**A Delta Smelt Individual-Based Life Cycle Model in the R Statistical Environment**

**Data description**

Five physical and biological variables, representing observed Delta conditions during 1995-2014, drive simulated population dynamics: prey density, Old and Middle River flow, delta smelt distribution, water temperature, and Secchi depth. All variables except Old and Middle River flow had dimensions year *y*, month *m*, and spatial strata *s*. Old and Middle River flow was a *y* x *m* matrix, and prey densities included a fourth dimension *p* indexing prey type.

*Prey density.* Prey density estimates for 12 zooplankton prey types *p* and years 1989-2015 were developed by Wim Kimmerer (Appendix A) and provided by Derek Hilts (Bay-Delta FWO; 24 September 2019). Daily log prey carbon densities, estimated from positive catches, and proportion zero catch were summarized from the Interagency Ecological Program’s zooplankton survey and the CA Department of Fish and Wildlife (CDFW) 20-mm survey. Monthly means of the daily values were used to simulate prey densities.

*Old and Middle River flow.* was the monthly average of the daily sum of tidally filtered flows from two adjacent channels of the San Joaquin River basin, Old and Middle Rivers. *OMR* data were available from US Geological Survey (USGS) streamflow databases (https://waterdata.usgs.gov). Much of the daily streamflow gauge data was missing. If data for one river was missing, daily *OMR* was predicted from a linear model of the flow in the remaining river, and if data for both rivers was missing, the model proposed by Andrews et al. (2016) was used to estimate missing *OMR* from San Joaquin River flows and exports from South Delta water diversions. Daily San Joaquin River flow and export data were available from the Dayflow database (https://data.cnra.ca.gov/dataset/dayflow).

*Fish distribution.* Observed delta smelt distributions were developed from 20-mm, Midwater Trawl, and Spring Kodiak Surveys. Townet Survey data were not used, because tow volume data were unavailable before 2003, use of Townet data would only leverage one additional month of distribution information (August), and in most year-months with a concurrent 20-mm Survey sample, delta smelt were observed in fewer spatial strata in the Townet Survey (in 7 of 13 years). Fish survey data were made available online by CDFW (ftp://ftp.dfg.ca.gov/). Monthly observed catch densities (catch/volume sampled) were assigned to the 12 spatial strata, and observed densities were expanded to total population abundance by multiplying by strata volumes. Strata volumes were estimated by Derek Hilts using DSM2. Estimated population abundances in each spatial stratum were then converted to proportions of the total abundance, which were interpreted as observed occupancy probabilities.

Surveys were not completed in all months, and in some months, no fish were observed (mostly February, March, and August). Distributions from the prior month were assumed when observations were missing. A high level of zero-inflation, resulting in zero observed occupancy of many spatial strata, was noted beginning in spring of 2015; therefore, the terminal cohort of observed spatial distributions was 2014 (last observed in April 2015).

Delta smelt densities in the South Delta stratum were depleted by entrainment, resulting in a negative bias and underestimation of the proportion of the population in this region of the Delta. Observed South Delta densities were therefore divided by estimates of proportional entrainment loss (Smith et al. 2021) to correct for bias.

*Temperature.* Water temperature data for years 1990-2010 were summarized from DSM2 hydrodynamic simulations by Derek Hilts (Fig. 2). The terminal year of the DSM2 *Temp* dataset, 2010, limited the number of years available for the delta smelt model by at least 4 years. A second set of water temperature data was therefore summarized from all available online data collected by the CDFW and FWS during Delta fish monitoring programs (ftp://ftp.dfg.ca.gov/; https://www.fws.gov/lodi/). The second water temperature dataset spanned years 1959-2020, but prior to 2011, data for some year-month-strata combinations were missing or sparse, with only a single sample.

The water temperatures that would have been simulated by DSM2 for missing years, 2011-2014, were predicted using a general linear model of DSM2 monthly means as a function of season, spatial strata, and monthly mean temperatures measured by fish monitoring programs . The best model of was identified using backwards selection, starting with a full model, having an effect for each stratum and season, and eliminating non-significant spatial effects (acceptance level < 0.05), one coefficient at a time, before eliminating non-significant seasonal effects. The model was fit to data from years 1990-2010, and used to predict from measured for years 2011-2014. See Appendix B for details of the model to predict .

