Criteria Assessment Protocol Workgroup (CAP) Meeting

Wednesday, October 9, 2024
1:00-2:30PM

Meeting Materials

This meeting was be recorded for internal use to assure the accuracy of meeting notes.

MINUTES

NEXT STEPS:

- ✓ Invite Mark Fernandez (TetraTech) to a future CAP WG meeting to discuss his work.
- ✓ Leah Ettema (EPA) will share publications about the no allowable exceedance line.

PARTICIPANTS:

August Goldfischer (CRC), Cindy Johnson (VA DEQ), Amanda Shaver (VA DEQ), Clifton Bell (Brown & Caldwell), Tish Robertson (VA DEQ), Matt Stover (MDE), Becky Monahan (MDE), Sophia Grossweiler (MDE), Melinda Cutler (MDE), Michael Macon (MDE), Kenneth A. Moore (VIMS), Peter Tango (USGS), Rebecca Murphy (UMCES), Joseph Morina (VA DEQ), Leah Ettema (EPA), David Parrish (VIMS), Mark Trice (MD DNR), Carl Friedrichs (VIMS), Richard Tian (UMCES), Elgin Perry (Independent Statistician), Zhaoying (Angie) Wei (UMCSE), William Reay (VIMS)

1:00 PM **Welcome, Introductions & Announcements – Peter Tango** (USGS), Chair

Upcoming Conferences, Meetings, Workshops and Webinars:

- Potomac River Conference – October 17, 2024, Lorton, Virginia.
- Watershed Forum – October 18-20, 2024, Shepherdstown, West Virginia.
- 12th US Symposium on Harmful Algae – October 27-November 1, 2024, Portland, Maine.
- Maryland Water Monitoring Conference – November 21, 2024, Linthicum, Maryland.

- They are still accepting posters. Early bird registration due by October 25th
  - 14th National Monitoring Conference – March 10-12, 2025, Green Bay, Wisconsin.

Mark Trice (MD DNR): We had our first convening of folks who do hypoxia reporting and modeling. At the end of this month, we'll put out the year end hypoxia summary. From the monitoring results side and from the VIMS side, we saw that hypoxia was above average starting out the year and started out earlier than normal. Then we saw below average hypoxia volumes through the end of August. Then our recent late September cruise showed hypoxia was above average.

Peter Tango (USGS): I'll have to follow up with NOAA colleagues for collaboration and coordination on their hypoxia publication work.

Matt Stover (MDE) introduced Michael Macon who just started working with MDE. Michael has a background in water resource management. He worked with Water Rights in CA with the state water resource control board, and was a MD sea grant science and policy fellow with MDE, working on ecological restoration permitting and connecting climate pollution reduction plan to the Chesapeake Bay and TMDLs.

1:10 PM **Questions of interest from our Partners – Start to address and tee up for future meetings**

1. Background on how a particular DO criterion was measured during the original criterion development work, what endpoint it protects (e.g., spawning), and how it was intended to be assessed.

2. Further questions – can we add to this list as topics for this and future meetings?
    a. Discuss ideas for how we would use discrete data to assess this criterion, if at all.
    b. Discuss ideas for how we would use continuous data (i.e., profilers, arrays, surface Conmon devices) to assess this criterion.
    c. Discuss how many stations we would need spatially to assess a particular DO criterion/Designated Use combination
    d. Discuss how many measurements at any individual station (minimum data threshold) we would need to temporally assess a particular DO criterion/Designated Use combination

Presentation:

US EPA 2003 Criterion derivation publication was reflective of decades of previous work. Hundreds of scientists were consulted in the development of criteria for chlorophyll, dissolved oxygen (DO), submerged aquatic vegetation (SAV), clarity. Strong foundations of criteria derivation and application were published in 1992. Restoration goals were suggested providing for sufficient DO to support survival, growth and reproduction of living resources. It reflects what was known about the science targeting DO concentration, including durations and spatial extents and reflecting different temporal needs of living resources to get that spectrum we use. 17 pages of studies were summarized in the 1991 Habitat Requirements for Chesapeake Bay Living Resources. Between the 1991 and 1992 work to the 2003 criteria seasonality and geographic considerations were detailed in more depth. There was also a lot of work looking at life stage considerations that was included in 2003 criteria development.

