Appendix B. Updated Watershed Flows and Water Quality

Canadian Watersheds

Continuous daily flows from the new Canadian watersheds added to the model domain (Figure B1) were estimated either directly from flow gages at or near the mouth of the watershed. If gages were not located at the mouth of the watershed, the flow was scaled to the mouth of the watershed using the ratio of the total watershed area to the gaged area. Similarly, flows for un-gaged watersheds were estimated by scaling the flow of the nearest gaged adjacent watershed of a similar size using the same approach, i.e., an area-weighted scaler.

Flow gages within the British Columbia (BC) study area are maintained by Water Survey of Canada (WSC) (https://wateroffice.ec.gc.ca/search/real_time_e.html). The BC study area consists of watersheds draining to either the Strait of Georgia (SOG) or the Strait of Juan de Fuca (SJF), and contains 139 WSC flow gages. Despite the vast numbers of gages within the BC watersheds, 90% of the gages either did not have any data for years 1999-2016, or were located within tributaries. Gages were selected for a given watershed according to 1) proximity to the mouth of the stream and 2) availability of data. If data gaps were present in flow records for a given watershed, then measures were taken to impute missing values, as described in the next section. Figure B1 shows the location of the final selected gages that we used, and Table B1 lists the names of the rivers/watersheds for which freshwater flows were added to the model.

Figure B1. Map of British Columbia watersheds, including WSC streamflow gages (yellow scatter points), used in the study.

Filling gaps in missing flow data

Continuous flow data in BC watersheds were missing for several of the gaged watersheds. Several methods were used to fill in these gaps, and the method used for a given flow gage depended on availability of nearby gages, the magnitude of data missing, and the range of data available for the gage of interest in a particular situation. These methods included:

- Carrying forward the last non missing value, also known as last observation carried forward (LOCF) – only used in cases where 1-2 days of data were missing.

- Linear or multiple linear regression

    - Gages of interest were fit with all gages that contained data for the missing time period (Figure B2). Heteroscedascity and normality regression assumptions were

checked and log transformation was applied if needed (Gotelli and Ellison 2013). Correlation between predictor variables were assessed using variance inflation factors (VIF) with the R Statistical Software VIF function from the car package. To reduce standard errors of estimates, regressors with VIF values greater than 4 were dropped from the model (Hair et al. 2016). A stepwise process based on Akaike Information Criterion (AIC) values was then used to determine the minimally adequate regression model. Regression equations used to impute missing flow data are shown in Table B1.

- Time series linear interpolation/ Seasonal linear regression.
    - Under circumstances in which regression could not be used (no inter-attribute correlations present) and gaps of data were missing, univariate imputation methods including time series linear or seasonal linear interpolation were used. Univariate imputation methods, unlike their multivariate counterpart, depend entirely on the characteristics of an individual time series. In particular, in order to impute missing values using univariate methods it is important that present values in a time series be related to past values, this is known as autocorrelation (Morritz et al. 2015). Autocorrelation is measured formerly by regressing a time series with a lagged version of itself (current time period vs previous time period). Using plots of autocorrelation over time can be used to identify important characteristics including seasonal patterns or monotonic trends (increasing or decreasing).

      Autocorrelation plots were generated in R Statistical Software to identify any trends or seasonal components for time series. Time series which either lacked seasonality or had data gaps of less than 3 months were linearly interpolated, otherwise seasonal linear interpolation was used (Chandrasekaran et al. 2016). To perform linear or seasonal linear interpolation, the R functions na.interpolation and na.seasplit from the imputeTS package were used respectively.

- Autoregressive Integrated Moving Average (ARIMA) forecasting/hindcasting with Fourier regression
    - ARIMA was used during circumstances in which historical data was missing (hindcasting) or where future data was needed (forecasting). ARIMA is a method of forecasting with widespread use in literature that consists of a combination of:
        - autoregressive models (linear combinations of present values with previous values)
        - integrated differencing (difference of current value with previous values needed to make a time series stationary)
        - moving average models (average of error terms for a given period) (Murat et al. 2018).

      In situations where time series are seasonal the ARIMA model does not adequately predict periodic shifts and therefore, additional regressors are needed. To account for

seasonal variation ARIMA models can be refit as a regression model with a select number of Fourier terms (Murat et al. 2018).

