Estimates of Steelhead in Dungeness River

Using Sonar

Kevin See[1,*]

August 03, 2022

Contents

1 Goals **2**

2 Methods **2**

2.1 Expanding 30 min to 60 min . . . 2

2.2 Excluding Bull Trout . . . 3

2.3 Missing Data . . . 3

3 Results **7**

3.1 Expanding 30 min to 60 min . . . 7

3.2 Excluding Bull Trout . . . 10

3.3 Missing Data . . . 10

4 Discussion Points **17**

4.1 Expanding 30 min to 60 min . . . 17

4.2 Excluding Bull Trout . . . 17

4.3 Missing Data . . . 18

5 Decisions to be Made **18**

[1] Washington Department of Fish & Wildlife

[*] Correspondence: Kevin See <Kevin.See@dfw.wa.gov>

1 Goals

To estimate the number of adult steelhead spawners in the Dungeness river, with appropriate uncertainty, based upon a sonar system set up in the lower river. This involves several sub-tasks:

1. Expand 30 min observations into 60 min estimates. Most hours only the first 30 min of video was analyzed, but for a subset of hours the entire hour was analyzed. Develop a crosswalk between the first 30 min and the entire hour, and then predict the total counts from all the hours when only the first 30 min are available, with uncertainty from this crosswalk.
2. Exclude bull trout from images based on species composition data.
3. Account for outages. There are periods when the sonar unit was not functioning for a variety of reasons. Develop a method to interpolate across those periods of missing data.

2 Methods

2.1 Expanding 30 min to 60 min

We started by examining the counts during the first 30 minutes and the entire hour. First, we filtered for records with a confidence level of 1 (extremely confident) and a length greater than 67 cm to ensure we were only comparing records of steelhead. For each hour with both halves recorded and for each half hour, we summed the fish determined to be moving upstream, and those moving downstream, and calculated the number of net upstream fish (upstream - downstream). We then added the two half hours together to provide a number of the net upstream fish for that hour.

We compared the first half hour with the entire hour at several temporal scales. We started with a single hour, then also summed net upstream fish in 6 hour blocks, and finally summed the net upstream fish by date. The data was structured such that hours where both half hours were examined were usually consecutive at least up to the day scale, meaning each 6 hour block and each day had nearly identical amounts of time with two half hours to other periods at the same temporal scale.

For each temporal scale grouping and direction (up, down and net upstream), we fit a linear model with the counts of fish in the first 30 minutes as the independent variable and the total count of fish in the entire hour as the dependent variable. In each model, we fixed the intercept at 0, to ensure that if no fish were counted in the first half hour, we would expand that to zero fish for the entire hour. We focused on the estimated slope, hypothesizing that it should be 2.

We also performed 20-fold cross validation. For each cross-validation, we created a training dataset by sampling all the data from 70% of the days with full hours reviewed, and held out the remaining 30% as a test dataset. We fit the same linear regression to each training dataset, and used it to predict the total net upstream steelhead for each day in the test dataset. We then summarized the mean bias, root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percent error (MAPE) for each temporal scale. To calculate these, first we calculated the error for each day, e_t, as the predicted minus the observed.

$$e_t = pred_t - obs_t$$

The mean bias is simply the average e_t. RMSE is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

MAE is

$$MAE = \frac{1}{n}\sum |e_t|$$

MAPE is the average of percent errors, and is defined as:

$$MAPE = \frac{1}{n} \sum \frac{|e_t|}{obs_t}$$

2.2 Excluding Bull Trout

Bull trout are swimming past the sonar unit as well as steelhead, and we need to parse which fish identified by the sonar are steelhead, and exclude any bull trout. The largest bull trout sampled in any species composition data was 67 cm, so we are assuming any fish larger than 67 cm detected on the sonar is a steelhead. It then remains to filter the fish equal to or less than 67 cm long from the sonar and determine what proportion of those are steelhead, and what are bull trout.

We only have species composition data for one year, 2021. It was collected weekly, using tangle nets just upstream of the sonar location. For every fish caught, we know the date and fork length of that fish. Based on this data, we have also determined that the steelhead run on the Dungeness is over by June 1. Therefore, we have only made predictions for fish detected prior to June 1st.

We have several options to model the probability of any particular fish, less than or equal to 67 cm, being a steelhead. we could model that probability as a factor of date (perhaps with a quadratic term to capture non-linearity), or as a factor of fork length, or both. If we use date, we can interpret the probability of being a steelhead as the proportion of all fish detected on that date that are steelhead. If we use fork length (or date and fork length), we can assign a probability to every fish detected by sonar, and assume that all fish with a probability greater than some threshold (probably 50%) are steelhead.

We chose to use fork length and the Juilan day of capture. From the species composition netting, there are 31 fish to use in this model. These can be seen in Figures 1 and 2. Since we only care about differentiating steelhead, we grouped resident rainbows with bull trout, and then fit a binomial GAM with a logit link, using splines of fork length and Julian day to predict the probability of a fish being a steelhead. We did not restrict the dataset to fish with fork lengths less than 67 cm, because larger fish have information about the shape of the logistic curve.

After fitting this GAM, we predicted the probability of being a steelhead for all fish observed on the sonar that were smaller than or equal to 67 cm, based on their length and Julian day of observation. Any fish with a probability of 50% or greater we assigned to be a steelhead. We then applied the same model (Section 2.1) to expand 30 minute counts for small fish to full hour counts. We added counts or estimates of large fish to small fish for each time period to estimate total net steelhead moving upstream for each time period. Estimates of large and small fish within the same time period were assumed to be independent when calculating the standard error.

2.3 Missing Data

There are periods when the sonar was not functioning, for a variety of reasons. Rather than ignore those time periods, and assume that no steelhead were passing then, we would prefer to impute net upstream fish for those missing values.

The first step is to expand the estimates of net upstream fish for periods when the sonar only partially operated (e.g. 14 hours out of a 24 hour day). We did this by dividing the estimate for that period by the percent of time the sonar was operational in that period. This assumes that fish are behaving similarly for that entire period.

The next step is to deal with those periods when the sonar was not operating at all, where we have truly missing data. Table 1 shows how much data was missing for each year, depending on how the periods were constructed (e.g. hourly, hourly blocks, daily).

