TETRA TECH

To:	Breck Sullivan
Cc:	Vanessa Van Note
From:	Jon Harcum
Date:	June 6, 2022
Subject:	CBP TD #6 Statistical Support: preliminary dissolved oxygen data review

This technical memorandum documents Tetra Tech's preliminary review of dissolved oxygen (DO) data that might be used to inform data pre-processing steps for the 4D DO model currently under development by the Chesapeake Bay Program (CBP). The purpose of this memorandum is to promote a discussion and feedback on next steps. I am appreciative of informal conversations with Dr. Elgin Perry which informed some aspects of this memorandum.

Summary

Data were compiled from the CBP DataHub as described in the *Data Pre-processing* section of this memorandum. The resulting data set includes approximately 675,000 DO records of which 515,000 are associated with the 144 stations most often used for tidal trend analyses.

Next data at two stations, CB5.3 and EE2.2, were examined in more detail. These stations were selected since some potential issues had been previously identified by Dr. Perry. Cluster analysis was used to identify profiles that might merit further investigation. Figure 1 is an example cluster analysis of data from October and April at Station EE2.2 using six (6) clusters. With a focus on looking for anomalous events, it might be appropriate to further examine those clusters with only one or two profiles. This approach is demonstrated in Figure 2 which highlights profiles using four (4) and six (6) clusters at both Stations CB5.3 and EE2.2. Red lines in Figure 2 indicate clusters with a single profile. Two red lines on a single panel indicate that two clusters had one profile each. Blue lines indicate clusters with two profiles. Gray lines indicate all remaining profiles irrespective of cluster.

Some of the profiles highlighted in Figure 2 as well as several other profiles were selected for further review. These reviews are described in the *Selected Profile Review* section of this memorandum and includes Figure 3 through Figure 12. It was hoped that strategies to identify anomalous data could be developed and generally applied to the remaining stations in a consistent, automated fashion. While a "clear" error was found (Station CB5.3 on September 15, 2003), the remaining profiles include a mixture of situations that range from weakly supported to plausible results. There is a need for a broader discussion of this preliminary data review to develop a defensible protocol for identifying and reviewing these data.

The next section of this memorandum identifies selected issues to initiate a broader discussion.

Discussion

This section provides selected issues to initiate a broader discussion about reviewing the available DO data in light of overall project goals.

1. There are approximately 160,000 DO observations at stations that are not part of the 144 primary CBP stations. (The 144 CBP stations include about 515,000 DO observations.)
    - Is there enough already known about the "additional data" to determine whether these data should be included in the 4D project?
    - How would these data be used (calibration, validation)?
    - Are there minimum data requirements for data inclusion (e.g., length of record, total number of observations, number of profiles)?
    - Proposed next step: Brief characterization of station locations and data richness.
2. "Clear" errors such as the event at Station CB5.3 on September 15, 2003 (Figure 3), is low hanging fruit for data cleaning; however, the number of observations and model impact is likely small.
    - What are the follow up steps beyond removing these observations?
    - Proposed next step: Remove observations of this type from the data set and create a summary report that could be provided to data integrity workgroup.
3. Total depths and maximum sample depths varied among profiles. (See Figure 3, Figure 8, and Figure 11 for examples.)
    - Does this issue affect data usability? Are there other CBP staff/stakeholders that can weigh in on this issue?
4. Cluster analysis was used to identify DO profiles that are least like the other profiles for a given month. This analysis used the R package, dtwclust, and a divisive analysis hierarchical control. The advantage of this approach is its repeatability and scalability to the other stations although application to more shallow stations has not been vetted.
    - Are there opportunities to refine DO profile identification? (Scaling up the 4-cluster analysis (12 of 896 profiles, 1.3%) would result in more than 800 profiles requiring follow up if applied to the 144 CBP stations.)
    - Are there opportunities to refine the follow-up review process? This preliminary review did not yield a tractable path. Balancing false positives (exclusion of "good" data that are simply extreme events) versus false negatives (inclusion of anomalous data).
    - Is there value in running the cluster analysis to minimally provide a mechanism to visually review all DO used for calibration and validation?

