Chesapeake Bay Program
Science. Restoration. Partnership.

Criteria Assessment Protocol Workgroup (CAP) Meeting

Tuesday, August 13, 2024
9:00AM-4:00PM

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

Next Steps:

- ✓ The CAP WG will start meeting every other month instead of quarterly.
- ✓ CBP staff will meet with MDE staff to discuss the 3D and 4D interpolators more in depth and have a conversation about MDE's needs.
- ✓ CBP staff will meet with VA DEQ staff to discuss the 3D and 4D interpolators in more depth and have a conversation about VA DEQ's needs.
- ✓ CAP WG discuss data drift when working with continuous sonds.
- ✓ Agencies will share any additional questions they have that we did not address during this meeting that can be addressed at future CAP WG meetings. Questions can be shared with Peter Tango (ptango@chesapeakebay.net) and August Goldfischer (agoldfischer@chesapeakebay.net) for incorporation into future agendas.
- ✓ The CAP WG will have future discussions on the definition of instantaneous minimum.

Participants:

Alex Gunnerson (USGS), Amanda Shaver (VA DEQ), Angie Wei (UMCES), August Goldfischer (CRC), Becky Monahan (MDE), Breck Sullivan (USGS), Bryant Thomas (VA DEQ), Carl Friedrichs (VIMS), Cindy Johnson (VA DEQ), Claire Buchanan (ICPRB), Clifton Bell (Brown & Caldwell), Dustin Shull (PA DEP), Elgin Perry (independent statistician), Gabriel Duran (CRC), Gary Shenk (USGS), Guido Yayacto (MDE), Jillian Adair (EPA), Joseph Morina (VA DEQ), Juan Vicenty-Gonzalez (EPA), Kelly Gable (EPA), Leah Ettema (EPA), Lew Linker (EPA), Mark Trice (MD DNR), Matthew Stover (MDE), Peter Tango (USGS), Richard Tian (UMCES), Sophia Grossweiler (MDE), Suzanne Trevena (EPA), Tish Robertson (VA DEQ), Tom Parham (MD DNR)

Minutes:

9:00 AM **Arrival, coffee, mingling**

9:30 AM **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.
- 14th National Monitoring Conference – March 10-12, 2025, Green Bay, Wisconsin.

Meeting Objectives:

- Continue building common understanding in our community on evolving and improving our Chesapeake Bay monitoring and regulatory assessment frameworks.
- Take stock of Dissolved Oxygen (D.O.) criteria assessment options that may help address all Chesapeake Bay D.O. criteria durations with existing tools.
- Continue building community understanding of the new tidal Chesapeake Bay water quality interpolator with opportunities for input on interface needs.

9:45 AM Session 1: Monitoring and assessment framework

- 9:45 EPA Requirements and expectations for 303(d)/305(b) assessment – Leah Ettema, EPA

Presentation:

Leah provided an overview of EPA's expectations and Clean Water Act (CWA) expectations for Integrated Reports (IR). One of the main uses of protocols is for state's IRs. The Integrated part of the Report refers to CWA sections 303(d), 305(b) and 314. Leah went through the text of sections 303(d) and 305(b). The 314 section covers reporting on publicly owned lakes. EPA suggests states reports on all of their waters by putting all water body use parameter combinations into certain categories – for example, Chesapeake Bay segment CBH.7, open water dissolved oxygen; or deep channel dissolved oxygen. Every single one of those water body use parameter combinations for the entire state should be placed into one of the five reporting categories for waterbody condition. The 303(d) list is *not* the same as the impaired waters list; the 303(d) list is Category 5, while the impaired waters list is Category 4 and Category 5. From an environmental perspective, they're both impaired, but administratively, one already has a Total Maximum Daily Load (TMDL) (Category 5), and the other (Category 4) does not. For the Bay segments, there is a TMDL in place, so all of the assessment units are not on the 303(d) list, they're either in Category 4 with a TMDL, or attaining. That affects the administrative procedures from EPA's standpoint.

Leah also noted that the integrated reporting process is separate from the TMDL process. Determining that an impaired water body with a TMDL is attaining is a good news story, and states will report it in their IR, but that doesn't remove any TMDL requirements. It is a state decision on whether to keep or withdraw the TMDL, and that is a separate administrative process to change a waterbody category from 4 to 2.