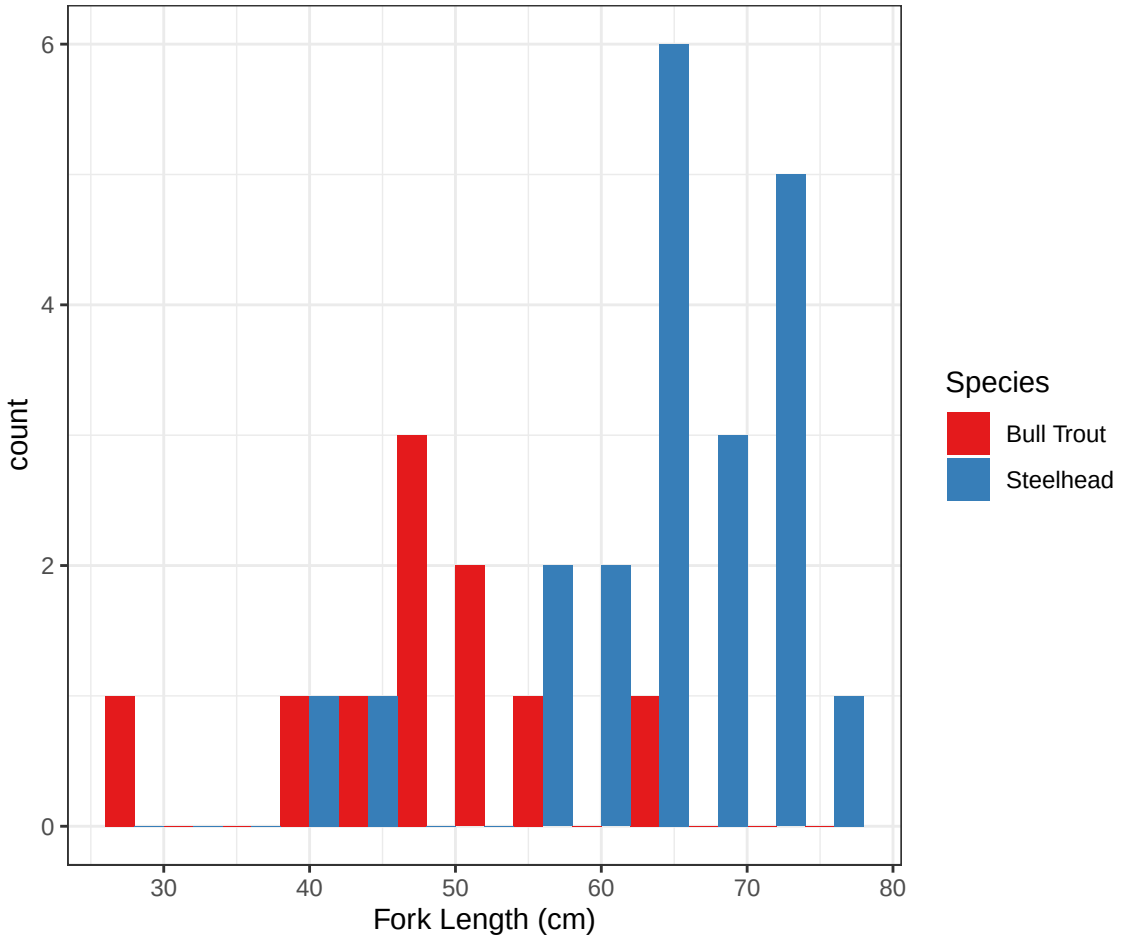

Figure 1: Histogram of forklengths, colored by species.

Table 1: Table showing how many periods are in each group of data, and how many of those periods are NAs (missing values).

Time Scale	Year	n Periods	n NAs	% NA
Hour	2019	2112	128	6.1
6 Hour Block	2019	352	20	5.7
Day	2019	88	5	5.7
Hour	2020	2597	646	24.9
6 Hour Block	2020	433	100	23.1
Day	2020	109	22	20.2
Hour	2021	2867	125	4.4
6 Hour Block	2021	478	10	2.1
Day	2021	120	0	0.0

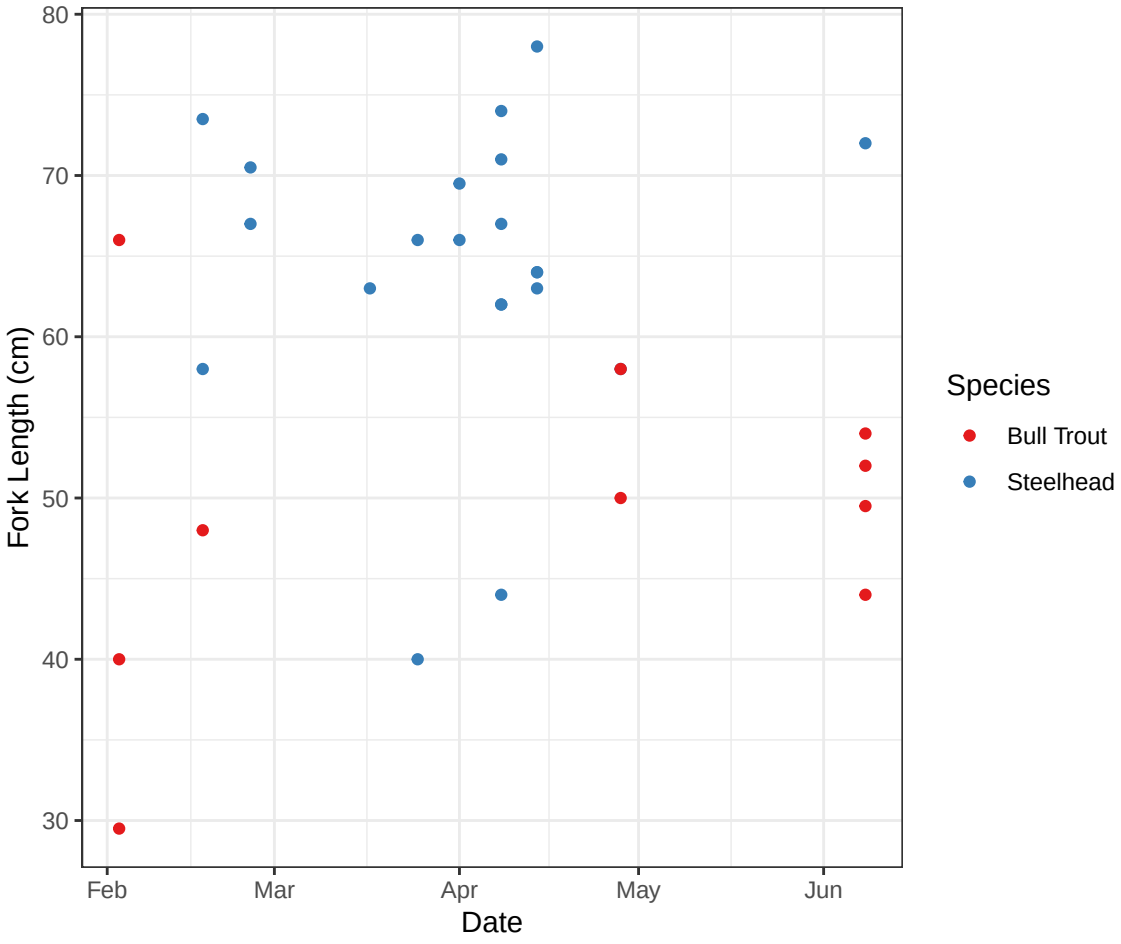

Figure 2: Scatterplot of date of capture and forklength, colored by species.