In the 2003 documentation, the approach to deriving DO criteria is laid out in more detail. The foundational studies are still strong foundations. California has some work going on in San Francisco Bay looking for development of criteria there.

The 2003 documentation not only put the criteria in place but also put guidance for consistent Bay wide application of criteria across jurisdictional boundaries. This document reflects what we've used for decades with the Cumulative Frequency Distribution, and describing the interpolation approach with the Bay interpolator and inverse distance weighting algorithm. The text also reflects that there are applications to point, transect and remote sensing data. This is considering the chlorophyll assessments; with the work VIMs has done on SAV coverage and now greater exploration of use of remote sensing data.

Recommended implementation procedures are discussed. Under the DO criteria it says states can waive assessment of these criteria until monitoring at the required temporal scales is implemented, or apply statistical methods to estimate probable attainment. What EPA has pointed out is we do have recommended methods. There may be other methods. When we talk about probable attainment and estimates of attainment, one of the other ways we approached that has been using the multimetric water quality indicator. That was based on data derived relationships and rules. It is a reflection of probable attainment when we use that as an indicator, but the indicator itself hasn't been adopted as a methodology for supporting water quality standards attainment. It's meant to be guidance in terms of understanding and tracking another way we can look at the health of the Bay.

Background on assessments – USEPA 2003 provided a recommended approach to assessing criterion attainment using interpolation and the CFD. The TMDL planning targets are built upon applying the CFD approach in a bay-wide common method between monitoring and modeled data assessments in a partnership-wide consistent manner. We have strived to have a cross-jurisdictional common approach to the criteria assessment and other watershed agreement outcome assessment in the bay and across the watershed. We recognize EPA does not approve assessment procedures but evaluates assessment results considering protocols used to make the assessment decisions.

Going forward, we will continue the path of the summer 2024 meeting and build our list of questions on the criteria and their assessment. We will continue to vet our Questions and Answers on the criteria and their assessment. Coincident with ongoing methods developments, CAP WG will document a common community understanding on the issues. Our documentation historically gets summarized in a Technical Document, reviewed, and published through EPA as guidance to support updates to criteria or their assessment and assessment framework. We will likely need some homework and longer meetings for addressing these topics.

<u>Discussion:</u>

Tish Robertson (VA DEQ): Can you provide some clarity on the very first bullet in the agenda ("discuss ideas for how we would use discrete data to assess this criterion, if at all"). What criterion are we talking about and why would we not use discrete data?

Peter Tango (USGS): I suspect we're being asked for any given criterion season application, is there a reason we can't use the discrete data to assess that particular criterion.

Matt Stover (MDE): I wrote that. I was thinking in terms of if you only have one or two samples per 30-day period is that enough to assess a 1-day or 7-day average. That's if you didn't have any con-mon data. I look at this as evaluating each criterion on its own in different scenarios of monitoring density both spatially and temporally.

Tish: Relating that idea back to the passage Peter had highlighted about using statistical approaches. If you don't have the most robust data temporally you could use an alternative that allows you to make an inference about the probability of attainment. If you're using discrete data in that way, the sky is the limit.

Peter: I think we've touched on that a few times in the last few years, if the 30-day mean is being evaluated on a single sample, how are we not equally applying that on a 7-day or 1-day mean estimate. And that it's an instantaneous minimum (IM) measure, why shouldn't we use that for IMs. I don't have all those answers, but they are valid questions when we think about the probability of attainment when we think about having data but maybe not at the most robust level.

Leah Ettema (EPA): If we're potentially considering multiple lines of evidence using discrete and continuous data it might help to understand the variability associated with each data set. Are they coming from the same distributions? It could help understand what it's telling us as we're looking at both data sets.

Matt: Going back to what Tish mentioned, if we only have one data point in a 30-day period, should we even be doing a 30-day mean assessment. Tish has provided some good rationale for using the discrete data for analyzing the higher frequency criteria but what about the opposite question, too?

Melinda Cutler (MDE): Maybe this is related to the conversation where we're talking about minimum data requirements. For data that's informing our criteria like the 7-day and 30-day, I've tried to do a little bit of digging and it seems to be a lot of toxicology studies based on 24-hour or 96-hour study periods. I wonder how studies like that led to 7-day averages, how was data collected in the lab, what were the assumptions there and can that inform how we should interpret the criteria.