ARIMA forecasts were performed for this study using the R Statistical Software forecast package functions Auto.arima and Fourier. Both the standard ARIMA model and the number of Fourier terms to be fit to the model were selected by minimizing the AIC (Hyndman, 2018). Despite limited literature regarding the use of this method, generally in favor of Seasonal ARIMA (SARIMA), the advantage of this method over SARIMA is that it allows any length for the seasonal period (Murat et al. 2018), and is therefore, beneficial for daily forecasts.

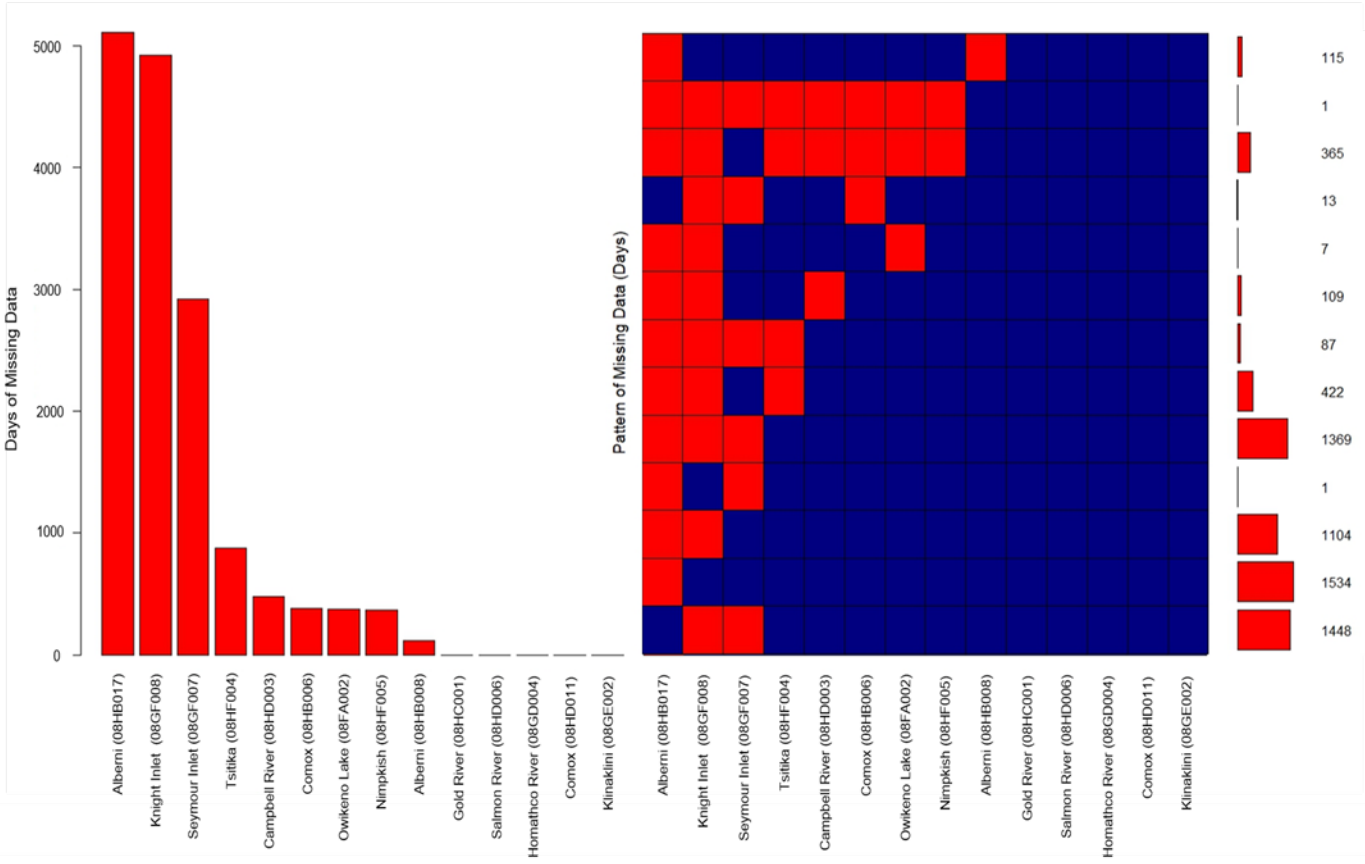

Figure B2. Figure on the left shows the absolute number of days of flow measurements missing from BC study gages for the years 1999-2016, while the figure on the right displays the pattern of data that is missing (red) or present (blue) during this time period. More specifically, the figure on the right provides a map of gages that have data at a particular time interval to impute flow values for gages with missing data. For example, in the first four years (1999-2003) 1448 days of flow data are missing from both Knight Inlet and Seymour Inlet, while from 2003-2007 only Alberni (08HB017) is missing data (1534 days).