To interpolate across those periods of missing data, we employed time-series models. Using the `forecast` package in R, we fit an ARIMA (auto-regressive integrated moving average) model, and let the `auto.arima` function determine the model with the best order (number of auto-regressive, moving average and difference steps) for each year and time-scale combination. We used the uncertainty from this model (σ^2) for all predictions.

We examined several forms of interpolation across the missing data, including a Kalman filter, linear regression and moving average. The Kalman filter uses the ARIMA structure to estimate the missing data. A linear regression essentially draws a straight line from the data point prior to the first missing data and the data point after the last missing data point for each gap in the time series. A moving average approach uses two non-missing values prior to the missing data point, and two non-missing values after, weights them exponentially by their distance from the missing data point, and calculates the weighted mean.

3 Results

3.1 Expanding 30 min to 60 min

The expansion factor (i.e. slope) changes depending on the temporal scale that data is summarized on. Figure 3 shows the various regressions, comparing them with the 1-1 line and the expected slope of two. None of the temporal scales produced a slope of two, but the longer the temporal scale the closer it got to that expected value.

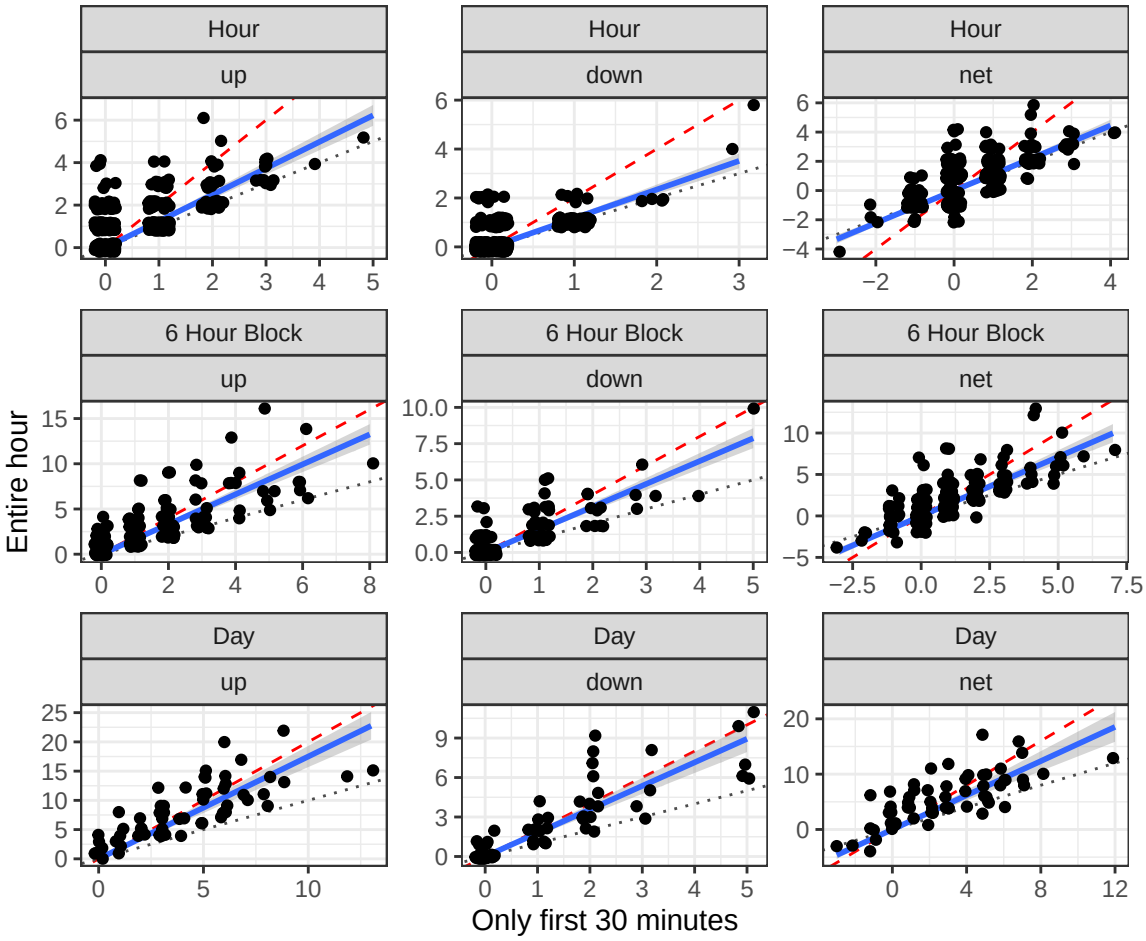

Figure 3: Scatter plots showing the net counts of fish moving upstream, using hours with both the first 30 minutes and second 30 minutes. The counts are summarized by hour, six hour blocks and entire day (24 hours) in the different facets. The dashed red line is has a slope of 2 (expected value), the dotted grey line has a slope of 1, and the blue line is the linear regression fit to that data, with 95% confidence intervals.

Table 2 shows the summary of linear models fit to data summarized at various time scales. We summarized the estimated number of steelhead larger than 67 cm at the day scale (summing estimates at smaller temporal scales) and plotted the time-series in Figure 4 to show the differences caused by summarizing data at different time scales.

Table 2: Results of fitting linear models with up, down and net upstream fish as the response and the number of fish counted in that direction in the first 30 minutes as the covariate with no intercept.

Time Scale	Direction	Slope	SE	95% CI	R2
Hour	up	1.24	0.05	1.15-1.34	0.62
Hour	down	1.17	0.05	1.07-1.27	0.58
Hour	net	1.11	0.05	1.01-1.21	0.56
6 Hour Block	up	1.66	0.07	1.51-1.8	0.77
6 Hour Block	down	1.58	0.07	1.44-1.71	0.77
6 Hour Block	net	1.42	0.08	1.27-1.58	0.67
Day	up	1.75	0.09	1.57-1.93	0.86
Day	down	1.79	0.10	1.59-1.99	0.84
Day	net	1.54	0.11	1.32-1.77	0.76

Table 3: Results of k-fold cross validation. Best result of each performance indicator is in bold.

