Appendix F, Modeling

Attachment X Delta Smelt Life Cycle Model with Entrainment

1.1 Delta Smelt

1.1.1 Delta Smelt LCME

Polansky et al. (2021) developed a hierarchical stage-structured state-space life cycle model for Delta Smelt to identify factors with the strongest statistical support for having influence on the species' recruitment and survival. This modeling approach is useful as an ecological modeling tool because it can separate descriptions of state and observation processes and permit the integration of disparate data sets. This Delta Smelt life cycle model was later expanded from four to seven life stages with a component that separately describes the entrainment process at the Delta export facilities (Smith et al. 2021). This model produces expected values for larval recruitment and survival at the subsequent life stages. The most statistically supported model variant in Smith et al. (2021) used means of December-June Old and Middle River (OMR) values and June-August outflow aggregated from monthly values or longer timescales; therefore, CalSim output for the scenarios/alternatives can be directly incorporated into the model framework. The most statistically supported model in Smith et al. (2021) also included food/prey metric term during the months of January to March. By using the relationship between zooplankton density and salinity, CalSim-predicted X2 values were then used to estimate the expected change in the food/prey metric for January-March months across alternatives. Reclamation used this model to calculate expected annual population growth rate (λ; the abundance of current year divided by abundance from previous year) for alternative flow scenarios by using CalSim output and subsequent zooplankton model. The metric of interest will be geometric mean of λ for a specified time frame (e.g., 1995-2015), which will be compared across alternatives. For the purpose of this text, Smith et al.'s (2021) model will be referred to as the Delta Smelt Life Cycle Model with Entrainment (LCME).

1.1.1.1 Methods, assumptions

The Delta Smelt LCME was run based on flow inputs from CalSim 3. The approach followed the Collaborative Science and Adaptive Management Program (CSAMP) Delta Smelt Structured Decision Making (SDM) process, where historical years (1995-2015) were adjusted according to a CalSim 3 scenario and the geometric mean λ was calculated for each scenario. There is an expectation that zooplankton abundance (i.e., prey item for Delta Smelt) would change based on flow (Kimmerer and Rose 2018), and as such, a zooplankton submodel constructed for the Delta Coordination Group and CSAMP Delta Smelt SDM was applied to the CalSim 3 scenarios. For the zooplankton term, an upper and lower 95% confidence intervals were calculated and applied into the analysis to better understand sensitivity of the model output to variation in zooplankton abundance. We did not update any other model inputs (turbidity, temperature, and predators) due to the complexity and lack of predictive models associated with the other values. Furthermore, it is unclear whether flow changes at a project operations scale meaningfully affect the functioning of the Bay-Delta food web. What is of interest in this analysis is to determine how much the expected long-term abundance of delta smelt might change based on the proposed changes in water management.

Main model

Monthly flow data were pulled from CalSim 3 dss files through R and were summed or averaged depending on the variables for the LCME. Old and Middle river flow variables were either extracted directly from CalSim 3 as monthly average value in cfs, or averaged if the timespan covers two months (Table 1). Sum of Delta outflow from June to August were calculated by multiplying the CalSim 3 predicted monthly Net Delta Outflow Index (NDOI-ADD + NDOI-MIN) by the number of days for each month and then added together. The total values in cfs per day were then converted to acre-feet (1 cfs = 1.983 acre-feet per day). Methods and findings of the original application of LCME can be found in Smith et al. (2021). The list of LCME flow variables that were acquired from the CalSim 3 runs can be found in Table 1. R script and data used for the model can be found at (https://github.com/BDO-Science/DeltaSmelt_LCM).

Table 1. List of covariates used in LCME that were replaced with values from CalSim 3 for each alternative. The Old and Middle River covariates imply entrainment as the mortality mechanism (Smith 2019; Smith et al. 2020). The Delta outflow covariate implies foraging habitat suitability as a suite of mechanisms that align better when outflow is elevated (Smith and Nobriga 2023). The covariates are listed in the order they affect a given cohort in the model.

Life stage	Covariate	Unit	Covariate summary details
Early post-larval (May)	April-May Old and Middle river flow	cfs	Mean of the daily sum of tidally filtered flows in the Old and Middle rivers during April to May
Late post-larval (June)	June-August Outflow	Acre-foot	Sum of the volume of water moving past a point near the confluence of the Sacramento and San Joaquin rivers, near Pittsburg, California, during June to August
Late post-larval (June)	June Old and Middle river flow	cfs	Mean of the daily sum of tidally filtered flows in the Old and Middle rivers during June
Early subadult (October-November)	December-January Old and Middle river flow	cfs	Mean of the daily sum of tidally filtered flows in the Old and Middle rivers during December to January
Late subadult (January-February)	February Old and Middle river flow	cfs	Mean of the daily sum of tidally filtered flows in the Old and Middle rivers during February
Early adult (March)	March Old and Middle river flow	cfs	Mean of the daily sum of tidally filtered flows in the Old and Middle rivers during March

Assumptions related to the model calibration and new flow inputs

- The LCME was parameterized using Old and Middle River (OMR) flow values derived from the USGS gages and Delta Outflow estimates from DAYFLOW (https://data.cnra.ca.gov/dataset/dayflow), which may differ to some extent with how CalSim 3 calculates these values (OMR and Delta Outflow).

- The LCME separately accounts for the influence of OMR and turbidity on delta smelt entrainment. However, the CalSim 3 runs had assumptions built into them about how frequently turbidity triggers that affect OMR would occur. This confounds the turbidity effect on entrainment with the OMR effect in a way that the LCME cannot account for. This may lead to a negative bias in the predicted effect of entrainment; in other words, it may be underestimated somewhat.

- The only flow data included in the published LCME (Smith et al. 2021) are OMR and June-August Delta Outflow. In essence, the LCME assumes that these are the most influential flow variables associated with Delta Smelt recruitment and survival. This assumption was supported by Polansky et al. (2021), which is why these flow variables were carried forward and re-tested in the Smith et al. (2021) model.

- This analysis consisted of the years 1995 to 2015, so it is unclear how representative model predictions of Delta Smelt population trajectory will be when simulating scenarios that include environmental conditions outside the range of observations the model was fit to. In addition, it is unclear how model parameter estimates and predictions of Delta Smelt population may be affected by climate change impacts and the ongoing and proposed supplementation efforts.

Zooplankton model

To calculate zooplankton abundance/density changes related to changes in flow associated with the CalSim 3 scenarios, Reclamation leveraged the zooplankton abundance estimation process used in the CSAMP Delta Smelt SDM group. To replicate the zooplankton abundance calculation used in the CSAMP Delta Smelt SDM process, estimated X2 values for each month were first retrieved from CalSim 3 dss files. These monthly X2 values were then converted into salinity values for each region defined in the Delta Smelt Individual-Based Model (IBM) (Rose et al. 2013) using a generalized linear model developed by Compass (see Attachment 1).

Similar to the CSAMP Delta Smelt SDM process, Generalized Additive Models (GAMs) were constructed to predict the zooplankton density Delta Smelt were expected to spatially overlap with given a salinity level for each IBM region and zooplankton taxon (see Attachment 2). Predictor variables for each GAM were the tensor product smooth of the interaction between salinity and day of year, as well as random effects for year and station (when more than one station exists in the dataset). To produce the monthly model output, the 15th of each month was used as the data input.

Once salinity values were calculated for each Delta Smelt IBM region, month, and scenario, expected zooplankton densities were then estimated for every zooplankton taxon, month, region, and scenario using output from the GAMs. Upper and lower 95% confidence intervals from these predictions were calculated through 1,000 independent draws from the model distribution, similar to a bootstrapping process. Just as was done in the CSAMP Delta Smelt SDM process, for each alternative, the initial output was scalar values of the taxa-specific zooplankton density under the particular management conditions divided by the same prediction under baseline/historical conditions. However, because 0 values were present in the baseline, it resulted in infinite values for the scalar calculations. These infinite values were replaced with the maximum finite scalar calculated from model predictions for a specific alternative, taxon, region, and month (across years). When this step still yielded no finite scalar value, the maximum finite scalar value from a given alternative, taxon, and month was used instead.

Because the Delta Smelt IBM and LCME differ in how regions are defined and how zooplankton taxa are grouped, additional conversions were needed. The Delta will continue to be managed as a freshwater ecosystem (i.e., not expected to vary much in terms of salinity) in the near future, and as such, any IBM regions upstream of the Confluence were ignored, and likewise LCME North and South regions were left as is (Figure 1). The Far West LCME region only overlapped with the SW Suisun IBM region and thus, the SW Suisun IBM region results were used to define zooplankton changes in the Far West LCME region. To calculate zooplankton changes in the West LCME region, the following IBM regions were used: NW Suisun, NE Suisun, SE Suisun, Suisun Marsh, and the Confluence. Results from the five IBM regions within the West LCME region were aggregated by multiplying each IBM region’s value with the proportion of the region’s water volume relative to the total water volume across all five regions. The calculations were as follows:

- Far West LCME region: SW Suisun IBM region (Figure 1)
- West LCME region: (Confluence IBM region x 0.233) + (Suisun Marsh IBM region x 0.174) + (NE Suisun IBM region x 0.110) + (SE Suisun IBM region x 0.220) + (NW Suisun IBM region x 0.264)

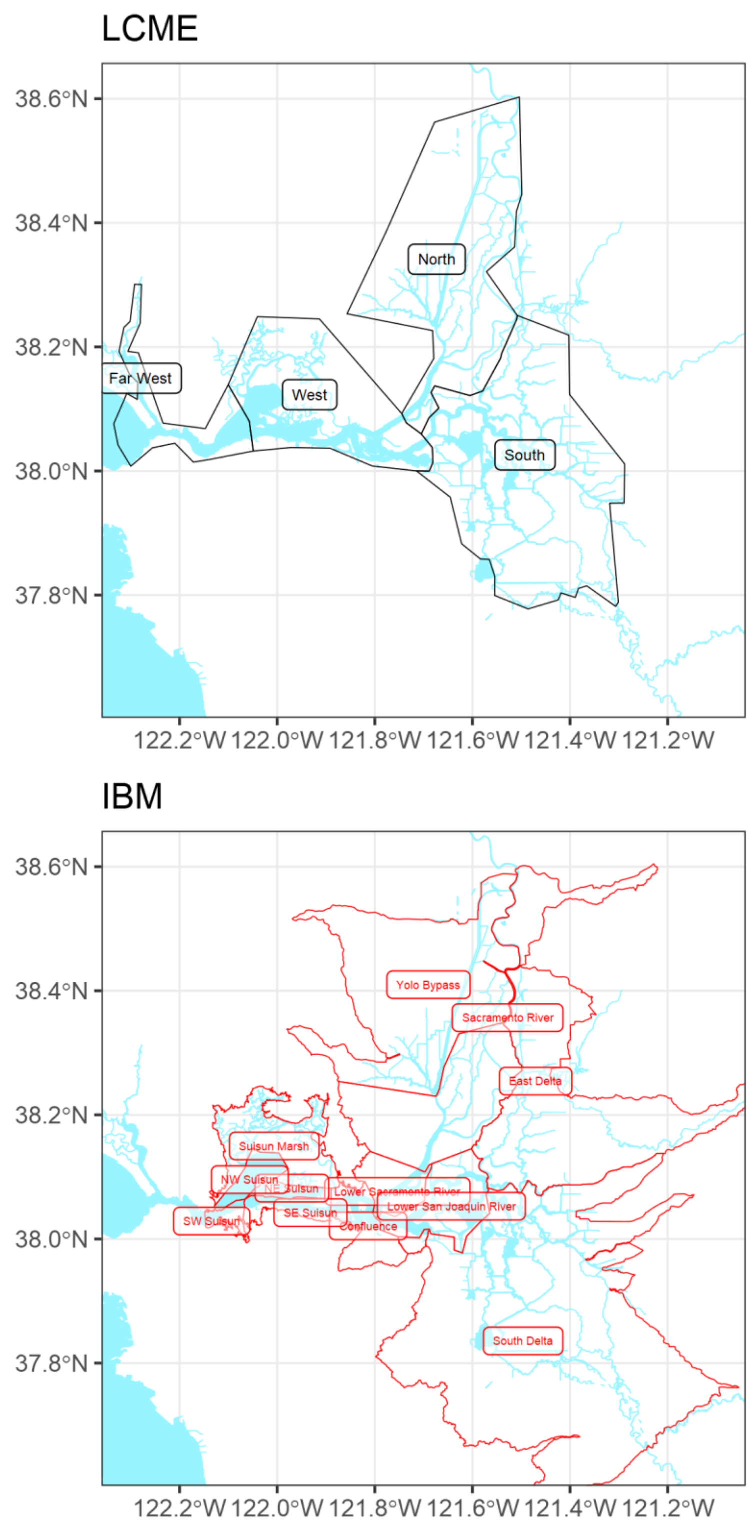


Figure 1. Map of the San Francisco Bay-Delta with LCME regions shown in black (top) and IBM regions shown in red (bottom).

The LCME uses aggregate zooplankton biomass per volume values calculated by summing a number of different zooplankton species and life stages (see Attachment 3), whereas IBM taxa were more specific, often down to species. Therefore, the proportion of each zooplankton taxa that make up the aggregate zooplankton groups in the LCME data input had to be first estimated for each month and LCME region using raw data provided by the primary authors of the LCME ("ZooMysid_74_19_df.csv"). Using these proportions, the final scalar multiplier values were acquired for the Far West and West LCME regions and zooplankton aggregate groups. In other words, the multiplier scalar values were applied based on the proportion of the particular taxon that make up the prey biomass for a given month and LCME region. For example, if *Pseudodiaptomus forbesi* adults are expected to be twice as abundant and *Eurytemora affinis* adults are expected to be three times as abundant under an alternative, and the two species make up 50% of the biomass each, the final multiplier scalar values will be 2.5 (i.e., [2 x 0.5] + [3 x 0.5]).

These final scalar multipliers were then applied to the LCME aggregated zooplankton dataset ("ZooMysid_74_19_df_median.csv") for the median estimate and the lower and upper 95% confidence interval values (Figure 2). These predictions were then capped at the maximum value that was observed in the LCME aggregated zooplankton dataset ("ZooMysid_74_19_df_median.csv") for the region and month using only data from 1995 to 2019. Lastly, the prey covariates (see Table 3) were acquired by calculating the mean across the four LCME regions.

Table 2. List of taxa analyzed using GAM and the equivalent LCME taxa used to calculate the proportion of each taxon that make up the prey biomass at a given month and LCME region.

GAM response variable	**Taxon definition**	**LCME taxon used to calculate proportion of prey biomass for each month and LCME region**
acartela	*Acartiella sinensis* (copepod) adults	*Acartiella sinensis* (copepod) adults
eurytem	*Eurytemora affinis* (copepod) adults	*Eurytemora affinis* (copepod) adults
pdiapfor	*Pseudodiaptomus forbesi* (copepod) adults	*Pseudodiaptomus forbesi* (copepod) adults
othcalad	Other calanoid copepod adults	Other calanoid adults + *Sinocalanus doerrii* (copepod) adults
othcaljuv	Other calanoid copepodites	Calanoid copepodids + Other calanoid copepodids + *Eurytemora affinis* copepodids + *Sinocalanus doerrii* copepodids + *Pseudodiaptomus* spp. Copepodids + *Acartiella sinensis* copepodids + *Acartia* spp. Copepodids + Diaptomidae copepodids + *Tortanus* spp. copepodids
limno	*Limnoithona* spp. copepods (all stages)	*Limnoithona* spp. + *Limnoithona sinensis* + *Limnoithona tetraspina*
othcyc	Other cyclopoid copepods (all stages)	*Acanthocyclops vernalis*
allcopnaup	Copepod nauplii (all spp.)	Copepod nauplii + Other copepod nauplii + *Eurytemora affinis* nauplii + *Sinocalanus doerrii* nauplii + *Pseudodiaptomus* spp. nauplii
daphnia	*Daphnia* spp. (cladocerans)	*Daphnia* spp. (cladocerans)
othclad	Other cladocerans	*Bosmina longirostris* + *Diaphanosoma* spp. + Other cladocera
other	All other taxa	N/A (model was not used)
mysid	*Hyperacanthomysis longirostris*	*Hyperacanthomysis longirostris* + *Neomysis mercedis*

As zooplankton covariates for natural mortality were only supported for adult life stages (Smith et al. 2021), only zooplankton modeling results from the months of February and March were used as data input for the LCME (Table 3). In other words, a flow effect on delta smelt's food supply is only supported statistically in February-March. The most parsimonious mechanistic explanation is that prey available to adult fish early in the spawning season had a population-scale effect, perhaps by affecting how many eggs could be produced or affecting how many adults survived to spawn a second time. R script and data used for the salinity and zooplankton models can be found at https://github.com/BDO-Science/DeltaSmelt_LCM.

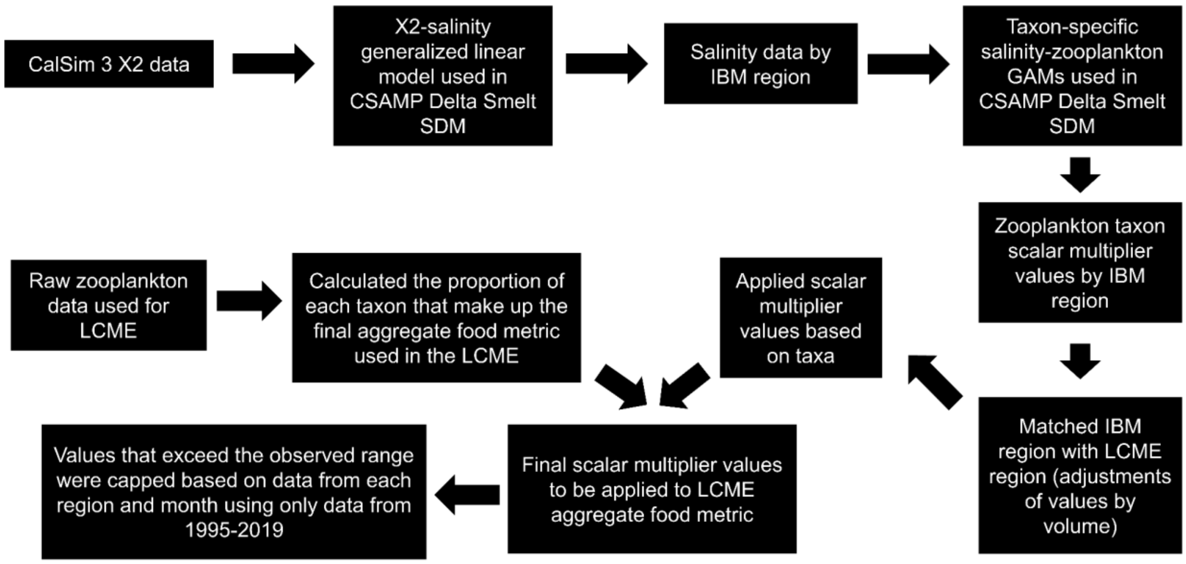


Figure 2. Summary of steps taken to generate estimates of the zooplankton prey density metric for each alternative.

Table 3. List of covariates used in LCME that were replaced with new values based on CalSim 3 and zooplankton model for each alternative. The mechanism implied by these prey density covariates is related to food limitation of adult spawners that may affect the number or quality of eggs produced or the number of repeat spawns the fish are able to complete before dying.

Life stage	Covariate	Unit	Covariate summary details
Late subadult (January-February)	Food metric for February	Microgram carbon per meter^3	Mean carbon-weighted density of adult calanoid copepods, cyclopoid copepods, cladocerans, and mysid shrimp observed during February zooplankton surveys
Early adult (March)	Food metric for March	Microgram carbon per meter^3	Mean carbon-weighted density of adult calanoid copepods, cyclopoid copepods, cladocerans, and mysid shrimp observed during March zooplankton surveys

Assumptions related to the model calibration and new flow inputs

- The zooplankton modeling workflow used salinity to estimate changes in zooplankton biomass related to flow. There are several mechanisms by which a correlation between flow and zooplankton biomass may arise that are not based on salinity *per se* such as transport from upstream, estuarine circulation, etc.
- The use of salinity as a covariate also meant that predicted zooplankton biomass at a particular region is static anywhere and everywhere salinity is ≤0.1 ppt salinity, even with additional Delta outflow.
- The original purpose of the salinity and zooplankton modeling was to adjust zooplankton data input for the Delta Smelt IBM (Rose et al. 2013; Kimmerer and Rose 2018). As such, there were limitations when the data were converted for the purpose of Delta Smelt LCME (e.g., some missing species and/or life stages in the aggregate LCME zooplankton groups).

1.1.1.2 Results

The general statistical prohibition against extrapolation suggests that model predictions are more uncertain when explanatory variables are outside the range of observations to which the model was fit. To visually inspect when the predicted flows and food were outside the observed/empirical range for the LCME, output from CalSim 3 and the zooplankton model were plotted against the empirical data (i.e., data used to estimate parameters in the LCME). See Figures 3-5 below. Most CalSim-predicted flows and zooplankton predictions were not outside the range of observations to which LCME was fit, but some alternatives did include out-of-range values. EXP1 included much lower June-August Delta Outflows than observed and higher (more positive) OMR values than observed in some years. EXP3 OMR values were similar to EXP1, but EXP3 June-August Delta Outflows were within the observed range. Alt1, the components of Alt2, and Alt4 contained some April-May OMR values that were more negative than the observed range. Alt1 also contained OMR values more negative than the observed data for the months of December-January and March. Overall, CalSim-predicted June-August Outflow values were generally lower than the DAYFLOW estimates under Wet or Above Normal years (Figures 3, 6). Predicted prey biomass for all alternatives was within the observed range (Figure 5). However, for certain years higher prey biomass than the empirical data were predicted for all alternatives (Figure 3). As a result, mean predicted prey biomass across all alternatives were also higher than the observed data (Figures 5-7).

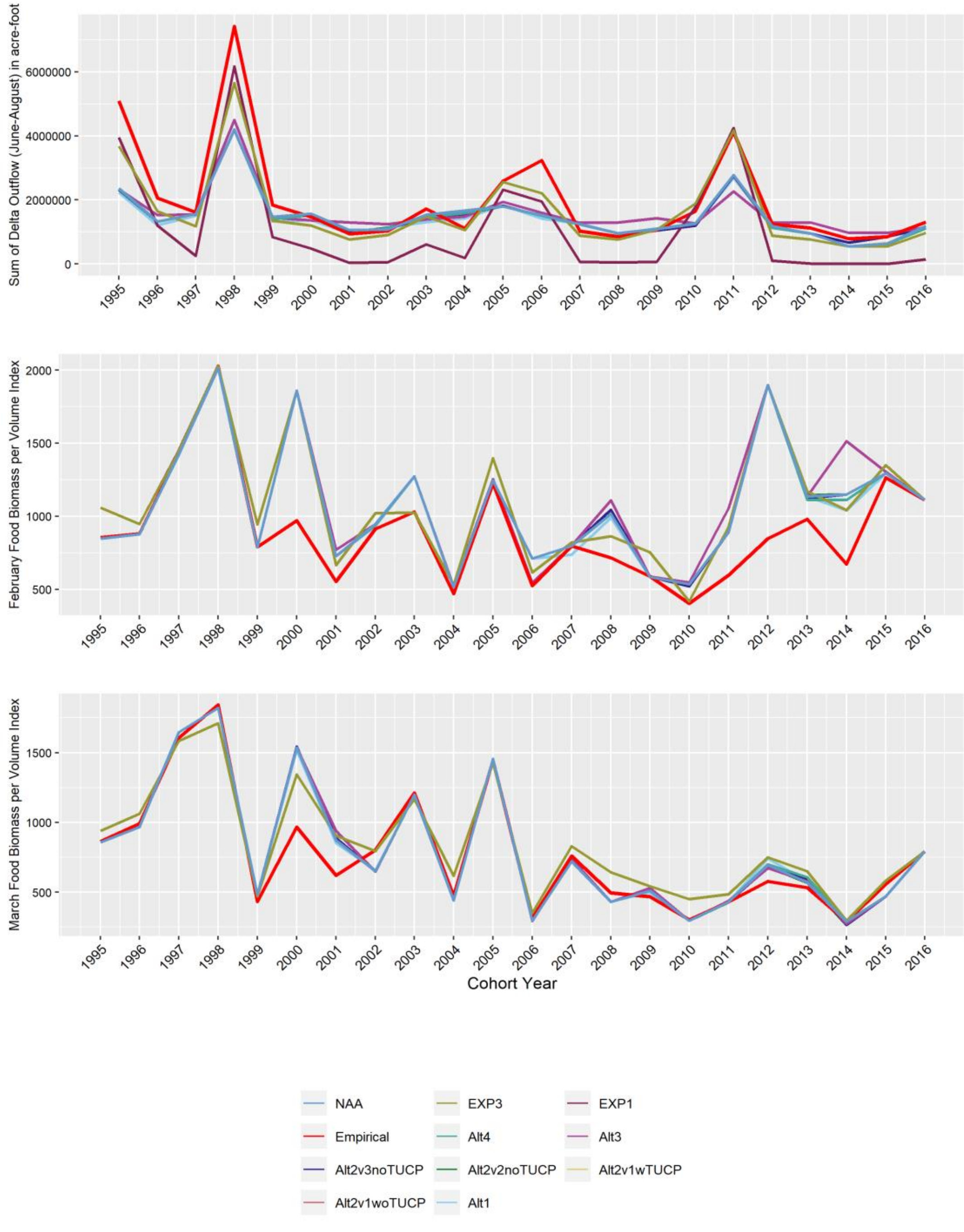


Figure 3. Annual time series of outflow and prey metric data based on CalSim3 data and salinity-zooplankton model relative to the original dataset used to build the Delta Smelt LCME (labeled as "Empirical"). From top to bottom: June-August sum of Delta outflow, February, and March prey metric (biomass per volume) data composed of adult copepods, cladocerans, and mysids. Note that the x-axis represents Delta Smelt cohort year (e.g., February and March prey metric for cohort year 2012 represents data for February and March of 2013).