*Secchi depth.* Secchi depth data were used as an index of turbidity. Like , Secchi depths for years 1959-2020 were summarized from all CDFW and FWS databases available online (Fig. 2). Means for each year-month-stratum combination were summarized, but data were not available for all strata in all year-months. Missing data were estimated from general linear models of the remaining Secchi data in other spatial strata. The best general linear model for each stratum was selected using backwards selection, beginning with the full model, having a separate coefficient for each spatial stratum, and eliminating non-significant coefficients, one at a time.

**Missing temperature and Secchi depth data**

General linear models were developed to predicting missing temperature and turbidity data. The objective of these models was prediction, not inference; therefore, covariate effects were not selected based any particular mechanistic link (e.g., spatial proximity), and the relative effects within each model were not compared. Only explanatory power and model performance was considered.

*Temperature.*DSM2 monthly mean for years *y*, 2011-2014, months *m*, and spatial strata *s* were predicted as a function of season, spatial strata, and monthly mean temperatures measured by fish monitoring programs , using a general linear model. from DSM2 represented a water column mean, while from monitoring programs represented surface measurements. The relationship between and was expected to vary seasonally, with the onset of thermal stratification as water warmed, and the effect of thermal stratification was expected to vary spatially, as water depth, tidal influence, and stratification vary spatially. Factorial seasonal and strata effects accounted for this spatiotemporal variation in the model to predict missing .

(1)

represented coefficients of the general linear model, the quantity represented a unique intercept for each *s*, and scaled fish monitoring temperatures to DSM2 temperatures. Beginning with a full model, having four seasonal effects , backwards selection was used to combine seasons until *n.season* groupings remained, with coefficient p-values < 0.05. After selecting seasonal effects, the same process was used to eliminate strata-specific effects. Beginning with a full model, having 12 strata-specific effects , backwards selection was used to combine strata until groupings remained, with coefficient p-values < 0.05.

*Secchi depth.*Missing Secchi depth measurements were also predicted from a general linear model. Since turbidity stratification was not expected to occur and no secondary measurements of were available from independent sources, missing were predicted using the remaining measurements in other strata.

(2) ,

for *i* in the set [Yolo, East Delta, Northeast Suisun]. Backwards model selection was used to eliminate strata-specific effects until strata effects remained, with coefficient p-values < 0.05.

Models (1) and (2) were fit using the glm() function in R (R 2018), and the best models indicated by model selection were evaluated graphically using general diagnostic plots, residual, q-q norm, standardized residual, and influence, for violations of basic linear model assumptions. Residuals were expected to be randomly distributed around zero. qq-norm plots were expected to be relatively linear, and the slope of standardized residuals was expected to be zero.

**Results**

The best model of included separate spring and summer effects, but winter and fall were grouped (Table 1). Most strata were grouped, but South Delta, Confluence, and Southwest Suisun strata each received unique coefficients, resulting in unique intercepts for these three strata. Diagnostics did not indicate severe violations of general linear model assumptions. Model residuals appeared to be normally distributed, and no extreme outliers or leverage points were detected (Fig. 1). The model explained 95.8% of the variation in the data, indicating high explanatory power.

All missing data were from years prior to 2011. Of 240 year-month combinations, 30 were missing, 7 were missing, and 1 were missing. The month of August was especially problematic for , with zero samples from 1998 to 2010.

The best model to fill in missing varied by stratum (Table 2). The best model to predict included Sacramento, South Delta, Lower Sacramento, and Lower San Joaquin effects. Some were missing in the same time periods as the missing Yolo data, so East Delta data were not used to predict Yolo Secchi depth. The best model to predict included Sacramento, South Delta, Southeast Suisun, and Southwest Suisun effects, and the best model to predict included Sacramento, Southeast Suisun, Northeast Suisun, and Northwest Suisun effects. Diagnostics revealed exactly one extreme outlier in the datasets to predict for each stratum, so these outliers were removed prior to fitting the final model. After removing outliers, most residuals appeared to be normally distributed (Figs. 2-4), and no high leverage points were detected. Residuals were somewhat higher at the highest predicted values of greater than 175 cm Secchi depth, suggesting that a log-linear model might fit the data better. Modeled effects on delta smelt did not vary at Secchi depths greater than 84 cm, so an improved model of the highest Secchi depths was not expected to change the IBMR dynamics.