Table B1. Flow gage relationships to estimate missing data.

Gage	Flow Gage Relationship	R^2
Alberni (08HB017)	$Q_{08HB017}=10^{(0.506*\log_{10}(Q_{08HB008}) + (0.382*\log_{10}(Q_{08HB006}) +0.663)}$	0.92
Seymour (08GF007)	$Q_{08GF007}=10^{(0.858*\log_{10}(Q_{08FA002}) + (0.362*\log_{10}(Q_{08HC001}) -0.971)}$	0.66
Knight Inlet (08GF008)	$Q_{08GF008}=10^{(0.794*\log_{10}(Q_{08FA002}) + (0.281*\log_{10}(Q_{08GD004}) -0.591)}$	0.8
Alberni (08HB008)	$Q_{08HB008}=10^{(0.861*\log_{10}(Q_{08HD003}) + (0.692*\log_{10}(Q_{08HD006}) -1.39)}$	0.8
Nimpkish (08HF005)	$Q_{08HF005}=10^{(0.9*\log_{10}(Q_{08HC001}) +0.036)}$	0.95
Comox (08HB006)	$Q_{08HB006}=10^{(0.502*\log_{10}(Q_{08HD011}) +1.06)}$	0.69

United States Watersheds

Streamflow entering the Salish Sea from Puget Sound watersheds were originally estimated as described in Mohamedali et al., (2011). This involved estimating flows by scaling flow from the most downstream monitoring stations to the mouth of the river using the ratio of total area weighted precipitation to total gaged area weighted precipitation. Flow data was obtained from The U.S. Geological Survey (USGS), which maintains continuous stream gages on several streams and on most of the large rivers within the Puget Sound area. Permanent USGS gaging stations capture approximately 69% of the watershed tributary to the main study area, which includes all watersheds tributary to Puget Sound (south of Deception Pass).

1. For this effort, the global change in methodology from the analysis described in Mohamedali et al., (2011) to estimate freshwater flows was the source of precipitation data used to calculate the scalers. The original method used an older annual-average precipitation data. We have now updated these scalers using a more recent dataset from Oregon State University's PRISM Climate Group (http://www.prism.oregonstate.edu/). The 30-year annual average precipitation normal from the 1981-2010 time-period were downloaded as raster datasets and intersected with shapefiles of gaged and un-gaged portions of the watersheds to calculate area-weighted precipitation scalers using GIS analysis. These scalers were then used to extrapolate flow to the mouth of each watershed, and estimate flow for un-gaged watersheds as described in Mohamedali et al., (2011). In using this method it was assumed that although annual rainfall will fluctuate with time that precipitation scalers, which are the ratio of total precipitation to precipitation in gaged areas of a given watershed, will be approximately constant. To test this, 30 year normal precipitation data, annual average precipitation data for 2006-2016, and monthly precipitation data for 2006-2016 were obtained from the Oregon State University PRISM Climate group. Analysis revealed negligible differences between 10 year and 30 year annual average scalers and further, showed monthly scalers on average to yield higher RMSE values when compared with observed flow data. Based on these

results it was decided that the use of 30 year normal precipitation data was a good approach. Based on this approach, the time-series flow data for all watersheds is now extended from 1999 through June 2017.

2. Flow for the Green River was originally calculated by adding flow measured at the Green River USGS gage near Auburn (USGS 12113000) to a fraction of the flow from the Sammamish River. We were unable to track the original reason of adding Sammamish River flow to the Green River. After consulted with a King County's lead hydrologist (personal communication with Curtis DeGasperi, November 6, 2017), we decided to updated the Green River flow so that it is based solely on flow measured at Auburn. This flow is then scaled by precipitation and watershed area to represent flow at the mouth of the Green River.