Time Scale	Direction	Avg. Bias	RMSE	MAE	MAPE
Day	up	**-1.07**	**3.71**	**2.98**	0.42
6 Hour Block	up	-1.50	3.74	2.98	**0.41**
Hour	up	-3.22	4.51	3.40	0.44
Day	down	**-0.53**	**1.92**	**1.49**	0.45
6 Hour Block	down	-0.96	1.98	1.51	**0.44**
Hour	down	-1.68	2.35	1.76	0.46
Day	net	**-1.18**	**3.64**	2.83	0.49
6 Hour Block	net	-1.56	3.65	**2.80**	0.48
Hour	net	-2.51	4.03	3.01	**0.47**

Table 4: Estimates of total net upstream fish larger than 67 cm, using only extremely confident observations. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Time Scale	Year	Up - Down Estimate	Up - Down SE	Net Estimate	Net SE	Difference
Hour	2019	344	1.42	305	1.27	40
6 Hour Block	2019	439	2.88	376	2.47	63
Day	2019	445	5.14	403	4.23	42
Hour	2020	312	0.93	297	0.78	16
6 Hour Block	2020	349	1.84	324	1.48	25
Day	2020	350	3.31	334	2.50	16
Hour	2021	340	1.46	301	1.26	39
6 Hour Block	2021	422	2.94	363	2.42	59
Day	2021	420	5.17	386	4.21	34

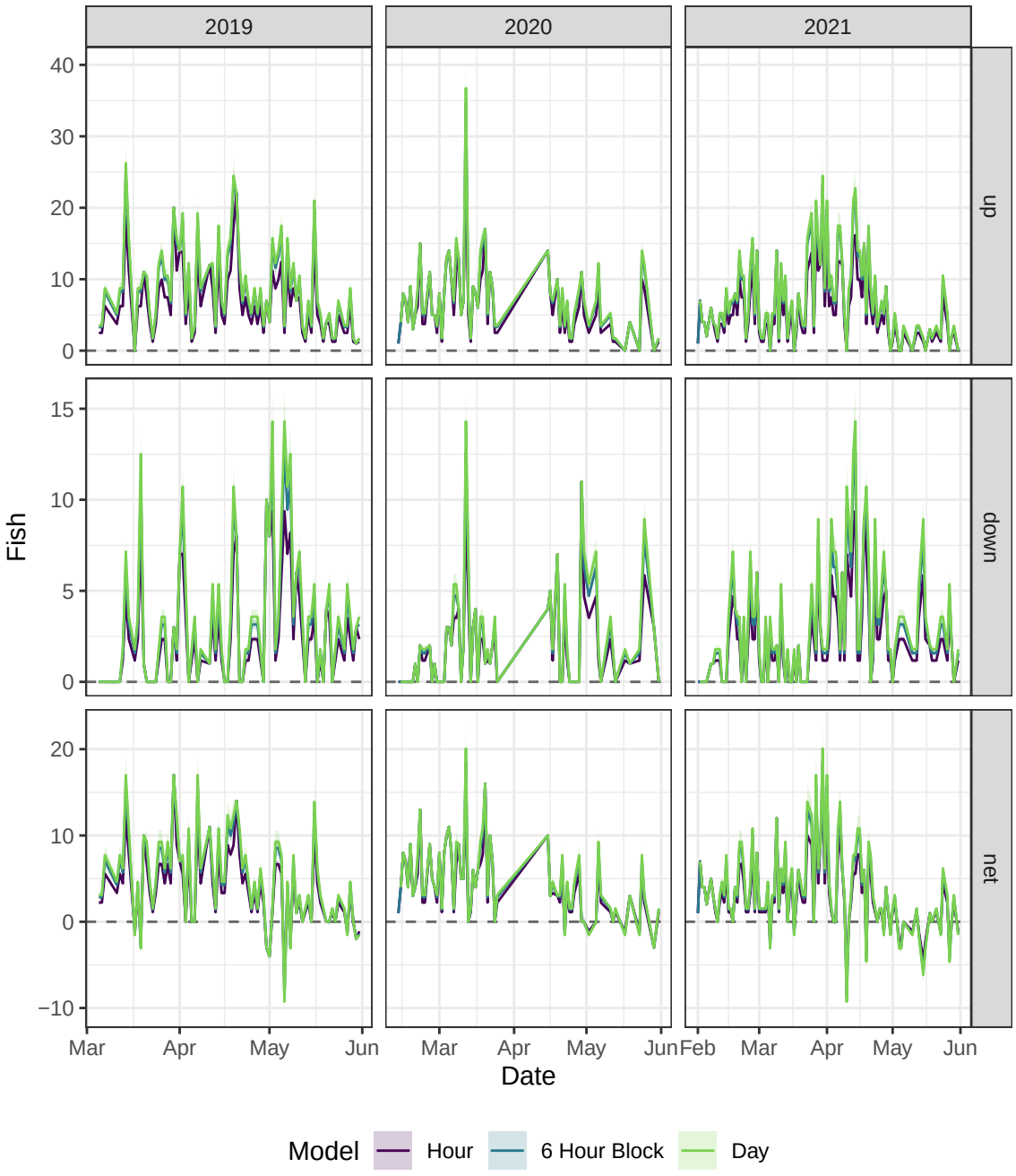

Figure 4: Time-series of estimates based on available data, using only extremely confident observations of fish greater than 67 cm, faceted by year and direction (upstream, downstream or net upstream). Colors correspond to which regression model was used to expand the 30 minutes observations. Any uncertainty shown is derived from the linear regression model.

3.2 Excluding Bull Trout

Figure 5 shows the fitted GAM that predicts the probability of being a steelhead based on a fish's length and date of capture. Note that the mean probability of being a steelhead for a 67 cm long fish, averaged across the entire season, would be 90.9% of being a steelhead with this model. Also note that a fish 60 cm long would not be considered a steelhead if observed at the very beginning of the season or after the beginning of May, but would if observed in late March or April.