The explanatory power of models to predict varied, with East Delta and Northeast Suisun models explaining 79.5% and 82.7% of variation, respectively. The best model to predict missing explained only 61.6% of variation. Many of the Yolo observations, or y values, were limited to a single replicate sample per month, versus and average of 30 Secchi depth samples within other year-month-strata combinations. This limitation may have been a source of observation error in the Yolo model.

**Discussion**

The models developed here to predict missing and leave much room for improvement.   
-The best way to improve the estimation of data used by IBMR would be to complete DSM2 simulations for the entire 1995-2014 time series.

-Backwards selection using coefficient p-values was used to identify the best and general linear models, but a comparison of all possible models, using AIC would be a superior method. The primary advantage would be increasing the number of strata groupings to consider. The set of all possible strata combinations is very large. For example, just consideration of all 1- and 2-strata groupings would result in a total of 78 candidate models to consider.

-Replacing the linear models with a more flexible functional form, such as a generalized additive model (GAM), could improve predictions by accounting for non-linear relationships.

-Some hydrodynamic models of the San Francisco Estuary are capable of simulating turbidity fields. Application of these models to estimate the IBMR dataset would be especially useful for the Yolo spatial strata, which was limited by a low number of within-month (replicate) samples and a larger number of unsampled year-month combinations, relative to other strata.

**Data files**

Filenames for datasets with raw data are listed below. Each dataset, except fish distribution and data are saved at the daily time scale; however the IBMR requires monthly-scaled data. Each dataset must, therefore, be summarized into the spatiotemporal dimensions of the IBMR, using a set of data summary functions that are included with the model code.

|  |  |  |  |
| --- | --- | --- | --- |
| Data type | Description | File name | Comments |
| Prey density | log(carbon density) of positive catches | ‘Zoop.acartela\_\_.txt’  ‘Zoop.allcopnaup.txt’  ‘Zoop.daphnia\_\_\_.txt’  ‘Zoop.eurytem\_\_\_.txt’  ‘Zoop.limno\_\_\_\_\_.txt’  ‘Zoop.othcalad\_\_.txt’  ‘Zoop.othcaljuv\_.txt’  ‘Zoop.othclad\_\_\_.txt’  ‘Zoop.othcyc\_\_\_\_.txt’  ‘Zoop.other\_\_\_\_\_.txt’ | One file for each prey type |
| proportion zero catches | ‘Zoop.P0.acartela\_\_.txt’ ‘Zoop.P0.allcopnaup.txt’ ‘Zoop.P0.daphnia\_\_\_.txt’ ‘Zoop.P0.eurytem\_\_\_.txt’ ‘Zoop.P0.limno\_\_\_\_\_.txt’ ‘Zoop.P0.othcalad\_\_.txt’ ‘Zoop.P0.othcaljuv\_.txt’ ‘Zoop.P0.othclad\_\_\_.txt’ ‘Zoop.P0.othcyc\_\_\_\_.txt’ ‘Zoop.P0.other\_\_\_\_\_.txt’ | One file for each prey type |
| OMR | Old and Middle River flow | ‘OMR\_daily.txt’ |  |
| Fish distribution |  | ‘DS\_distribution\_JanDec\_1995\_2015.txt’ |  |
| Temperature | DSM2 | ‘temp.txt’ |  |
| Fish survey | ‘delta\_water\_quality\_data\_with\_strata\_v2.csv’ |  |
| Secchi depth | Fish survey | ‘delta\_water\_quality\_data\_with\_strata\_v2.csv’ |  |