Cluster Analysis

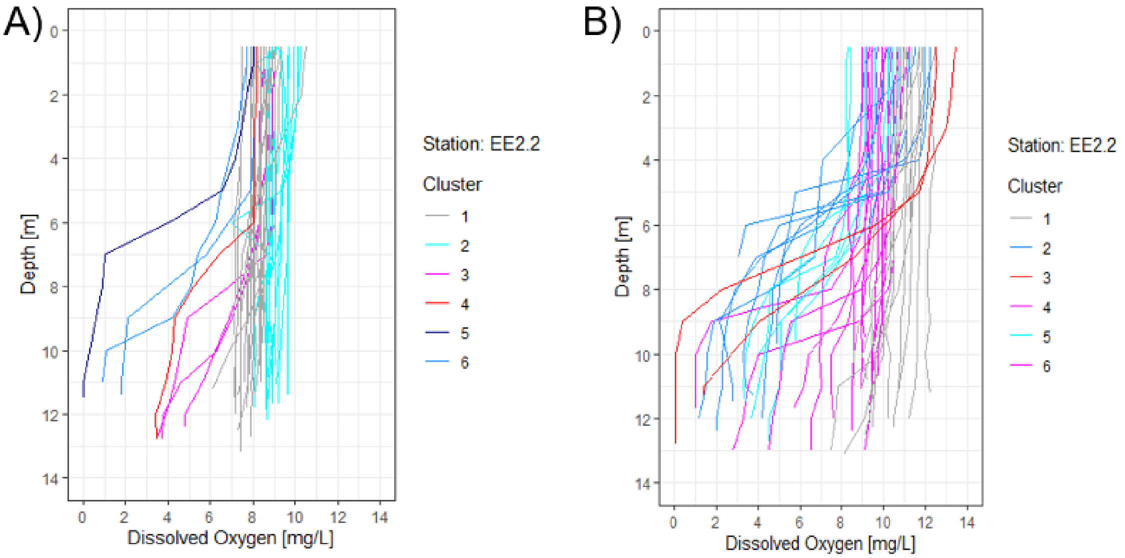


Figure 1. Cluster analysis of 1990-2019 DO data from Station EE2.2 using 6 clusters. Panel A: October, Panel B: April.

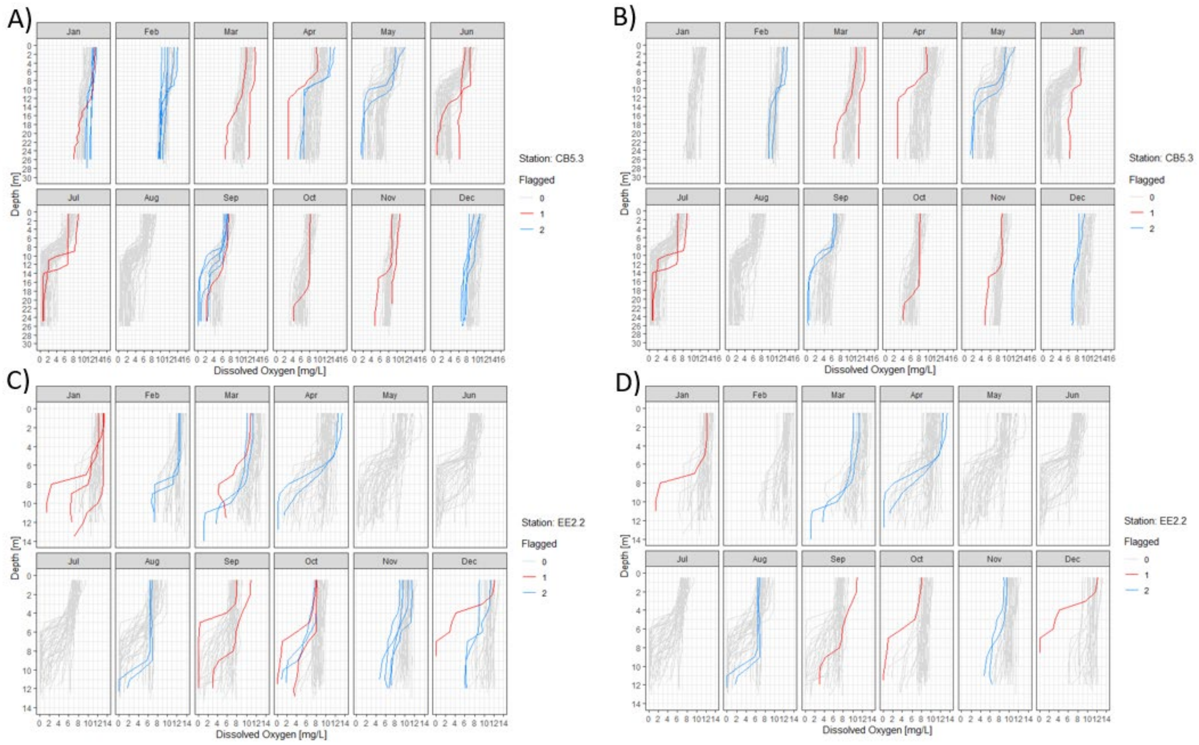


Figure 2. Cluster analysis of DO data at Stations CB5.3 and EE2.2. Panel A: CB5.3—6 clusters, Panel B: CB5.3—4 clusters, Panel C: EE2.2—6 clusters, Panel D: EE2.2—4 clusters. Red lines indicate clusters with a single event. Blue lines indicate clusters with two events. Gray lines indicate all remaining events irrespective of cluster.

Selected Profile Review

Station CB5.3

Figure 3 presents dissolved oxygen profiles for Station CB5.3 from 1990-2019 with each panel representing a different month. Four profiles are highlighted for discussion.

- **1) September 15, 2003.** One observation of 0.49 mg/L is reported among values above and below of 4.9 mg/L. We generally see observations of this type as “low hanging” fruit and propose to delete observations of this type.
- **2) April 20, 1998 (identified in cluster analysis).** The DO profile on this day represents an “earlier than normal” reduced dissolved oxygen environment that is more typical of May. Nevertheless, a more detailed examination of the water quality profiles from March 23—October 13, 1998 (see Figure 4) show a nominal progression of salinity and dissolved oxygen over time.
- **3) June 19, 1995 and 4) November 18, 1996.** The maximum sample depths associated with these June and November events are 29 and 21 m, respectively. Theses depths are unusual for this station (see Figure 5) where the typical maximum sample depth is typically 25—26 m and the reported total depth is 26—28 m. One possible explanation is that both observations are from the early part of the data record and cruises might have been more likely to have drifted to deeper or more shallow channels. When viewed with other events in the same time period (+/- a few months), both events appear plausible (see Figure 6 and Figure 7).