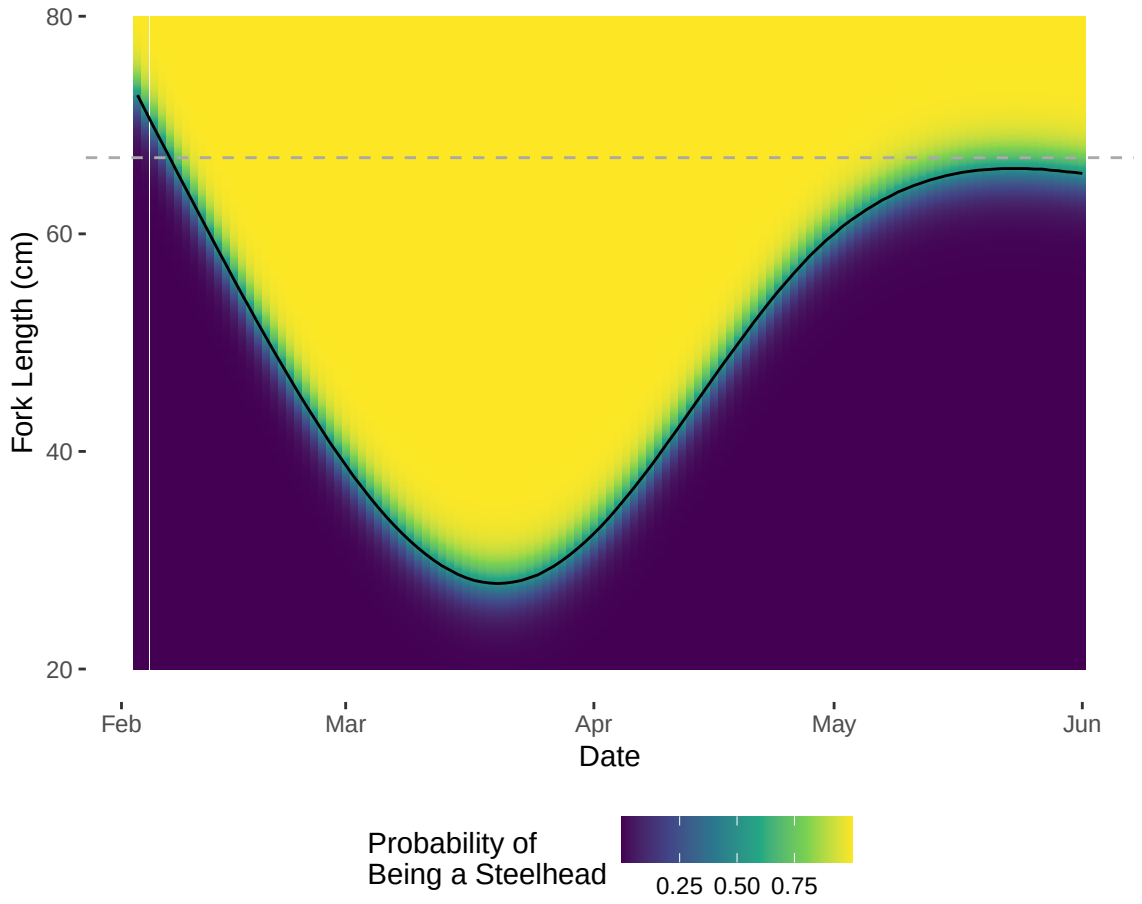

Figure 5: The color depicts the probability of fish being a steelhead given the date of capture and fork length. Fish above the black line would have a greater than 50% probability of being a steelhead. The dotted line shows the 67 cm threshold for which fish we will be applying this model to.

Applying this model and rule-set to all the observed fish smaller than or equal to 67 cm, including the predictive model to expand 30 minute counts to full hour counts, a number of additional steelhead are added to our estimate each year (Table 5). The total estimates (including all fish larger than 67 cm, as well as fish less than or equal to 67 cm that are predicted to be steelhead) are shown in Table 6.

3.3 Missing Data

Figure 6 shows the periods when the sonar array was not operating, and Figure 7 shows how that impacts the time-series of fish counts. Note the large period in 2020 when the sonar was shut down due to COVID-19.

Table 5: Estimates (SE) of steelhead smaller than 67 cm, using only extremely confident observations, by direction of movement. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Direction	Time Scale	2019	2020	2021
up	Hour	147 (0.5)	379 (0.7)	213 (0.7)
up	6 Hour Block	183 (0.9)	431 (1.4)	272 (1.4)
up	Day	192 (1.6)	442 (2.6)	285 (2.2)
down	Hour	46 (0.3)	174 (0.5)	157 (0.6)
down	6 Hour Block	57 (0.4)	198 (0.7)	203 (1)
down	Day	63 (0.7)	210 (1.4)	227 (1.9)
net	Hour	91 (0.6)	191 (0.7)	44 (0.8)
net	6 Hour Block	110 (1.1)	212 (1.4)	53 (1.4)
net	Day	117 (1.9)	221 (2.3)	57 (2.3)

Table 6: Estimates of total steelhead, using only extremely confident observations. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Time Scale	Year	Up - Down Estimate	Up - Down SE	Net Estimate	Net SE	Difference
Hour	2019	445	1.54	395	1.40	50
6 Hour Block	2019	565	3.05	486	2.69	79
Day	2019	574	5.44	520	4.66	53
Hour	2020	517	1.24	488	1.06	29
6 Hour Block	2020	581	2.41	536	2.01	45
Day	2020	582	4.43	555	3.37	27
Hour	2021	396	1.74	345	1.51	51
6 Hour Block	2021	491	3.40	416	2.82	75
Day	2021	479	5.95	443	4.80	36

As the temporal scale on which counts are aggregated increases, the amount of missing data decreases. For example, if three hours are missing within a day, we can expand the rest of the day's counts by the percent of time the sonar was operational, so that day will not be "missing" at the day time-scale, although those three hours still are if we are operating on an hour time-scale.