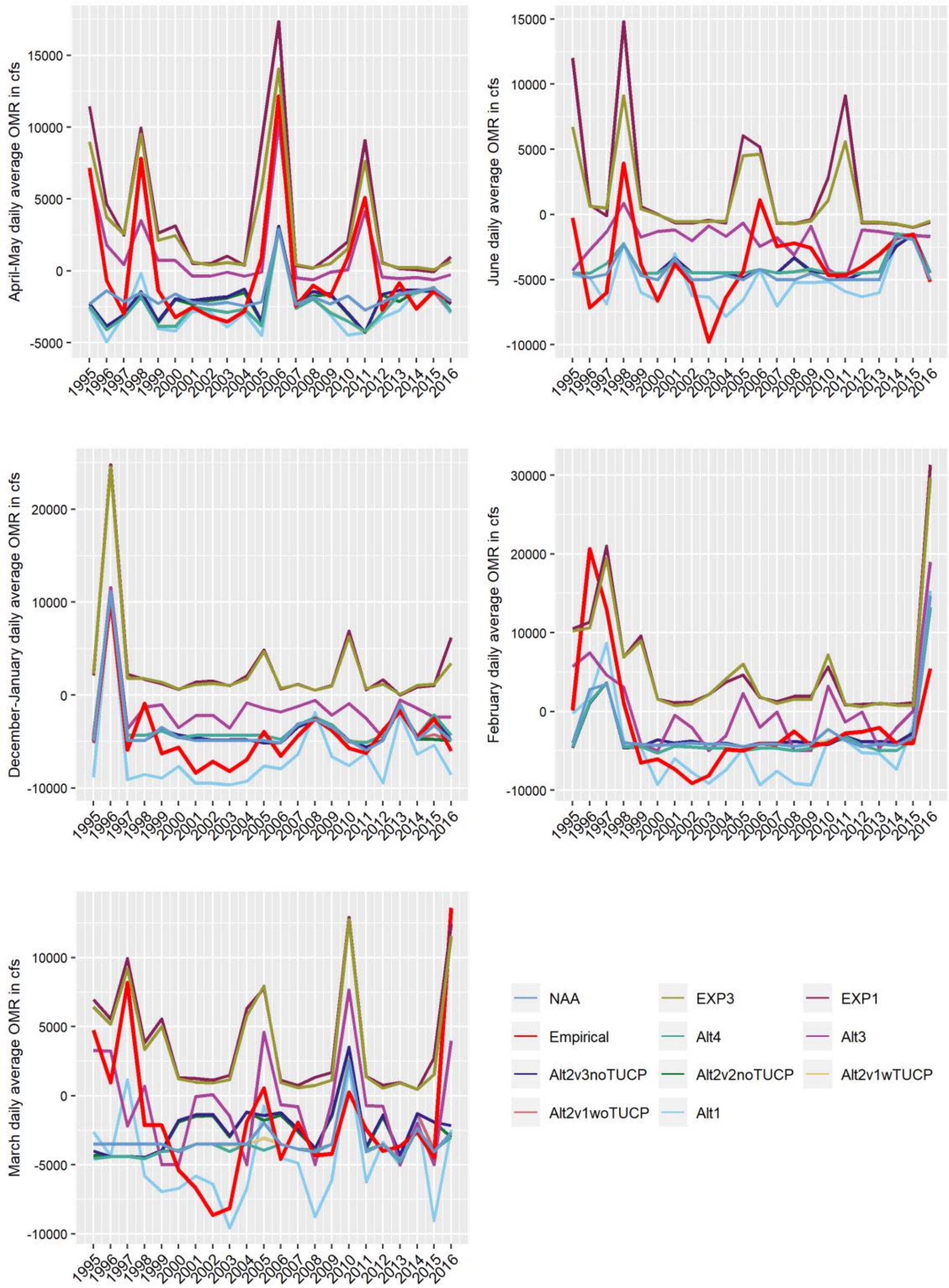


Figure 4. Annual time series of monthly average OMR flow data for input to the LCME produced from CalSim3 relative to the original LCME dataset (labeled as “Empirical”). (e.g., February and March OMR values for cohort year 2012 represents data for February and March of 2013).

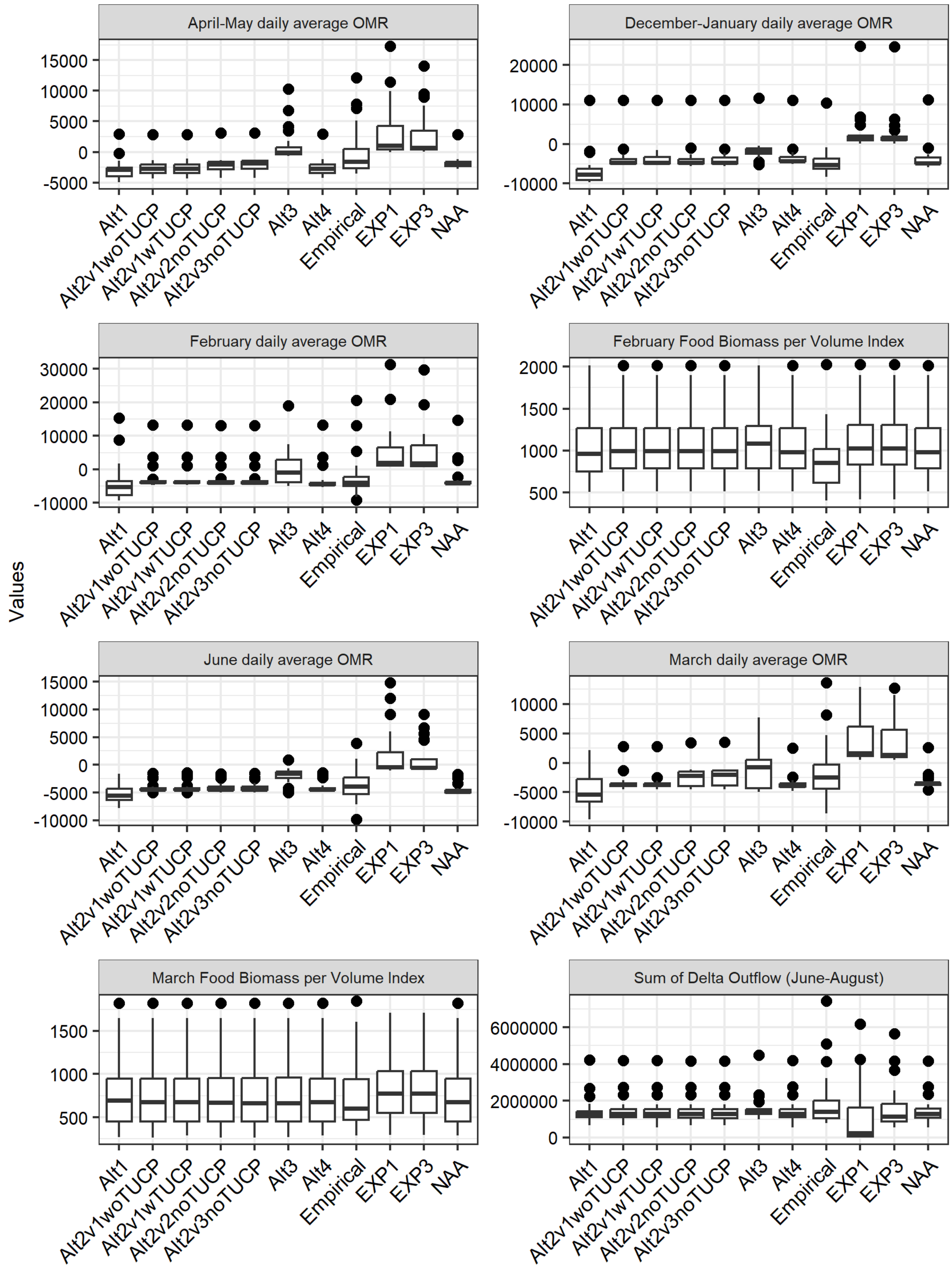


Figure 5. Box plot of covariate values for cohort year 1995 to 2015 sorted by alternative.

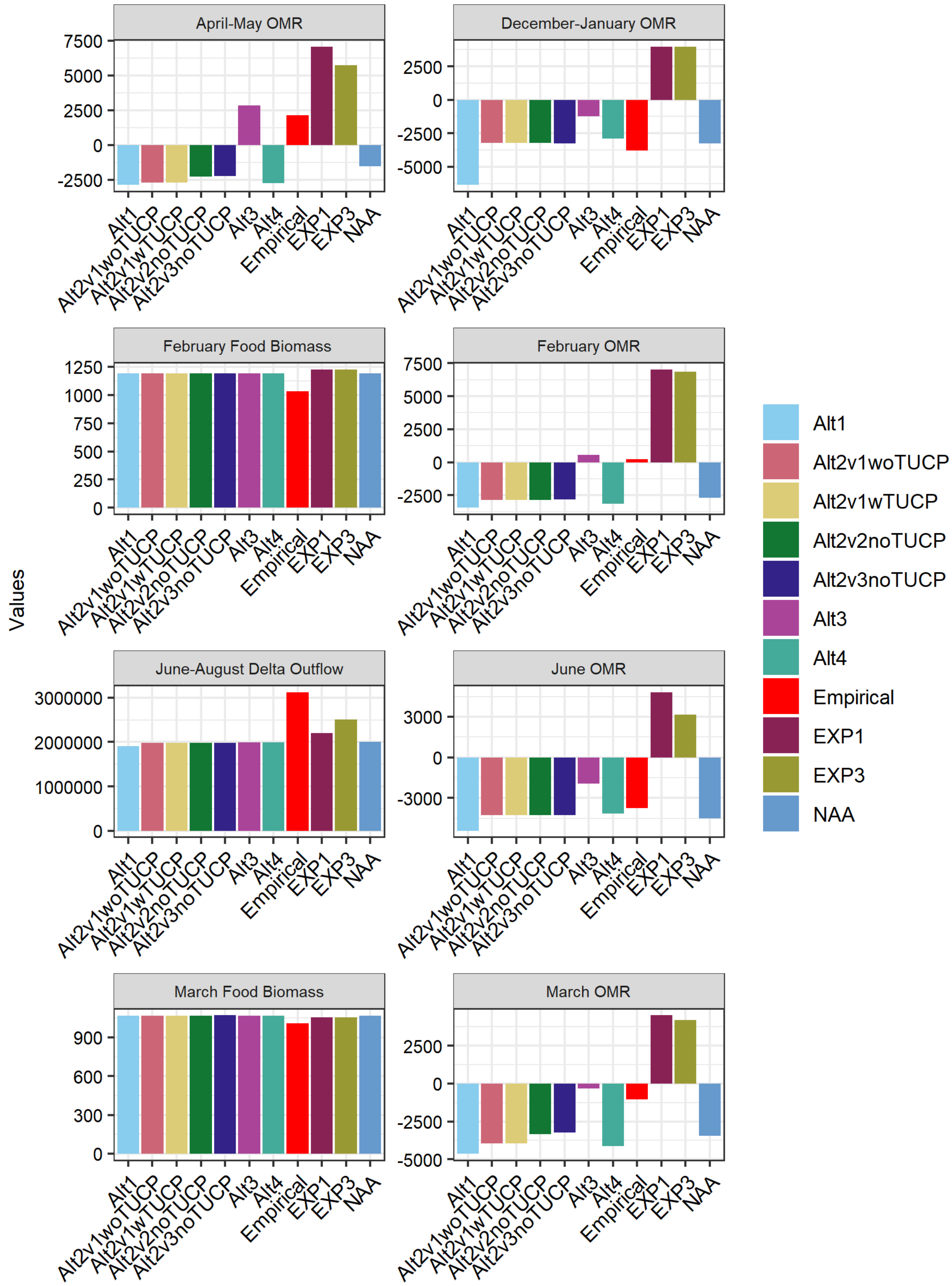


Figure 6. Mean covariate values used in the LCME for Wet and Above Normal year types. Note that cohort year was matched with the water year that the cohort was born in (e.g., cohort year 1995 = water year 1995).

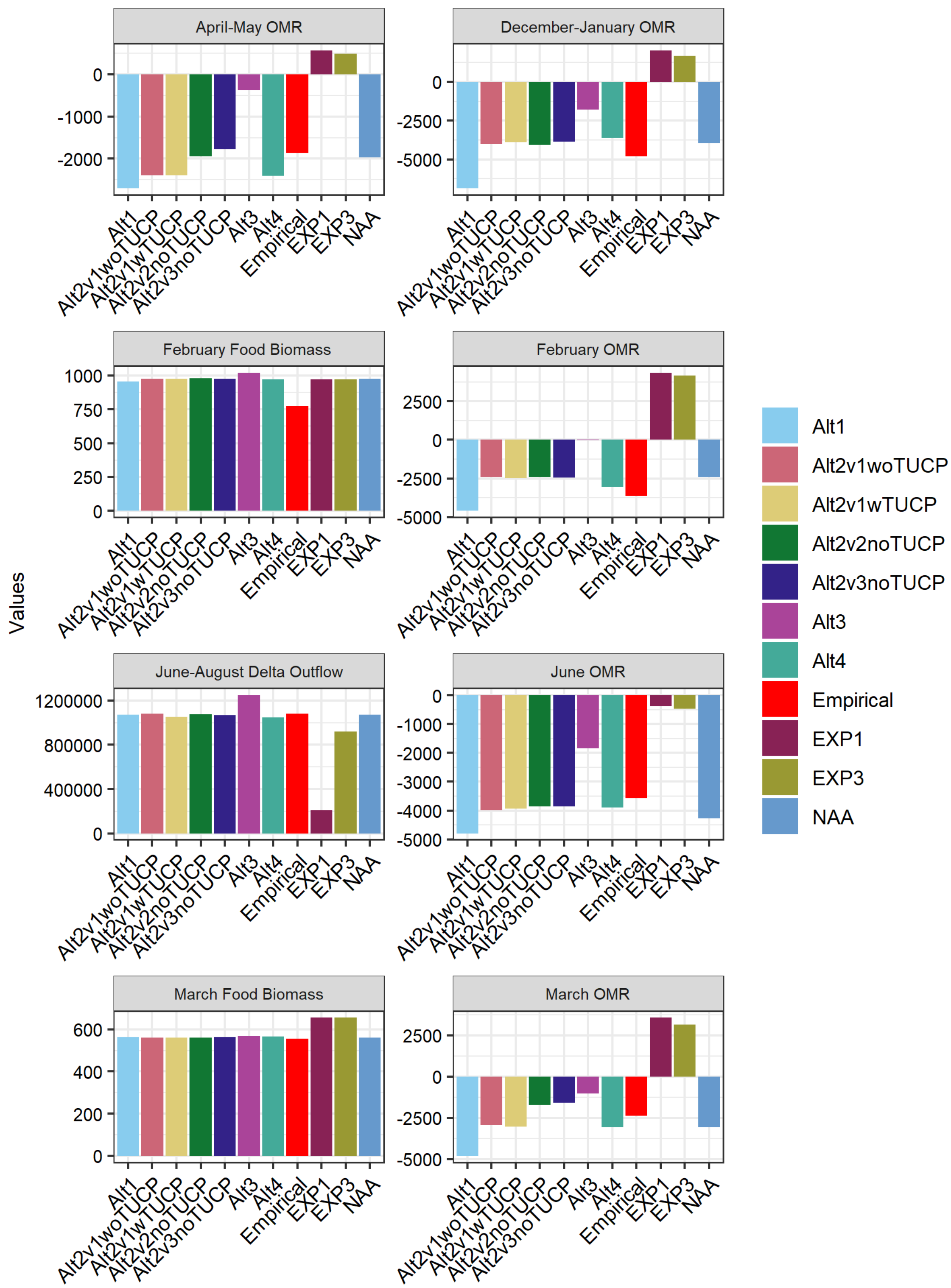


Figure 7. Mean covariate values used in the LCME for Below Normal, Dry, or Critically Dry year Note that cohort year was matched with the water year that the cohort was born in (e.g., cohort year 1995 = water year 1995).

Estimates of population growth rate (λ; the abundance of current year divided by abundance from previous year) are provided for each cohort year and alternative (Table 4; Figure 8). Generally, dry years showed lower geometric mean λ than wet years (Table 5), and wet years occurred with greater frequency at the beginning of the time series (1995-1999) compared to the end of the time series (2006-2015).

Summarized across all years by calculating the geometric mean of λ for the full 21-year time series (1995-2015), predicted flow and zooplankton conditions associated with EXP3 resulted in the highest mean λ, followed by Alt3 (Figure 9). Meanwhile predicted conditions associated with Alt1 resulted in the lowest value of mean λ. All other alternatives resulted in mean λ between 0.95 and 0.97 (Table 5). Relative to the no action alternative (NAA), Alt3 and EXP3 mean projected λ were the highest among all alternatives, and Alt1 was the lowest (Figure 9). Decomposition of mean λ into time series plots of % change of population growth rate for a given alternative divided by the population growth rate of NAA demonstrated that EXP3- and Alt3-projected λ were greater than NAA in most years (Figure 8). Alt1 projections differed from NAA primarily in the first half of the time series (1995-2005) and were very similar to NAA projections in the latter half of the time series (2006-2015). EXP1-projected λ were relatively greater than NAA in wet years, but less than NAA-projected λ in all other years.

NAA, the various versions of Alt2, and Alt4 performed similarly to the empirical data. While these CalSim-generated scenarios/alternatives resulted in higher λ than the empirical data during dry years, they also resulted in lower λ than the empirical data during wet years (Table 5). The CalSim-generated scenarios/alternatives (NAA, the various versions of Alt2, and Alt4) may have produced higher λ during dry years due to the more positive OMR values for multiple months and higher zooplankton estimates in February (Figure 6). Meanwhile, these same CalSim-generated scenarios/alternatives (NAA, the various versions of Alt2, and Alt4) may have produced lower λ than the empirical data during wet years because of the lower June-August Delta Outflow values and more negative OMR values for certain months (Figure 7).

Table 4. Predicted population growth rate (λ; abundance of current year divided by abundance from previous year) for each cohort year by alternatives. λ for cohort year 1995 was calculated by using a static abundance estimate for cohort year 1994. Empirical indicates the observed data used by the LCME.

Year	Empirical	Alt1	Alt2v1 wTUCP	Alt2v1 woTUCP	Alt2v2 noTUCP	Alt2v3 noTUCP	Alt3	Alt4	EXP1	EXP3	NAA	Sacramento Valley Water Year Index
1995	3.56	1.63	1.78	1.79	1.80	1.78	2.25	1.84	4.25	4.05	1.86	Wet
1996	1.37	0.66	0.64	0.64	0.65	0.65	1.06	0.64	1.04	1.27	0.73	Wet
1997	0.68	0.38	0.56	0.57	0.57	0.57	0.78	0.59	0.57	0.99	0.59	Wet
1998	4.78	1.75	1.70	1.67	1.68	1.68	3.03	1.73	5.15	4.88	1.82	Wet
1999	0.79	0.56	0.69	0.70	0.68	0.70	0.89	0.69	0.94	1.27	0.79	Wet
2000	0.69	0.45	0.83	0.83	0.88	0.89	0.81	0.78	0.79	1.17	0.90	Above Normal
2001	0.11	0.12	0.30	0.30	0.33	0.34	0.50	0.30	0.32	0.53	0.31	Dry
2002	0.55	0.68	0.94	0.96	1.04	1.03	1.24	0.94	0.69	1.16	0.93	Dry
2003	0.87	0.71	1.45	1.46	1.54	1.54	1.58	1.45	1.36	2.12	1.51	Above Normal
2004	0.44	0.46	0.84	0.84	0.87	0.87	0.97	0.83	0.53	0.91	0.87	Below Normal
2005	1.94	1.04	1.27	1.27	1.34	1.36	1.58	1.28	2.66	2.85	1.31	Above Normal
2006	3.37	2.04	2.31	2.35	2.40	2.45	2.58	2.29	3.34	3.67	2.41	Wet
2007	0.51	0.33	0.57	0.57	0.58	0.57	0.83	0.53	0.46	0.77	0.57	Dry
2008	0.95	1.00	1.10	1.10	1.17	1.18	1.50	1.07	0.70	1.11	1.09	Critically Dry
2009	0.64	0.45	0.68	0.68	0.68	0.67	0.76	0.66	0.54	1.02	0.67	Dry
2010	1.26	1.31	1.45	1.45	1.47	1.43	1.66	1.46	1.87	1.97	1.48	Below Normal
2011	3.65	3.14	3.26	3.24	3.24	3.25	3.13	3.28	5.57	5.47	3.23	Wet
2012	0.95	0.74	0.98	0.99	1.00	1.00	1.15	1.00	0.67	1.10	1.02	Below Normal
2013	0.90	0.87	0.88	0.88	0.87	0.86	1.04	0.87	0.50	0.83	0.87	Dry
2014	0.43	0.43	0.48	0.52	0.51	0.52	0.71	0.46	0.38	0.57	0.47	Critically Dry
2015	0.66	0.65	0.56	0.63	0.63	0.63	0.74	0.56	0.41	0.60	0.56	Critically Dry

Table 5. Geometric mean of predicted population growth rate (λ) across all years and binned into wetter and drier years for all alternatives. Empirical scenario indicates the LCME fit to observed data, while all alternative models represent simulations using CalSim output.