Station EE2.2

Figure 8 presents dissolved oxygen profiles for Station EE2.2 from 1990-2019 with each panel representing a different month. Several profiles are highlighted for discussion.

- **1) March 5, 1997 (identified in cluster analysis).** This event was identified by Dr. Perry as an event with a large residual error when compared to preliminary model results. As shown in Figure 9, the low March DO is supported by corresponding low February DO although the stratification is occurring at a shallower depth. We also note that this event is associated with a relatively deeper maximum sample depth of 14 m albeit not atypical (see Figure 10).
- **2) January 11, 2006 (identified in cluster analysis).** This event was also identified by Dr. Perry as an event with a large residual error when compared to preliminary model results. This low DO event is notable given the January time frame. The preceding December event does show a slight decrease in DO relative to the surface waters, but the next event (February 8) does not have samples below 8 m (see Figure 11).
- **3) December 9, 2019 (identified in cluster analysis).** Like the previous example, this is another low DO event in the winter. Figure 12 displays water quality profiles from the preceding events.
- **4) February 14, 1996 and 5) April 3, 1990.** These events have maximum sample depths of 14 and 7.2 m, respectively but are otherwise unremarkable.

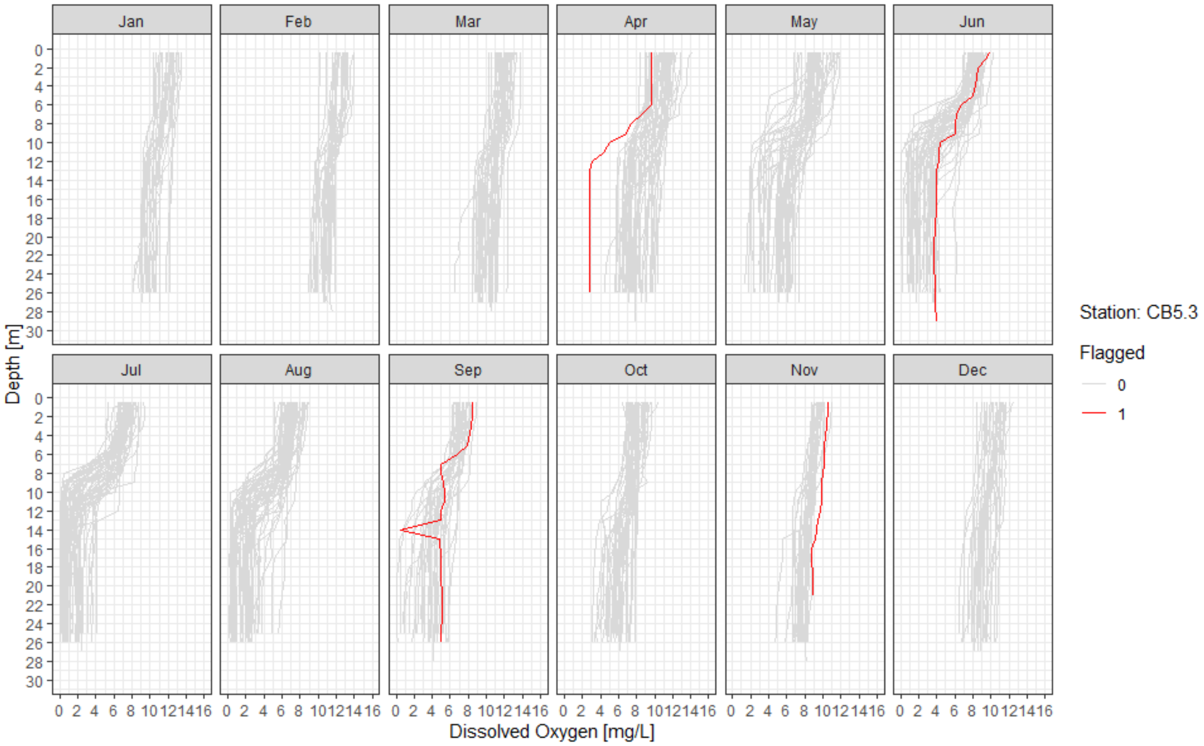


Figure 3. Dissolved oxygen profiles for Station CB5.3 from 1990-2019 separated by month.