3. For Skokomish River it was found that the flow at the USGS gage at Potlach (USGS 12061500, see Figure 5) was not reported for river stage in excess of 16.5 feet since it was outside the range of the rating curve. So, an alternate method was used for these missing data by summing the gage flow at the North (USGS 12059500) and South (USGS 12060500) forks and adding to that South fork flows scaled by a ratio of ungagged area (below the gages of North and South forks) to the area of South fork as per the following equation (personal communication Mark Mastin, USGS, December 22, 2015):

   $$Q_{Skokomish} = Q_{Skokomish\_NF} + Q_{Skokomish\_SF} + Q_{Skokomish\_SF} * (A_{un\_gaged}/A_{Skokomish\_SF})$$

   Note that the flow gage used for North Fork gage is below Lake Cushman and Lake Kokanee so it excludes any flow diversion from Lake Cushman for power generation by Tacoma Public Utilities (this was discussed in the previous section).

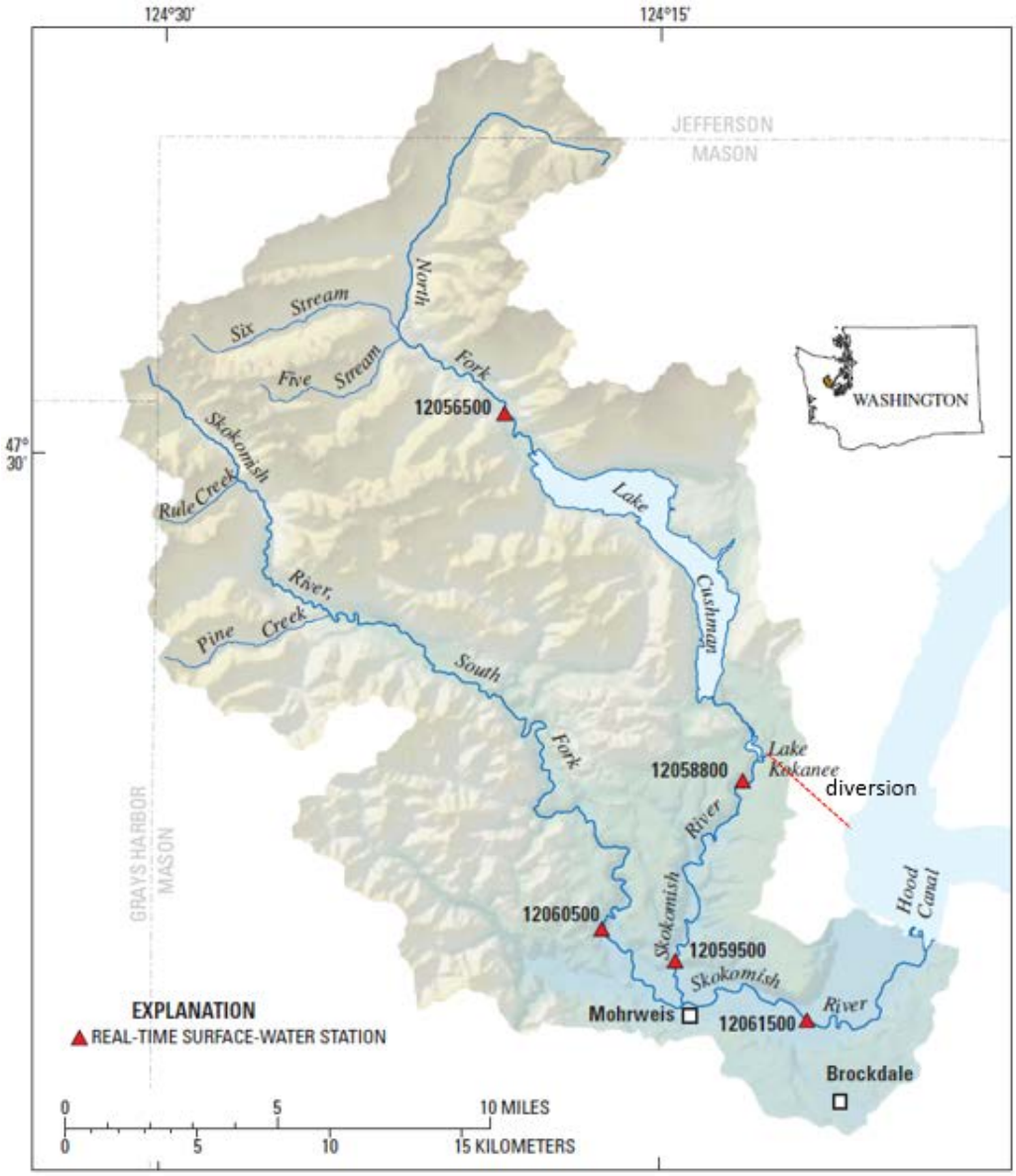

Figure B3. Skokomish River flow gages and diversion for power generation