Peter: I was not in the particular group that laid that out. I don't know the specifics of how they created the spectrum when taking about the 7-day, 30-day. That is a question that deserves some additional insight.

Cliffton Bell (Brown & Caldwell): Those studies have quite a range of durations. Some of them are mentioned in the 2003 criteria document. You'd have to go back to that literature to see what each study did. For the short term criterion the minimum was 2 hours (effects on short nose sturgeon). A lot were 6-12 hour studies too that ended up being converted back to IMs. You might not have known when within those 2 hours or 6 hours that short term effect occurs. It highlights the conservative nature of how a lot of the criteria were derived. In general you went on the low side of averaging period compared to the actual averaging period of studies in the literature. That was intentionally done to be somewhat conservative.

Peter: If I go back to 1992 table and then look at 2003 table there were different considerations that were species specific, life stage specific, season specific considerations. For the IMs, you'll notice for example the deep channel we have 1 [mg/l]. If you look at deep water and open water, they are 2.7 and 3.2. There is a little bit of variety on the protections. That was reflective of which species – down in the channel we're trying to get some of those benthos that can survive in tough conditions in a little bit of oxygen, compared to something like open water where we're focused on species with greater sensitivities so the IM criterion I'm seeing there is something that's reflective of different sensitivities. If you go back to looking at table some of the considerations have been which species were considered in each of those habitats that were targeted for protections. It's not like our deep channel is being targeted for meeting striped bass criteria in the middle of summer when we're only getting 1 mg/l as our target. That's the additional insight I can offer. If you do go back to those studies we would welcome hearing about those details.

1:35 Discussion: Opening discussion on Computational Definitions for Dissolved Oxygen Criteria

> We need functional definitions for our means computations with high temporal density data. With 10-15-minute data (or similar, e.g., hourly), what is our procedure for computing instantaneous minimums (IMs), 1-day mean, 7-day mean, and 30-day mean?
>
> > Identify considerations for minimum data requirements, how to address gaps in data records with continuous data.

Presentation:

CAP had looked for ways to assess high frequency data. The group came up with 3 options. Tish was instrumental in this work. She tested some of these methods on data for some of the VA systems. The 3 options on how to use the con-mon for an IM were:

> Rule 1: No more than 10 percent of days during a single season with an exceedance— 9 total of 12 days can have a single exceedance. This translates into about 30 minutes x 12 or 5 hours total per season, and given 2880 hours in a summer season, about 0.17 percent of the summer season.

Rule 2: No more than 1 day with 10 percent time (>2.5 hours) exceedance during a single season. This translates into 3 or more hours or about 0.1 percent of the summer season.

Rule 2-Alternate. No more than two consecutive days with 10 percent time (>2.5 hours) exceedance during a single season. This translates into 6 or more hours or about 0.2 percent of the summer season.

We were discussing that the way the 30-day mean is looked at now in the CFD, it's a 10% curve. If you take 10% of instantaneous measurements, if they're the 15-minute samples, there are many more thousands of samples at a 10% allowable exceedance than perhaps what would be reasonable for an organism to expect to survive. The Murderkill River in DE has one day out of a summer season that they allow for a viable exceedance to meet their IM values, otherwise they're considered out of attainment. What we came up with here is in line with that body of work. Within what we have published, these are options that are referenceable and are not already applied. Tish's testing showed each of these 3 rules showed nearly identical results with slight differences.

Discussion:

Tish: On the last CAP call I mentioned I'd been thinking about the rule we came up for the IM on the con-mon. At the time it made perfect sense. I think running through actual monitoring data I started having some trepidation on the protection side. Because we were saying we could allow an unlimited number of days where we're exceeding the IM as long as there's not 2 days consecutive days in a row. I think we need another constraint on it. It's my job to think of worst-case scenario – which would be we have an entire season where we exceed the IM every other day. I was just thinking in terms of a straight percentage of days we were allowed to exceed. 10% maybe. Seems to me we need some way to keep from having 100 days where we're exceeding the high.