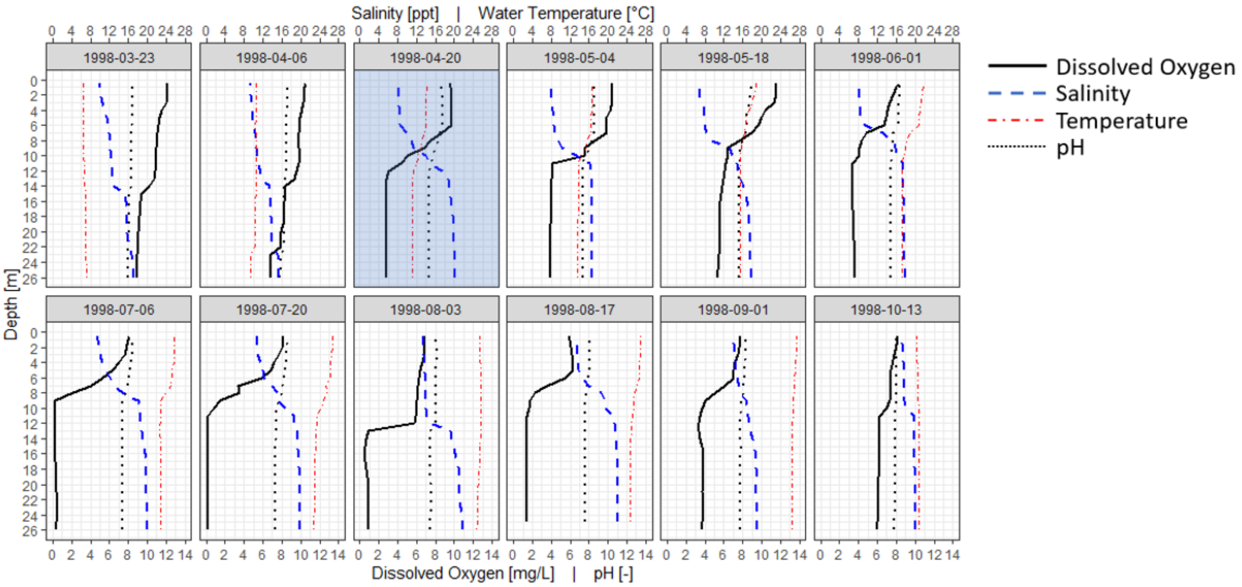


Figure 4. Water quality profiles at Station CB5.3 for selected events from March 23—October 13, 1998. The April 20, 1998 event (highlighted panel) demonstrates an earlier than normal reduced dissolved oxygen environment and corresponding stratification. Note, the top x-axis presents the scale for salinity and water temperature and the bottom x-axis presents the scale for dissolved oxygen and pH.

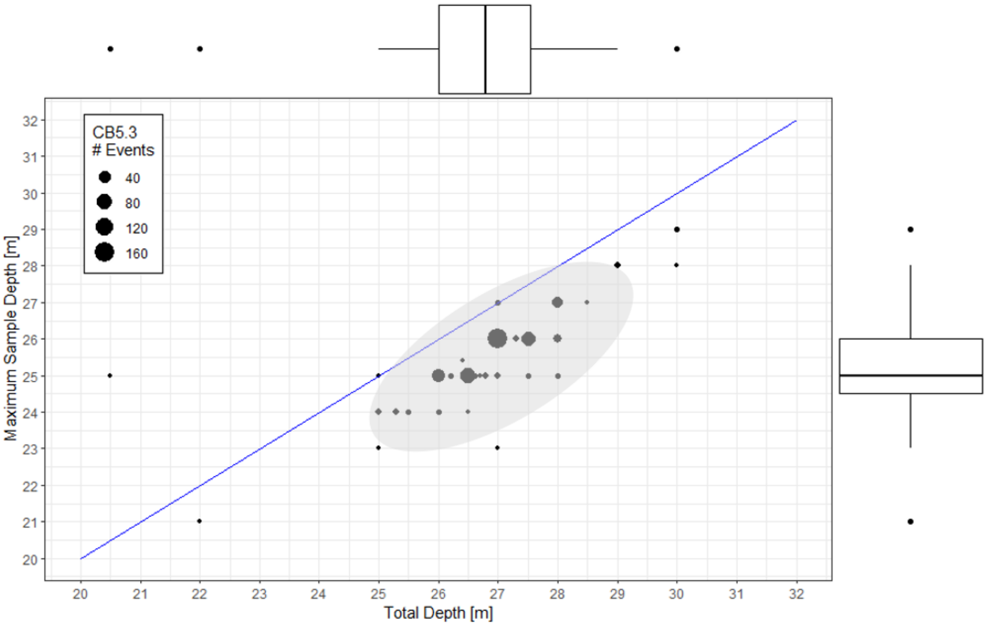


Figure 5. Comparison of the maximum sample depth associated with a sampling event and corresponding reported total depth at Station CB5.3. Symbol size represents the number of sampling events. The maximum sample depth would generally be expected to below the blue one-to-one line.