Category	**Alt1**	**Alt2v1 woTUCP**	**Alt2v1 wTUCP**	**Alt2v2 noTUCP**	**Alt2v3 noTUCP**	**Alt3**	**Alt4**	**EXP1**	**EXP3**	**Empirical**	**NAA**
1995-2015	0.72	0.95	0.94	0.98	0.98	1.20	0.94	1.01	1.41	0.96	0.97
Below Normal, Dry, or Critically Dry years	0.54	0.75	0.74	0.77	0.77	0.95	0.72	0.57	0.90	0.58	0.74
Wet and Above Normal years	0.98	1.24	1.24	1.27	1.28	1.55	1.25	1.91	2.32	1.68	1.32

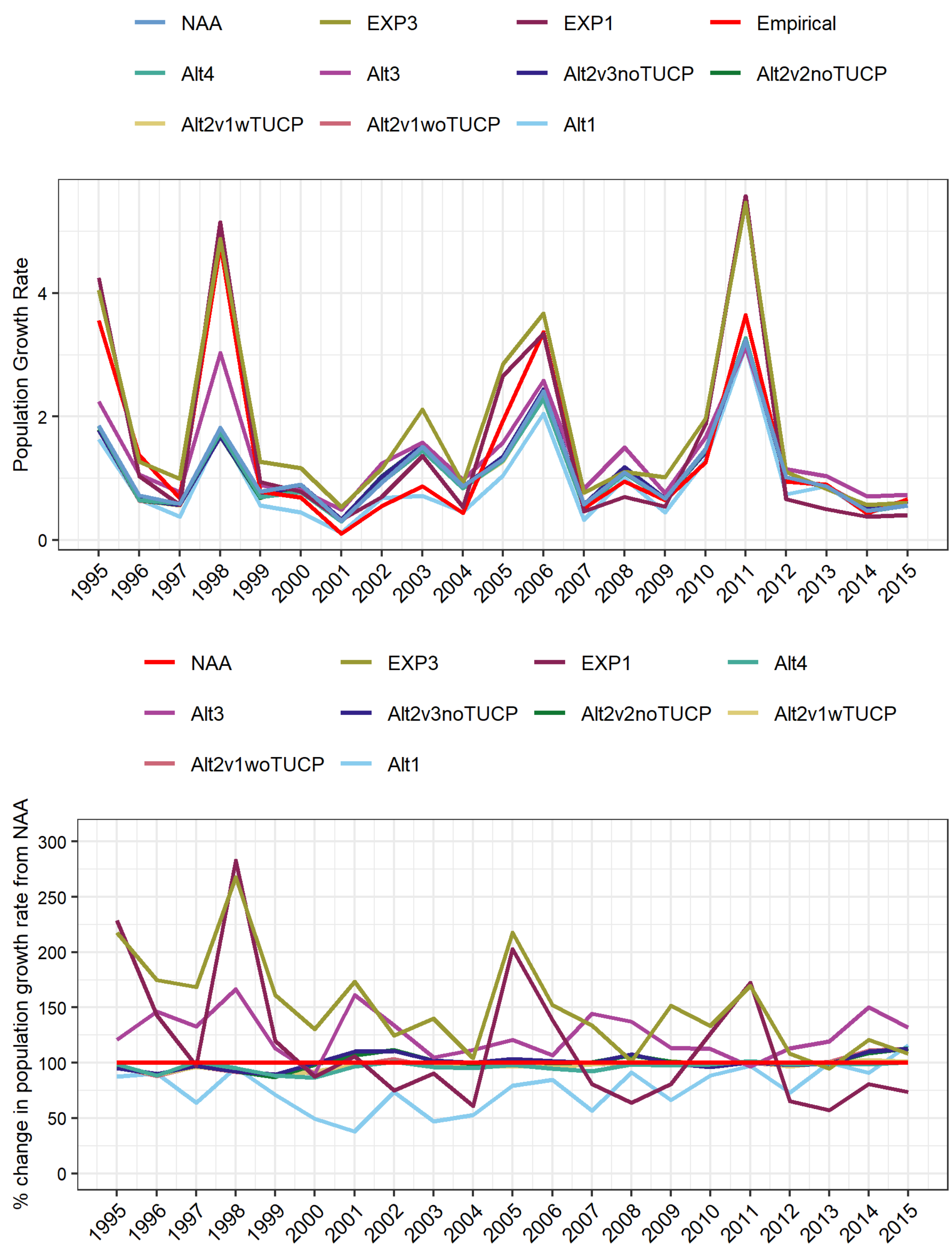


Figure 8. Annual time series of delta smelt population growth rate. Top: Line plot of population growth rate (λ) across alternatives as seen in Table 4. Bottom: Line plot showing % change calculated as λ for a given alternative divided by estimated population growth rate for NAA (no action alternative); no change from NAA = 100. Note the color change for NAA in the bottom figure.

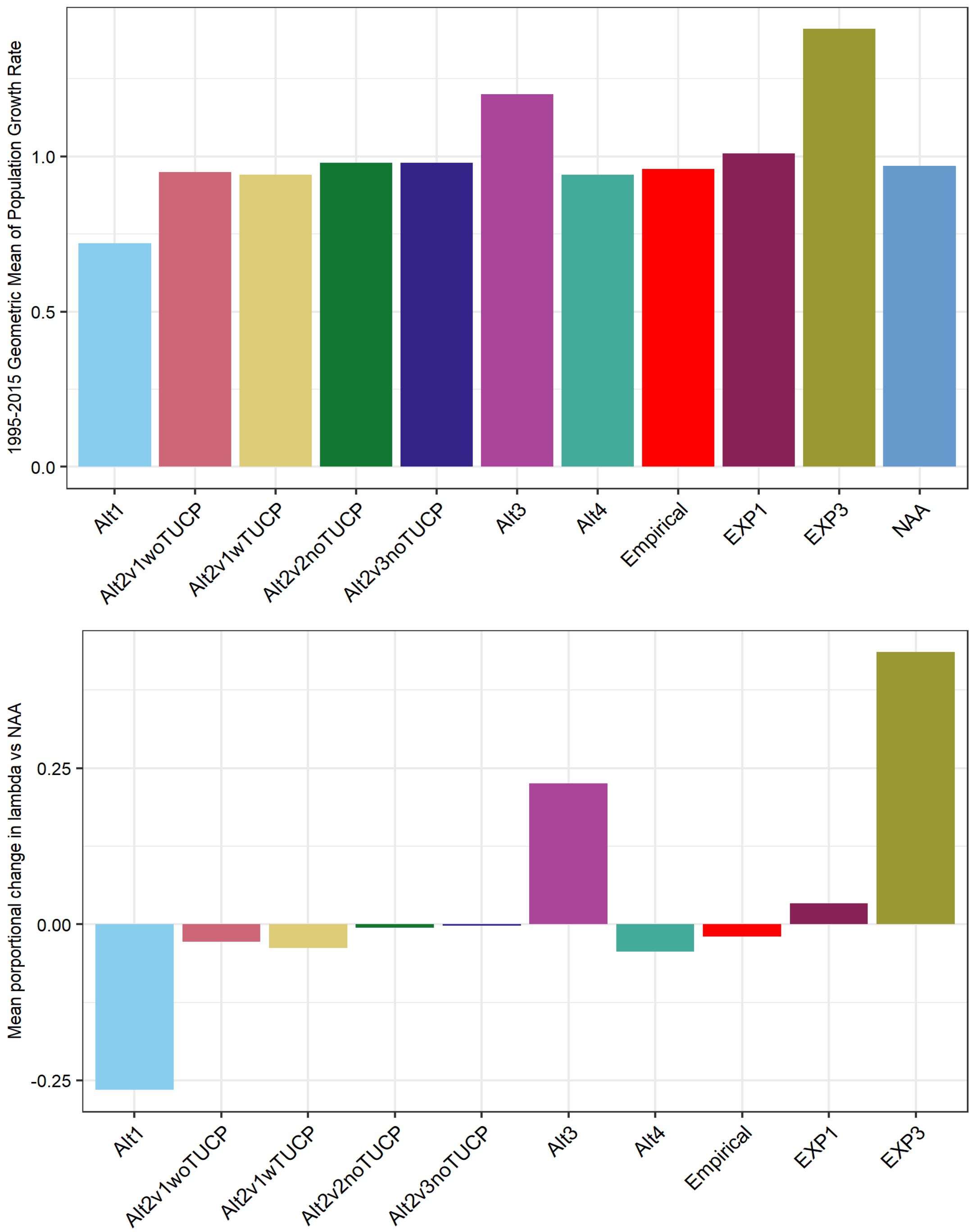


Figure 9. Mean population growth rates aggregated across the years. Top: Bar plot demonstrating the geometric mean of population growth rate (lambda) from 1995 to 2015 for the various alternatives as seen in Table 5. Bottom: Bar plot demonstrating the relative difference in geometric mean of population growth rate (1995-2015) for each alternative compared to the no action alternative ($[\lambda_{alternative} - \lambda_{no\ action}]/\ \lambda_{no\ action}$). Negative numbers indicate alternatives that result in poorer conditions for delta smelt and positive numbers indicate alternatives that are predicted to improve conditions.

1.1.1.3 Key Takeaways

- Geometric mean of population growth rate from 1995 to 2015 only showed considerable differences from the observed data and/or NAA for EXP3, Alt1, and Alt3 scenarios, where EXP3 and Alt3 performed better than most scenarios/alternatives (i.e., higher λ) and Alt1 performed worse than most alternatives (i.e., lower λ).
- EXP3 and Alt3 scenarios likely produced in higher λ due to more positive OMR flows for most months and the relatively high June-August Delta Outflow during dry years (Figures 5, 6).
- Alt1 scenario likely produced lower λ relative to most scenarios due to the more negative OMR flows during most months (Figure 5, 6).
- NAA, all components of Alt2, and Alt4 did not produce considerably higher λ than the empirical data despite OMR restrictions that should reduce entrainment. This may be due to either how flow is calculated from CalSim 3 or the apparent trade-off between OMR flow and summer Delta outflow that somehow occurred between these alternatives and the empirical data (Figures 3, 4).

1.1.1.4 BA Takeaways

The Delta Smelt Life Cycle Model with Entrainment Analysis (LCME), Appendix F, Attachment X produces estimated values for larval recruitment and survival at the subsequent life stages (Smith et al. 2021). The most statistically supported model used means of December-June Old and Middle River (OMR) values, June-August outflow aggregated from monthly values or longer timescales, and aggregated food/prey metric from January to March. The model is used to calculate expected annual population growth rate (λ; the abundance of current year divided by abundance from previous year) as a performance measure of Delta seasonal flow operations influence on OMR and outflow over a twenty year time period (1995-2015).

The general statistical prohibition against extrapolation suggests that model predictions are more uncertain when explanatory variables are outside the range of observations to which the model was fit. Most CalSim-predicted flows and zooplankton predictions were not outside the range of observations to which the Delta smelt LCME was fit, but some alternatives did include out-of-range values. EXP1 included much lower June-August Delta Outflows than observed and higher (more positive) OMR values than observed in some years. EXP3 OMR values were similar to EXP1, but EXP3 June-August Delta Outflows were within the observed range. The multiple components of the PA also contained some April-May OMR values that were more negative than the observed range. Overall, CalSim-predicted June-August Outflow values were generally lower than the empirical data under Wet or Above Normal years. Predicted prey biomass for all alternatives was within the observed range, except for certain years where higher prey biomass were predicted than the empirical data for all alternatives.

The geometric mean of the expected population growth across years (1995-2015), λ, for the PA components ranged from 0.95 to 0.98 (Table 6). The means of the expected population growth rate varied more widely across water year types, and showed positive growth rates under wetter meteorology and negative growth rates under drier meteorology. Note that wetter years also occurred with greater frequency at the beginning of the time series (1995-1999) compared to the end of the time series (2006-2015). Predicted flow and zooplankton conditions associated with EXP3 resulted in the highest geometric mean λ (1.41), whereas NAA and the various components of PA produced geometric mean λ similar to the empirical data (0.95-0.98 vs. 0.96). While NAA and the various components of the PA resulted in higher λ than the empirical data during drier years, they also resulted in lower λ than the empirical data during wetter years (Table 6).

NAA and the various components of the PA may have produced higher λ during drier years due to the more positive OMR values for multiple months and higher zooplankton estimates in February. Meanwhile, NAA and the PA components may have produced lower λ than the empirical data during wetter years because of the lower June-August Delta Outflow values and more negative OMR values for some months. NAA and the PA components did not produce higher λ despite OMR restrictions that should reduce entrainment of Delta smelt. This may be due to the apparent trade-off between OMR flow and summer Delta outflow that somehow occurred between PA components and the empirical data (Figures BA2, BA3).

Table 6. Geometric mean of predicted population growth rate (λ) across all years and binned into wetter and drier years for all alternatives. Empirical scenario indicates the LCME fit to observed data, while all alternative models represent simulations using CalSim output.

Category	**EXP1**	**EXP3**	**NAA**	**PAwoTUCP woVA**	**PAwoTUCP DeltaVA**	**PAwoTUCP SystemwideVA**	**Empirical**
1995-2015	1.01	1.41	0.97	0.95	0.98	0.98	0.96
Below Normal, Dry, or Critically Dry years	0.57	0.90	0.74	0.75	0.77	0.77	0.58
Wet and Above Normal years	1.91	2.32	1.32	1.24	1.27	1.28	1.68

1.1.1.5 EIS Takeaways

The Delta Smelt Life Cycle Model with Entrainment Analysis (LCME), Appendix F, Attachment X produces estimated values for larval recruitment and survival at the subsequent life stages (Smith et al. 2021). The most statistically supported model used means of December-June Old and Middle River (OMR) values, June-August outflow aggregated from monthly values or longer timescales, and aggregated food/prey metric from January to March. The model is used to calculate expected annual population growth rate (λ; the abundance of current year divided by abundance from previous year) as a performance measure of Delta seasonal flow operations influence on OMR and outflow over a twenty year time period (1995-2015).

The general statistical prohibition against extrapolation suggests that model predictions are more uncertain when explanatory variables are outside the range of observations to which the model was fit. Most CalSim-predicted flows and zooplankton predictions were not outside the range of observations to which the Delta smelt LCME was fit, but some alternatives did include out-of-range values. Alternative 1, the multiple phases of Alternative 2, and Alternative 4 contained some April-May OMR values that were more negative than the observed range. Alternative 1 also contained OMR values more negative than the observed data for the months of December-January and March. Overall, CalSim-predicted June-August Outflow values were generally lower than the empirical data under Wet or Above Normal years. Predicted prey biomass for all alternatives was within the observed range, except for certain years where higher prey biomass were predicted than the empirical data for all alternatives.

The geometric mean of the expected population growth across years (1995-2015), λ, for the Alternative 2 phases ranged from 0.94 to 0.98 (Table 7). The means of the expected population growth rate varied more widely across water year types, and showed positive growth rates under wetter meteorology and negative growth rates under drier meteorology. Note that wetter years also occurred with greater frequency at the beginning of the time series (1995-1999) compared to the end of the time series (2006-2015). Predicted flow and zooplankton conditions associated with Alternative 3 resulted in the highest geometric mean λ (1.20), whereas conditions associated with Alternative 1 resulted in the lowest geometric mean λ (0.72). No Action Alternative, the various phases of Alternative 2, and Alternative 4 produced geometric mean λ similar to the empirical data (0.94-0.98 vs. 0.96). While No Action Alternative, the various phases of Alternative 2, and Alternative 4 resulted in higher λ than the empirical data during drier years, they also resulted in lower λ than the empirical data during wetter years (Table 7).

No Action Alternative, the various phases of Alternative 2, and Alternative 4 may have produced higher λ during drier years relative to the empirical data due to the more positive OMR values for multiple months and higher zooplankton estimates in February. Meanwhile, No Action Alternative, the various phases of Alternative 2, and Alternative 4 may have produced lower λ than the empirical data during wetter years because of the lower June-August Delta Outflow values and more negative OMR values for some months. No Action Alternative, the various phases of Alternative 2, and Alternative 4 did not produce higher λ despite OMR restrictions that should reduce entrainment of Delta smelt. This may be due to the apparent trade-off between OMR flow and summer Delta outflow that somehow occurred between these alternatives and the empirical data.

Table 7. Geometric mean of predicted population growth rate (λ) across all years and binned into wetter and drier years for all alternatives. Empirical scenario indicates the LCME fit to observed data, while all alternative models represent simulations using CalSim output.

Category	Alt1	Alt2wo TUCPwoVA	Alt2wTUCP woVA	Alt2woTUCP DeltaVA	Alt2woTUCP SystemwideVA	Alt3	Alt4	Empirical	NAA
1995-2015	0.72	0.95	0.94	0.98	0.98	1.20	0.94	0.96	0.97
Below Normal, Dry, or Critically Dry years	0.54	0.75	0.74	0.77	0.77	0.95	0.72	0.58	0.74
Wet and Above Normal years	0.98	1.24	1.24	1.27	1.28	1.55	1.25	1.68	1.32

1.2 References

Kimmerer, W. J. , and K. A. Rose. 2018. Individual-Based Modeling of Delta Smelt Population Dynamics in the Upper San Francisco Estuary III. Effects of Entrainment Mortality and Changes in Prey. *Trans Am Fish Soc.* 147(1):223–243. https://doi.org/10.1002/tafs.10015.

Polansky, L., K. B. Newman, and L. Mitchell. 2021. Improving inference for nonlinear state-space models of animal population dynamics given biased sequential life stage data. *Biometrics.* 77(1):352–361. https://doi.org/10.1111/biom.13267.

Rose, K. A., W. J. Kimmerer, K. P. Edwards, and W. A. Bennett. 2013. Individual-based modeling of delta smelt population dynamics in the upper San Francisco estuary: I. model description and baseline results. *Trans Am Fish Soc.* 142(5):1238–1259. https://doi.org/10.1080/00028487.2013.799518.

Smith, W. E. 2019. Integration of transport, survival, and sampling efficiency in a model of south delta entrainment. *San Fr Estuary Watershed Sci.* 17(4). https://doi.org/10.15447/sfews.2019v17iss4art4.

Smith, W. E., K. B. Newman, and L. Mitchell. 2020. A bayesian hierarchical model of postlarval delta smelt entrainment: Integrating transport, length composition, and sampling efficiency in estimates of loss. *Can J Fish Aquat Sci.* 77(5):789–813. https://doi.org/10.1139/cjfas-2019-0148.

Smith, W. E., and M. L. Nobriga. 2023. A bioenergetics-based index of habitat suitability: spatial dynamics of foraging constraints and food limitation for a rare estuarine fish. *Trans Am Fish Soc.* https://doi.org/10.1002/tafs.10427.

Smith, W. E., L. Polansky, and M. L. Nobriga. 2021. Disentangling risks to an endangered fish: Using a state-space life cycle model to separate natural mortality from anthropogenic losses. *Can J Fish Aquat Sci.* 78(8):1008–1029. https://doi.org/10.1139/cjfas-2020-0251.

compass

Memo

To: CSAMP Delta Smelt SDM Technical Working Group (TWG)

From: Brian Crawford and Sally Rudd, Compass Resource Management

Date: March 11, 2022

Re: Methods for predicting changes in salinity due to X2/outflow management actions

1 Background

This analysis was motivated by the interest to predict changes in salinity (PSU) in the Delta as a function of management actions that alter outflow and the location of X2 (i.e., the distance from the Golden Gate to the point in the Delta where daily average salinity is 2 ppt). Both salinity and X2 location influence the distribution of Delta Smelt across subregions in a submodel (Smith 2021b) within the Individual-based Model in R (IBMR) programmed by Will Smith (Smith 2021a). The distribution submodel and IBMR will serve as tools for predicting Delta Smelt population outcomes, given actions and portfolios evaluated in Round 1 of the SDM analysis.

Depending on the ambient conditions, additional outflow and/or X2 management are actions that can help achieve environmental and biological goals, which relate to improving dynamic habitat for Delta Smelt and fish distribution. There are several options for the timing and intensity of additional outflow and/or X2 management to increase dynamic habitat. The portfolios included in Round 1 of the SDM evaluation evaluate the current management actions for outflow and X2 management and compare performance of the current approach with alternatives. The objective of this analysis was to quantify the relationship between historical salinity and X2 data, as well as other factors, to predict subsequent changes in salinity when evaluating Delta Smelt outcomes under outflow/X2 management alternatives.

2 Methods and results for modeling historical salinity conditions

2.1 Datasets

Salinity (PSU) data was taken from fish and water quality monitoring data on CSAMP's GitHub page, which synthesizes 11 data sources (listed below). The original dataset, which included observed salinity values per survey at a given location, was linked to the 12 subregions used in the IBMR and summarized into mean year-month-subregion values. Because some year-month-subregion combinations did not have salinity data, missing values were filled in with predicted values using data from other subregions fit to generalized linear models following the same methods described for missing temperature data in Smith (2021a: Appendix B).

- Data sources: "EMP" (Environmental Monitoring Program); "STN" (Summer Townet Survey); "FMWT" (Fall Midwater Trawl); "EDSM" (Enhanced Delta Smelt Monitoring); "DJFMP" (Delta Juvenile Fish Monitoring Program); "20mm" (20mm Survey); "SKT" (Spring Kodiak Trawl); "Baystudy" (Bay Study); "USGS" (USGS San Francisco Bay Surveys); "USBR" (United States Bureau of Reclamation Sacramento Deepwater Ship Channel data); and "Suisun" (Suisun Marsh Fish Study)

Historical daily X2 location data was obtained from the Dayflow dataset (available on https://data.cnra.ca.gov/dataset/dayflow), which includes data between October 1996 and 2020. For X2 data prior to October 1996, we used values consistent with the IBMR that were calculated by previous researchers according to an equation provided on the Dayflow documentation website (Will Smith, pers. comm.).

2.2 Modeling historical salinity

We built a global generalized linear regression model with a gamma distribution where mean year-month-subregion salinity was the response variable influenced by month, $month^2$, X2 location, $X2^2$, subregion, and an X2 x subregion interaction effect. For the month covariate, we centered the data and added the quadratic term to account for potential cyclical, seasonal patterns in covariates. We also standardized X2 prior to model fitting. We fit the model to a final dataset of salinity and X2 conditions between January 1995 and December 2014 (n = 2880) to align with the modeling timeframe of the IBMR. We performed AIC_c-based model selection for a specific set of candidate models including the global model and simpler models where predictor effects were dropped (Burnham and Anderson 2002).

The global model outperformed all others by >350 ΔAIC_c (Table 1), and this was confirmed by backwards step-wise regression that tested all parameter combinations within the global model. The model showed adequate fit to the data (Figure 1) and explained 91.5% of the variation in salinity. Therefore, we used the global model to predict changes in salinity (see Section 3).

Results are presented in Figure 2 and Table 2. Overall, salinity was low and did not vary with X2 for all subregions east of the Confluence. Salinity increased marginally with higher X2 in the Confluence and Suisun Marsh, and salinity increased more greatly with higher X2 in the Suisun Bay subregions.

R code and associated datasets to replicate these methods are available in a zipped folder provided to the Delta Smelt TWG.

3 Predicting salinity changes due to X2 management

We predicted mean year-month-subregion changes in salinity using the salinity model, given three management alternatives that simulated changing historical monthly X2 values between 1995 and 2014:

1) No Fall X2 management: X2 in August, September, October, and November in 2011 (the only year in our model timeframe when X2 management occurred) was set to 80, 84, 87, and 85, respectively to simulate this management action not occurring that year (these were the X2 values for these months in 2000, which was chosen because 2000 and 2011 had similar X2 positions in early August);
2) Fall X2 as per 2008 BiOp: X2 in September and October was set to 74 in Wet (1995 to 2000, 2006) and 81 in Above Normal (2003) water year types that occurred before 2007 – i.e., the year that the 2008 BiOp actions (such as Fall X2 and OMR management) began to be implemented; and
3) Fall X2 as per 2020 ITP/BiOp: X2 in September and October was set to 80 in Wet (1995 to 2000, 2006, 2011) and Above Normal (2003) water year types.

When predicting new salinity values in X2 management scenarios, we ensured predicted values did not fall outside of the range of observed salinities. All values that were initially predicted to be greater than the maximum observed salinity value were reassigned as the maximum observed salinity value.

Overall, Figure 2 illustrates how predicted salinity changed, on average, in these scenarios. For example, changing X2 from 71 to 80 generally resulted in no change of salinity east of the Confluence, a marginal increase in salinity in the Confluence and Suisun Marsh, and a greater increase in salinity in the four

Suisun Bay subregions. We note that these changes represent average expected changes, relative to observed values. Because observed year-month-subregion salinity conditions were somewhat variable (see black points in Figure 1) and values predicted from X2 management were deterministic (see red lines in Figure 1), it is possible that relative changes between observed and predicted salinity did not always follow these patterns for certain year-month-subregions.

