few of the non-running days in the middle of the run. In general though it did not learn to classify non-running days during the middle and later parts of the run.

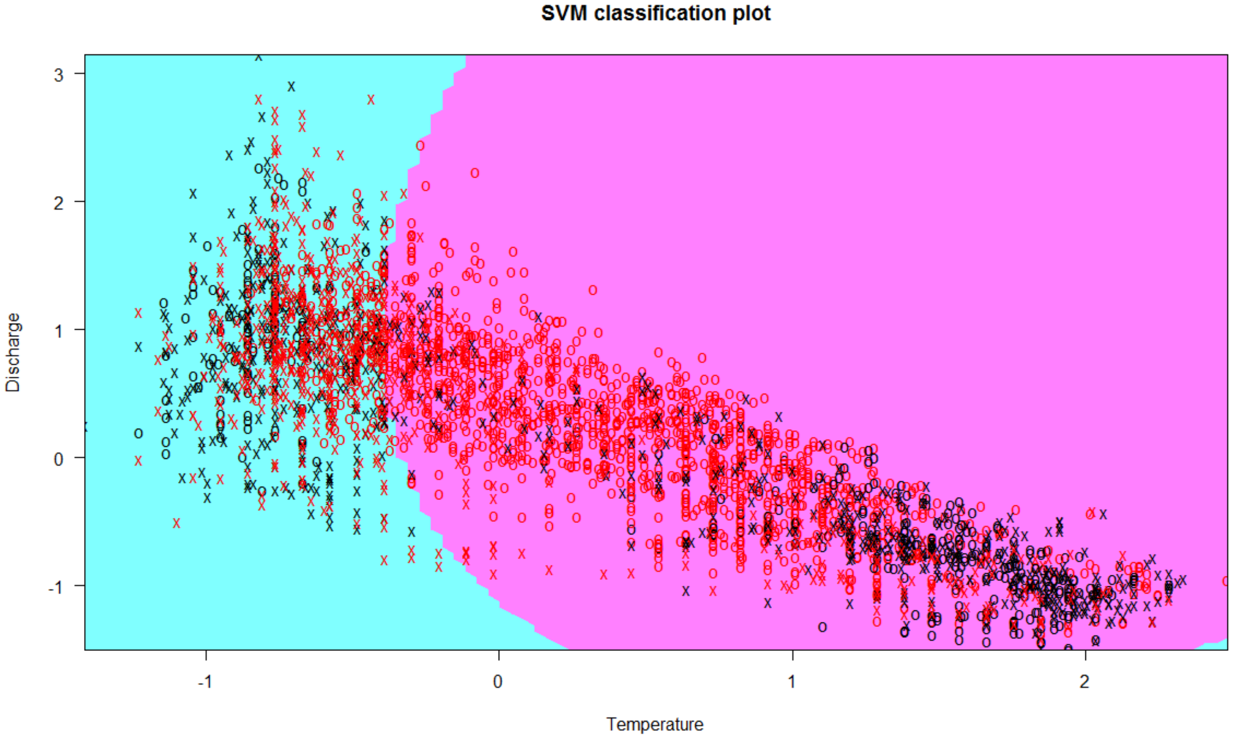


Figure 8: This is the training data used to test on the 2005 run. The purple region is the region classified as running days and the teal region classified as non-running days. The red points are the actual running days. The black points are the actual non-running days. Points marked with an “x” are support vectors.

Metric	Average Performance	95% Conf.
Accuracy – Baseline	83.23%	8.24%
All Vars	85.26%	7.84%
Recall Run – Baseline	70.51%	29.82%
All Vars	69.51%	19.74%
Recall Not – Baseline	89.70%	13.94%
All Vars	92.95%	9.12%
Precision Run – Baseline	78.88%	24.48%
All Vars	82.34%	21.66%
Precision Not – Baseline	86.40%	12.24%
All Vars	86.65%	9.06%

Table 2: Table of performance metrics for the baseline model (temperature and discharge) and the model with all features.