**References**

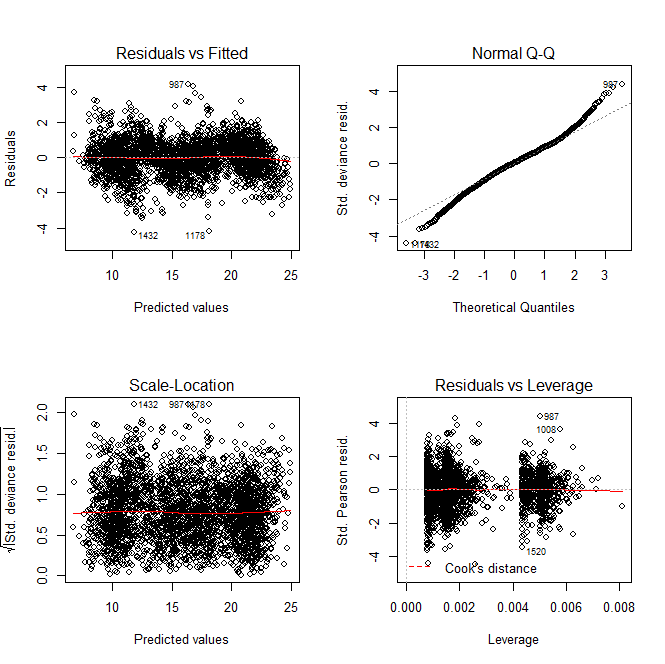
R Core Team. 2018. R version 3.5.0: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

**Table 1.** Results from the best model to predict DSM2 mean water column temperatures from fish monitoring measurements of surface temperature .

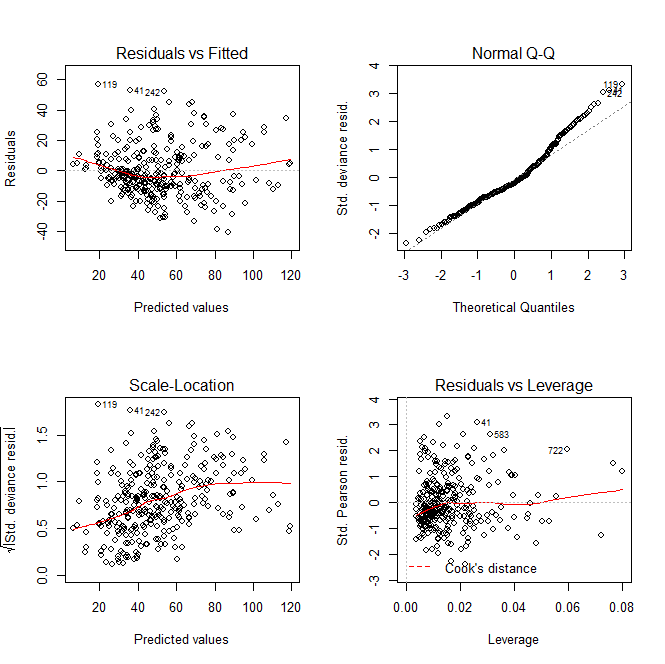
|  |  |  |
| --- | --- | --- |
| Coefficient | Estimate | p-value |
| Intercept | 1.370 | < 2x10-16 |
| Spring | 0.529 | < 2x10-16 |
| Summer | 1.649 | < 2x10-16 |
| Southwest Suisun | -0.197 | 0.002 |
| South Delta | -0.579 | < 2x10-16 |
| Confluence | -0.280 | 1.1x10-5 |
|  | 0.891 | < 2x10-16 |
| % null deviance explained | 95.8% | |

**Table 2.** Results from the best models to predict missing Secchi depth data in Yolo, East Delta, and Northeast Suisun spatial strata from the remaining data in other strata.