4. Lake Washington flows at Ballard Locks were previously estimated by Roberts et al. (2008) based on a modified regression equation originally developed by Lincoln (1977). Robert et al. (2008) also recommended that these flows be updated when data became available. Flow data are now available for Ballard Locks from the US Corps of Engineers (USCE), which more accurately represents flows from Lake Washington. However, this data set is for the period of September 2007-present. A comparison of the flows (September 2007 – July 2017) from the previous regression and that obtained from the USCE for Ballard Locks shows that the regression overestimated the flows by about 30% on an average (Figure 6). So for prior years (before 2007 and for missing months in 2007) a factor of 0.7 was applied to flows obtained from regression.

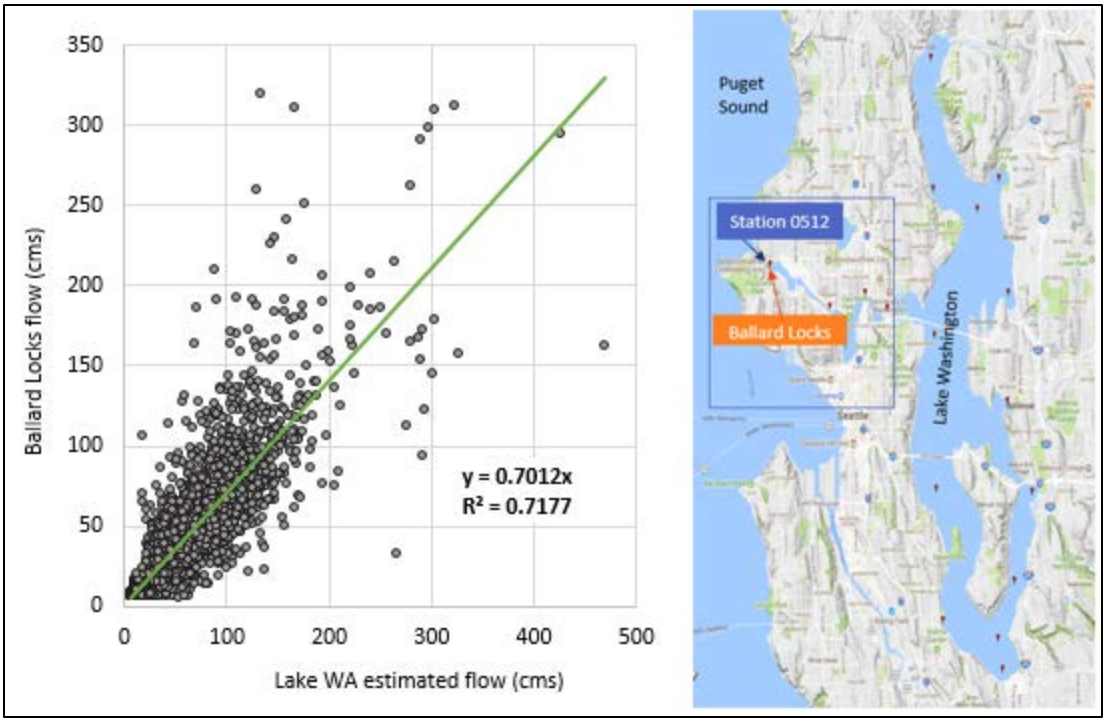

Figure B4. Comparison of USCE flow data at Ballard Locks to estimated flows based on modified regression of Lincoln (1977).

Water Quality data for discharge at Ballard Locks were obtained from King County for station 0512 (Figure 4). Only surface water data were considered as representative of outflow concentrations as per personal communication with Curtis DeGasperi of King County (October 18, 2017). Previous estimates of water quality were based on limited data gathered in the ship canal by Ecology during 1993-94 period (Roberts et al. 2008).