The data used in states' IRs is available through the Water Quality Portal (after being uploaded through WQX). States make their decisions, and these decisions are housed in the EPA's assessment database Attains that contains every single assessment unit in the state, the name, the parameter, the designated use, the cycle first listed, the IR category, and comments. The public can see this information on How's My Waterway. States Report on state assessment units in the IR, which don't necessarily align with the Chesapeake Bay segmentation that the 3D Interpolator outputs are for. Much of EPA's expectations is outlined in the IR guidance, which is issued every 2 years. There is little discussion about characterizing the spatial and temporal extend, how much data you need an assessment unit. The most applicable information says the methodology should describe the decision logic used to determine the temporal and spatial extent samples represent.

There are few regulations outlining the expectation to use all readily available data, but there is text outlining the expectation in 130.7. It also says you have to provide a rationale for not using any data. EPA IR guidance expands on this, including that lack of a State-approved Quality Assurance Project Plan (QAPP) should not be used as the basis for rejecting data, and the state should provide specific technical reasons for not using data. This is a hard requirement to meet since there is a lot of data available. All of this rationale is to bolster the ability to make a sound, science-based decision, and to mitigate litigation risk. The IR guidance also outlines that models are considered readily available data that needs to be considered. So the output of the 3D interpolator or any other assessment methodology needs to be considered when making IR decisions. The Bay is not the only area where there could be multiple lines of data considered to make a decision. EPA would expect each line of evidence to be evaluated, and if one line is relied on more than another, there is a description for why it is used in making these decisions, and that often related to evaluations of uncertainty or quality with each of the lines of evidence. If one is more spatially or temporarily representative, that could be a reason to rely more on that dataset. IR guidance also says assessment methodology should be consistent with water quality standards and sound science and statistics. The Bay criteria outlines recommended implementation procedures. EPA doesn't approve assessment methodologies, only the 303(d) list and standards.

If a state's final decision aligns with EPA's final decision, EPA will approve the list. EPA approves or disapproves the final 303(d) list and removal of waters from the list. Movements from different categories are not technically approved by EPA as they are outside of EPA's approval authority, but EPA will review and comment on them.

Discussion:
Joe Morina: For data that's not quality assured, EPA says states have to come up with an argument for why that shouldn't be used – from my perspective if we can't control the quality of the data that would fall in that category of not scientifically valid.
Leah Ettema: We would need a rationale for that specific data set for why it's not quality. Did they not provide the protocols that they used? Were the protocols different from what your state would use, making the data not representative of the stream you're trying to assess? Having a QAPP doesn't guarantee quality data – a QAPP is a plan for collection, and documentation. But just because you don't have that doesn't mean the data you collect isn't high quality. The EPA's expectation is states provide specific information about the dataset that's being reviewed.
Bryant Thomas: Many of the things you talked about are things we worked through together with our IR submittals. This is a great way to start the day. Your comments and remarks set the stage well for the last item on the agenda, the roles and responsibilities in Bay criteria assessment.
Richard Tian: For category 3, there's not enough data for assessment. In that case, what does EPA do? You require the state to collect data, or just to leave it as is?
Leah Ettema: This is generally where states have some data, but not enough confidence in that data to make a decision. Perhaps there's only a few samples collected over the years. So this is a way for them to say there's some data but not enough to make a final decision.
Gary Shenk: Within EPA region 3, you're looking at standards assessment, but also EPA region 3 approves TMDLs, so is there a connection with the folks approving the TMDLs? Do they look at water quality standards assessments as part of that, or is it separate?