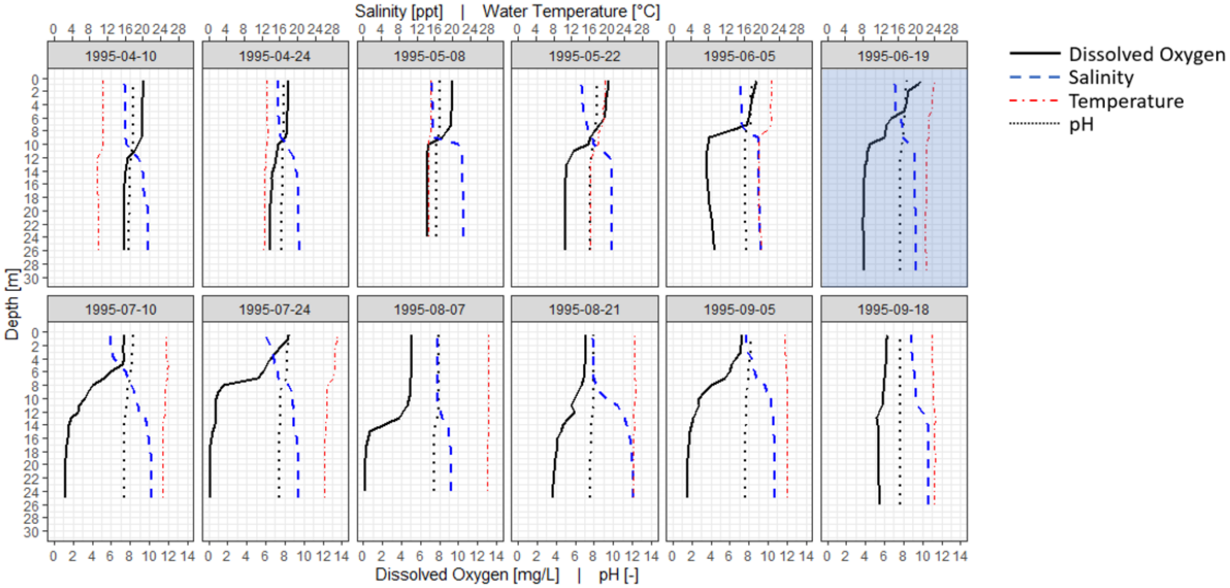


Figure 6. Water quality profiles at Station CB5.3 for selected events from April 10—September 18, 1995. The June 19, 1995 event (highlighted panel) includes samples from as deep as 29 m.

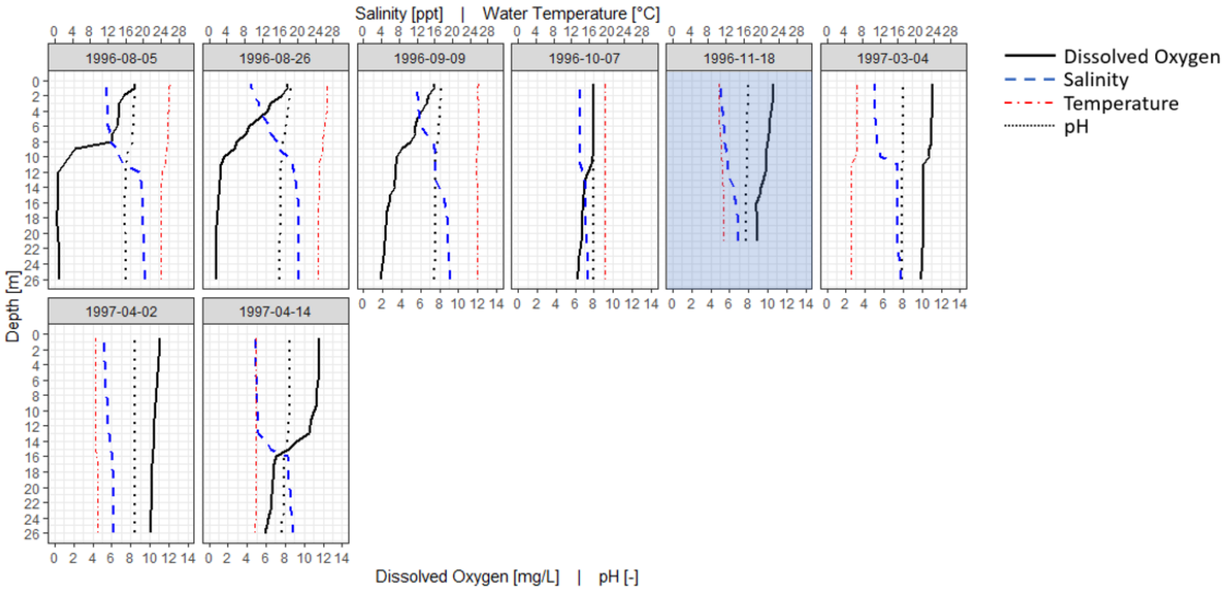


Figure 7. Water quality profiles at Station CB5.3 for selected events from August 5, 1996—April 14, 1997. The November 18, 1996 event (highlighted panel) includes samples only to 21 m.