Data are available at this station for ammonia (NH_4) and nitrate (NO_3) on filtered sample. Total persulfate nitrogen (TPN) data is available for unfiltered sample, so it includes algal nitrogen. Therefore, TPN minus dissolved inorganic nitrogen (DIN = NH_4 + NO_3) will give us total organic nitrogen (TON), which is then split equally between particulate and dissolved fractions of organic nitrogen as input to the model. Total organic carbon (TOC) data is available on unfiltered sample so it is inclusive of algal carbon. However, the dissolved organic carbon (DOC) data is available on filtered sample. Therefore Chlorophyll data at station 0512 was used to calculate algal carbon based on carbon to chlorophyll ratio for freshwater algae as reported and used in the Budd Inlet Capitol Lake model (Roberts et al. 2012). Since freshwater algae would die in marine environment and release all the carbon into the particulate and dissolved carbon pool, the fraction of dissolved algal carbon was added to the filtered DOC to give the total DOC for station 0512. The fraction of algal dissolved carbon was obtained from total algal carbon based on fractions used in the Budd Inlet and Capitol Lake model for freshwater algae (Roberts et al. 2012). Daily time-series of data were generated through multiple linear regression relating data to flow and time of the year as outlined in Mohammedali et al (2011).

5. The flow at four new watersheds inflow locations along Washington Coast were obtained from the nearest USGS gaging stations. Willapa River flows were obtained from USGS gage 12013500 near Willapa, WA. Likewise, Chehalis River flows were obtained from USGS gage 12031000 in Porter, WA. Columbia River flows were obtained from USGS gage 14105700 in Dalles, OR. Tidally corrected flows for Willamette River were obtained from USGS gage 14211720 in Portland, OR. These four coastal inflows were included in the hydrodynamic model (Khangaonkar et. al 2018). However, in the water quality model, these inflows were set at ambient marine water quality concentrations through allowance within the modeling software. The water quality of these inflows is unlikely to affect conditions within the Salish Sea, so this simplification was deemed appropriate.

6. Part of the flow from North Fork Skokomish River is diverted to Lake Cushman which feeds into the Cushman No. 1 and 2 power houses on as needed basis. The discharge from the power houses eventually discharges into Hood Canal at the 'great bend' and was previously not included in the model (the model did have the Skokomish River flow, based on monitoring downstream of this diversion, entering Hood Canal further south). Flow data for this discharge was provided to us by Tacoma Public Utilities, which operates the power plant. No water quality data are available for this discharge, therefore the water quality was assumed to be the same as that of Skokomish River.

Other Water Quality Updates

1. Freshwater inputs of dissolved organic carbon (DOC) were previously split 50/50 between labile and refractory fractions. However, there is no universal distinction between labile and refractory DOC. Given the fact that almost 70% of BOD is consumed in the first 5 days during a BOD test for municipal wastewater (Metcalf and Eddy, 1991), this split was updated to 90/10 and will be used moving forward in the SSM.

2. Particulate organic carbon (POC) from freshwater sources were previously split into three categories, based on dissolution rates, fast (decay rate of 0.08 /day), slow (decay rate of 0.02 /day) and refractory (no decay) as described in the South and Central Puget Sound model (Ahmed et al. 2014). In the Salish Sea Model, POC has two components, labile and refractory. The dissolution rates for the two components were 0.03/day and 0.0015/day, respectively. Initially, for the SSM, slow and refractory components were lumped together into the refractory fraction, while the fast fraction was considered labile. However, judging by the dissolution rate it was deemed more appropriate to lump the previous fast and slow fractions used in the South and Central Puget Sound model into the labile POC for the Salish Sea model, and retain the refractory fraction as refractory POC for the SSM.

References

Chandrasekaran, S., Zaefferer, M., Moritz, S., Stork, J., Friese, M., Fischbach, A. and Bartz-Beielstein, T. 2016. *Data Preprocessing: A New Algorithm for Univariate Imputation Designed Specifically for Industrial Needs*. Bibliothek der Technischen Hochschule Köln.

Gotelli, N.J and Ellison, G.N. 2013. A primer of ecological statistics (2nd ed.). *Sinauer, Sunderland, Massachusetts, USA*.

Hyndman, R.J., and Athanasopoulos, G. 2018. *Forecasting: Principles And Practice*. OTexts, pp.112.

Miles, J., and Shevlin, M. 2001. *Applying Regression And Correlation: A Guide For Students And Researchers*. Sage.

Mohamedali, T. M. Roberts, B. Sackmann, and A. Kolosseus. 2011a. Puget Sound dissolved oxygen model nutrient load summary for 1999-2008. Washington Department of Ecology. Olympia, WA. Publication No. 11-03-057. https://fortress.wa.gov/ecy/publications/SummaryPages/1103057.html.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M. and Stork, J., 2015. Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924*.

Murat, M., Malinowska, I., Gos, M. and Krzyszczak, J., 2018. Forecasting daily meteorological time series using ARIMA and regression models. *International Agrophysics 32*(2): 253–264.