Leah Ettema: When we take an action, EPA is reviewing the regulations for each action. There are separate water quality standards regulations, separate 303(d) listing regulations, separate TMDL requirements regulations. When a TMDL is submitted for us to review that's the framework we're reviewing it under. We have state coordinators who do all 3 things. In our region, most are familiar with the entire states' protocols. But in our review, assessment methodology would not be reviewed or considered in our review of a TMDL. They're two separate administrative processes.
Bryant Thomas: Currently I don't work in TMDLs, but I worked in one of our regional offices. Where there is a criterion, in the TMDL arena we do look at the criterion and expression of the criterion (duration, frequency, magnitude), we do also consider the assessment methodologies that may be involved. Sometimes we've had phased TMDLs from the implementation standpoint to get to achieving standards through the assessment process where there may be an allowable deviation from 100% compliance. From a programmatic standpoint we do look at the criteria and assessment procedures. That's from Virginia's perspective, not necessarily in EPA's review.
Clifton Bell: I can think of a situation in South Carolina where on the TMDL side South Carolina was trying to give a certain allowance based on their assessment methodology but it wasn't written into the standard. I think EPA Region 4 had issues with it. I don't think it was completely resolved. Can you speak to that in terms of the extent of exceedance allowed on the assessment side, how explicitly can you consider that on the TMDL side?
Leah Ettema: I'm not the best person to answer this in my region, I'm not a TMDL reviewer. But TMDLs have to demonstrate that they can achieve standards. From a regional standpoint, it's reasonable to consider the exceedance frequency that states use in assessment methodologies. A bit depends on the model. I am only the TMDL coordinator for West Virginia, but I know when they do their fecal monitoring, they incorporate that exceedance frequency into their models when they're predicting whether a segment would be impaired or not. It can be appropriate but it is a TMDL specific determination on how you connect standards assessment and TMDL.

- 10:15 MDE presentation on regulatory needs and perspectives

Presentation:
Matthew Stover shared MDE's perspective on the IR first. He considers going from standards to implementation as a cascading level of impacts, and the IR plays an important role by identifying impairments, and the impairments have real world implications for National Pollutant Discharge Elimination System (NPDES) permits, water quality certifications, and voluntary actions. States are not required to monitor every water body segment in a 2-year span. MD does use cycling strategies, which helps them put resources in focus areas. That's something to keep in mind when considering where to put new con-mons and monitoring resources.

Every state tries to use the most up to date information. Something we use in the IR is the assessment methodology, which is what we work on most at CAP – how we get from a criterion without a ton of details to an assessment of an impairment or whether we're exceeding water quality criteria. That involves the statistics, spatial description and any other logical framework involved. That's what MDE thinks the CAP WG should spend most of its time on developing. All the technical addendums are essentially assessment methodologies for DO, water clarity, Submerged Aquatic Vegetation (SAV) and Chlorophyll-a.
Matt then shared the timeline of MDE's IR submission. The submission deadline is April 1 on every even numbered year, which is preceded by a public comment period and the writing of the report.

It’s a lengthy process especially with a limited number of staff and a lot of data received from a large variety of constituents, and that data has to be quality assured. Matt said it would be ideal if they could have Bay Dissolved Oxygen (DO) stoplight assessments, which are a critical part of their IR, in July in the odd-numbered year preceding the IR submission.

Matt shared MDE’s general perspectives. MDE thinks the regulatory needs of the Bay states should be the highest priority for the CAP, BORG, Hypoxia Collaborative and Integrated Monitoring Networks workgroups as a whole. This is because we’re currently unable to comprehensively show measured (not modeled) progress in how nutrient reductions are impacting DO in the Bay. MDE sees that as critically important to show some degree of success in our Bay restoration efforts as this helps us to keep momentum in driving restoration. It is a key part of messaging which goes to constituents who are providing funding. Assessing all applicable DO criteria in all designated uses in the Bay is the top priority for MDE.

MDE thinks that while making revisions to the DO water quality criteria to address impacts from climate change is a worthy goal and something to address, it’s not the most pressing need we have in the CAP WG or any other standards focused WG. We’re still grappling with the challenges of figuring out what loads will help us meet our current criteria. Some of the additional loads required to meet the Bay TMDL under climate change scenarios are evolving in terms of the science. Until we can figure out how to assess the current criteria, we shouldn’t be changing the goal posts. MDE recommends tabling this for now and coming back to it, addressing foundational questions first such as the short duration criteria, developing minimum data requirements for spatial and temporal coverage, and coming up with the plan to guide placement of monitoring resources.

MDE’s perspective on monitoring is to narrow focus on assessing DO criteria that apply to every designated use in the Bay and that should be done collectively, with regulatory needs taking primacy over other goals (such as research). Where interests align assessment goals (e.g., regulatory and research), it makes sense to select locations with co-benefits.

MDE is excited about the hypoxia collaborative funding proposal, especially once they got more of the details. It’s exactly what MD needs: contractor assistance to help coordinate these larger efforts that involve coming to consensus on where to place con-mons. MDE doesn’t want to have 92 meetings and is on-board to develop a logical framework that can be applied to different types of segments.