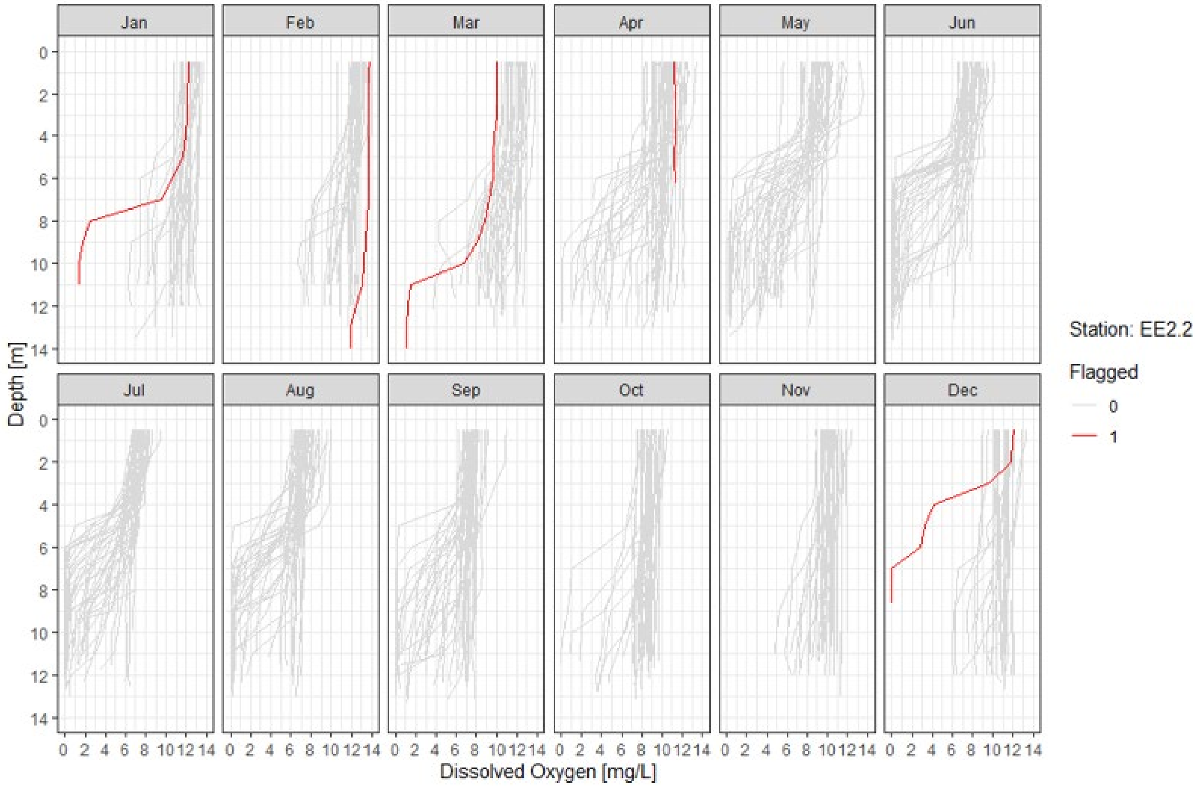


Figure 8. Dissolved oxygen profiles for Station EE2.2 from 1990-2019 separated by month.

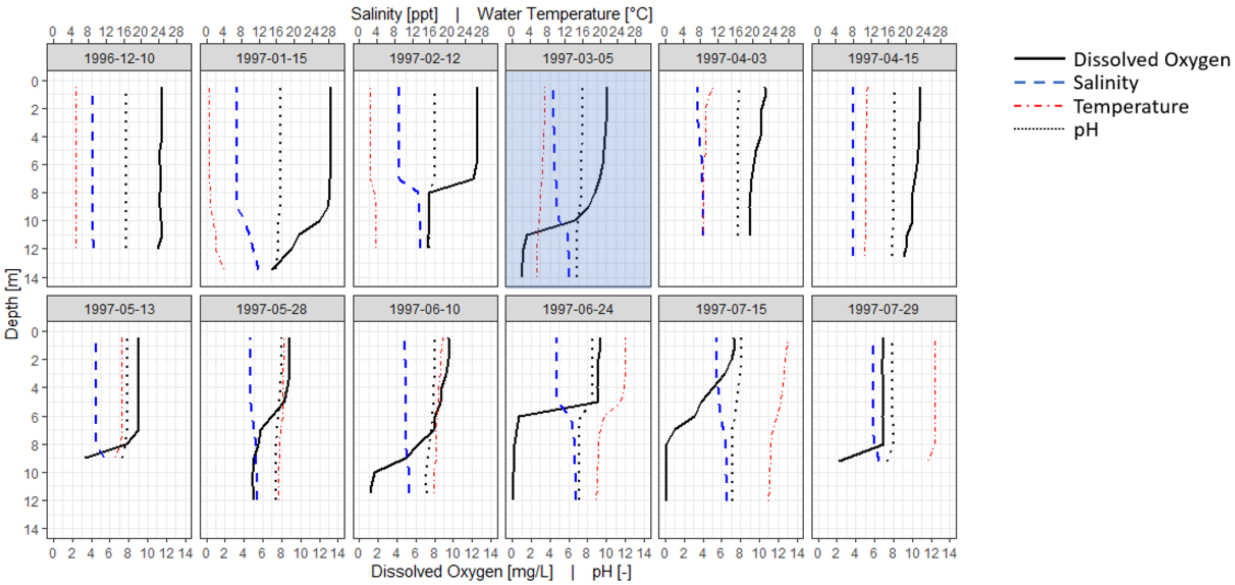


Figure 9. Water quality profiles at Station EE2.2 for selected events from December 10, 1996—July 29, 1997. The March 5, 1997 event (highlighted panel) includes samples to 14 m and includes lower than typical DO observations at the deepest sampling depths.