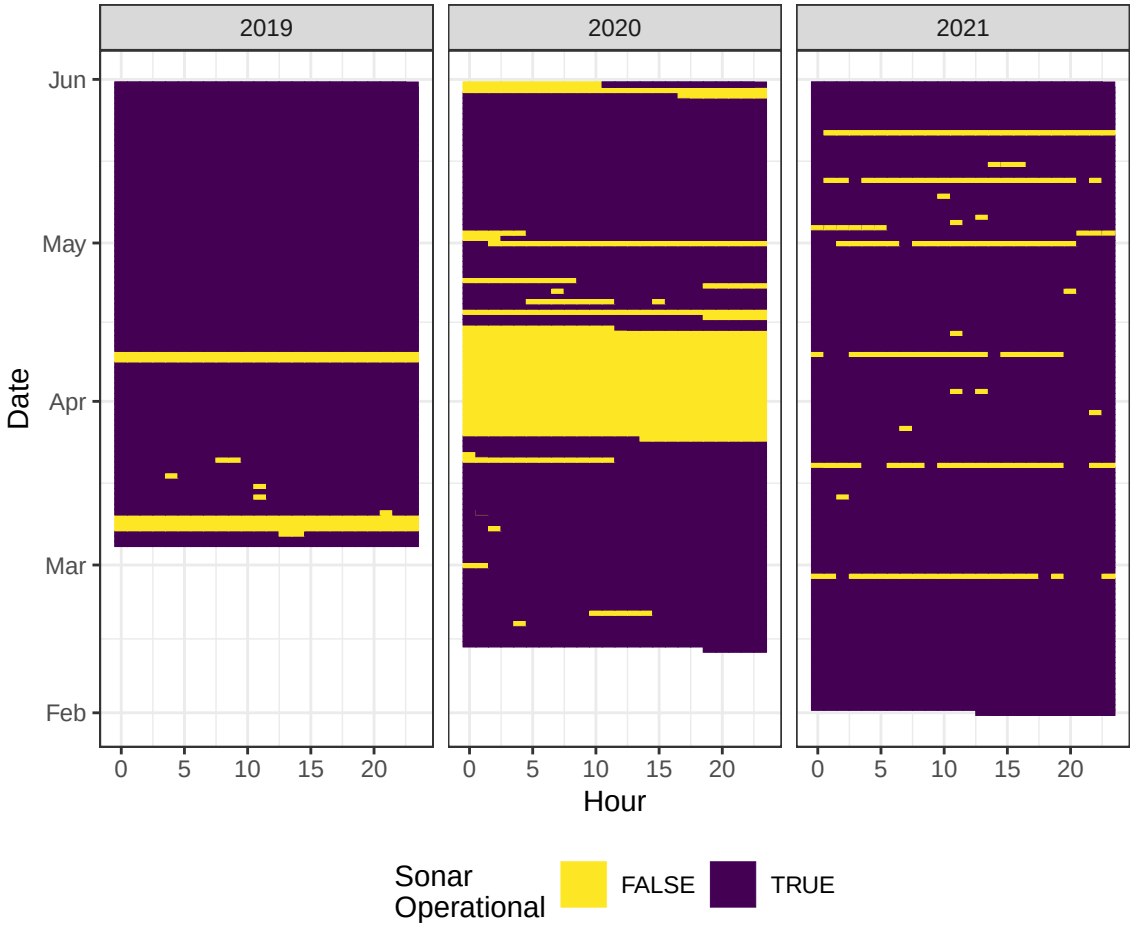

Figure 6: Purple depicts hours when the sonar was working, while yellow indicates the sonar was not functioning.

After interpolating across the missing data, Table 7 displays how many fish were added to each year's estimate, based on the temporal scale and the interpolation method. Table 8 shows estimates of all steelhead, by year, direction and time-scale, split by what interpolation method was used. Table 9 provides final estimates of total net upstream steelhead each year, including fish smaller than 67 cm and periods of missing data, split out by the temporal scale the data was summarized on and the interpolation model used for missing data.

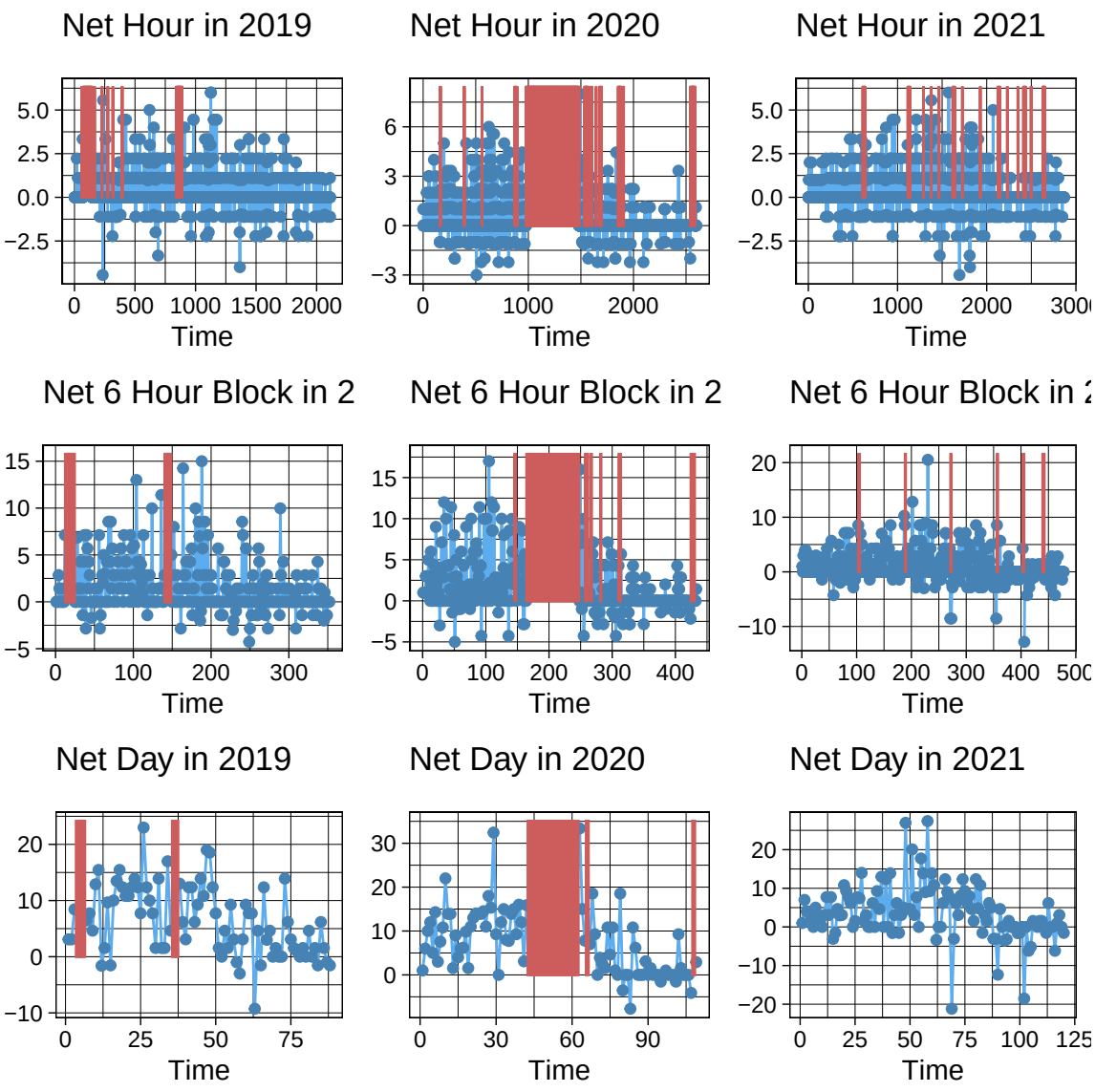

Figure 7: Time series of net upstream fish in blue, with missing data highlighted in red.

Table 7: Estimates (SE) of how many steelhead by direction are added to the totals from periods with wholly missing data. Interpolation methods include the Kalman filter, linear model, and moving average. There are three estimates per year, corresponding to the three different temporal scales. Blank cells indicate no interpolation was necessary for that year / temporal scale combination.

