**Fall Status and Trends Report Data Set Assembly Methods**

**All Seasons: Net Delta Outflow**

The R code file called [IEP\_Status&Trends\_Flow.R](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/IEP_Status%26Trends_Flow.R) that builds the outflow plots depends on the CSV file [flow\_1929-10-01\_2018-09-30.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/flow_1929-10-01_2018-09-30.csv). The R file and CSV files are on the GitHub repository. The process of creating the CSV files is described below.

There are CSV files for these data located on the DWR webpage for [Dayflow](https://water.ca.gov/Programs/Environmental-Services/Compliance-Monitoring-And-Assessment/Dayflow-Data). The files aren’t particularly consistent in format across years, which makes rounding up the data a hassle. I threw together an R code file that grabs the date and net Delta outflow (i.e., “OUT”, “OUT1”, or “OUT2”) columns, but the code isn’t that great. Jason Dubois has a GitHub [page](https://github.com/jasondubois/rdayflow) where he has worked on this too, although I think something about the DWR page has changed since he wrote his code, so it doesn’t work currently.

**All Seasons: Water Quality**

The R code file called [IEP\_Status&Trends\_WaterQuality.R](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/IEP_Status%26Trends_WaterQuality.R) that builds the water quality plots (i.e., Secchi depth, ammonium, nitrate/nitrite, chlorophyll-a) depends on two CSV files: [WQ\_Discrete\_1975-2017.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/WQ_Discrete_1975-2017.csv), [wq\_stations.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/wq_stations.csv). The R file and two CSV files are on the GitHub repository. The process of creating the two CSV files is described below.

All of the EMP water quality monitoring data are stored on [Water Data Library](http://wdl.water.ca.gov/waterdatalibrary/waterquality/index.cfm) (WDL), but only a subset of those data are currently publicly available. I had to log into WDL with my DWR credentials and follow these steps to get the complete data set:

WQ Admin > Internal Water Quality Reports : Water Quality Report by Project >

EMP Monitoring: January 1975 to December 1995

EMP Discrete Water Quality: January 1996 to December 2017

The data are stored in two main chunks, and I also downloaded them in two chunks. I only downloaded data for the parameters that I thought we would need: all the field measurements plus Nitrate/Nitrite, Chlorophyll-a, and Ammonia. The data I got from WDL were in worksheets (.xlsx). Before bringing the data into R, I copied the same subset of columns from each of these two worksheets into a single CSV file called [WQ\_Discrete\_1975-2017.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/WQ_Discrete_1975-2017.csv), which is on the GitHub repository. Below are the names of the worksheet columns I copied and corresponding CSV file names I used:

Collection Date = date

Station Name = name

Station Number = id

Analyte = parameter

Result = value

Units = units

Any other manipulations to this data frame I made in the R code file [IEP\_Status&Trends\_WaterQuality.R](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/IEP_Status%26Trends_WaterQuality.R).

I paired the water quality data with the GPS coordinates of the stations using this excel file, called [Discrete\_Stations.xlsx](https://emp.baydeltalive.com/projects/11285/page), which is available on the Data and Metadata [page](https://emp.baydeltalive.com/projects/11285/page) for Water Quality in the San Francisco Estuary on the California Estuaries Portal. I needed these station location data, so I could group the data by region (i.e., Delta, Suisun, San Pablo). Annoyingly, there are no perfectly shared columns between the water quality data and the station location data. The water quality data set has columns for Station Name and Station Number, while the station location data set has a column for Site\_Code, which is an abbreviation of the Station Name. Ultimately, I created a CSV file called [wq\_stations.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/wq_stations.csv), which is on the GitHub repository. This CSV file has four columns. Three of these are from the ‘Discrete\_Stations.xlsx’ file: Site\_Code (site), Latitude (lat), and Longitude (long). The last column is the Station Number (id) from the ‘WQ\_Discrete\_1975-2017.csv’, which I matched to the site codes manually.