Peter: I agree. In my reflections back to Murderkill River in DE, they were clear about not allowing periods to repeat. Much more restrictive than 10% of days in a season.

Tish: With the 4-D interpolator, we will be implementing the IM criteria for all the designated uses, we probably need to for con-mon data set a method that maps onto how we would do it with 4-D interpolator. If we have a rule for consecutive days for actual empirical monitoring data, why wouldn't we have that for the 4-D interpolator. We should be consistent. I just don't know if we would have a way of predicting whether you have an excessive number of consecutive days of exceedance as well as a frequency of exceedance for the interpolator. If it's not do-able then, maybe we shouldn't have that condition on monitoring data on the continuous monitoring side and try to have something that's consistent.

Elgin: It is do-able with the way we're putting this 4-D interpolator together. Until we get it built and start validating how well it performs I can't speak to how realistic the assessments will be. For now, the machinery we're planning to build will allow for the assessment you're talking about.

Leah: Regarding the TetraTech study – they were answering a very similar question. How many days of data do you need to calculate a 30-day, 7-day mean minimum various statistics. They used continuous data and subsampled that down, and compared 1-day mean results to 7-day mean, compared 2 days to 7 days using the same data set comparison. Normally if you're comparing data sets a common test to use is the t-test. If it's significant you can say the data sets are different. Here we want to know if they're equivalent rather than different. So TetraTech did an equivalence test which flips that hypothesis. Mark Fernandez was the one who worked on it and would be the best person to describe that more. For this analysis they used rivers and streams data because that was the question they were asking. I think it's one option to consider for answering that question for the Bay data.

Peter: Maybe Mark would be amenable to joining us at a future meeting.

Leah: One other consideration: the minimum criteria here is expressed as an IM whereas in a lot of other states it's just a minimum without that duration specified or it's a daily minimum. For IM, EPA has guidance that means no instance in the water body should be below crit meaning there's no allowable exceedance frequency for IM when the criteria are expressed in that way. EPA has overlisted in NY for having 2% exceedance in DO because the criteria was expressed in instantaneous. From a scientific side I don't know if that's reasonable or original intent. I don't know if there's insights from how the instantaneous criteria were derived to explain why exceedances would be allowed, that might be helpful.

Peter: The longer you measure something, the more likely you are to encounter an anomaly. The improbability of not having an improbable event seems unlikely.

Matt: This is why we've been digging into the original studies done in the laboratory because those were the underpinnings of the criteria – and we're not questioning the criteria but how do we interpret those criteria. For those initial lab studies, how were they measuring DO in respect to those species? Were they doing it on a continuous basis or hourly basis? We run into this in the toxics world, or maybe for other parameters we can measure continuously. A lot of times the 304 a recommended criteria from EPA were developed on discrete data. The more you measure/the more you're capable of measuring, the more you're likely to find exceedances. We have to rethink how we're assessing criteria attainment because the original criteria were derived often times with much less frequent monitoring data. That's our thinking around the IM – it's designed to protect benthics which if they were exposed to levels below certain DO they would die. Should we be allowing any exceedances or more than one? I don't know. Without knowing more about those original studies it makes it difficult to interpret those criteria as they were originally designed and intended.

Peter: I agree and I respect what Tish did in VA with chlorophyll criteria and using 90th percentiles with means to be reflective of data distributions and not prescriptive down to a do or die threshold of exceedance. Thanks to everyone for the helpful insights and discussion to set the stage. Leah, if you have a publication that points to the no allowable exceedance line, please do share that.

Leah: I'll pass it along.

2:00 PM **Presentation: Water Clarity Assessment Approaches, and Update on Satellites and Turbidity** – *David Parrish (VIMS), Carl Friedrichs (VIMS)*

Presentation:

This work touches on work from many different people done over many years, including Elgin Perry, Ken Moore, Carl Friedrichs and Willy Reay. CBNERRs' monitoring program has been built around YSI sonds, which provides measurements of temperature, chlorophyll, turbidity, salinity, specific conductivity, pH, Dissolved Oxygen and Colored Dissolved Organic Matter (CDOM). The instruments are deployed at fixed stations (in a fixed location near the bottom of the water column in shallow water areas, collecting measurements every 15 minutes); at the Dataflow platform (mounted on the hull of a boat, with data collected every 2-3 seconds); and on a few of the CBIBs Buoys (at the water surface collecting data every 6 minutes).