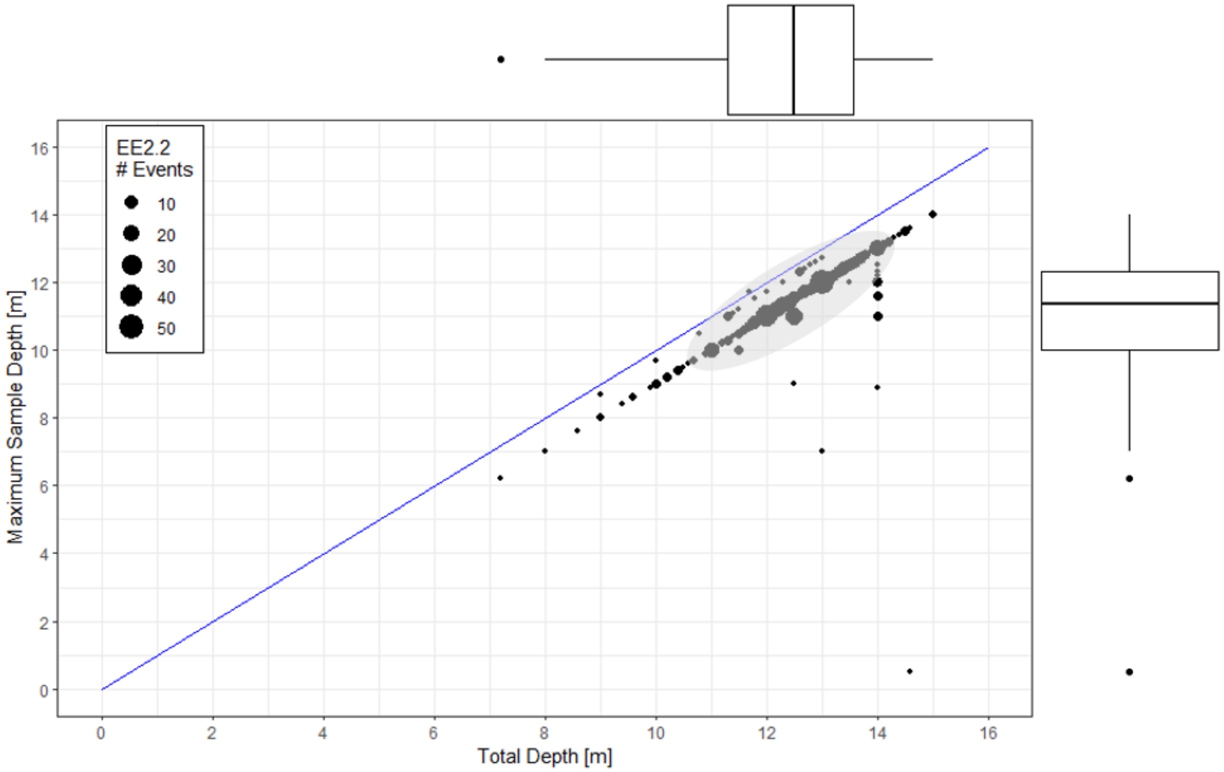


Figure 10. Comparison of the maximum sample depth associated with a DO measurement and corresponding reported total depth at Station EE2.2. Symbol size represents the number of sampling events. The maximum sample depth would generally be expected to at or below the blue one-to-one line.

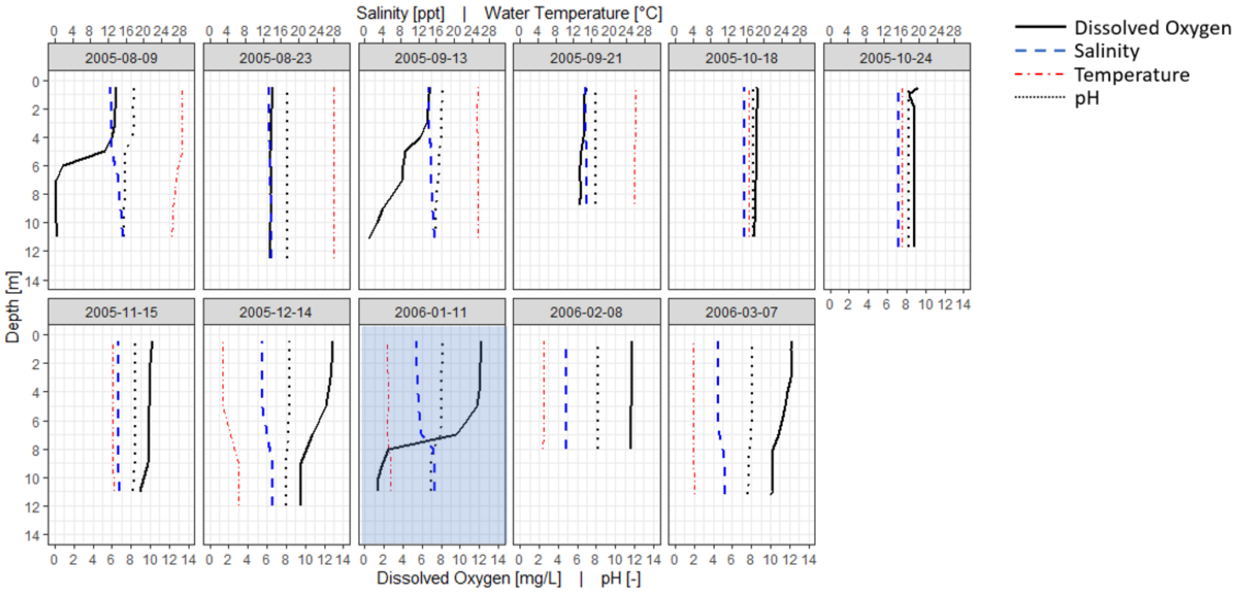


Figure 11. Water quality profiles at Station EE2.2 for selected events from August 9, 2005—March 7, 2006. The January 11, 2006 event (highlighted panel) includes lower than typical DO observations at the deepest sampling depths.