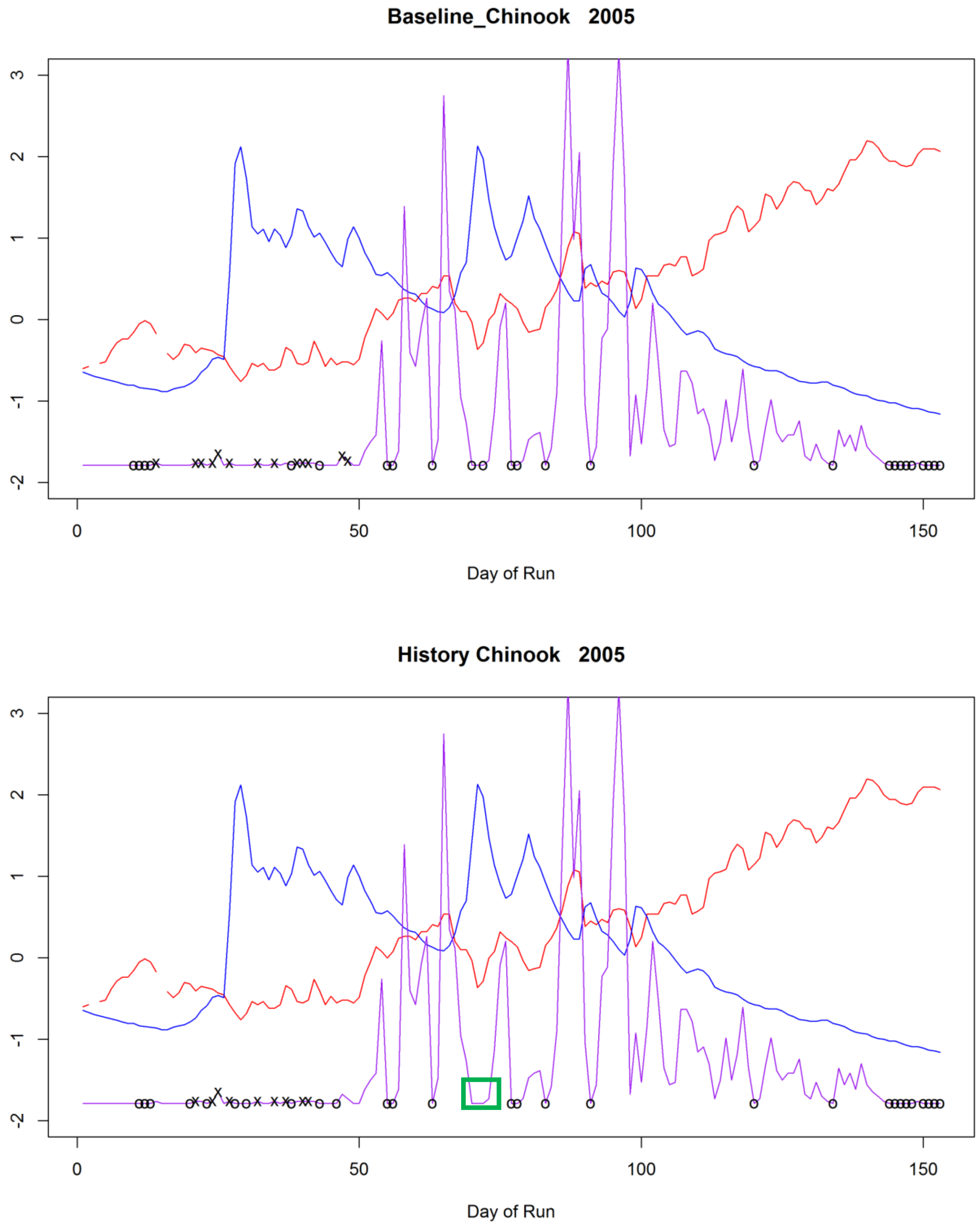


Figure 9: Test for 2005 cross-validation of the two models baseline (top) and all features (bottom). The “x” indicates an incorrectly classified run day and “o” indicates an incorrectly classified non-run day.

Conclusions

Through our investigation we have defined the run conditions for spring Chinook salmon on the North Umpqua River and developed a prediction model for the median day of the run. The wild Chinook salmon pass through the Winchester Dam in large numbers when water temperatures are between 55.1˚F and 73.2˚F and discharge is between 1850 cu ft/sec and 7855 cu ft/sec. Our method of predicting the median day of the run uses average April water temperature, average April snow level at Diamond Lake, and the average precipitation in Winchester, OR during March. This method achieves an r^2 correlation of .7134 and 95% confidence interval of ± 8.02 days.

Bibliography

Cederholm, C. Jeff et al. “Pacific Salmon Carcasses: Essential Contributions Of Nutrients and Energy for Aquatic and Terrestrial Ecosystems.” *Fisheries* 24.10 (1999): 6–15. Web.

Dean Runyan Associates. Fish, Hunting, Wildlife Viewing, and Shellfishing in Oregon, 2008. Portland: Oregon Depeartment of Fish and Wildlife, Travel Oregon, 2009. Document.

Gresh, Ted, Jim Lichatowich, and Peter Schoonmaker. “An Estimation Of Historic and Current Levels of Salmon Production in the Northeast Pacific Ecosystem: Evidence of a Nutrient Deficit in the Freshwater Systems of the Pacific Northwest.” *Fisheries* 25.1 (2000): 15–21. Web.

Keefer, Matthew L., Christopher A. Peery, and Christopher C. Caudill. “Migration Timing Of Columbia River Spring Chinook Salmon: Effects of Temperature, River Discharge, and Ocean Environment.” *Transactions of the American Fisheries Society* 137.4 (2008): 1120–1133. Web.

The Research Group, LLC. "Oregon Commercial Fishing Industry in 2013, Briefing Report." Briefing Report. 2014. Document.