Table 1. AICc, ΔAICc, and Akaike weights for generalized linear regression candidate model predicting salinity in the Delta (between 1995 and 2014; n = 2880).

Model	K[a]	AIC_c[a]	ΔAIC_c[a]	w_i[a]
X2 + $X2^2$ + Month + $Month^2$ + Subregion + Subregion*X2	28	1354	0	1.0
X2 + Month + Subregion + Subregion*X2	26	1735	381	0.0
X2 + Subregion	14	3476	2122	0.0
X2 + Month + Subregion	15	3477	2124	0.0
X2	3	8501	7147	0.0
Null	2	9498	8144	0.0

[a] K = no. of parameters, AIC_c = Akaike's Information Criterion, ΔAIC_c = difference in AIC_c from the best model, and w_i = Akaike wt.

Table 2. Parameter estimates from the generalized linear model predicting salinity as a function of X2 position, $X2^2$, month, $month^2$, subregion, and a subregion x X2 interaction in the Delta (between 1995 and 2014; n = 2880).

	Main effects		Interaction between subregion and X2	
	Estimate	SE	Estimate	SE
Intercept	-2.029***	0.041		
X2	0.111**	0.04		
$X2^2$	0.061***	0.01		
Month	-0.012**	0.004		
$Month^2$	0.016***	0.001		
Sacramento River	-0.472***	0.053	0.076	0.053
South Delta	0.122*	0.053	0.131*	0.053
East Delta	-0.721***	0.053	0.104	0.053
Lower Sacramento	0.655***	0.053	0.789***	0.053
Lower San Joaquin	0.372***	0.053	0.517***	0.053
Confluence	1.64***	0.053	1.149***	0.053
SE Suisun	2.566***	0.053	1.256***	0.053
NE Suisun	2.68***	0.053	1.278***	0.053
Suisun Marsh	2.774***	0.053	0.741***	0.053
SW Suisun	3.736***	0.053	0.861***	0.053
NW Suisun	3.273***	0.053	1.15***	0.053

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Figure 1. Diagnostic plots for the best model to predict mean subregion-month salinity in the Delta.

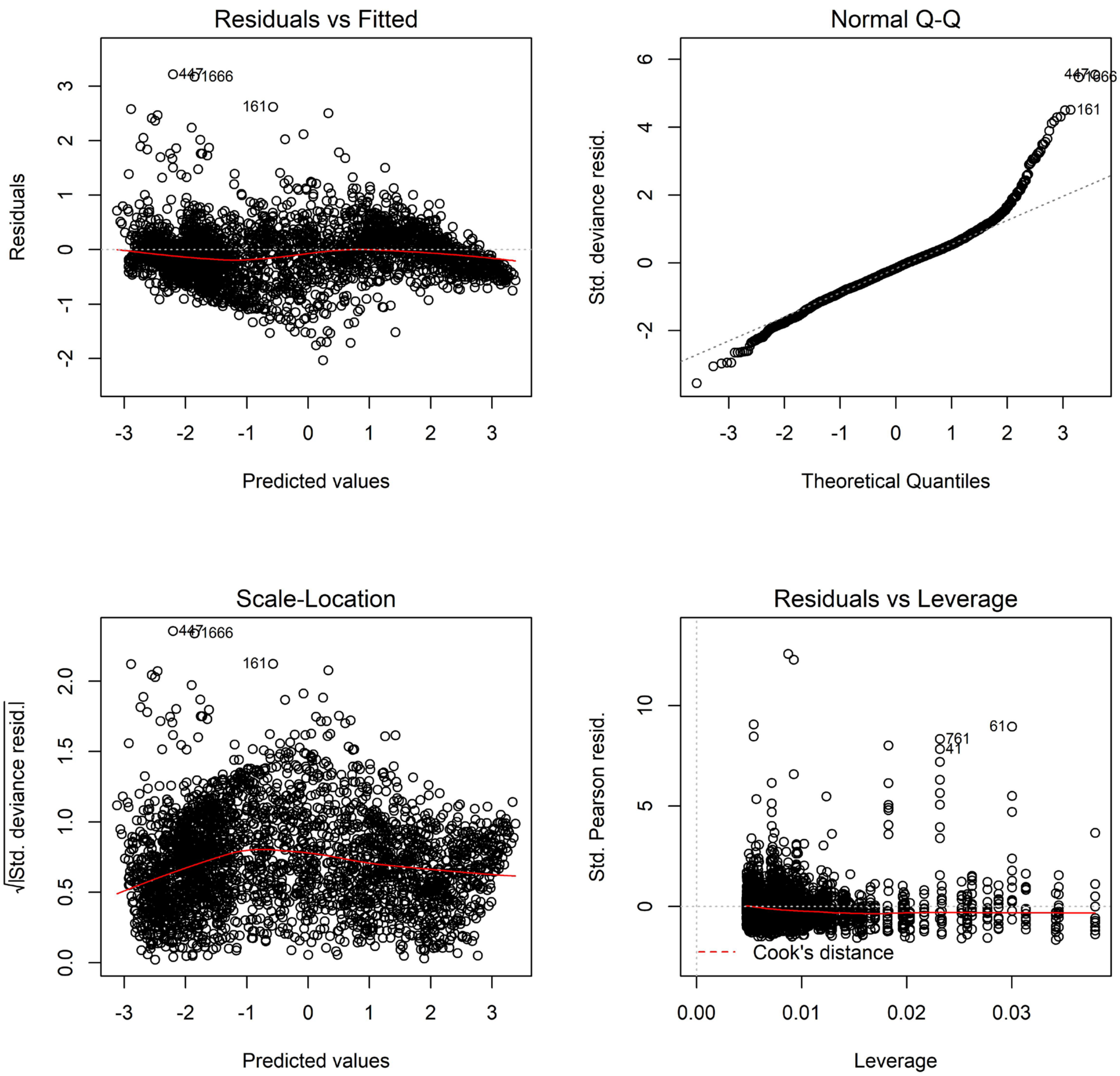
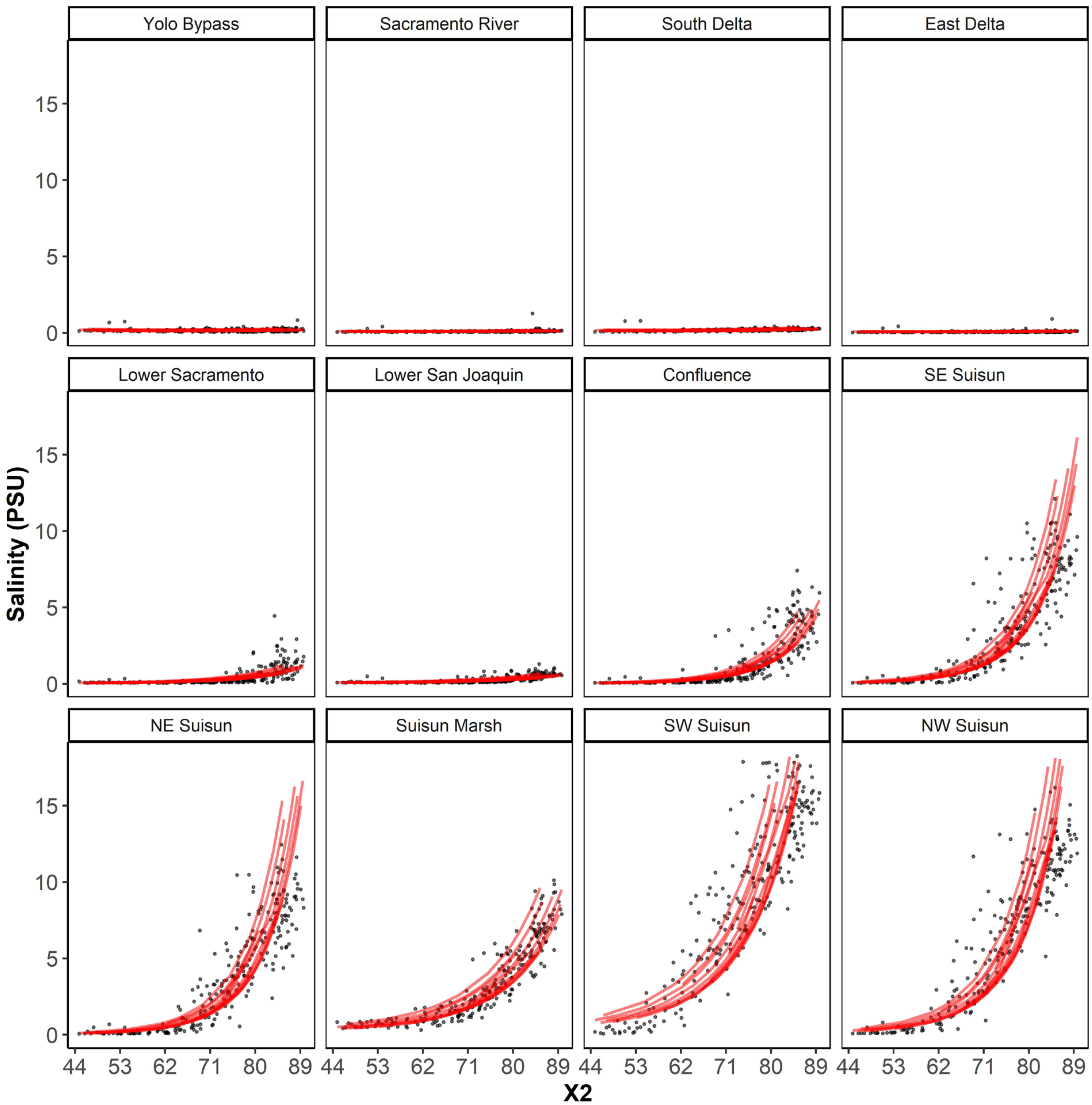


Figure 2. Salinity vs. X2 by subregion. Black points are the mean subregion-month conditions from Dayflow data (X2) and water quality monitoring data (salinity) in the Delta (between 1995 and 2014; n = 2880). The red lines are predicted values from a generalized linear model (GLM) with a gamma distribution (for when data is like the salinity data – always positive and skewed). There is a red line for each month.

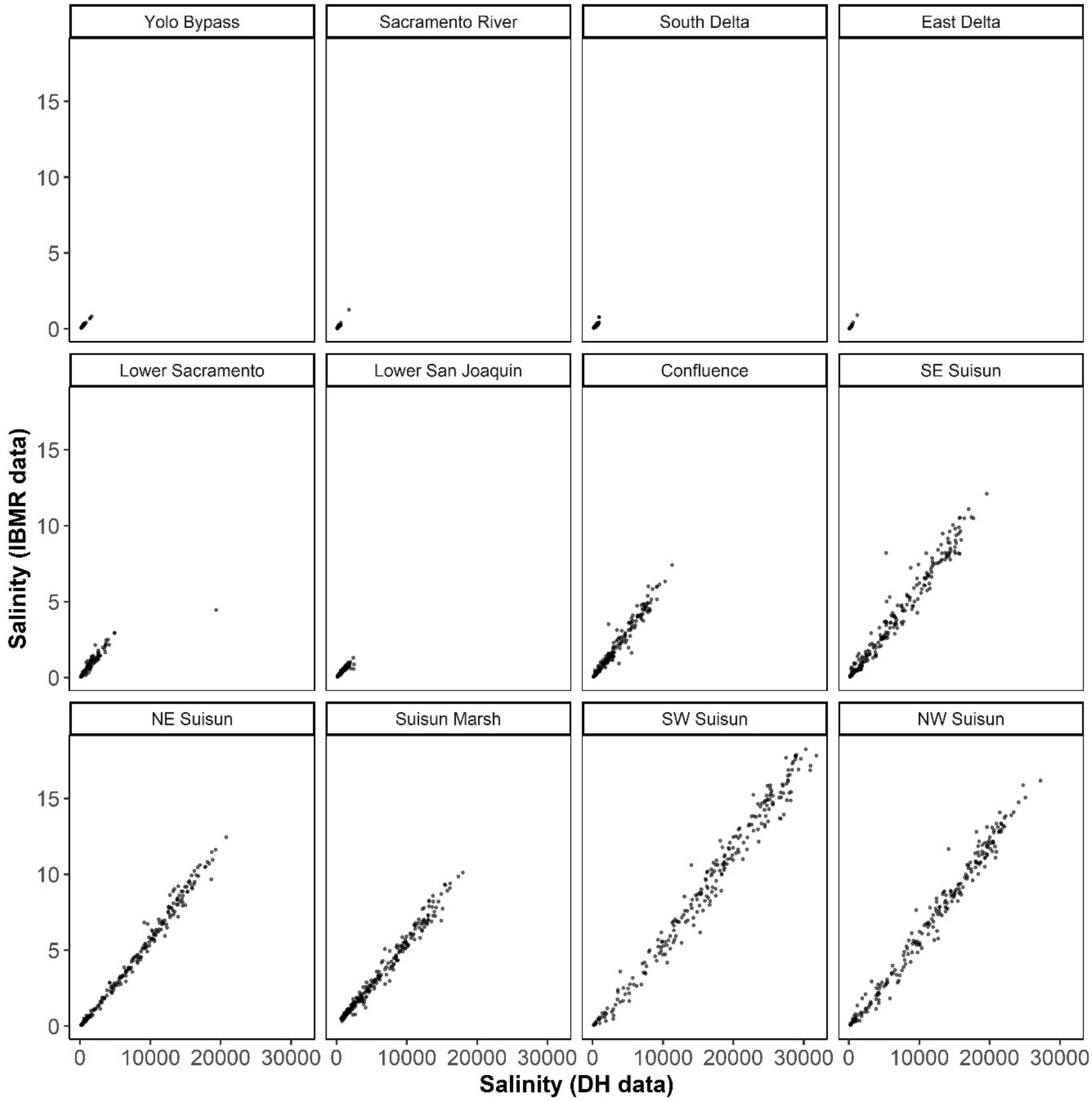


4 Additional Quality Assurance / Quality Control (QA/QC) work

As part of this analysis, Compass performed QA/QC of the IBMR input datasets for salinity, turbidity, temperature, and prey. Steps included:

- Reviewing original water quality datasets (from monitoring program data) that were being used to generate mean year-month-strata values for the IBMR for dynamic habitat attributes.
- Assessing agreement between mean values for dynamic habitat attributes from code used to generate input datasets for the IBMR ("IBMR data") and values generated within the Dynamic Habitat Tool developed by Compass and the Technical Working Group ("DH data"). Through this QA/QC exercise, a few issues were discovered and resolved with the IBMR code. These included showing stronger support for using salinity values derived from monitoring programs, relative to DSM2-derived values. After adjustments to the IBMR code, the IBMR data and DH data showed strong agreement, with only slight variation in values occurring because the IBMR took mean values across all samples from a year-month-strata whereas the DH data calculated daily means first and then means for year-months.

Figure 3. Comparison of mean year-month-subregion salinity between the IBMR data (units: PSU) and Dynamic Habitat data (units: microS/cm) in the Delta (between 1995 and 2014; n = 2880).



5 References

Burnham K, Anderson D. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

Smith, W.E., 2021a. A delta smelt Individual-Based Life Cycle Model in the R statistical environment (Technical Note). Prepared for CSAMP Delta Smelt SDM Technical Working Group (TWG).

Smith, W.E., 2021b. Environmental conditions driving habitat use: A model of the spatial distribution of delta smelt (Technical Note). Prepared for CSAMP Delta Smelt SDM Technical Working Group (TWG).

Memo

To: CSAMP Delta Smelt SDM Technical Working Group (TWG)

From: Brian Crawford and Sally Rudd, Compass Resource Management; Brian Mahardja, US Bureau of Reclamation; Sam Bashevkin, Delta Science Program (currently State Water Board)

Date: November 1, 2022

Re: Methods for predicting changes in zooplankton density due to salinity changes from management actions

1 Background

This analysis was motivated by the interest to predict changes in zooplankton density in the Delta as a function of management actions that alter salinity – either via outflow and the location of X2 (i.e., the distance from the Golden Gate to the point in the Delta where daily average bottom salinity is 2 PSU) or operations of the Suisun Marsh Salinity Control Gates (SMSCG). Zooplankton density influences several demographic parameters in the different Delta Smelt population models being applied in this SDM evaluation, and the Individual-based Model in R (IBMR) programmed by Will Smith (Smith 2022) accounts for taxa-specific densities and effects across 12 prey taxa.

Sam Bashevkin (Delta Science Program [now with the State Water Board]) developed a previous model predicting taxa-specific changes in zooplankton from changes in salinity for use by the Delta Coordination Group (DCG) to inform management decisions for the SMSCG. Sam, working with Compass and Brian Mahardja (Reclamation), adapted that model for the current analysis, where the objective was to predict taxa-specific changes in zooplankton across relevant Bay-Delta subregions due to expected changes in salinity from X2 and SMSCG actions as part of Round 1 of the SDM evaluation.

2 Methods for modeling zooplankton-salinity relationships

2.1 Datasets

Compass supplied Sam with data for 1) salinity (PSU) under baseline conditions between 1995 and 2014 to correspond with the SDM modeling evaluation time frame, and 2) salinity predicted for five management actions. Sam compiled zooplankton densities (mgC/m^3) for 12 taxa used in the IBMR under baseline conditions from data accessed in the zooper package in R, available on the IEP's GitHub site (Bashevkin et al. 2022a). All values were means for a given subregion, month, and year. Baseline values for salinity and zooplankton density were taken directly from the IBMR (see methods in Smith 2022), which originally were generated from fish and water quality monitoring data (Bashevkin et al. 2022b; available on CSAMP's GitHub page), which synthesizes 11 data sources (listed below).

- Data sources: "EMP" (Environmental Monitoring Program); "STN" (Summer Townet Survey); "FMWT" (Fall Midwater Trawl); "EDSM" (Enhanced Delta Smelt Monitoring); "DJFMP" (Delta Juvenile Fish Monitoring Program); "20mm" (20mm Survey); "SKT" (Spring Kodiak Trawl); "Baystudy" (Bay Study); "USGS" (USGS San Francisco Bay Surveys); "USBR" (United States Bureau of Reclamation Sacramento Deepwater Ship Channel data); and "Suisun" (Suisun Marsh Fish Study)

The predicted salinity values for X2 management actions were generated from a separate model relating changes in X2 location/outflow to subregion-specific salinity. See the Action Specification Sheet for a brief description and the Compass Technical Memo for complete details. Predicted salinity values in Suisun

Marsh for the SMSCG action were generated from criteria informed by experts and the SMSCG Monitoring Plan for 2020. See the Action Specification Sheet for a description of methods. All methods predicting the effects of management action on salinity were informed and reviewed by the Technical Working Group (TWG).

The zooplankton dataset used to construct food-salinity relationship comes from the monitoring datasets integrated in the zooper R package (Bashevkin et al. 2022a). From the zooper dataset, we used data from the EMP, FMWT, 20mm, and STN surveys.

2.2 Modeling food-salinity relationships

As part of the DCG structured decision-making process, Sam developed generalized additive models (GAMs) using data in the zooper package to estimate change in zooplankton biomass (by taxonomic group) based on changes in salinity at Suisun Marsh. This modeling effort was done to better understand expected changes in zooplankton as part of the SMSCG action for Delta Smelt (Sommer et al. 2020). Models were fit in the statistical programming language R (R Core Team 2021), using the package mgcv (Wood 2011; Wood et al. 2016). Sam adapted the original DCG model to estimate zooplankton density and salinity relationships by taxon, subregion, month, and management action (model code available on GitHub) for the relevant salinity management actions in Round 1 of CSAMP's SDM evaluation. In effect, Sam's modeling effort was used to predict changes in zooplankton density for the Confluence, Suisun Marsh, and Suisun Bay subregions. Previous analysis showed no measurable effects of X2/outflow actions on salinity – and therefore, no expected changes in zooplankton from salinity – in subregions east of the Confluence (see Compass Technical Memo). IBMR subregions upstream of the Confluence were also not analyzed because there is no management action under consideration that would increase salinity in the Delta.

After initial inspection of the dataset, the NE Suisun and SE Suisun regions were combined due to limited available data in the NE Suisun region. For this analysis, data at the sampling event level and from the year 1995 and after were used. GAMs were constructed for each subregion and taxonomic group with natural log of biomass per unit effort + 1 as the response variable. Predictor variables for each GAM were the tensor product smooth of the interaction between salinity and day of year, as well as random effects for year and station (when more than one station exists in the dataset). To produce the monthly model output for the Delta Smelt IBMR, the 15th of each month was used as the data input.

Generalized additive models are appropriate for predicting the influence of a set of predictor variables on a response variable (i.e., zooplankton density) when the shape of those relationships cannot be assumed to be linear. The use of historical data to predict future or hypothetical management actions is a valid approach in the absence of appropriately powered experimental datasets, especially when the mechanisms underlying the relationship (flow, salinity, and zooplankton taxa) are not entirely clear.

Post-processing steps:

We performed three post-processing steps on model predictions to align model outputs to be usable as inputs for the IBMR and other Delta Smelt models. Due to the Bayesian interpretation of GAMs fit with mgcv (Miller 2021), model predictions were produced as a full Bayesian posterior (similar to bootstrapping) consisting of 1000 independent draws from the posterior distribution, for each model prediction. This enabled the propagation of uncertainty from the model through the calculated quantities. Each step described below was performed on each of the 1000 draws. At the end of the process, the 2.5%, 50% (median), and 97.5% percentiles were extracted to represent the central tendency and uncertainty in the final percentiles.

First, model predictions (from the GAMs) were converted to scalars representing the change in zooplankton density from baseline. For each action, subregion, and month between 1995 and 2014, we calculated the scalar as the predicted taxa-specific zooplankton density under management conditions / the same prediction under baseline conditions. This scalar change will be multiplied to baseline prey density values in Delta Smelt population models. Scalar values can range from 0 to infinity, where a scalar of 0=prey taxon declines to 0 in that month/year/subregion, values between 0 and 1 mean prey density decreases with management, 1=no change from baseline, and values greater than 1 mean prey density increased with management.

Second, the model occasionally predicted negative values for zooplankton density. This was possible due to the log(x+1) transformation which created the possibility for back-transformed values as low as –1. When this occurred, those negative values were changed to 0. However, when 0s were introduced for baseline predictions, it resulted in infinite values for the scalars We replaced these infinities with the maximum finite scalar calculated from model predictions for a specific action, taxon, subregion, and month (across years). When this step still yielded no finite scalar value for a given action, taxon, subregion, and month, we used the maximum finite scalar value from +/- one month.

Third, one taxon from the IBMR set of taxa– *Pseudodiaptomus forbesi* copepodites "pdiapjuv" – was not predicted in the model since one survey (20mm) did not start counting this category until 1998. Therefore, we assumed scalar changes for pdiapjuv were the same as those predicted for *Pseudodiaptomus forbesi* (copepod) adults "pdiapfor."

3 Results and discussion

The taxa-specific models showed adequate fit to the data (see Model fitting section on GitHub). Predicted patterns between zooplankton density and salinity varied by taxon, subregion, and month (Figures 1-3). Within the salinity range modeled, relationships were fairly linear for some taxa, subregions, and months but exhibited nonlinear patterns occasionally where peaks of zooplankton density occurred at intermediate salinity values.

We present the scalars generated from model predictions (scenario / baseline) for *Acartiella sinensis* across years by month and subregion (Figure 4). Results from other taxa can be seen on the Results section on GitHub. Bands on the figures indicate 95% CIs around median model predictions. For some months, years, subregions, and taxa, there was a relatively small difference between lower and upper 95% estimated scalars; for other months, years, subregions, and taxa, there was greater uncertainty (e.g., see October in SW Suisun subregion in Figure 4).