For assessment, MDE would still like to have CBP’s technical assistance to run the 3D interpolator. They want it by July 1, and want to have it more up-to-date. For the most recent IR in 2024, the 3 years assessed were 2018-2020. For the 2026 IR, it would be ideal to have 2022-2024 assessed. Right now MDE doesn’t have the capability to run the interpolator themselves, but would like to learn how to run it and more in-depth of how it works. There are a lot of Tier 3 data resources that can be used in running the assessments, and MDE would like to use as much of the data as possible (and as they are required to).

MD would also like CBP to run the data from Fishing Bay through the 3D Interpolator so that we can compare the results from the alternative assessment methods MD is developing. As part of this work, MDE would like to meet with CBP staff to better understand the decision rules in using the 3D Interpolator, how different datasets are included, and the limitations of this method.

MDE recommends the future directions of the CAP, BORG and Hypoxia collaborative focus on:

- Developing detailed decision rules for an assessment methodology for the short duration criteria (and even the 30 day mean DO criteria).
- Answering the following questions:
  - How to interpret DO criteria (e.g., Does instant. min = never to exceed? One exceedance in 3 yrs? 10% exceedance? 1% exceedance?)
  - How should we calculate a 7-day mean?
  - What is the minimum spatial and temporal data threshold to be met to assess a segment-designated use combination?
  - What zones (e.g., trib of trib, shallow water, etc) within each designated use should have data for assessment?

MDE would like CBP's assistance in addressing these fundamental criteria interpretation questions and helping us to develop and evaluate options for assessment methodologies including those that do not require the use of a model or synthesized data, and find agreement on best options.

Matt shared MDE's major concerns with the 4D interpolator:

- There isn't an established assessment methodology for many of the short duration DO criteria and the 30-day mean DO criteria assessment could also use reconsideration. How can we develop a tool for assessment when we haven't laid the groundwork for how the assessment should be done?
- Should we pause development of the 4D until we answer these fundamental questions?
- Rarifying ConMon data down to daily measurements reduces its temporal measurement advantage.
- Synthesizing the temporal aspect of the data rather than using high frequency monitoring data.
- Difficulty understanding how it works and concerned that it will be too complex for States to understand, run, and to independently verify the results. Ultimately MDE needs to be able to explain results to their leadership and to the public and don't feel able to do that currently.

MDE plans to explore alternative assessment methodologies to the 3D Interpolator and the 4D Interpolator. They'd like to consider simpler solutions than the 4D Interpolator, and recommend pausing development of the 4D Interpolator until we address fundamental questions with how to assess the DO criteria.

Discussion:

Peter Tango: Thank you for your presentation. We can certainly meet more frequently with all of these topics of discussion.

Lew Linker: I second postponing a reassessment of climate change water quality standards for the Chesapeake Bay TMDL. I was speaking with Council in Region 3 and they had warned that Federal actions are now more easily challenged. The 2010 TMDL was the last Federal action in the CBP. What would trigger challenges to the Chesapeake TMDL are water quality standards replacements

or updating. Therefore EPA is unlikely to support a reassessment of water quality standards for implementation in the Chesapeake TMDL. Assessment for a scientific analysis is welcome.
Bryant Thomas: The topics Matt brings up are all worth additional conversation. I think focusing on methodology; VA's needs and suggestions may be about who does what, but I think consistency in approaches and methodologies is a good goal to work towards.
Gary Shenk: I understand you (Matt) may feel like all of these things have been rolling down the tracks quickly without opportunity for questions and conversations. Perhaps it would be helpful for us to have a conversation with MDE, myself, Peter, Richard, so we can get on the same page about what we're trying to accomplish, and what your needs are.
Matt Stover: We'd love to do that.
Gary Shenk: We can do that for DEQ as well.
Kelly Gable: Following up on Lew's comment, for clarity, there is nothing in the CWA that specifically states or requires that TMDL needs to be modified if X happens. CWA doesn't speak to modifications of TMDL. Of course TMDLs can be modified or new TMDLs can be established, but there's nothing in the CWA that says if A, then B. There's not a requirement it happens in a certain circumstance.
Lew Linker: Thank you for that clarification.