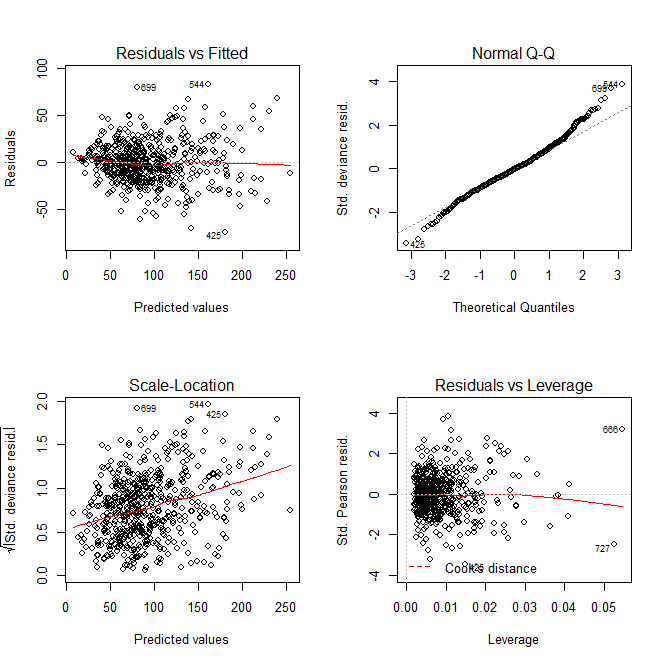
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Coefficient | Yolo | |  | East Delta | |  | Northeast Suisun | |
| Estimate | p-value |  | Estimate | p-value |  | Estimate | p-value |
| Intercept | -10.219 | 0.0013 |  | -7.967 | 0.0075 |  | -2.303 | 0.0054 |
| Sacramento | 0.219 | 2.8x10-9 |  | 0.804 | < 2x10-16 |  | -0.041 | 1.57x10-5 |
| South Delta | 0.310 | 7.6x10-10 |  | 0.411 | 1.4x10-14 |  | -- | -- |
| East Delta | -- | -- |  | -- | -- |  | -- | -- |
| Lower Sacramento | 0.545 | 2.1x10-9 |  | -- | -- |  | -- | -- |
| Lower San Joaquin | -0.183 | 0.011 |  | -- | -- |  | -- | -- |
| Confluence | -- | -- |  | -- | -- |  | -- | -- |
| Southeast Suisun | -- | -- |  | 0.227 | 0.00011 |  | 0.269 | < 2x10-16 |
| Northeast Suisun | -- | -- |  | -- | -- |  | 0.309 | 3.4x10-15 |
| Suisun Marsh | -- | -- |  | -- | -- |  | -- | -- |
| Southwest Suisun | -- | -- |  | 0.227 | 0.041 |  | -- | -- |
| Northwest Suisun | -- | -- |  | -- | -- |  | 0.462 | < 2x10-16 |
| % null deviance explained | 61.6% | |  | 79.5% | |  | 82.7% | |



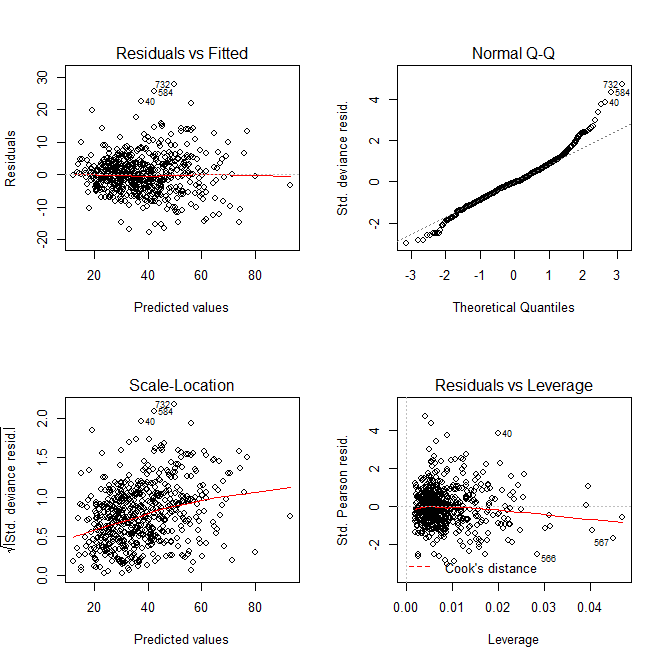
**Figure 1.** Diagnostic plots for the best model to predict DSM2 mean water column temperatures from fish monitoring measurements of surface temperatures.



**Figure 2.** Diagnostic plots for the best model to predict missing Yolo strata Secchi data from data in the remaining spatial strata during the same time period.



**Figure 3.** Diagnostic plots for the best model to predict missing East Delta strata Secchi data from data in the remaining spatial strata during the same time period.



**Figure 4.** Diagnostic plots for the best model to predict missing Northeast Suisun strata Secchi data from data in the remaining spatial strata during the same time period.