Lastly, we captured the degree of uncertainty of scalars generated from model predictions across taxa and scenarios. We calculated pairwise differences between upper and lower 95% CI estimates for each taxon, subregion, month, year, and management scenario. We then summarized these differences by taxa and scenario (on a log 10 scale, Figure 5). Lower difference values indicate less uncertainty in scalars for a given taxon and action; higher difference values indicate greater uncertainty in scalars. Although there was substantial uncertainty in scalars for specific taxa and actions, we will account for this uncertainty by running separate scenarios in the IBMR using the lower and upper 95% CI predicted scalars for each action. This will provide predicted Delta Smelt population outcomes under "low" and "high bookend" effects of these actions on zooplankton density.

This approach makes a number of assumptions that should be considered when interpreting the results, although most of these caveats would apply to any modeling effort using the same type of data and are not unique to GAMs. Data used to construct the model may include some years from a different ecological regime. Winder and Jassby (2011) found a shift in the zooplankton community in the 1990s and we also used data from years prior to the Pelagic Organism Decline shift in the early 2000s (Sommer et al. 2007). Furthermore, the scalars only represent changes in zooplankton biomass that may be expected based on changes in salinity. They do not represent any other flow-related impacts on zooplankton biomass. The models were also fit to historical data and thus are only accurate within the realm of historical variability and will be less accurate near the fringes of historical variability.

Figure 1. Predicted zooplankton densities (log scale: mgC/m^3) by salinity (PSU), subregion (rows), and month (columns) for 12 taxa used in the IBMR. Bands represent 95% credible intervals of estimated relationships between zooplankton and salinity.

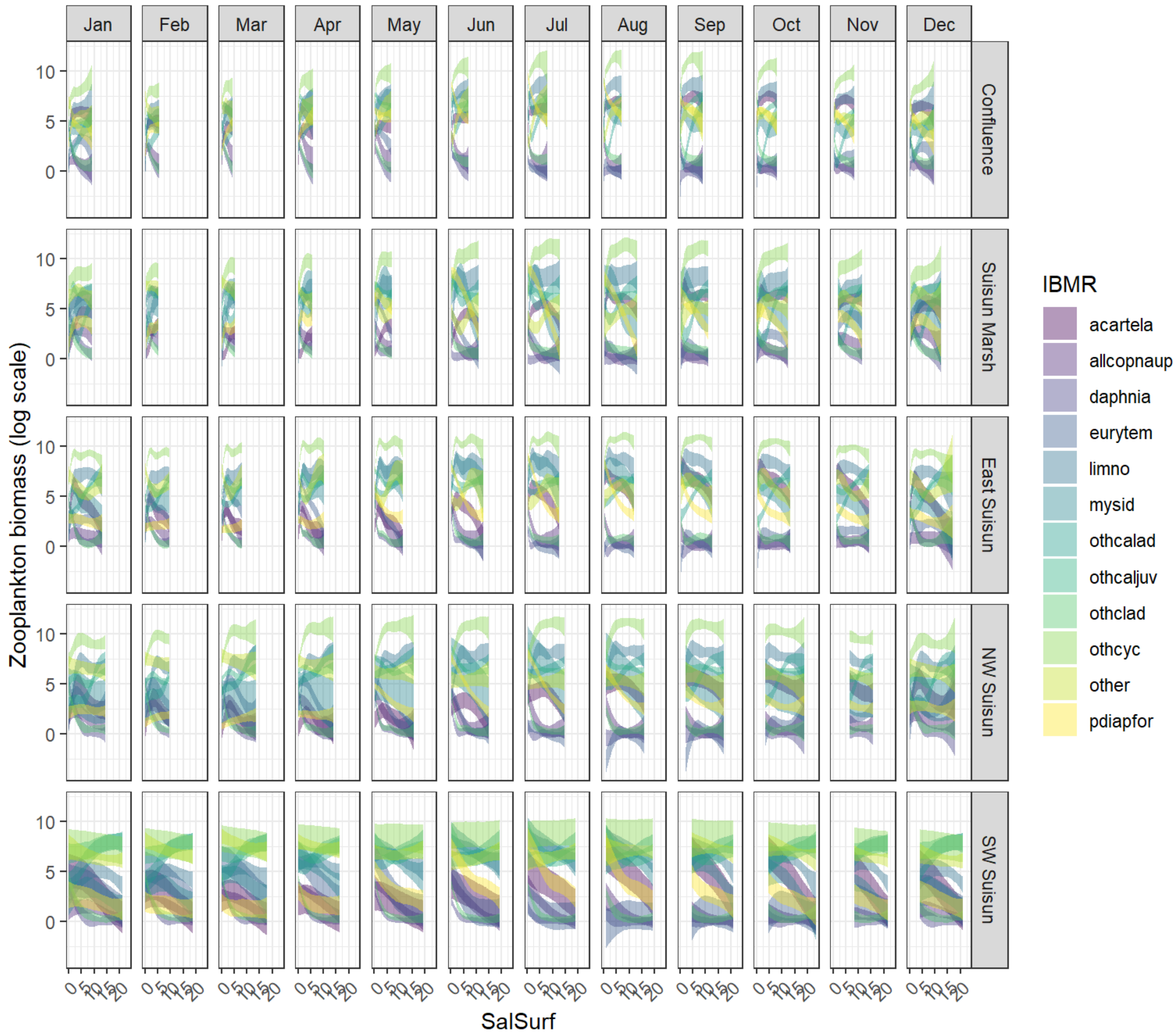


Figure 2. Predicted zooplankton densities (log scale: mgC/m³) by salinity (PSU), subregion (rows), and month (columns) for *Acartiella sinensis*. Bands represent 95% credible intervals of estimated relationships between zooplankton and salinity.

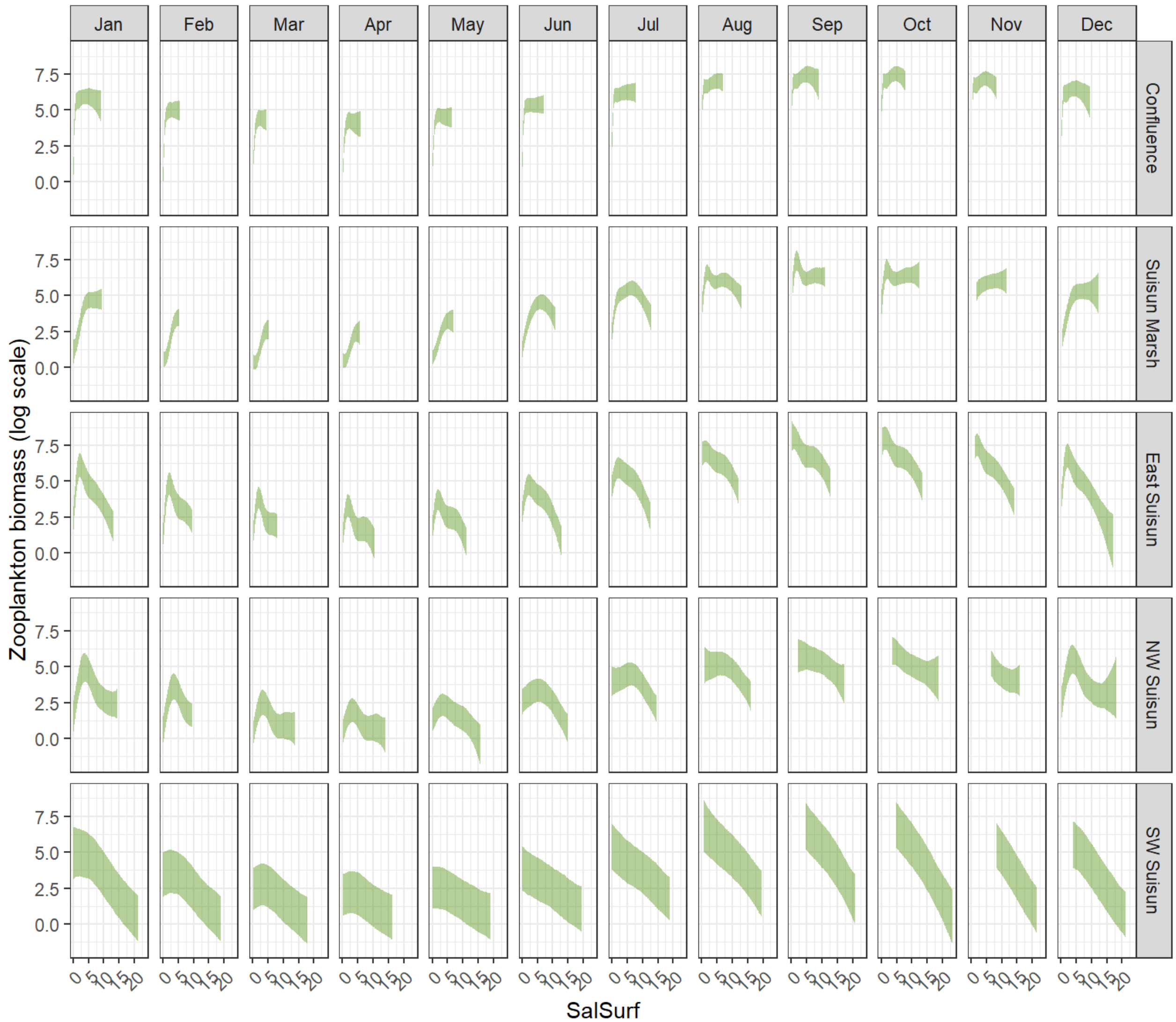


Figure 3. Predicted zooplankton densities (log scale: mgC/m³) by salinity (PSU), subregion (rows), and month (columns) for *Pseudodiaptomus forbesi*. Bands represent 95% credible intervals of estimated relationships between zooplankton and salinity.

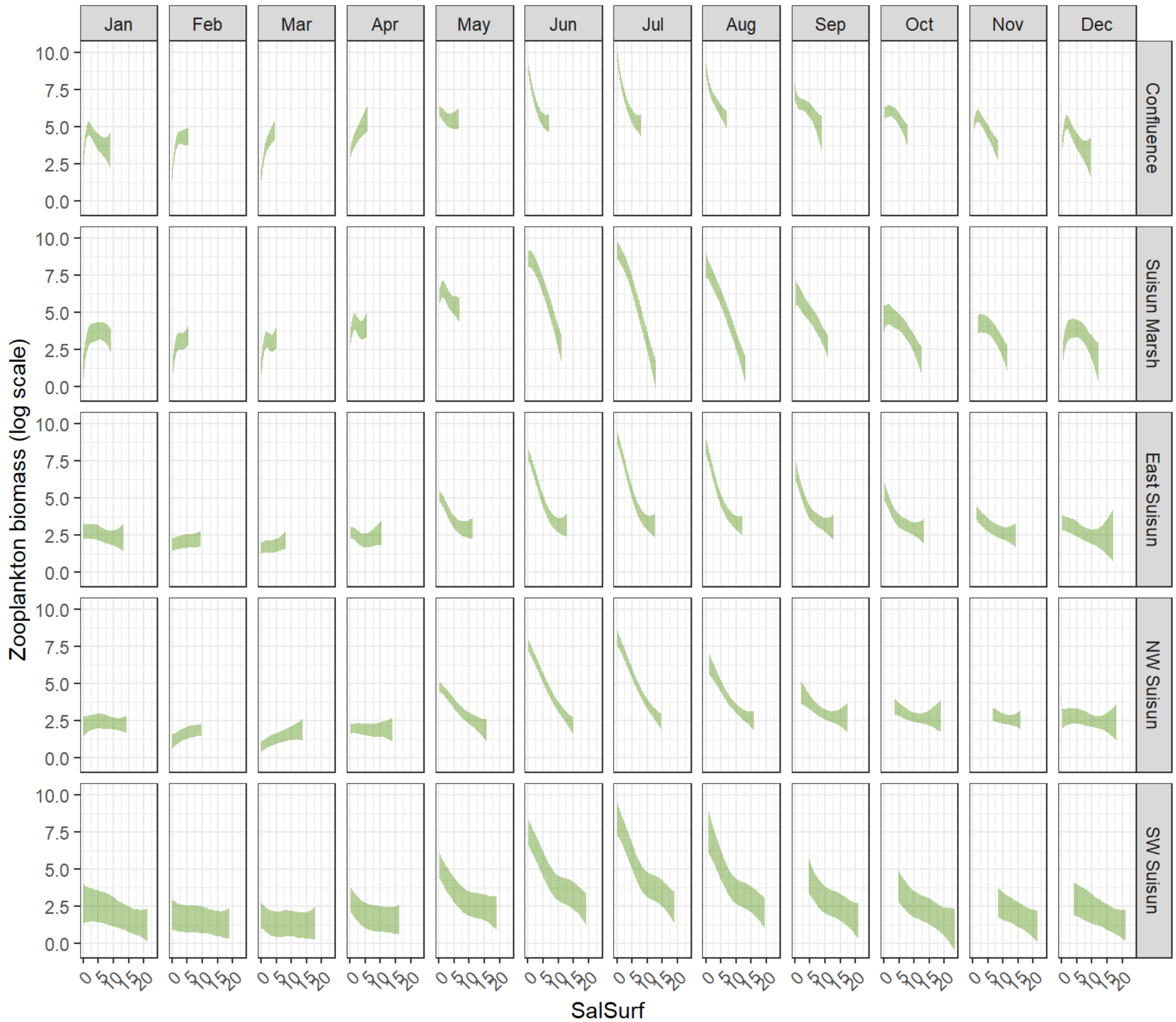


Figure 4. Scalars (log scale) generated from model predictions (scenario / baseline) of zooplankton densities for *Acartiella sinensis* across years by month and subregion for the Fall X2 high bookend management action. Bands represent 95% credible intervals around scalars that incorporated uncertainty of model predictions.

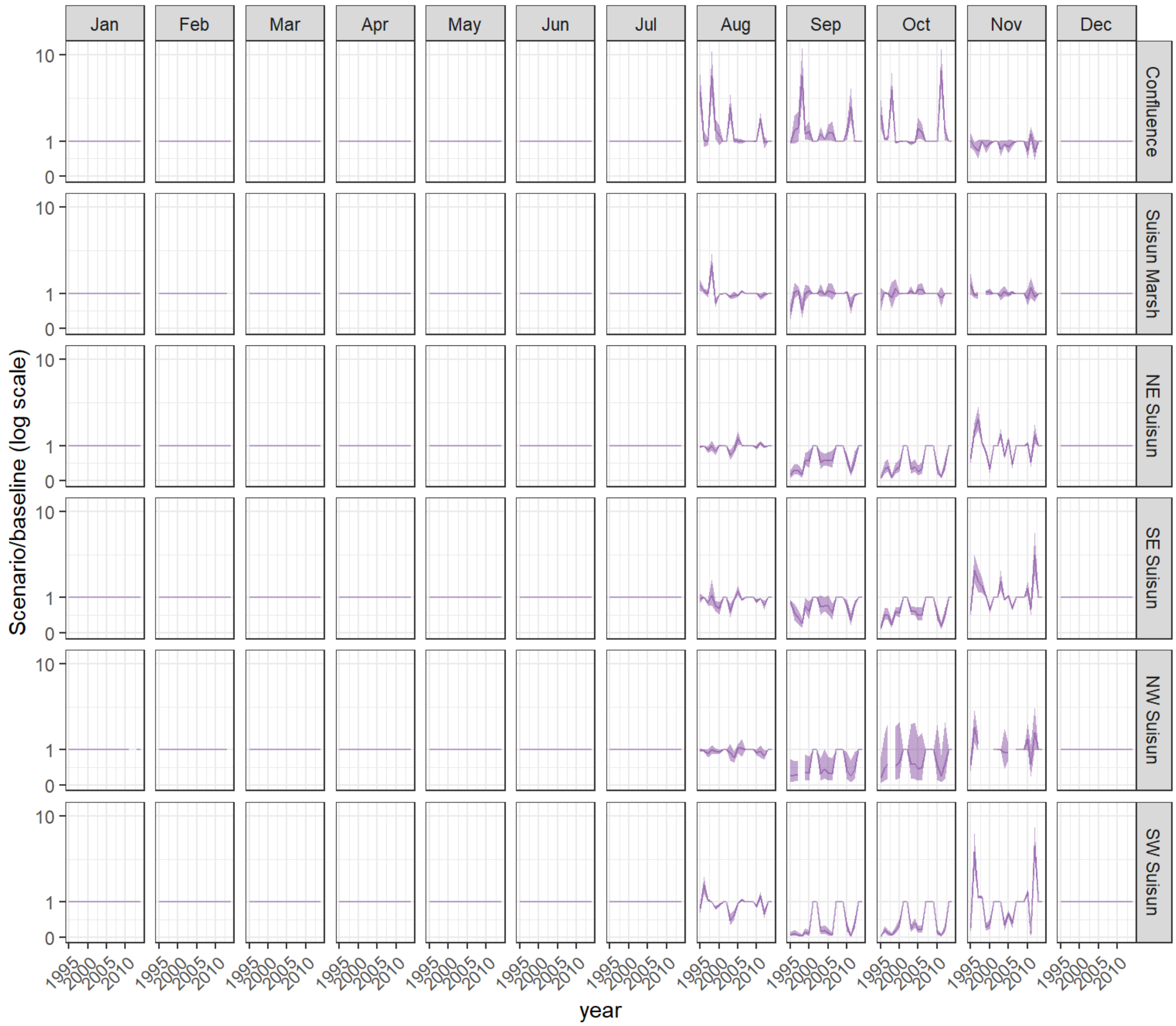
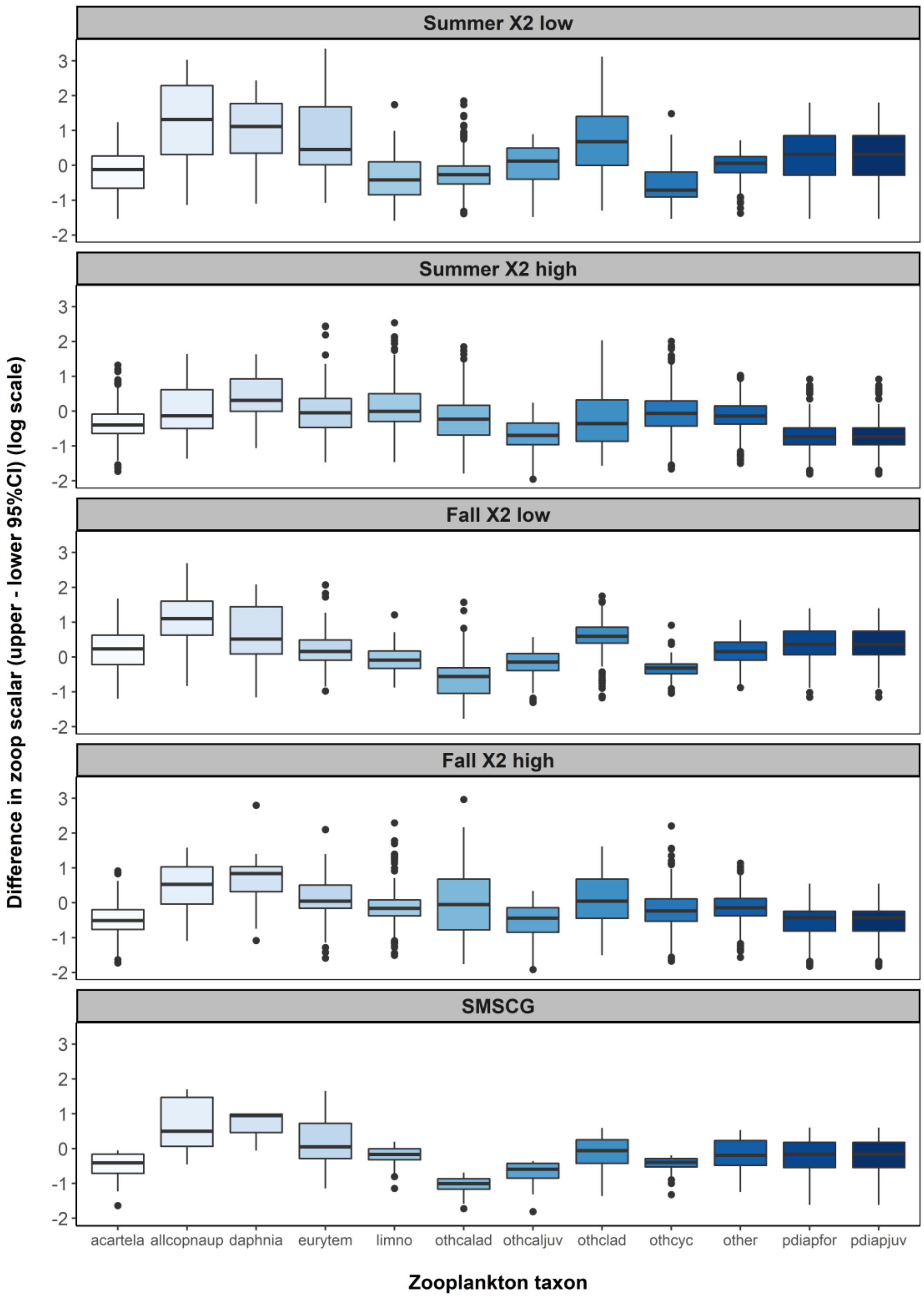


Figure 5. Difference in scalars (scenario / baseline zooplankton densities) generated from the upper and lower 95% credible interval estimates from the Bashevkin zooplankton-salinity model, by zooplankton taxa and management scenario. Pairwise differences were calculated between upper and lower 95% CI estimates for each taxon, subregion, month, year, and action before being summarized into boxplots. Lower values indicate less difference (uncertainty) between upper and lower predicted scalars for a given taxon and action; higher values indicate greater uncertainty between upper and lower predicted scalars.



4 References

Bashevkin, S.M., R. Hartman, M. Thomas, A. Barros, C.E. Burdi, A. Hennessy, T. Tempel, and K. Kayfetz. 2022a. Five decades (1972–2020) of zooplankton monitoring in the upper San Francisco Estuary. PLOS ONE 17: e0265402. doi:10.1371/journal.pone.0265402.

Bashevkin, S.M., S.E. Perry, and E.B. Stumpner. 2022b. Six decades (1959-2021) of water quality in the upper San Francisco Estuary: an integrated database of 15 discrete monitoring surveys in the Sacramento San Joaquin Delta, Suisun Bay, Suisun Marsh, and San Francisco Bay ver 5. Environmental Data Initiative. https://doi.org/10.6073/pasta/c9b3da65a8c89cbfa6fc28d26f938c22.

Miller, D.L. 2021. Bayesian views of generalized additive modelling. arXiv. https://doi.org/10.48550/arXiv.1902.01330

R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Smith, W.E., 2022. A delta smelt Individual-Based Life Cycle Model in the R statistical environment (Technical Note). Prepared for CSAMP Delta Smelt SDM Technical Working Group (TWG). 15 July 2022.

Sommer T., C. Armor, R. Baxter, L. Brown, M. Chotkowski, S. Culberson, F. Feyrer, M. Gingras, B. Herbold, W. Kimmerer, et al. 2007. The Collapse of Pelagic Fishes in the Upper San Francisco Estuary. Fisheries. 32(6):270–277.

Sommer, T., R. Hartman, M. Koller, M. Koohafkan, J.L. Conrad, M. MacWilliams, A. Bever, C. Burdi, and M. P. Beakes. 2020. Evaluation of a large-scale flow manipulation to the upper San Francisco Estuary: Response of habitat conditions for an endangered native fish. Plos ONE 15(10). doi:https://doi.org/10.1371/journal.pone.0234673.

Winder, M., and A.D. Jassby. 2011. Shifts in Zooplankton Community Structure: Implications for Food Web Processes in the Upper San Francisco Estuary. Estuaries and Coasts. 34(4):675–690. doi:10.1007/s12237-010-9342-x.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Series B Stat. Methodol. 73:3–36. doi:10.1111/j.1467-9868.2010.00749.x.

Wood, S.N., N. Pya, and B. Säfken. 2016. Smoothing parameter and model selection for general smooth models. J. Am. Stat. Assoc. 111: 1548–1563. doi:10.1080/01621459.2016. 1180986.