- 10:45 MD's Enhanced Monitoring - Fishing Bay case study – Becky Monahan, MDE; Sophia Grossweiler, MDE
    - Sharing concerns/requests using these data for assessment

Presentation:
MDE is pursuing a pilot project and case study in Fishing Bay. This is an update of the work since 2020. Billions have been invested in Bay restoration, nitrogen and phosphorous are improving, we need to show results and yet all of our tidal waters are either listed as impaired or having insufficient information to assess for DO. We know that not every segment is impaired, we have the data to show that DO is improving. Yet we're not seeing it in our assessments. It comes down to not being able to assess all of the criteria.
The goals of our pilot project:

- Develop a process to monitor and assess all DO criteria for all uses within a Bay segment – the most important part of the project.
- Demonstrate restoration success story or at least show a segment in good health.
- Apply the lessons learned from this project in the future segments.

General Steps of the project were:

- Pick candidate segments.
- Develop a 3 year monitoring plan.
- Execute the monitoring plan.
- Assess the data using all available tools.

Fishing Bay Mesohaline (FSBMH) was selected for the pilot because it met Open Water (OW) DO Criteria for Summer and non-Summer; nutrient Indicator trends are improving (TP, TN, TSS, and DO); it met its SAV restoration goal; had no major logistical barriers (none were foreseen, but they did encounter some); was a simple (turned out to not be as simple as thought) pilot since the only designated uses present: OW and Migratory Spawning and Nursery (MSN); was smaller and shallower in depth; and is currently not assessed as impaired.

MDE partnered with MD DNR to conduct the pilot project. Collaboratively they planned the logistics and what kinds of sampling can be used to assess each criterion within each zone and where is there overlap, whether they would use discrete, con-mon or profilers or a mix of all. Ultimately they decided on doing a combination of discrete and continuous monitoring, covering all zones and all uses. Becky showed a map of all of the monitoring stations and how they decided on the stations. They collaborated with MDE's shellfish group which has been collecting DO data at certain shellfish stations for a couple years now. They submit their data through the Chesapeake Monitoring Cooperative's data portal. The last round of interpolator results included this data as well as other Tier 3 data.

The shellfish group collected discrete monitoring data. To get continuous data and a profile, NOAA lent DNR a profiler. DNR is maintaining that equipment. MDE purchased 4 additional continuous monitors as well. 4 were needed for rotation and cleaning. DNR placed the sonds.

Sampling began on April 25th, 2022 and will continue through the end of 2024. Due to some issues encountered along the way, actual monitoring sites had to be changed from originally envisioned monitoring sites. Becky shared the lessons learned from this project. Equipment challenges were one issue. The profiler needed maintenance. There was a lot of biofouling that impacted the sonds and probes. There was a lot of data drift and had to take out huge chunks of the data due to this. DNR is doing the monitoring and they had other projects. Becky suggested data drift is a future topic for the CAP WG since it is an issue when placing continuous sounds. MDE purchased additional sonds with a central wiper to keep the sonds cleaner from biofouling. They just placed these sonds this summer. It's providing an extra week.

Additionally, some stations were moved. They've thought about putting it by a moored station. They did that for one continuous monitor, but then moved it because it was too close to another continuous monitor station. That is a challenge for analyzing the data. Now this particular station is further out at the mouth of the Bay. DNR talked with the watermen and asked where to put the sonds where people won't move them.

They started to look at some of the data coming in. That's how they noticed the biofouling. They also noticed the 1404200 station was showing the lowest DO. They're trying to figure out if it's natural or not. Is it because it's a discrete station? For this summer, they added stations next to it to see if the low DO is going into the Fishing Bay larger segment. MDE shellfish will be collecting DO there. They also recently asked them to collect BOD and DOC. They're wondering if it's a black water stream or brown water stream. It's close to black water national wildlife refuge. There's also a wetland area.

The next step is to assess the data using all the available tools. One of those tools is the interpolator. However one issue is all the criteria can't be assessed. [the audio cut out here]

The continuous monitoring based assessment of the instantaneous minimum is listed in the 2017 technical addendum. We know we'll need continuous data to get to that assessment. Discussions on how we do that with a spatial and temporal array of data is a great use of CAP WG. This is why we'd like to have Fishing Bay data run through the interpolator if possible. I don't know if the interpolator can handle continuous monitoring data. We'd like to run it through the interpolator so if we decide to do a 10% rule we know how it compares to other stations and segments.