One of the first uses of the Dataflow platform in VA was to assess water clarity. During these cruises, there will be 5 verification station stops on any given day. At these stations, LI-COR profiles are collected and used to calculate light attenuation. These are compared with monitoring data off the Dataflow platform. These paired measurements are used to model Kd from the monitoring data off the instruments, to get a multi-parameter linear regression using turbidity, chlorophyll and salinity off of the sond. This model is used for water clarity assessment.

Separately, they also interpolate the Dataflow data to give a bank to bank surface prediction of each of the parameters that go into the linear model (turbidity, chlorphyll and salinity). The method used is Kriging, which takes measurements from Dataflow and using those to predict turbidity outside the Dataflow cruise track. This can be done for the other parameters as well. Those results can be fed through the linear regression to get an estimate of Kd for the surface of the Chesapeake Bay segment of interest. Kd surface can be compared to Kd thresholds for the given application depth. Areas out to 2m are assessed; there's a separate Kd threshold for 0-1m and 1-2m, and for the polyhaline and mesolaline segments have a separate threshold from oligohaline and tidal fresh segments. Using the appropriate Kd threshold they can assess the Kd estimate from the Dataflow on that day compared to the Kd threshold. That results in a binary grid of attainment and nonattainment for that cruise. The cruises are then averaged across the year to get a water clarity acreage of attainment for the segment in a given year. That is compared to the water clarity acres goal for the segment.

Maryland assess water clarity in almost the same way. Some of the challenges with this methodology include:

- Segments with low goals can easily pass the water clarity acres goal by having one cruise with good clarity.
- Segments with high SAV goals and moderate/high SAV, may not pass water clarity acres due to insufficient remaining shallow water habitat due to the 2.5 multiplication factor.
- Spatial interpolation and modelling error is not accounted for in methodology.
- Spatial and temporal monitoring constraints limit data coverage (1 cruise per month).
- Sampling can be biased to good weather

Some opportunities include:

- Opportunity for development of Kd models by analyzing at a larger verification dataset (space and time) instead of focusing on current segment could improve models.
- Opportunity to integrate other existing datasets (ex. fixed stations, satellite)

In Virginia CBNERR has been monitoring and assessing water clarity for over 20 years. 2021 marked the first complete lap around VA waters of the Bay. The Rappahannock was revisited for the last 3 years (2021-2024). In 2024 they worked with DEQ to think about how to move forward with this monitoring program and what the priorities are. They're changing a couple of things going forward. Typically they rotate fixed stations where they're using Dataflow to monitor water clarity. They've been using 5 fixed stations in the region. They'll leave 3 of the instruments in the Rappahannock, which will turn into long term stations. In 2024 they started monitoring in the Mobjack and York polyhaline and mesohaline with the Dataflow platform and monthly sampling. In 2025 they will continue with this sampling design, plus adding 2 sentinel long term stations in the James. Currently they are thinking about where to place those.

As they move forward, they're thinking about how to integrate additional datasets with the monitoring data to further capture spatial and temporal water quality variability in the surface waters. They've been working on anchoring satellite imagery with their monitoring institute data with the idea being that Dataflow can provide thousands of verification measurements for a satellite image on a single day, and fixed station can provide hundreds of verification measurements in a year. They're looking at how to leverage all of these data sets to develop predictive models of water quality.

To explore this idea they started with a data set from Planet. They have these microsatellites that are small enough to fit on a bookshelf. They collect high resolution satellite imagery – 8 band imagery at 3m resolution. One of the reasons this data set was selected was it is near daily coverage of the Bay.

The methodology is: Acquire imagery from Planet (since 2022); do atmospheric correction (ACOLITE)-> 8 surface reflectance bands; match surface reflectance to CBNERR-VA fixed station

and Dataflow turbidity measurements; and develop predictive models using paired data sets to estimate turbidity from 8 bands and depth. The current data set they're working with is in the York River estuary. They've paired 5 Dataflow samples from 2024 and 2023 with satellite images, and paired it at 2 of the fixed stations and 1 buoy. After pairing the data they pulled the data across the stations and did exploratory modeling. Everything is working fairly well. They found that they can simplify the number of bands needed to predict turbidity. Things are fitting well even with simple linear modeling.