DSM TN 1. Delta Smelt Life Cycle Model Data Files

Lara Mitchell and Ken Newman

U.S. Fish and Wildlife Service

July 28, 2020

DRAFT

Many data sets have been assembled in support of the Delta Smelt Life Cycle Model. These data sets, referred to here as raw data sets, have been used to create clean data sets designed specifically for model fitting purposes. The clean data sets are produced using a series of R scripts that take the raw data files as input, carry out various data cleaning procedures, and save the resulting clean data sets as both CSV files and R objects with the same root file names. For brevity, only the CSV file names are presented here. This document provides descriptions of the raw data sets, clean data sets, and R cleaning scripts, organized by data type. All files and procedures described here are subject to change.

Acronyms

DSLCM - Delta Smelt Life Cycle Model

CDFW - California Department of Fish and Wildlife
EMP - Environmental Monitoring Program
FWS - U.S. Fish and Wildlife Service
IEP - Interagency Ecological Program for the San Francisco Estuary

Bay - Bay Study Midwater Trawl, conducted by CDFW
Chipps - Chipps Survey, conducted by FWS
FMWT - Fall Midwater Trawl Survey, conducted by CDFW
SKT - Spring Kodiak Trawl Survey, conducted by CDFW
STN - Summer Townet Survey, conducted by CDFW
Twentymm - 20mm Survey, conducted by CDFW

CCFB - Clifton Court Forebay
CVP - Central Valley Project water management project
SWP - State Water Project water management project

1 Fish Survey Data

The R script DataCleaner_FishSurveys.r creates separate clean catch data sets for the Bay, FMWT, SKT, STN, and Twentymm fish surveys. For each survey, the script merges four raw data sets, categorized as station, catch, length, and tide data, to produce a standardized data set containing delta smelt catch, age, and fork length information, select environmental variables, and tide information. The clean data sets are "updated" versions of the raw catch data sets, designed to have one record per unique combination of sampling date and sampling station.

The script DataCleaner_Chipps.r creates a clean catch data set for the Chipps survey. The Chipps clean data set is similar to those of the other surveys except that each record represents a unique date-time as opposed to a unique date-station.

1.1 Raw Data

1.1.1 Station Data

The raw station data sets are stored in the Excel files listed below, with the corresponding survey name or abbreviation included in the file name. Copies of these files were saved in CSV format for reading in to R. Within a data set, each row represents a different sampling station for the corresponding survey. The columns provide the three digit station code (defined by CDFW), the latitude and longitude of the station in decimal degrees, and the region and subregion in which the station falls (see documentation on the DSLCM for details on how region and subregion are defined). Part of the STN station file is shown in Table 1. The other station files are similar. There is no station file for the Chipps survey because it takes place at only one location: Chipps Island.

<u>Raw Data Files</u>

Bay_Stations_coords.xlsx
FMWT_Stations_coords.xlsx
SKT_Stations_coords.xlsx
STN_Stations_coords.xlsx
Twentymm_Stations_coords.xlsx

Table 1: An example of a station data set.

Station	LatDD	LonDD	Region	SubRegion
323	38.05	-122.28	Far West	East San Pablo Bay
328	38.06	-122.35	Far West	Mid San Pablo Bay
329	38.06	-122.30	Far West	East San Pablo Bay
334	38.08	-122.34	Far West	Mid San Pablo Bay
335	38.07	-122.32	Far West	East San Pablo Bay
336	38.06	-122.28	Far West	East San Pablo Bay

1.1.2 Fish Survey Catch Data

The raw fish survey catch data sets are stored in the Excel files listed below, with the corresponding survey name included in the file name. Copies of these files were saved in CSV format for reading in to R.

<u>Raw Data Files</u>

BayStudy_MWT_1980-2014_FishMatrix.xlsx
Chipps_Catch_1976_2011.xlsx
Chipps_Catch_2011_2016.xlsx
FMWT_1967-2015_Catch Matrix updated.xlsx
Mitchell_SKT_2016Update.xlsx
LMithcell_DatReq_STN_2016.xls
Mitchell_20_mm_CatchMatrix_1995_2016.xlsx

The raw Bay catch file was downloaded from the CDFW ftp site in February 2016.

The Chipps catch files were provided by Jonathan Speegle (FWS) on February 19, 2016. The FMWT file was provided by Sarah Finstad (CDFW) on February 12, 2016.

The SKT file was provided by Lauren Damon on October 27, 2016.

The STN file was provided by Felipe La Luz (CDFW) in October 2016. Note that the STN data file appears to only contain core index stations (plus station 340, which has been sampled since 1978), while other data files contain both index and non-index stations.

The Twentymm catch file was provided by Lauren Damon (CDFW) on September 9, 2016.

Each catch data set describes the fish species composition of the survey on a per-tow basis. In the case of Bay, FMWT, SKT, STN, and Twentymm, each record in the data set corresponds to a single tow and contains information on when and where the tow took place, what species were caught, how many individuals of each species were caught, and what the physical conditions were like at the time of sampling. Table 2 describes a set of data fields common to many of the catch files; some of these field names vary between surveys (for example, Time vs. TimeStart). Further details on how the data are collected or calculated are available through the CDFW website.

The first Chipps Excel file contains two worksheets, labeled "Chipps Island Trawls" and "Chipps Island Larval DSM remove." The first worksheet consists of a data set containing count and length data for multiple fish species. Each record describes the number of individuals of a given species and size caught in a given tow on a given date. Descriptions of the fields of interest are given in Table 3, with further details available through the Lodi FWS website. If no organisms were caught at a given date-time, the record appears in the data set with a blank value in the Organism field. It should be noted that Chipps is carried out at one location and hence does not sample from a range of stations like the other surveys. Between 1976 and roughly 1996, larval delta smelt (defined as delta smelt less than 25 mm in fork length) were counted and recorded as part of the Chipps survey. The second worksheet consists of the same data as in the first worksheet except with pre-1996 records identified as "larval delta smelt" removed. Some uncertainty remains about whether any records in this data set, in particular those without length information, still include larval delta smelt. The second Chipps Excel file contains data from later years not included in the first file.

Table 2: A partial summary of the data fields in the SKT, FMWT, Bay, Twentymm, and STN raw catch files. ✓ indicates that the field is present, X that it is absent.

Field Name	Description	Survey				
		SKT	FMWT	Bay	Twenty-mm	STN
Date	Date of tow.	✓	✓	✓	✓	✓
TimeStart	Time at start of tow.	✓	✓	✓	✓	✓
Survey	A number describing the progression of the survey on a biweekly or monthly basis.	✓	✓	✓	✓	✓
Station	Station number.	✓	✓	✓	✓	✓
Tow	For Bay: an indication of tow "quality". For the other surveys: the unique tow number at a given station, on a given date.	X	X	✓	✓	✓
Volume	Estimate of water volume sampled (m^3).	✓	✓	✓	✓	✓
TowDirection	Tow direction code: 1=with current, 2=against current, 3=unknown (during slack).	✓	✓	✓	X	X
Secchi	Secchi depth (For Chipps: m; other surveys: cm).	✓	✓	✓	✓	✓
CondSurf	Specific conductivity of the first foot of water from the surface (µS).	✓	✓	X	✓	✓
CondBott	Specific conductivity of the first foot of water from the bottom (µS).	X	✓	X	✓	✓
TempSurf	Water temperature (°C).	✓	✓	✓	✓	✓
Tide	Tide codes (Bay: 1=flood, 2=ebb, 3=low slack, 4=high slack; other surveys: 1=high slack, 2=ebb, 3=low slack, 4=flood).	✓	✓	✓	✓	✓
Depth	Depth of water at the station (Bay: m; other surveys: ft).	✓	✓	✓	✓	✓
SalinSurf	Salinity (ppt) for first meter of water column.	X	X	✓	X	X
delta.smelt	Number of delta smelt in the tow.	✓	✓	✓	✓	✓

Table 3: A partial summary of the data fields in the Chipps raw data set.

Field Name	Description
SampleDate	Date of tow.
TimeStart	Time at start of tow.
TowNumber	The unique tow number on a given date.
TowDirection	Tow direction code: U = upstream, D = downstream.
Secchi	Secchi depth (m).
WaterTemp	Water temperature (°C).
Volume	Estimate of volume sampled (m^3).
Organism	Organism code (DSM = delta smelt).
ForkLength	Fork length (mm).
Count	Number of fish in the given tow that have the given organ-ism code and fork length.

1.1.3 Delta Smelt Length Data

The raw length data files contain fork length measurements on delta smelt caught in the fish surveys. For each tow, the number of smelt measured for length is usually (but not always) equal to the total number caught. The Excel files containing the raw length data are listed below, with the corresponding survey name included in the file name. Copies of these files were saved in CSV format for reading in to R. All delta smelt fork lengths are in millimeters.

Raw Data Files

Bay_DSM_Lengths_1980_2014.xlsx
FMWT_DSM_Lengths_1967_2015.xlsx
Mitchell_SKT_2016Update_DSM_Lengths.xlsx
STN_DSM_Lengths_1959_2016.xlsx
Mitchell_20-mm_DS_Lengths_1995_2016.xlsx

The Bay length file was generated by Lara Mitchell using a copy of the Bay Study Access data base obtained from the CDFW ftp site in February 2016. This file contains fork length measurements for delta smelt caught by the Bay midwater trawl. Each record represents a unique tow and fork length combination, with the field Frequency giving the total number of delta smelt represented by that record.

The FMWT length file was provided by Sarah Finstad on January 15, 2016. It has a structure that is similar to the Bay file, except that the frequency column is labelled LengthFrequency.

The STN length file was provided by Felipe La Luz by in October 2016. It has the same structure as the FMWT length file, and also uses the field name LengthFrequency. Fork length measurements are not available prior to 1973.

The Twentymm length file was provided by Lauren Damon on September 9, 2016. It has the same structure as the FMWT file, but uses the frequency field name CountOfLength.

The SKT length file was provided by Lauren Damon on October 27, 2016. It has a separate record for each individual delta smelt caught in the survey. In addition to fork length data, it contains sex and reproductive information.

Chipps delta smelt fork length information is represented in the ForkLength field of the Chipps raw catch data file, rather than in a separate length file.

1.1.4 Tide Data

The tide data sets listed below were created by Chandra Chilmakuri (CM2H-Hill), and contain tidal information from the times and locations where fish surveys took place. The primary data fields are summarized in Table 4. There exists one record per date-station in the case of Bay, FMWT, SKT, STN, and Twentymm, and one record per date-time in the case of Chipps. Versions of these files were saved in CSV format for reading in to R.

Raw Data Files

Bay_Tide_Vars.xlsx
Chipps_Tow_Tide_Vars.xlsx
FMWT_Tide_Vars.xlsx
SKT_Tide_Vars.xlsx
STN_Tide_Vars.xlsx
Twentymm_Tide_Vars.xlsx

Table 4: A partial summary of the data fields in the Bay, Chipps, FMWT, SKT, STN, and Twentymm raw tide data sets.

Field Name	Description
Date	Fish survey sample date.
TimeStart	Fish survey sample time.
Region	Fish survey region designation.
Station	Fish survye station number.
TideStage	Tide level (in feet) relative to NGVD29.
HighType	Closest peak high tide: HH = High High, LH = Low High.
Time-to-High-Min	Difference between sampling time and the closest peak high tide time (min).
LowType	Closest peak low tide: HL = High Low, LL = Low Low.
Time-to-Low-Min	Difference between sampling time and the closest peak low tide time (min).
TideVelocity	Instantaneous Velocity (ft/s).
Ebb-Type	Closest peak ebb velocity: HE = High Ebb, LE = Low Ebb.
Time-to-Ebb-Min	Difference between sampling time and the closest peak ebb velocity time (min).
FloodType	Closest peak flood velocity: HF = High Flood, LF = Low Flood.
Time-to-Flood-Min	Difference between sampling time and the closest peak flood velocity time (min).
Time-to-Slack-Min	Difference between sampling time and the closest slack velocity time (min).

1.2 Clean Data

The six clean catch data files are listed below. The field names and units used in the clean catch files are described below that. Additional columns are added when a survey conducts replicate tows; see the section on aggregating replicate tows.

Clean Data Files

Bay_80_14.csv
Chipps_78_15.csv
FMWT_67_15.csv
SKT_02_15.csv
STN_59_16.csv
Twentymm_95_16.csv

Clean Data Field Names

Date - Sample date.
Year - Sample year.
Month - Sample month.
Survey - Survey number.
Station - Station code.
TimeStart - Sample time.
Volume - Volume of water sampled (m^3).
TowDirection - Tow direction string (With Current, Against Current, or Neither).
Region - Sampling region, as defined in the DSLCM.
SubRegion - Sampling subregion.
Lat - Latitude (degree decimal).
Lon - Longitude (degree decimal).
Secchi - Secchi depth (cm).
CondSurf - Surface conductivity (µS).
TempSurf - Surface temperature (°C).
SalinSurf - Surface salinity (ppt).
CondBott - Bottom conductivity (µS).
Tide - Tide string (High Slack, Ebb, Low Slack, or Flood).
Depth - Depth to bottom (ft).
Inland_silverside - Number of inland silverside caught.
Striped_bass_age0 - Number of age 0 striped bass caught.
Striped_bass_age1_plus - Number of age 1+ striped bass caught.
Striped_bass_all - Total number of striped bass caught.
Longfin_Smelt - Total number of longfin smelt caught.
Threadfin_Shad - Total number of threadfin shad caught.
Tridentiger_spp - Total number of Tridentiger gobies caught.
delta.smelt - Number of delta smelt caught.
delta.smelt.age0 - Number of age 0 delta smelt caught.
delta.smelt.age1 - Number of age 1 delta smelt caught.
Age0_n_L - Number of age 0 delta smelt measured for fork length.
Age0_L_bar - Age 0 delta smelt mean fork length (mm).
Age0_s_L - Age 0 delta smelt fork length standard deviation (mm).
Age1_n_L - Number of age 1 delta smelt measured for fork length.
Age1_L_bar - Age 1 delta smelt mean fork length (mm).
Age1_s_L - Age 1 delta smelt mean fork length standard deviation (mm).

Age0_L_min - Minimum age 0 delta smelt fork length (mm).
Age0_L_max - Maximum age 0 delta smelt fork length (mm).
Age1_L_min - Minimum age 1 delta smelt fork length (mm).
Age1_L_max - Maximum age 0 delta smelt fork length (mm).
TideStage - Tide level (converted from ft to m).
HighType - Closest peak high tide: HH = High High, LH = Low High.
Time-to-High-Min - Difference between sampling time and the closest peak high tide time (min).
LowType - Closest peak low tide: HL = High Low, LL = Low Low.
Time-to-Low-Min - Difference between sampling time and the closest peak low tide time (min).
TideVelocity - Instantaneous Velocity (converted from ft/s to m/s).
EbbType - Closest peak ebb velocity: HE = High Ebb, LE = Low Ebb.
Time-to-Ebb-Min - Difference between sampling time and the closest peak ebb velocity time (min).
FloodType - Closest peak flood velocity: HF = High Flood, LF = Low Flood.
Time-to-Flood-Min - Difference between sampling time and the closest peak flood velocity time (min).
Time-to-Slack-Min - Difference between sampling time and the closest slack velocity time (min).
Cable.Out - Length of cable let out during tow (ft). Used to calculate EstimatedTowDepth_ft.
EstimatedTowDepth_ft - Estimated maximum depth that the trawl reached during a tow (ft).
Age0_age_in_days - Pseudo age (in days) of an age 0 delta smelt given its catch date and an assumed "cohort-wide" hatch date of March 1st.
Age0_pgt - Estimated probability of the trawl catching an age 0 delta smelt on the given sample date, given an assumed population length distribution.
Age1_age_in_days - Pseudo age (in days) of an age 1 delta smelt given its catch date and an assumed "cohort-wide" hatch date of March 1st.
Age1_pgt - Estimated probability of the trawl catching an age 1 delta smelt on the given sample date, given an assumed population length distribution.

General Procedure:

The general procedure for producing a clean catch data set is described here. Details specific to a given fish survey are included as necessary.

1. Merge Station Data

First, the station and catch data set are merged by station code. Records with stations that are not included in the station data set, and records with stations that are located outside of the four DSLCM regions (Far West, West, North, and South), are removed from the merged data. Next, fields are renamed as necessary so that the merged file has the standardized field names shown above. Fields not originally represented in the catch file are added and filled in with the value NA.

2. Make Survey-Specific Changes:

Bay

The field SalinSurf is used to calculate CondSurf using the following conversion equation: $Conductivity = 178500(1 - e^{-0.01*Salinity})$. This equation may need correcting (Wim Kimmerer, personal communication). The Depth field is converted from meters to feet. The numerical levels of Tide and the numerical levels of TowDirection are changed to descriptive strings for ease of interpretation (see below). The mapping for Tide is: 1 = "Flood", 2 = "Ebb", 3 = "Low Slack", 4 = "High Slack". The mapping for TowDirection the mapping is: 1 = "With Current", 2 = "Against Current", 3 = "Neither".

FMWT

The numerical levels of Tide and the numerical levels of TowDirection are changed to descriptive strings for ease of interpretation. The mapping for Tide is: 1 = "High Slack", 2 = "Ebb", 3 = "Low Slack", 4 = "Flood". The mapping for TowDirection is: 1 = "With Current", 2 = "Against Current", 3 = "Neither".

SKT

Records with survey numbers greater than or equal to 6 are removed from the clean data set. These are special, non-routine surveys. The record from 3/9/2004, survey 3, station 610, at 13:30 has indeterminate Tide and TowDirection values (indicated by 0's). The previous 7 tows from that date have numerical values of 2 for both fields, so the 0's are replaced with 2's. After this, the numerical levels of Tide and the numerical levels of TowDirection are changed to descriptive strings for ease of interpretation. The mapping for Tide is 1 = "High Slack", 2 = "Ebb", 3 = "Low Slack", 4 = "Flood". The mapping for TowDirection is: 1 = "With Current", 2 = "Against Current", 3 = "Neither".

STN

Tow volumes prior to 2003 are unavailable, but an average volume of 735 m^3 has been provided by CDFW in the raw catch file, and this value is also used in the clean data file. Some values of TimeStart are missing; these are left as blank strings. Infrequent 4th tows, indicated by a Tow value of 4, are removed per advice from Julio Adib-Samii (personal communication). The numerical levels of Tide are changed to descriptive strings for ease of interpretation. The mapping for Tide is: 1 = "High Slack", 2 = "Ebb", 3 = "Low Slack", 4 = "Flood". Note that many environmental field variables are imputed in the STN data set (see Table 6).

Twentymm

Some values of TimeStart are missing; these are left as blank strings. Records with survey numbers greater than or equal to 10 are removed from the clean data set. These are special, non-routine surveys. Missing values of the field Depth are filled in with the value 32 (provided by Trishelle Morris) rather than an average value. The numerical levels of Tide are changed to descriptive strings for ease of interpretation. The mapping for Tide is: 1 = "High Slack", 2 = "Ebb", 3 = "Low Slack", 4 = "Flood".

Chipps

The clean Chipps catch data set is structured to have one record for every unique date-startTime combination. The field Secchi is converted from meters to centimeters. The field Region is filled in with the value "West", SubRegion is filled in with "Honker Bay". Station is filled in "Chipps", Lat is filled in with 38.055, and Lon is filled in with -121.9109. Some records have a ForkLength value of 0. These are changed to NA before length and age statistics are calculated. Delta smelt records with fork lengths less than 25 mm or greater than 100 mm are reclassified as "Other Smelt" and hence not used in constructing the Chipps clean data set. The decision to remove fish less than 25 mm is based on a meeting with Matt Dekar, Joseph Kirsch, Jonathan Speegle, and Pat Brandes on September 16, 2013. The decision to remove fish greater than 100 mm is based on the hypothesis that larger delta smelt may have been misidentified in the past (William Bennett, personal communications). See Table 5 for a summary of the records removed based on fork length.

Table 5: Frequency of Chipps delta smelt records removed with fork length > 100 mm or < 25 mm, by year and month.

Month	Year: 1979	1981	1982	1983	1985	1989	1990	1991	1992	1994	1995	1996	1998	1999	2002	2003	2004	2013	Total
January	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
March	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
April	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
May	0	2	0	2	2	1	1	0	1	0	0	1	1	0	0	0	0	0	11
June	1	0	2	0	0	2	0	1	0	0	1	0	0	0	0	0	0	0	7
July	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2
November	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
December	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	4
Total	1	2	2	2	2	3	1	1	1	1	3	1	2	1	1	1	1	2	28

(a) Length > 100mm.

Month	Year: 1979	1983	1984	1993	1994	1995	Total
May	0	0	0	1	0	20	21
June	4	2	7	3	4	157	177
July	0	0	0	0	0	26	26
Total	4	2	7	4	4	203	224

(b) Length < 25 mm.

3. Include Predator and Competitor Fields

Fields containing counts of striped bass (Morone saxatilis), inland silverside (Menidia beryllina), longfin smelt, threadfin shad, and Tridentiger goby are included in the clean data set. Some raw catch files have separate fields for age 0, 1, 2, and/or 3 striped bass. In the clean data set, there are separate fields for age 0 striped bass, age 1+ striped bass, and total striped bass.

4. Impute Volume and Environmental Fields

At this point, attempts are made to fill in values of Volume, Secchi, CondSurf, TempSurf, SalinSurf, CondBott, and Depth that are either missing or physically unrealistic. For depth, values of 0 are considered physically unrealistic and replaced. Similarly, secchi values of 0 are replaced. For FMWT, volumes less than 3000 are also replaced (Sarah Finstad, personal communication).

The general procedure is to try substituting with mean values calculated by date-station, then by date-subregion, then date-region, year-month-subregion, year-month-region, month-region, and finally by year- region. Averages calculated by date-station are tried first because this method uses data from records that are close to the missing record in both time and space. This method can be used when multiple tows were carried out at a single date-station. For records that are still missing or physically unrealistic after this, an attempt is made to substitute mean values calculated per date-subregion. In this case, the available data are still close to the missing record in time, but cover a wider geographic range. Next, mean values calculated per date-region are substituted, when available. If at this point values are still missing, the time frame is expanded to the same month as the missing record (within the same year), and substitutions are carried out using means calculated per year-month-subregion, then per year-month-region. Finally, means calculated per month-region (across years) and year-region (across months) are tried.

Missing CondBott values are first filled in using predicted values from a linear model for bottom conductivity as a function of the corresponding surface conductivity. After this, any missing values are filled in using the procedure described above with the seven alternative average values.

For Chipps, missing volumes are filled in with mean volumes calculated by year-month. In cases where no data are available from the same year-month as the missing value, volumes from the two adjacent months are used to calculate the mean.

Table 6: Number of missing or physically unrealistic values imputed by fish survey (row) and imputation method (column). Fields not shown either did not need to have values imputed, or did not exist in the raw catch file to begin with. Dashes indicate that all missing or invalid values were successfully replaced.

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month	Linear Model
Bay									
CondSurf	0	51	17	16	26	183	-	-	-
TempSurf	0	53	22	17	26	171	-	-	-
SalinSurf	0	51	17	16	26	183	-	-	-
Chipps									
Volume	0	0	0	0	0	0	0	79	-
Secchi	788	0	0	381	-	-	-	-	-
TempSurf	162	0	0	159	-	-	-	-	-
FMWT									
Volume	300	513	70	56	40	6701	22	-	-
Secchi	1	191	54	33	28	16	-	-	-
CondSurf	1	58	43	15	69	16	-	-	-
TempSurf	0	41	23	39	57	7	-	-	-
CondBott	0	0	0	0	0	0	0	0	15870
Depth	0	11	9	4	24	4583	22	-	-
SKT									
Secchi	2	1	7	-	-	-	-	-	-
CondSurf	2	-	-	-	-	-	-	-	-
TempSurf	2	-	-	-	-	-	-	-	-
Depth	0	7	4	-	-	-	-	-	-
STN									
Volume	1	-	-	-	-	-	-	-	-
Secchi	0	55	61	239	7	3267	-	-	-
CondSurf	0	11	25	84	41	3327	-	-	-
TempSurf	0	36	39	255	47	4494	-	-	-
CondBott	0	0	0	0	0	0	0	0	11026
Depth	0	42	44	162	60	4290	-	-	-
Twentymm									
Volume	10	-	-	-	-	-	-	-	-
Secchi	0	15	69	101	21	-	-	-	-
CondSurf	0	9	37	108	-	-	-	-	-
TempSurf	0	15	22	60	-	-	-	-	-
CondBott	373	-	-	-	-	-	-	-	-

5. Create Selectivity and Catchability Fields

The last five fields of the clean catch data set are intended to be used for analyzing gear selectivity for delta smelt. They are based on preliminary analyses and are subject to change.