Time Scale	Direction	Year	Kalman	Linear	MA
Hour	down	2019	0 (5.1)	1 (5.1)	1 (5.1)
6 Hour Block	down	2019	0 (7.1)	6 (7.1)	4 (7.1)
Day	down	2019	2 (8.6)	6 (8.6)	8 (8.6)
Hour	net	2019	27 (8.8)	94 (8.8)	83 (8.8)
6 Hour Block	net	2019	38 (12.1)	74 (12.1)	48 (12.1)
Day	net	2019	34 (12.3)	40 (12.3)	37 (12.3)
Hour	up	2019	40 (9.4)	103 (9.4)	91 (9.4)
6 Hour Block	up	2019	48 (15)	90 (15)	58 (15)
Day	up	2019	53 (14.4)	49 (14.4)	48 (14.4)
Hour	down	2020	157 (12.7)	3 (12.7)	20 (12.7)
6 Hour Block	down	2020	156 (18.9)	25 (18.9)	117 (18.9)
Day	down	2020	157 (20.8)	157 (20.8)	168 (20.8)
Hour	net	2020	244 (22.3)	53 (22.3)	31 (22.3)
6 Hour Block	net	2020	325 (28.5)	15 (28.5)	271 (28.5)
Day	net	2020	266 (29.8)	500 (29.8)	398 (29.8)
Hour	up	2020	168 (24.2)	58 (24.2)	52 (24.2)
6 Hour Block	up	2020	498 (34.7)	43 (34.7)	397 (34.7)
Day	up	2020	460 (38.7)	678 (38.7)	581 (38.7)
Hour	down	2021	20 (5.5)	58 (5.5)	60 (5.5)
6 Hour Block	down	2021	8 (6.3)	41 (6.3)	29 (6.3)
Day	down	2021	-	-	-
Hour	net	2021	5 (8.4)	8 (8.4)	2 (8.4)
6 Hour Block	net	2021	4 (8.4)	-1 (8.4)	2 (8.4)
Day	net	2021	-	-	-
Hour	up	2021	64 (8.1)	71 (8.1)	67 (8.1)
6 Hour Block	up	2021	7 (9)	41 (9)	33 (9)
Day	up	2021	-	-	-

Table 8: Estimates (SE) of total steelhead by direction, using only extremely confident observations, and after interpolating counts for periods of missing data. Interpolation methods include no interpolation, the Kalman filter, linear model, and moving average. There are three estimates per year, corresponding to the three different temporal scales.

Time Scale	Direction	Year	None	Kalman	Linear	MA
Hour	down	2019	248 (0.8)	248 (5.1)	249 (5.1)	249 (5.1)
6 Hour Block	down	2019	312 (1.2)	312 (7.2)	319 (7.2)	316 (7.2)
Day	down	2019	346 (2.4)	348 (8.9)	351 (8.9)	353 (8.9)
Hour	net	2019	395 (1.4)	422 (8.9)	489 (8.9)	479 (8.9)
6 Hour Block	net	2019	488 (2.7)	526 (12.4)	562 (12.4)	536 (12.4)
Day	net	2019	523 (4.7)	557 (13.2)	563 (13.2)	560 (13.2)
Hour	up	2019	693 (1.3)	733 (9.5)	796 (9.5)	785 (9.5)
6 Hour Block	up	2019	880 (2.8)	928 (15.3)	970 (15.3)	939 (15.3)
Day	up	2019	923 (4.9)	975 (15.2)	972 (15.2)	971 (15.2)
Hour	down	2020	303 (0.6)	460 (12.7)	306 (12.7)	322 (12.7)
6 Hour Block	down	2020	352 (1)	509 (19)	378 (19)	469 (19)
Day	down	2020	393 (2.1)	550 (20.9)	550 (20.9)	561 (20.9)
Hour	net	2020	488 (1.1)	733 (22.4)	541 (22.4)	519 (22.4)
6 Hour Block	net	2020	544 (2)	869 (28.5)	558 (28.5)	815 (28.5)
Day	net	2020	615 (3.8)	880 (30)	1114 (30)	1012 (30)
Hour	up	2020	820 (1.1)	988 (24.2)	878 (24.2)	872 (24.2)
6 Hour Block	up	2020	942 (2.2)	1440 (34.8)	985 (34.8)	1339 (34.8)
Day	up	2020	1040 (4.3)	1500 (38.9)	1718 (38.9)	1621 (38.9)
Hour	down	2021	399 (1)	418 (5.6)	457 (5.6)	459 (5.6)
6 Hour Block	down	2021	560 (2)	568 (6.6)	601 (6.6)	589 (6.6)
Day	down	2021	690 (4.6)	-	-	-
Hour	net	2021	345 (1.5)	350 (8.6)	354 (8.6)	347 (8.6)
6 Hour Block	net	2021	421 (3.2)	425 (8.9)	419 (8.9)	423 (8.9)
Day	net	2021	431 (5.8)	-	-	-
Hour	up	2021	795 (1.4)	859 (8.3)	866 (8.3)	861 (8.3)
6 Hour Block	up	2021	1059 (3.1)	1066 (9.5)	1100 (9.5)	1092 (9.5)
Day	up	2021	1152 (5.7)	-	-	-

Table 9: Estimates of total net upstream fish, using only extremely confident observations. There are three estimates per year, corresponding to the three different regression models for expanding 30 minute observations.

Time Scale	Year	Model	Up - Down Estimate	Up - Down SE	Net Estimate	Net SE	Difference
Hour	2019	None	445	1.54	395	1.40	50
Hour	2019	Kalman	485	10.79	422	8.94	63
Hour	2019	Linear	548	10.79	489	8.94	58
Hour	2019	MA	536	10.79	479	8.94	57
6 Hour Block	2019	None	568	3.06	488	2.70	80
6 Hour Block	2019	Kalman	616	16.93	526	12.38	89
6 Hour Block	2019	Linear	651	16.93	562	12.38	89
6 Hour Block	2019	MA	622	16.93	536	12.38	86
Day	2019	None	577	5.48	523	4.69	54
Day	2019	Kalman	627	17.62	557	13.20	70
Day	2019	Linear	621	17.62	563	13.20	57
Day	2019	MA	618	17.62	560	13.20	57
Hour	2020	None	517	1.24	488	1.06	29
Hour	2020	Kalman	528	27.38	733	22.36	-205
Hour	2020	Linear	572	27.38	541	22.36	31
Hour	2020	MA	550	27.38	519	22.36	30
6 Hour Block	2020	None	590	2.43	544	2.04	46
6 Hour Block	2020	Kalman	931	39.60	869	28.54	62
6 Hour Block	2020	Linear	608	39.60	558	28.54	49
6 Hour Block	2020	MA	870	39.60	815	28.54	55
Day	2020	None	647	4.74	615	3.84	32
Day	2020	Kalman	950	44.16	880	30.04	70
Day	2020	Linear	1168	44.16	1114	30.04	54
Day	2020	MA	1060	44.16	1012	30.04	48
Hour	2021	None	396	1.74	345	1.51	51
Hour	2021	Kalman	441	9.97	350	8.56	90
Hour	2021	Linear	409	9.97	354	8.56	55
Hour	2021	MA	402	9.97	347	8.56	55
6 Hour Block	2021	None	499	3.70	421	3.19	78
6 Hour Block	2021	Kalman	498	11.59	425	8.95	73
6 Hour Block	2021	Linear	499	11.59	419	8.95	80
6 Hour Block	2021	MA	503	11.59	423	8.95	80
Day	2021	None	462	7.34	431	5.84	31
Day	2021	Kalman	462	7.34	431	5.84	31
Day	2021	Linear	462	7.34	431	5.84	31
Day	2021	MA	462	7.34	431	5.84	31