They've also done some modeling with the Dataflow data – pairing the Dataflow data with the satellite on day of sampling. There is a lot of agreement. 5 of the cruises have been modeled. There's a lot of spatial autocorrelation in the datasets, which they addressed with some block cross validation to ensure they're validating with dissimilar data to the model training data sets.

Early data exploration shows promise for anchoring satellite imagery with dataflow and CMON monitoring platforms to estimate surface turbidity. Kd can be modelled using monitoring data paired with verification measurements – a similar idea to current assessment methodology. Using a hierarchical model we could potentially estimate a Kd surface whenever there is a clear satellite image (~ 100 estimates a year). Kd surface estimates could be compared to Kd thresholds as with the current methodology.

Next steps include:

- Continue expanding datasets.
- Merge dataflow and fixed station datasets for modeling effort.
- Add turbidity measurements from fixed station as a predictor in Turb ~ bands model.
- Tune model for extrapolation between monitoring datasets.
- Assess use of these models for water clarity standards assessment – compare results with previous methods.
- Leverage supporting datasets to interpolate in time b/w satellite estimates.

A lot of monitoring data is available on the VECOS platform. Many people have contributed to these datasets.

<u>Discussion:</u>

Mark Trice (MD DNR): You said you averaged the month and it indicated you created an average GIS surface from all the Kd layers for each month. We (in Maryland) actually get the water clarity acres for each month and average those.

Dave: That's a good clarification, we do the same thing (in Virginia).

Peter 2.5 water clarity acres is equivalent to one SAV real acre and there's a translation that happens in the assessments.

Mark: We had a similar kind of thing that happened in the Patuxent River. We worked with what used to be called the NOAA costal services center. It was an intense effort. DNR and MDE went out, rented YSI meters to put along the Patuxent, and did remote sensing. There was a report that was a match-up of various parameters. I'll try to find that. One of the first things I did at DNR was a remote sensing conference.

Estuarine and Watershed Monitoring Using Remote Sensing Technology

Ken Moore: There was a lot of consternation when we came up with water clarity acres. There was a table that was total available bottom out to 1 or 2m zone in a segment vs how much SAV was there. The average of all that came up with 2.5. A lot of us said now that we've moved forward with these satellite data, we should focus on historic areas that were vegetated and we can look at water clarity within those and use that as our standard not just as what proportion of that acreage did we need to 2.5 times. In some places we won't be able to, in some places we'll never be able to. What are your thoughts on that?

Dave: I think that's an interesting idea. I've also had thoughts on if we looked at that assessment domain, you're sort of restricting yourself by the 0-2m area you're looking at, where we have a lot of information about say, light conditions in a given segment. We have lots of data about that. When you restrict the assessment protocol to having to be in the 0-2m zone, you have to start doing these interpolations away from where we sample the data. I wonder if we were to shift the assessment methodology we could do it in a way that's based on how we characterize the variability in light conditions in the segment where we monitor as opposed to trying to make an estimate in a particular area.

Tish: Ken, do you think the 2.5 multiplier is pretty conservative? If we were to actually make a site-specific, segment-specific target for water clarity, do you think we would come up with higher goals or lower goals than we currently have?

Ken: I think it depends. In areas where the SAV is doing well, we have large coverage, it may be more conservative than needed. That would be a good test. I think that would be a thing to look at.

Tish: It makes sense to make the water clarity goal as conservative as possible since you may not have SAV at all in a segment, so you wouldn't want to have a water clarity goal that wouldn't be titrated to a condition where there is SAV. You'd expect SAV to do a significant amount of baffling in the water column. If you had SAV there, you wouldn't have to have such a high water clarity goal. But if you didn't have any and wanted to build it back I could see you wanting to be as conservative as possible. That would be my concern with coming up with a segment specific target. You want it to be tied into where you may not have SAV.

Ken: Keep in mind for the lower Bay water temperature which we had a sentence we put in about temperature and light requirements being inter-related. To some degree perhaps our Kd values are not strict enough now, 30 years later.

2:30 PM **Adjourn**