The field EstimatedTowDepth_ft gives an estimate of the maximum depth in feet reached by the trawl during a tow. For Twentymm, this is calculated as (3.937/25) * Cable.Out – 8.3, where Cable.Out is the number of feet of cable let out (from the raw catch data set) and it is estimated that for every 25 feet of cable let out, the trawl drops 3.937 feet. 8.3 is the average distance, in feet, from the block to the water surface for the Twentymm boats (Trishelle Morris, personal communication). For FMWT, the formula is (3.937/25) * Cable.Out – 6.67 (Sarah Finstad, personal communication). Bay is operated such that the net descends to close to the station depth, but not so deep that the net plows the substrate (Kathy Hieb, personal communication). For example, the tow depth at a 20 foot-deep station is roughly 18 to 19 feet. The net does not fish below 40 feet though, so if a station depth is greater than 40 feet, the tow depth will remain 40 feet (Kathy Hieb, personal communication). In the Bay clean data set, EstimatedTowDepth_ft is set equal to Depth except in cases where Depth is greater than 40 feet, in which case 40 feet is used instead.

Cable.Out values that are missing or 0 are imputed using the process described in step 2. For FMWT, one cable value of 15 is also replaced because it leads to a negative tow depth. For Twentymm, cable values of less than 53 feet are replaced because they lead to negative tow depths. See Table 7 for a summary of imputed Cable.Out values. In cases where EstimatedTowDepth_ft ends up being greater than Depth, the value of EstimatedTowDepth_ft is replaced by the value of Depth; see Table 8.

The fields Age0_age_in_days and Age1_age_in_days give the pseudo ages (in days) of age-0 and age-1 delta smelt based on its catch date and assuming a "cohort-wide" hatch date of March 1st. The fields Age0_pgt and Age1_pgt give estimates of the probabilities of catching age-0 or age-1 delta smelt on that date given an assumed population length distribution. These values are currently coming from the file prob.catch.bygear.dayD.df_3_7_2016.csv. Details on how these values are calculated will be coming soon.

Table 7: Number of missing values of Cable.Out imputed by fish survey (row) and imputation method (column). Note that when Cable.Out is imputed, EstimatedTowDepth_ft is also necessarily imputed.

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month
FMWT								
Cable.Out	1544	1522	327	35	22	13671	482	-
STN								
Cable.Out	0	67	158	173	66	2995	-	-
Twentymm								
Cable.Out	3	66	44	-	-	-	-	-

Table 8: Number of EstimatedTowDepth_ft values replaced with the corresponding value of Depth.

Survey	Number Replaced
FMWT	3434
STN	794
Twentymm	145

6. Merge Length Data

The raw length data set is used to calculate age and length-related fields in the clean data set. For SKT, which has a separate record for every fish, an age assignment key is used to assign an age (0 or 1) to each fish based on fork length and month-of-catch. The key, shown in Table 9, was developed by CDFW (Steve Slater, personal communication). The records are then aggregated by unique tow, and the fields Age0_n_L through Age1_L_max are calculated for each tow. This aggregated length data set is then merged with the clean data set. The fields delta.smelt.age0 and delta.smelt.age1 represent the total number of age 0 and age 1 delta smelt caught, and are calculated by multiplying the total delta smelt catch in the tow by the proportion of age 0 and age 1 individuals represented in the length data for that tow (e.g., Age0_n_L/(Age0_n_L + Age1_n_L) would be the proportion of age 0 smelt). The same process is used for Bay, FMWT, STN, and Twentymm, except first each record in the length data set is duplicated according to the frequency column in order to produce a data set with the same structure as the SKT length data set.

As described previously, the Chipps raw catch file has a separate record for each date-time-species-length combination. As part of the data cleaning process, length values of zero are first changed to NA. Ages are then assigned to each delta smelt record using the CDFW age-assignment key, and length statistics are calculated on a date-time basis and merged with the clean data set.

Table 9: Delta smelt age assignment key. The numbers indicate the cut-off length (in mm) used to distinguish between age 0 and age 1 fish in the given month. Individuals below the cut-off length are taken to be age 0; individuals at or above this length are taken to be age 1.

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
40	50	50	50	50	60	65	70	75	80	80	80

7. Impute Ages

When length information is not available to calculate the age fields delta.smelt.age0 and delta.smelt.age1, values are imputed using the following procedure. A mean proportion of age 0 smelt is calculated us- ing the same procedure described in step 2. That is, first an attempt is made to calculate a mean by date-station, then if that is not possible, an attempt is made to calculate a mean by date-subregion, and so on. The imputed value of delta.smelt.age0 is given by the product of the mean age 0 proportion and delta.smelt, rounded to the nearest integer. The imputed value of delta.smelt.age1 is given by delta.smelt - delta.smelt.age0. See Table 10 for a summary of imputed age information.

For FMWT, the following age 0 proportions are used whenever calculated values are not available, including years prior to 1975: January: 0, February: 0, March: 0, May: 0, September: 0.9, October: 1, November: 1, December: 1. Additionally, if the calculated values in September, October, or November are less than 0.9, the value 0.9 is used instead. All of these values were provided by Dave Contreras (personal communication).

8. Aggregate Replicate Tows

The Twentymm and STN surveys typically take three replicate tows at a given station on a given date. Some limited tow replication also takes place during SKT sampling. In the clean data sets, these replicate tows are aggregated to form one unique record per date-station. The value of delta.smelt for the aggregated record is given by summing the values of this field across the replicate tows; similarly for delta.smelt.age0, delta.smelt.age1, Volume, Inland_silverside, Striped_bass_age0, Striped_bass_age1_plus, Striped_bass_all, Longfin_Smelt, Threadfin_Shad, and Tridentiger_spp.

Table 10: Number of missing values of delta.smelt.age0 and delta.smelt.age1 imputed by fish survey (row) and imputation method (column). Note that when delta.smelt.4age0 is imputed, delta.smelt.age1 is also imputed.

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month
Chipps								
delta.smelt.age0/1	461	0	0	394	0	268	-	-
FMWT								
delta.smelt.age0/1	0	56	19	23	20	1153	6	-
STN								
delta.smelt.age0/1	16	15	14	57	4	1726	-	-
Twentymm								
delta.smelt.age0/1	6	-	-	-	-	-	-	-

The value of EstimatedTowDepth_ft for the aggregated record is given by the mean of the replicate tow depths. The fields Age0_n_L through Age1_L_max are recalculated at the date-station level using the length data set. All other fields, including TimeStart and TowDirection, are taken from the *first* tow record.

When tows are aggregated, additional fields are added to the clean data set in order to preserve catch infor- mation from the replicate tows. Let n be the maximum number of replicates conducted for any date-station combination. These additional fields have the same names as the aggregated fields except with .tow*i* ap- pended to the end, where $i = 1, \ldots, n$. For example, the fields delta.smelt.tow*i*, delta.smelt.age0.tow*i*, and delta.smelt.age1.tow*i* indicate the total number of delta smelt, the number of age 0 delta smelt, and the number of age 1 delta smelt caught in tow i, respectively.

9. Impute Lengths

After any replicate tows are aggregated, attempts are made to imput missing values of Age0_L_bar and Age1_L_bar. The process is the same as that used in step 2. See Table 11 for a summary of imputed length information. If values are unable to be imputed, they are left as NA. No attempts are made to impute sample sizes (Age0_n_L, Age1_n_L) or standard deviations (Age0_s_L, Age1_s_L).

10. Merge Tide Data

At this point, the tide data set is merged with the clean data set. The field TideStage is converted from feet to meters, and TideVelocity is converted from ft/s to m/s.

Table 11: Number of missing values of Age0_L_bar and Age1_L_bar) imputed by fish survey (row) and imputation method (column).

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month
Chipps								
Age0_L_bar	131	0	0	306	0	263	-	-
Age1_L_bar	312	0	0	183	0	164	-	-
FMWT								
Age0_L_bar	0	50	15	13	18	815	5	-
Age1_L_bar	0	9	3	6	2	590	2	-
STN								
Age0_L_bar	0	7	5	23	2	826	-	-
Age1_L_bar	0	0	2	11	0	121	-	-

2 Average Temperature and Secchi Data

The clean Bay, Chipps, FMWT, SKT, STN, and Twentymm data sets were used to calculate mean water temperature and secchi values for every combination of Year-Month-Region between January 1980 and December 2015. The Year, Month, Region, TempSurf, and Secchi fields of the six clean fish survey data sets were combined into one data frame which was then used to calculate the average temperature and secchi values. Hence, these averages are calculated across survey type, sampling date, and sampling location within a given Year-Month-Region. Missing mean values were imputed by averaging over averages with the same Month and Region (i.e., by averaging across years). 145 mean temperatures and 145 mean secchis were imputed. The resulting data set contains the fields Region, Month, Year, MeanTemperature, and MeanSecchi, and has a separate record for each Year-Month-Region combination. This data set is created with the script Create_FishSurvey_TempSecchi.r, and saved in the file Mean_Temp_Secchi.csv. The mean temperature and secchi values are also stored in individual 3D arrays in the files Mean_Temp_3Darray.R and Mean_Secchi_3Darray.R.

Data collected in the Mid San Pablo Bay subregion *were* included in these calculations, which is at odds with the next two sections. I want to discuss this with everyone before making any changes, though.

3 Predator and/or Competitor Indexes

The clean FMWT, SKT, STN, and Twentymm data sets were used to calculate indexes of abundance for age 0 striped bass, age 1+ striped bass, inland silverside, threadfin shad, and Tridentiger goby by year-month- region using a stratified ratio-expansion procedure. See the technical note "TN2_Design_Based_Estimates_of_Delta_Smelt_Abundance" for a general description of the procedure. Note that the resulting values are indexes of abundance because we did not try to account for gear selectivity or fish availability/catchability. The calculations are done by the script Create_FishSurvey_PredCompetitor.r, and the resulting data set is saved in the file FishSurvey_PredCompetitor_long.csv, which has fields Calendar_Year, Month, Region, Gear, Species, and Index. The abbreviated name SBAge0 is used for age 0 striped bass, SBAge1Plus is used for age 1+ striped bass, ISS is used for inland silverside, TFS is used for threadfin shad, and TriGoby is used for Tridentiger goby. The file FishSurvey_PredCompetitor_wide.csv contains a wide-formatted version of the data set with field names Calendar_Year, Month, Region, Index_SBAge0_TMM, Index_ISS_TMM, etc.

Data collected in the Mid San Pablo Bay subregion were not included in these calculations.

4 Mean Length Data

The script Create_FishSurvey_MeanLength.r uses fish survey length data to calculate mean fish lengths in a given year-month (calculated over the stations sampled in that year-month). It produces the data files FishSurvey_MeanLength.csv and FishSurvey_MeanLength_cohort.csv. Some of the average lengths are adjusted for gear selectivity.

Mean lengths are calculated separately for age-0 and age-1 delta smelt. A typical field in the data file looks like this: MeanLength_TMM_DSM_Apr0_adj, where TMM indicates that the lengths came from the 20mm Survey, DSM indicates that the species is delta smelt, Apr0 indicates that it is the mean length of age 0 delta smelt in April, and adj means that 20mm gear selectivity estimates were used to try to adjust for gear selectivity when calculating the mean. MeanLength_TMM_DSM_Apr0_unadj would be the version without gear selectivity adjustments.

Threadfin shad (TFS) and Tridentiger goby (TriGoby) mean lengths are now also included in the clean length files. A typical field name looks like this: MeanLength_STN_TFS_Jul. No attempts are made to account for gear selectivity or separate out juvenile and adults. In the data file organized by cohort year, mean TFS and TriGoby lengths from year-month y-m are assigned to cohort year y if m is in March to December and cohort year y - 1 otherwise.

Data collected in the Mid San Pablo Bay subregion were not included in these calculations.

5 Entrainment-Related Physical Variables

The R script DataCleaner_EntrainPhysicalVar.r creates a clean data set containing physical variable measurements, including delta flows and turbidity.

5.1 Raw Data

5.1.1 Dayflow Data

All of the files listed below, with the exception of Daily_Outflow_and_X2_1930-2011.xlsx, were downloaded from the Dayflow home page by Lara Mitchell; the date on which each file was retrieved is shown in parentheses. These files contain, among other fields, daily values of Sacramento River flows, San Joaquin River flows, SWP and CVP exports, delta outflow, and QWest flows. X2 values are only present in the files for water year 1997 and later. Flow values are in cubic feet per second (cfs) and X2 is in km. These raw data files were combined into a single clean data set spanning water years 1969 - 2016.

The file Daily_Outflow_and_X2_1930-2011.xlsx was created by Fred Feyrer and contains X2 values from October 1, 1929 to December 31, 2011, with values prior to water year 1997 calculated according to the X2 equation provided on the Dayflow documentation website. A copy of this file was provided by Ken Newman on March 15, 2016. A copy of the worksheet named "Daily Outflow and X2 1930-2011" was saved in csv format for reading in to R. All values of X2 in the clean data set prior to water year 1997 come from this file.

<u>Raw Data Files</u>

wy1970-1983.csv (downloaded on February 24, 2016)
wy1984-1996.csv (downloaded on February 24, 2016)
dayflowCalculations1997.csv (downloaded on February 24, 2016)
dayflowCalculations1998.csv (downloaded on February 24, 2016)
dayflowCalculations1999.csv (downloaded on February 24, 2016)
dayflowCalculations2000.csv (downloaded on February 24, 2016)
dayflowCalculations2001.csv (downloaded on February 24, 2016)
dayflowCalculations2002.csv (downloaded on February 24, 2016)
dayflowCalculations2003.csv (downloaded on February 24, 2016)
dayflowCalculations2004.csv (downloaded on February 24, 2016)
dayflowCalculations2005.csv (downloaded on February 24, 2016)
dayflowCalculations2006.csv (downloaded on February 24, 2016)
dayflowCalculations2007.csv (downloaded on February 24, 2016)
dayflowCalculations2008.csv (downloaded on February 24, 2016)
dayflowCalculations2009.csv (downloaded on February 24, 2016)
dayflowCalculations2010.csv (downloaded on February 24, 2016)
dayflowCalculations2011.csv (downloaded on February 24, 2016)
dayflowCalculations2012x.csv (downloaded on February 24, 2016)
dayflowCalculations2013x.csv (downloaded on February 24, 2016)
dayflowCalculations2014a.csv (downloaded on February 24, 2016)
dayflowCalculations2015.csv (downloaded on February 24, 2016)
dayflowCalculations2016.csv (downloaded on April 21, 2017)
Daily Outflow and X2 1930-2011.xlsx

5.1.2 OMR Data

The file OMR_Q_wy1980-cy2014.csv contains combined daily Old River and Middle River flows in cfs. Values from water years 1987 - 2014 are based on data obtained online from USGS by Pete Smith (USGS). Values from water years 1980 - 1986 were imputed by Pete Smith from a regression of combined flows on exports and San Joaquin River flows at Vernalis.

The file omr-2010-2017.csv contains combined daily Old River and Middle River flows in cfs. The data were obtained online by Lara Mitchell on April 21, 2017 from the same USGS website cited above. For this file, missing values were imputed using simple linear interpolation. The code for creating this file is located in the R script DataCleaner_EntrainPhysicalVar.r, below where clean physical variable files are saved in csv and RData format.

In cases where both files contained the same date, values from the first file were used.

5.1.3 Turbidity Data

The file daily_CCFB_turbidity_Mar88-Aug12.csv contains daily turbidity measurements (in ntu) from CCFB for the years 1988 - 2010. The data were obtained from the CDEC website by Pete Smith.

The file CCFB_Turbidity_Daily_2012_2017.txt contains daily CCFB turbidity data from 2012 to 2017. These data also come from CDEC, and were retrieved by Lara Mitchell on April 21, 2017.

Missing values in both files were imputed using a simple moving average. In cases where both files contained the same date, values from the first file were used.

5.2 Clean Data

5.2.1 Daily Physical Variable Data Set

The Dayflow, OMR, and turbidity data described above were used to construct the clean file Entrain_Physical_Daily_69_16.csv, the fields of which are described below. Each row of the file represents a unique date, with all dates between October 1, 1969 and September 30, 2016 represented.

Clean Data Field Names

Date - Unique date.
Year - Calendar year corresponding to Date.
Month - Month corresponding to Date.
Inflow - Total Delta inflow (converted to acre-feet per day). Source: Dayflow Data.
SacFlow - Sacramento River flow (converted to acre-feet per day). Source: Dayflow Data.
SJRFlow - San Joaquin River flow (converted to acre-feet per day). Source: Dayflow Data.
Outflow - Total Delta outflow at Chipps Island (converted to acre-feet per day). Source: Dayflow Data.
QWEST - San Joaquin River flow past Jersey Point (converted to acre-feet per day). Source: Dayflow Data.
SWP.Exports - State Water Project (SWP) exports (converted to acre-feet per day). Source: Dayflow Data.
CVP.Exports - Central Valley Project (CVP) exports (converted to acre-feet per day). Source: Dayflow Data.
Total.Exports - Total exports, including SWP, CVP, and others (converted to acre-feet per day). Source: Dayflow Data.
X2 - Distance from Golden Gate to 2ppt Salinity (km). Source: Dayflow Data.
OMR - Sum of Old River and Middle River flow (cfs). Source: OMR Data.
OMR.scale - OMR divided by the standard deviation of all daily OMR values.

CCFB.Turbidity - Clifton Court Forebay turbidity (ntu). Source: Turbidity Data.
CCFB.Turbidity.scale - CCFB.Turbidity divided by the standard deviation of all daily CCFB.Turbidity values.

5.2.2 Monthly Physical Variable Data Set

The file Entrain_Physical_Monthly_69_16.csv is a version of the daily file that is aggregated by year-month. Within a year-month, we sum over the fields Inflow, SacFlow, SJRFlow, Outflow, QWEST, SWP.Exports, CVP.Exports, and Total.Exports, and take the mean of the fields X2, OMR, OMR.scale, CCFB.Turbidity, and CCFB.Turbidity.scale. Field names remain the same between the daily and monthly files. The following values were substituted for the calculated values because of suspected errors in the daily data (Pete Smith, personal communication). The substituted values came from a regression analysis carried out by Pete Smith. Currently, changes are *not* made to the problematic daily data.

January 1991: 10.4 ntu
January 1994: 4.0 ntu
February 1994: 6.0 ntu
February 1995: 16.6 ntu
January 1996: 14.8 ntu
February 1996: 26.8 ntu
February 1997: 39.2 ntu
March 1997: 30.0 ntu

6 Prey Data

The R script DataCleaner_ZooMysid_median_vX.r creates a clean data set containing information on delta smelt prey items, including zooplankton and mysids.

6.1 Raw Data

6.1.1 Zooplankton and Mysid Data

Zooplankton and mysid data are collected through the CDFW Zooplankton Study, which is part of IEP's Environmental Monitoring Program. The study started in 1972 and uses three gear types: (1) a pump targeting microzooplankton less than 1 mm in length, (2) a modified Clarke-Bumpus (CB) net targeting mesozooplankton 0.5 – 3.0 mm in length, and (3) a macrozooplankton net targeting zooplankton 1 – 20 mm in length, including mysid shrimp.

The raw data files listed below contain data collected by the Zooplankton Study, and were used to create a clean delta smelt prey data set.

Raw Data Files

EMPCBMatricesMASTMay2017.xlsx
EMPMysidMatricesMASTMay2017.xlsx
CB.taxon.cutoffs.csv
Mysids.taxon.cutoffs.csv
ZPStations.csv

The first file contains data on zooplankton species including copepods, cladocerans, and rotifers, and was provided by April Hennessy on May 4, 2017. The worksheet named "1972-2014 CB BPUE Matrix" contains Carbon biomass-per-unit-volume (BPUV, micrograms of Carbon/m^3) estimates for a variety of taxa with each record corresponding to a unique combination of sample date and sampling station. This worksheet was used as the raw zooplankton file. The second file contains data on mysids, and was provided by April Hennessy on May 3, 2017. The worksheet named "MysidBPUEMatrix1972-2016" contains BPUV estimates for different taxa with each record corresponding to a unique combination of sample date and sampling station. This worksheet was used as the raw mysid file. The carbon biomass of an organism serves as an indicator of how "nutritious" the individual is: the higher the weight, the more nutritious (Wim Kim-merer, personal communication). Mysid weights are highly dependent upon individual size (Wim Kimmerer, personal communication). The zooplankton and mysid Excel files describe how the BPUV estimates were calculated.

The third and fourth files contain information on the years during which each species in the zooplankton and mysid files have been monitored. This information is used to distinguish cases of 0 catches from cases in which data were not collected. The fifth file indicates to which of the DSLCM regions each EMP sampling station belongs; see Figure 1 for a map of the EMP stations. The last three files were provided by Ken Newman in 2014.

Table 12 shows select fields from the raw zooplankton and mysid data sets that are used to calculate biomass metrics in the clean zooplankton and mysid data sets (see next section). Note that zooplankton are restricted to calanoid copepods, cyclopoid copepods, cladocerans. Different taxa have been collected at different times throughout the history of the survey, as indicated by the "Sampling Period" column in Table 12. Differences in collection periods are due, in part, to the fact that many of the species are non-indigenous to the bay-delta.

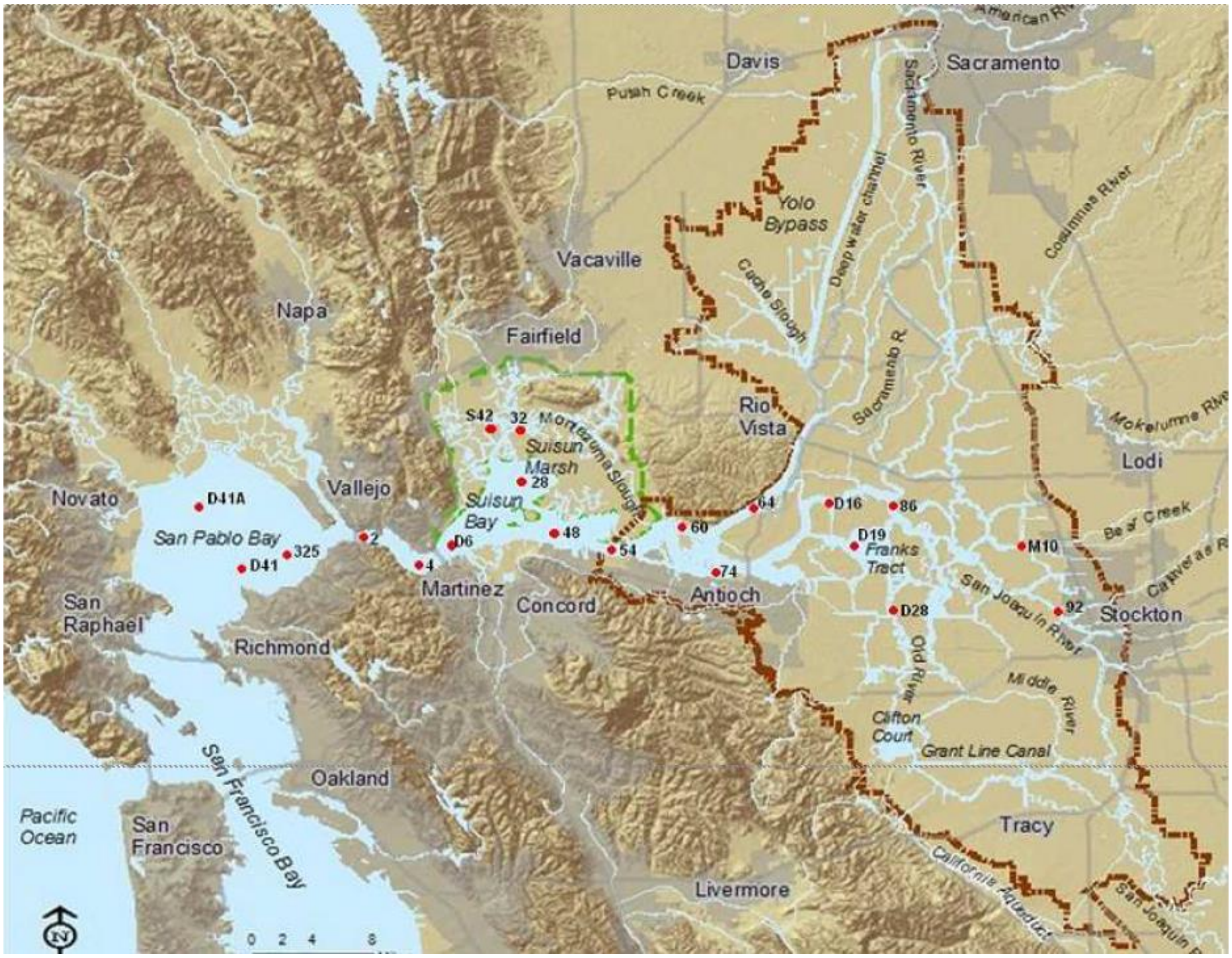


Figure 1: EMP Zooplankton Study sampling locations, shown as red dots (http://www.dfg.ca.gov/).