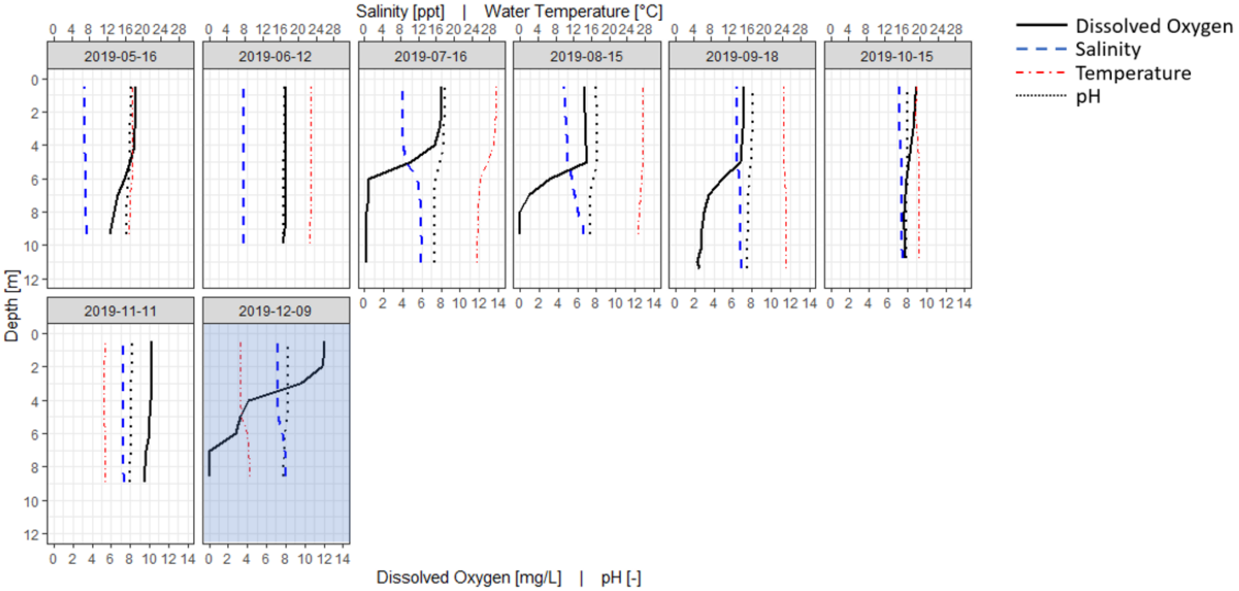
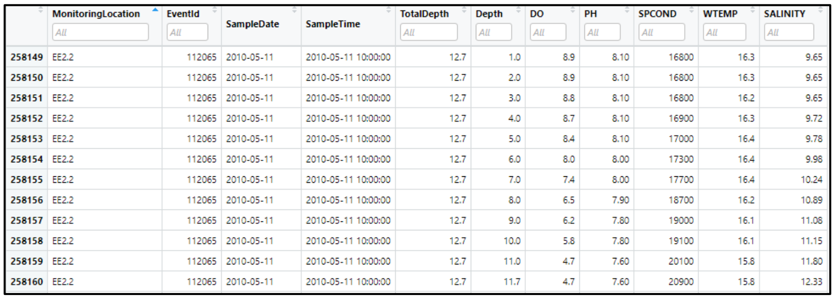


Figure 12. Water quality profiles at Station EE2.2 for selected events from May 16, 2019—December 9, 2019. The December 9, 2019 event (highlighted panel) includes lower than typical DO observations.

Data Pre-processing

Data from the CBP's DataHub (https://datahub.chesapeakebay.net/FileDownloads) were downloaded for the time period of 1990-2019. These data were filtered to the following parameters: dissolved oxygen [mg/L], pH [standard units], salinity [ppt], specific conductivity [µmhos/cm], and water temperature [ºC]. Preliminary inspection revealed that the same DO measurement value for the same depth was reported for multiple layers, e.g., a value of 7.5 mg/L for Layer "M" and Layer "AP" and a depth of 4 m. This duplication was addressed by removing one of the observations without regard to Layer. Approximately 86% of the DO data are associated with Method F04, 13% with Method F02, and the remainder with Method F02. DO was rarely reported for multiple methods at the same depth and was excluded from the processed data files, leaving EventId, MonitoringLocation, SampleDate, SampleTime, TotalDepth [m], and Depth [m] to characterize an event. There was a small number of records (less than 250) where different DO measurement values were reported for the same depth. Upon inspection, some of these cases (but not all) could be traced to sample replicates. Some of the replicate values did not align well with other DO measurements in the depth profile. In these cases, both observations were removed given the small number of cases in comparison to the overall large number (more than 500,000) of DO measurements. The resulting data set includes approximately 675,000 records of which 515,000 are associated with the 144 stations most often used for tidal trend analyses.



	MonitoringLocation	EventId	SampleDate	SampleTime	TotalDepth	Depth	DO	PH	SPCOND	WTEMP	SALINITY
	All	All			All	All	All	All	All	All	All
258149	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	1.0	8.9	8.10	16800	16.3	9.65
258150	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	2.0	8.9	8.10	16800	16.3	9.65
258151	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	3.0	8.8	8.10	16800	16.2	9.65
258152	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	4.0	8.7	8.10	16900	16.3	9.72
258153	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	5.0	8.4	8.10	17000	16.4	9.78
258154	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	6.0	8.0	8.00	17300	16.4	9.98
258155	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	7.0	7.4	8.00	17700	16.4	10.24
258156	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	8.0	6.5	7.90	18700	16.2	10.89
258157	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	9.0	6.2	7.80	19000	16.1	11.08
258158	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	10.0	5.8	7.80	19100	16.1	11.15
258159	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	11.0	4.7	7.60	20100	15.8	11.80
258160	EE2.2	112065	2010-05-11	2010-05-11 10:00:00	12.7	11.7	4.7	7.60	20900	15.8	12.33

Figure 13. Example data for Station EE2.2 on May 11, 2010.