**All Seasons: Zooplankton**

The R code file called [IEP\_Status&Trends\_Zooplankton.R](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/IEP_Status%26Trends_Zooplankton.R) that builds the zooplankton plots depends on six CSV files:

* [zoop\_cb.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/zoop_cb.csv)
* [zoop\_pump.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/zoop_pump.csv)
* [zoop\_individual\_mass.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/zoop_individual_mass.csv)
* [zoop\_mysid.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/zoop_mysid.csv)
* [zoop\_mysid\_mass.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/zoop_mysid_mass.csv)
* [zoop\_stations.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/zoop_stations.csv)

The R file and six CSV files are on the GitHub repository. The process of creating the CSV files is described below.

Some of the data are located on CDFW’s zooplankton FTP [site](ftp://ftp.dfg.ca.gov/IEP_Zooplankton/). On this site, there is an XLSX file for each of the three types of survey gear. All three sets of data are collected together on the monthly EMP water quality run.

The pump samples small zooplankton like rotifers: [1972-2018Pump Matrix.xlsx](ftp://ftp.dfg.ca.gov/IEP_Zooplankton/1972-2018Pump%20Matrix.xlsx)

The Clark-Bumpus net samples mid-sized zooplankton like cladocerans: [1972-2018CBMatrix.xlsx](ftp://ftp.dfg.ca.gov/IEP_Zooplankton/1972-2018CBMatrix.xlsx)

The mysid net samples large zooplankton, namely mysid shrimp: [1972-2018MysidMatrix.xlsx](ftp://ftp.dfg.ca.gov/IEP_Zooplankton/1972-2018MysidMatrix.xlsx)

For our plots, we need biomass estimates for all zooplankton taxa that make up the four key groups: mysids, cladocerans, cyclopoid copepods, calanoid copepods.

The ‘1972-2018MysidMatrix.xlsx’ file for the mysid net includes both counts (CPUE) and estimated biomass (BPUE). The files ‘zoop\_mysid.csv’ and ‘zoop\_mysid\_mass.csv’ are simply copies of the ‘MysidCPUEMatrix1972-2018’ and ‘MysidBPUEMatrix1972-2018’ tabs, respectively. For the final plots, only the BPUE data are used, but it is nice to have the CPUE data as well.

The files for the other two gear types contain CPUE only. The file ‘zoop\_cb.csv’ is a copy of the tab ‘CB CPUE Matrix 1972-2018’ from the ‘1972-2018CBMatrix.xlsx’ file, and similarly, the file ‘zoop\_pump.csv’ is a copy of the tab ‘Pump CPUE Matrix 1972-2018’ from the ‘1972-2018Pump Matrix.xlsx’ file.

I got a file directly from April Hennessy with estimated individual biomass values for copepods and cladocerans. I used this file to generate the file ‘zoop\_individual\_mass.csv’. These individual masses were then used to estimate BPUE from the CPUE data for both the pump and CB net.

Each of the XLSX files has a ‘Station Lookup’ tab, which I think is pretty much identical across XLSX files. The file ‘zoop\_stations.csv’ is a copy of this tab, though I simplified the column header names a bit and excluded column ‘Location’ which describes the location of each station.

**Fall Season: Fishes**

The R code file called [IEP\_Status&Trends\_Fishes\_Fall.R](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/IEP_Status%26Trends_Fishes_Fall.R) that builds the fish plots (i.e., Delta smelt, longfin smelt, striped bass, American shad, white sturgeon, fall-run Chinook) depends on three CSV files: [fmwt.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/fmwt.csv), [wst.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/wst.csv), [frch.csv](https://github.com/InteragencyEcologicalProgram/Status-and-Trends/blob/master/frch.csv). The R file and three CSV files are on the GitHub repository.

I received the files ‘fmwt.csv’ (Delta smelt, longfin smelt, striped bass, American shad) and ‘wst.csv’ (white sturgeon) from CDFW pretty much as they are now.

Louise got the fall-run chinook salmon data from [GrandTab](https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=84381&inline). It looks like she grabbed data from CHINOOK SALMON ESCAPEMENT – FALL RUN Summary I (pp 10-11).