It has been hypothesized that organisms sampled by the pump component of the EMP Zooplankton study may be too small for juvenile and adult delta smelt to actively target as prey (Matt Nobriga, personal communication). For this reason, the pump data are not being used at this time. Zooplankton data collected as part of the Twentymm fish survey are also not being used because they are temporally limited relative to the EMP study, and because there is tentative evidence for correlation with the EMP data (Steve Slater, personal communication).

6.2 Clean Data

The R script DataCleaner_ZooMysid_median_vX.r uses the raw zooplankton and mysid data sets to create a clean data file, called ZooMysid_74_16_df median.csv, containing measures of zooplankton and mysid biomass calculated by year-month-region for the years 1974 – 2016. The field names in the clean file are listed below.

Clean Data Field Names

Year - Sample year.
Month - Sample month.
Region - Sampling region, as defined in the DSLCM.
NJ BPUV - Prey metric composed of copepod nauplii and juveniles.
JA BPUV - Prey metric composed of copepod juveniles and adults.
JAC BPUV - Prey metric composed of copepod juveniles and adults, and cladocerans.
NJAC BPUV - Prey metric composed of copepod nauplii, juveniles, and adults, and cladocerans.
M BPUV - Prey metric composed of mysids.
JACM BPUV - Prey metric composed of copepod juveniles, copepod adults, cladocerans, and mysids.
NJACM BPUV - Prey metric composed of copepod nauplii, juveniles, and adults, cladocerans, and mysids.
ACM BPUV - Prey metric composed of copepod adults, cladocerans, and mysids.

The first step in creating the clean file is to remove any records from the raw zooplankton data set that fall outside of the four main regions and replace any 0 BPUV values outside of each field's sampling period with NA's. Next, any records that fall outside of the core sampling stations 1 and 2, surveys 3 – 11, and the years 1974+ are removed per a recommendation in the document "ReadMeZooplanktonStudyMatricesJune2015.doc." Then, eight measures of aggregated prey biomass are calculated from different combina- tions of zooplankton and mysid species at different life stages. Separate biomass estimates are calculated for each combination of year, month, and region, with the median BPUV being calculated across all sampling stations within the given region. The fields in the clean prey data set are described below, with the eight aggregated biomass field names ending in "BPUV." Details on the specific species used to construct each field are available in Table 12. Missing values were imputed by linearly interpolating across the year-month time series in a given region when data were available to do these calculations. Table 13 summarizes the imputation scheme for the NJ BPUV field. The other fields were handled similarly.

Table 12: A summary of select fields from the raw zooplankton data set (above the double line) and the raw mysid data set (below the double line), organized by taxon and, in some cases, life stage. *Field* gives the field name used in the raw data set, *Description* describes the species or group of species represented by the field, and *Sampling Period* shows the year range during which the fields have been used. Asterisks indicate "catch all" categories that exclude species that were explicitly being counted at the time. *Status* indicates whether a field represents native or introduced species. In the latter case, the last column gives the year the species are hypothesized to have been introduced, or the year in which they first became abundant (Orsi et al. 1983; Orsi 1999; Kimmerer et al. 1999).

Taxon	Life Stage	Field	Description	Sampling Period	Status	Intro Year
Copepod (Calanoid)	nauplius	COPNAUP	Copepod nauplii*	1972 – 1988		
		OTHCOPNAUP	Other copepod nauplii*	1989 – present		
		EURYNAUP	*Eurytemora affinis* nauplii	1989 – present	Introduced?	?
		SINONAUP	*Sinocalanus doerrii* nauplii	1989 – present	Introduced	1979
		PDIAPNAUP	*Pseudodiaptomus* spp. nauplii	2000 – present		
Copepod (Calanoid)	juvenile	CALJUV	Calanoid copepodids*	1972 – 1988		
		OTHCALJUV	Other calanoid copepodids*	1989 – present		
		EURYJUV	*Eurytemora affinis* copepodids	1989 – present	Introduced?	?
		SINOCALJUV	*Sinocalanus doerrii* copepodids	1989 – present	Introduced	1979
		PDIAPJUV	*Pseudodiaptomus* spp. copepodids	1990 – present		
		ASINEJUV	*Acartiella sinensis* copepodids	2006 – present	Introduced	
		ACARJUV	*Acartia* spp. copepodids	2006 – present	Native	NA
		DIAPTJUV	*Diaptomidae* copepodids (includes several genera)	2006 – present		
		TORTJUV	*Tortanus* spp. copepodids	2006 – present		
Copepod (Calanoid)	adult	EURYTEM	*Eurytemora affinis*	1972 – present	Introduced?	?
		OTHCALAD	Other Calanoid adults*	1972 – present		
		SINOCAL	*Sinocalanus doerrii*	1978 – present	Introduced	1979
		PDIAPFOR	*Pseudodiaptomus forbesi*	1988 – present	Introduced	1988
Copepod (Cyclopoid)	adult	AVERNAL	*Acanthocyclops vernalis*	1972 – present		
		LIMNOSPP	*Limnoithona* spp.	1979 – present		
		LIMNOSINE	*Limnoithona sinensis*	2007 – present	Introduced	1993
		LIMNOTET	*Limnoithona tetraspina*	2007 – present	Introduced	1994
Cladoceran		BOSMINA	*Bosmina longirostris*	1972 – present		
		DAPHNIA	*Daphnia* spp.	1972 – present		
		DIAPHAN	*Diaphanosoma* spp.	1972 – present		
		OTHCLADO	Other cladocera*	1972 – present		
Mysid		H longirostris	*Hyperacanthomysis longirostris* (formerly *Acanthomysis bowmani*)	1993 – present	Introduced	1993
		N mercedis	*Neomysis mercedis*	1972 – present	Native	NA

Table 13: A summary of the year-month-region combinations for which the field NJ BPUV was imputed. Values in the table represent region (FW = Far West; W = West; N = North; S = South; All = Far West, West, North, and South). A value of 0 means that no imputation was necessary.

	Month								
Year	January	February	March	May	July	August	October	November	December
1974	0	0	0	0	0	0	0	0	All
1975	All	All	FW	0	0	0	0	0	All
1976	All	All	0	0	0	0	0	0	All
1977	All	All	0	0	0	0	0	FW,N	All
1978	All	All	0	0	0	0	0	0	All
1979	All	All	0	0	0	0	0	0	All
1980	All	All	0	0	0	0	0	0	All
1981	All	All	FW	0	0	0	0	0	All
1982	All	All	0	0	0	0	0	0	All
1983	All	All	0	0	0	0	0	0	All
1984	All	All	0	0	0	0	0	0	All
1985	All	All	0	0	0	0	0	0	All
1986	All	All	N	0	0	0	0	0	All
1987	All	All	0	0	0	0	0	0	All
1988	All	All	FW	0	All	0	0	0	All
1989	All	All	0	0	0	0	0	FW	All
1990	All	All	0	0	0	0	0	0	All
1991	All	All	FW,N,S	0	0	0	0	0	All
1992	All	All	0	0	0	0	0	0	All
1993	All	All	0	0	0	0	0	0	All
1994	All	All	0	0	0	0	0	0	All
1995	All	All	0	0	0	0	FW	0	All
1996	All	All	0	0	0	0	0	0	All
1997	All	All	0	0	0	0	0	0	All
1998	All	All	0	0	0	0	0	0	All
1999	All	All	0	0	0	0	0	0	All
2000	All	All	0	0	0	0	0	0	All
2001	All	All	0	0	0	0	0	0	All
2002	All	All	0	0	0	0	N	0	All
2003	All	All	0	0	0	0	0	0	All
2004	All	All	0	0	0	0	0	0	All
2005	All	All	0	0	0	0	0	0	All
2006	All	All	0	0	0	0	0	0	All
2007	All	All	0	0	0	0	0	0	All
2008	All	All	0	0	0	0	0	0	All
2009	All	All	0	0	0	0	0	0	All
2010	All	All	0	FW	0	0	0	0	All
2011	All	All	0	0	0	0	N	0	All
2012	All	All	0	0	0	0	0	0	All
2013	All	All	0	0	0	0	0	0	All
2014	All	All	0	0	0	0	0	0	All
2015	All	All	0	0	FW	N	0	0	All
2016	All	All	0	0	0	N	0	0	0

7 Salvage Data

The R script DataCleaner_Salvage.r creates a clean data set containing data on delta smelt salvaged at the State Water Project or the Central Valley Project.

7.1 Raw Data

The files listed below were provided by Geir Aasen (CDFW) and contain count and fork length (mm) information on salvaged delta smelt. Further information on data collection is available on the salvage section of the CDFW ftp site.

Raw Data Files

ForkLengths-1979-1992.csv
ForkLengths-1993-2014.csv
Salvage-1979-1992.csv
Salvage-1993-2014.csv

7.2 Clean Data

The clean salvage files are listed below, and cover the years 1979 to 2014. The first file contains the total number of age 0 delta smelt salvaged at the SWP and CVP combined. The data are grouped by year (row) and month (column). The second file is structured similarly, and contains total age 1 salvage. The third file contains daily smelt salvage counts and mean fork lengths, partitioned by age group (0 or 1) and facility (SWP or CVP).

Clean Data Files

Salvage.Age0.Year.by.Month.csv
Salvage.Age1.Year.by.Month.csv
Salvage.Daily.csv
Salvage.Monthly.csv

8 Spawning Water Quality Index

The R script DataCleaner_WaterQuality.r creates an index reflecting the quality of water temperature for delta smelt spawning on a year-month basis.

8.1 Raw Data

The files listed below contain hourly water temperature measurements (°F) from five data collection stations in the Bay-Delta. The data were downloaded from the California Data Exchange Center (CDEC) website on April 26, 2016. According to the CDEC station metadata site, the Antioch (ANC), Pittsburg (PTS), Rio Vista (RIV), and San Andreas Landing (SAL) stations are operated by the U.S. Bureau of Reclamation, and the Martinez (MRZ) station is operated by the California Department of Water Resources. The rows in each file represent unique date-hour combinations. Some changes were made to these files immediately after they were downloaded. Namely, HTML formatting statements were removed, and missing temperatures, originally indicated by two dashes (- -), were replaced with the text NA. We note that there are other data collection stations represented on the CDEC website that could be considered beyond those considered here.

Raw Data Files

CDEC_Temp_ANC 3-1-99_to_4-20-16.txt
CDEC_Temp_MRZ 7-1-94_to_4-20-16.txt
CDEC_Temp_PTS 4-1-99_to_4-20-16.txt
CDEC_Temp_RIV 2-22-99_to_4-20-16.tx
DEC_Temp_SAL 2-23-99_to_4-20-16.txt

8.2 Clean Data

The clean spawning water quality index data file, SpawningWaterQualityIndex.csv, contains a water quality index value for each combination of year-month between January 1995 and March 2016, with each row corresponding to a unique year-month.

An outline of the procedure for producing spawning water quality index values is as follows: clean the hourly temperature data; use the clean hourly data to calculate mean daily temperatures; calculate the water quality index for a given month as a weighted sum of mean daily temperatures within that month, where higher weights are given to temperatures that are more favorable for delta smelt spawning.

The following procedure was carried out for each of the five data sets in order to produce clean hourly and mean daily temperature values. A data frame containing every hour of every date was created with temperature values of NA, then temperature values from the raw data set were filled in. This was done to ensure that every date-hour combination was represented in the clean data set even if any combinations were missing from the raw data set. A visual inspection of all five time series indicated that temperatures outside of the interval [40°F, 85°F] were probably not realistic, so any values outside of this range were replaced with NAs. Most of these invalid temperatures were rather extreme, e.g., 2000, and appeared to be the result equipment malfunction. Next, empirical lower and upper bounds were calculated for a given date-time, and any temperatures falling outside the range defined by the bounds were replaced with NAs. This was done to detect and remove potentially problematic points that fell outside of the overall visual pattern of the data. Bounds were constructed by splitting the temperature data by day (within a year) and hour, e.g., January 1 at 12:00 pm, and calculating $\theta \pm 2\sigma\theta$, where θ and $\sigma\theta$ are the calculated mean and standard deviation (ignoring any missing temperature values). For a given day-time combination, this gives a rough 95% confidence interval calculated across years. This method is simple and systematic, but is also very crude and, based on a visual inspection, probably overestimated the number of problematic temperatures. We alternatively considered comparing individual temperature measurements with a moving average, but this method was not always able to detect points that we thought should have been detected and removed. The reason for this was that some potentially problematic temperatures occurred in sequences, leading to problematic moving average values. At this point, day-time specific mean temperatures were recalculated (to exclude problematic values) and these means used to impute all missing (NA) temperature values. Finally, the cleaned hourly data set was used to calculate mean daily temperatures in both °F and °C.

We found that the mean daily temperatures across the five sites were highly correlated (see Figure 2). Martinez is further west than spawning is likely to occur, but because the Martinez data go back further in time than the other four data sets, and because water temperatures at the five stations were so correlated, we chose to use Martinez temperatures for calculating the spawning water quality index. The index for a given year-month combination is calculated as a weighted sum of the mean daily temperatures in °C, where the weights range from 0 to 1 and reflect how favorable a temperature is for delta smelt spawning with higher weight indicating higher favorability. The weighting function, shown in the top panel of Figure 3, is based on work by Wang (1986) suggesting that delta smelt spawn between 7 and 15°C, and on observations of aquaculture delta smelt spawning between 12 and 22°C (Lindberg et al. 1997; Bennett 2005). Spawning water quality indices were not calculated for incomplete year-months, i.e., year-month combinations that had missing days. In this case, the year-month was simply excluded from the final data set.

9 Utility Files and Functions

The R scripts described in this document use functions defined in the file DataCleaner_Utility.r.

References

J J Orsi, T E Bowman, D C Marelli, and A Hutchinson. Recent introduction of the planktonic calanoid copepod *Sinocalanus doerrii* (centropagidae) from mainland china to the sacramento-san joaquin estuary of california. *Journal of Plankton Research*, 5(3):357 – 375, 1983.

J J Orsi. Long-term trends in mysid shrimp and zooplankton. *Interagency Ecological Program Newsletter*, 12(2):13 – 15, 1999.

W Kimmerer, C Penalva, S Bollens, S Avent, and J Cordell. *Interagency Ecological Program Newsletter*, 12(2):16 – 21, 1999. URL http://www.water.ca.gov/iep/products/newsletter.cfm.

J C S Wang. Fishes of the sacramento-san joaquin estuary and adjacent waters, california: a guide to the early life histories. Technical report, IEP, California Dept. of Water Resources, Sacramento, CA, 1986.

J Lindberg, R Mage, B Bridges, and S Dorshov. Status of delta smelt culture. *Interagency Ecological Program Newsletter*, 10(Summer):21–22, 1997. URL http://www.water.ca.gov/iep/products/newsletter.cfm.

W A Bennett. Critical assessment of the delta smelt population in the San Francisco estuary, California. *San Francisco Estuary and Watershed Science*, 3(2), 2005. URL http://www.escholarship.org/uc/item/0725n5vk.

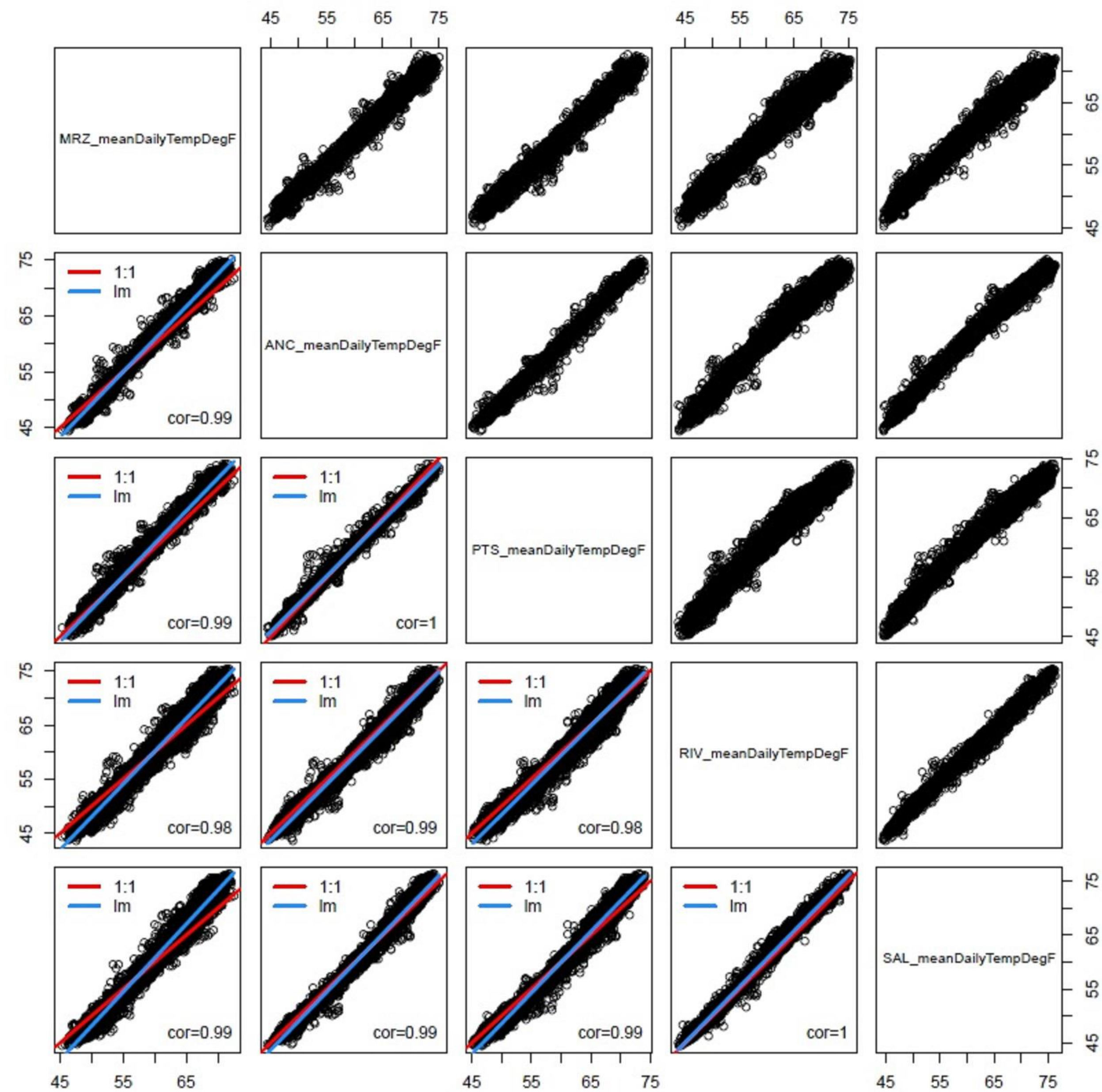


Figure 2: Correlation between average daily temperatures (°F) from five monitoring stations: MRZ, ANC, PTS, RIV, and SAL. “lm” stands for fitted linear model.

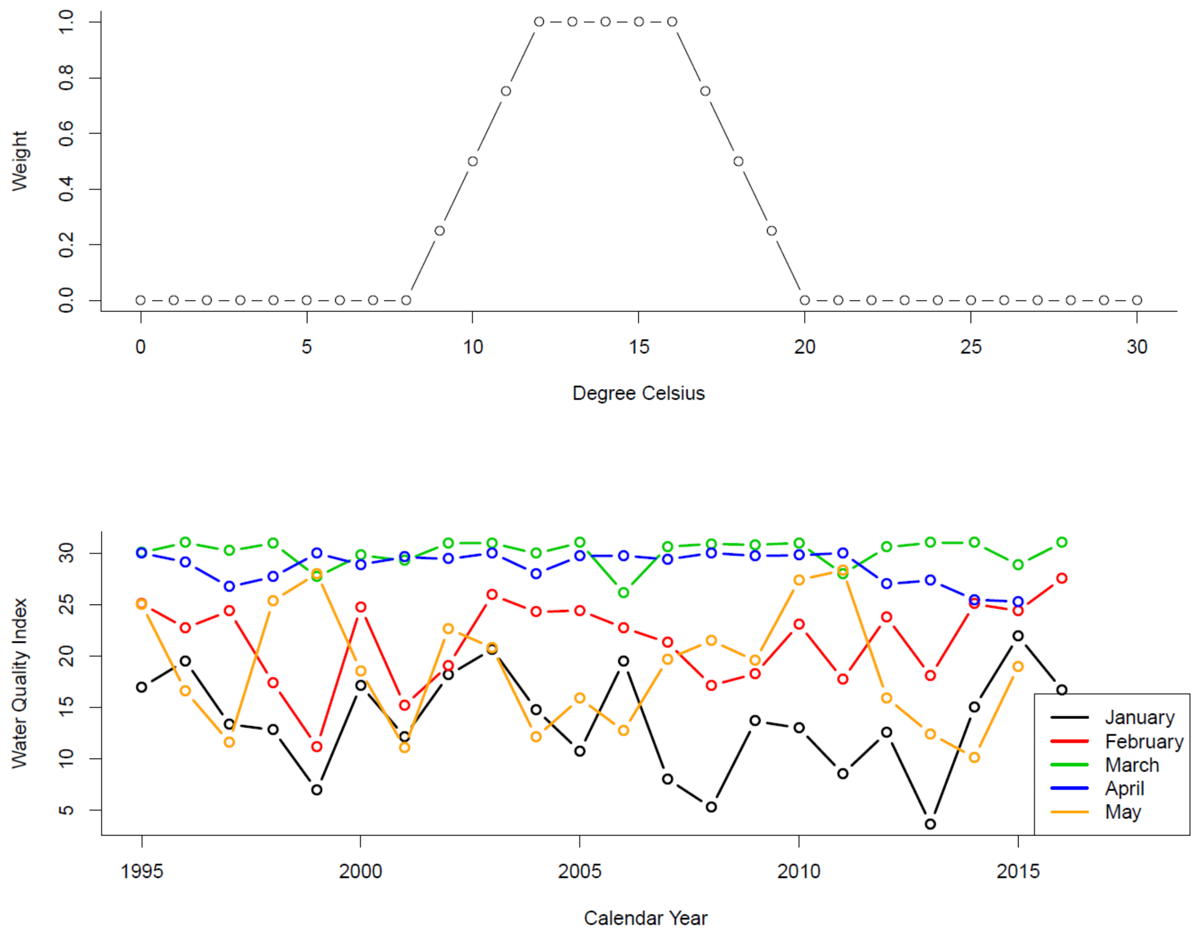


Figure 3: Spawning water quality index weighting function (top panel) and calculated index values by year and month (bottom panel).