4 Discussion Points

- What to do with rows where `data_recorded` is "Partial?" This includes 175 rows, or 1.8% of the data. Exclude and treat as missing data? Or is there a better way to parse this? Currently I've filtered it out and treated it as missing.
- What should we do with observations with confidence of 2 or 3? They are currently excluded completely.
- What to do about kelts? There is currently one observation from the tangle netting that was identified as a kelt (caught on June 8, 2021). With enough data, we could build another GAM model to describe the probability of a steelhead moving downstream as being a kelt, perhaps based on Julian day. Unfortunately, we don't have nearly enough data at the moment. Alternatively, we could assign a date and assume all downstream fish before that date are ***not*** kelts, and all downstream fish observed after that date ***are*** kelts. Given the paucity of data, we may be relying on expert opinion to determine this date.

4.1 Expanding 30 min to 60 min

- The regression between counts in the first hour and the entire hour shows a consistent expectation that the counts in the second part of the hour will be less than counts in the first part. This holds regardless of whether we aggregate data by hour, day or something in between.
- However, if we split the counts into upstream and downstream, the results show closer equality between the first and second half hours, especially if we aggregate counts into 6 hour blocks, or entire days.
  - I don't have a good explanation for why this occurs, but it seems worth utilizing (rather that using total net upstream regressions).
  - This could provide more control over how we account for kelts in the downstream data as well, providing another justification for this approach.
  - One option is to assume the slope is 2, but use the standard error from these regressions to provide an estimate of uncertainty. Another is to use the estimated slope as well. Given that we have no expectation that the first half hour should have higher counts (in either direction) than the second, I'd be inclined to assume a slope of two. That approach will also mean that our estimates don't change as we gather additional data in the future.
- Aggregating date to the 6 hour block or the day seems to make the most sense. The estimates from either of these are consistently higher than aggregating to the hour scale (due to higher regression slopes). However, the choice between the two will have impacts on our estimates.

4.2 Excluding Bull Trout

- Should we assume any fish smaller than 40 cm is a bull trout, since that was the smallest steelhead length recorded? Under the current method, fish caught towards the end of March are considered a steelhead if they are at least 27.9 cm. This does increase the number of small steelhead we're estimating.
- Since the last iteration, I realized I had forgotten to filter the species length data to exclude all hook and line samples, which I presume came from further upstream in the watershed. Excluding this data is meant to avoid any confounding with steelhead encounters and day of year. However, this does reduce the dataset used to fit the GAM. While the general shape of Figure 5 makes biological sense, it does predict most fish will be predicted to be steelhead between March and April, almost regardless of size. This has increased the estimates of small steelhead, but perhaps to a more realistic number.
- The current model includes both the Julian day of capture and the fish length, using a spline for both covariates. The fish length spline turned out to be pretty close to a straight line, with larger fish being more likely to be a steelhead. The Julian day of capture had a peak probability of being a steelhead

occurring in mid- to late-March, and tapering off on either side. Given there were a number of bull trout caught in the very beginning and towards the end of the sampling period, this shape makes sense.

- Incorporating another year of species composition data would make this species probability model more robust, especially for the spline related to the day of capture.

4.3 Missing Data

- Any of the interpolation models we tested (Kalman filter, linear regression or moving average) resulted in larger estimates of steelhead moving upstream (Table 8), but differed in which one provide the biggest increase depending on the time-scale and year.
- The uncertainty (e.g. standard error) grew when incorporating those missing data, which is appropriate. The uncertainty grew substantially in 2020, when there was a large period of missing data due to COVID restrictions.
- Depending on how big the missing data gaps are, and the time-scale we are aggregating data on, there were some years and time-scales with no missing data (e.g. 24 hour scale in 2021). The lack of missing data relies on using the percentage of hours when sonar was operational within each time-step to increase the estimates for any time-steps when the operational time was less than 100%.
  - The alternative to expanding time-steps when the sonar was partially operational is to remove all data from those time-steps and treat them as missing data.
  - Or develop a threshold of operating percent (e.g. 20%?) below which we exclude that data because it's too small a percentage to feel comfortable expanding it, and instead rely on extrapolations from nearby time periods with full data.
  - It's unclear to me whether that would have a substantial impact on the overall estimates or uncertainty.
  - I would recommend the operating threshold approach, maybe using 20% as a threshold. This would mean that for time-periods (6 hour blocks or days) when the sonar was operating for less than 20%, we would mark the data from that time period as `NA` and treat it as missing. Other time periods when the sonar was operating for more than 20% we would take the estimates for that operating time and expand them by the operating percentage. So for example, if we estimated 30 upstream fish during a time period when the sonar was only operating 60% of the time, we would expand by dividing 30 by 60% and end up with an estimate of 50 upstream fish.

5 Decisions to be Made

- Use separate upstream/downstream regressions, or net upstream?
  - *Recommend:* separate upstream/downstream regressions
- What time-scale shall we aggregate data on?
  - *Recommend:* Day
- Should we use estimated slope for expanding 30 min counts, or assume a slope of 2?
  - *Recommend:* Assume a slope of 2
- Should we impose a minimum length to consider a fish a steelhead?
  - *Recommend:* Yes. The smallest observed steelhead length was 40 cm. I propose we use that as a cut off to override the GAM predictions for small fish, ensuring that every fish smaller than this threshold is considered a non-steelhead.
- What to do with data with confidence 2 or 3?
  - *Recommend:* delete / drop it

- What time-series interpolation model should we use?
  - *Recommend:* Kalman filter
- What to do with time-periods when sonar only partially operated?
  - *Recommend:* If sonar operated for less than a threshold (e.g. 20%), treat that data as missing and use time-series interpolation model. For operational percentages greater than this threshold, expand the estimates by that percentage.
- Include 2022 data?
  - *Recommend:* Yes. Write up a report that covers 4 years of data: 2019-2022.
- How to account for kelts?
  - *Recommend:* **???**