Sophia:

Preliminary assessment. Thanks for VA for sharing their R scripts. They used those scripts with a couple modifications. They didn't include rounding rules. They assessed all applicable samples against IM criteria regardless of sample layer. Assessed profiler stations at max depth. Required at

least 75% monitored days to calculate 7-day and 30-day means (not a modification but something they carried forward from VA). The main objective was to get a station by station assessment first before assessing Fishing Bay as a whole.
They found that most of Fishing Bay does not have any exceedances.
For the stations that were moved multiple times each data chunk was treated like a separate station site, rather than combined together. Applicable criteria for Fishing Bay were Migratory fish spawning and nursery use and open water fish and shellfish use.
The general trend of the continuous monitoring stations, there were a couple of exceedances in the upper portion of Fishing Bay whereas the traveling station in the lower portion of Fishing Bay looked better. Exceedances were highest in the summer. Around 5-6% for instantaneous minimum and 7 day mean. The continuous monitoring station had trouble meeting the instantaneous minimum. This also speaks to biofouling and other issues we had with continuous station in summer 2022.
The general trend for discrete stations showed there was just one station that had the highest exceedance rate for open water 7 day mean (27%) in the June-January time frame. Several of the discrete stations were deep enough to get a depth profile. Readings were consistent across depths at a particular station. They looked if there would be a change in the assessment if samples were split by depth. There were generally 3 different depth samples. The sample sizes are similar and there were the same number of exceedances in each depth. The exceedance rates increase as criteria become more stringent. All exceedances were in the June 1 – January 31 time period.
Conmon profilers closest to 1404200 mostly showed no exceedances but they will be collecting additional data around the station to investigate how far the low DO concentrations extend.
Next steps for the project include developing discrete + continuous monitoring assessment methodology and fully assess the Fishing Bay segment. Additionally, MDE requests CBP assistance to run this data through the 3D and 4D interpolator and assist with comparisons of assessment methodologies.
Questions and concerns arising from this project include:

- Blackwater conditions and potential parallel with POCOH/POCTF
- How to assess space represented by one low DO station?
- How to include profiler data in the assessment?
- What should be the frequency of exceedance for IM and 7- day mean?
- Should we assess each station individually rather than aggregated by monitoring frequency?
- Should we continue data collection based on current results or other factors? MDE requests feedback on this.

They hope to apply this to other segments in the future and achieve similar results with less stations. They hope to apply lessons learned from Fishing Bay to the GIT funded proposal for monitoring design plan.

Discussion:
Lew Linker: This is encouraging to see this wealth of observed data. A priori we thought with the MBM schism the Fishing Bay was going to be a challenging area to calibrate, and important for shallow water. It will be useful to use this data. I think your team should consider the prodigious productivity of tidal wetlands. All of that organic matter decays and respires. It’s not uncommon to see other areas with tidal wetlands like the Pocomoke and Mataponi to have variances that allow

DO to go as low as 4.5 for open water. I think that's what we're seeing for 1404200. That station is upstream surrounded by tidal wetlands. A lot of consumption of oxygen is taking place by those tidal wetlands, export of DOM which is causing additional respiration in the water column. 1404100 is also up in the wetlands. The other stations have more mixing, open water and chance to bring that oxygen up. But in the tidal wetlands areas you'll see DO like what you're seeing. Variances are not uncommon in those situations and it's a way to acknowledge the natural conditions.
Peter Tango: In Sasoon Marsh in San Francisco Bay they looked at similar reflections for young of year habitat. The ponds upstream were managed for water fowl hunting. They would dam the water and when they released the water that had been sitting there those low DO slugs bolted downstream.
Leah Ettema: Thank you for presenting this data. I think all the questions you posed regarding observed data methodology are important, and if that's the path forward important for the CAP WG to weigh in on. Like you said, the 10% exceedance, we don't know what is appropriate. We could have another presentation on why EPA may or may not allow a 10% exceedance of something that's expressed as an instantaneous criteria. That's another policy discussion to be had.
Peter Tango: There's no general rule that's a percent exceedance relative to the criteria duration? That's ripe material for us to pursue?
Leah Ettema: Correct.
Elgin Perry: Echoing Leah's comments that this data is very valuable. Did you collect any DataFlow data as part of this study?
Becky Monahan: No.
Peter Tango: Did you collect temperature and salinity as well as DO? Which could provide insight into habitat differences.
Mark Trice: Yes.
Gary Shenk: Great having all this data. You were aggregating across continuous monitoring and looking at that vs having a model. I think one way to think about it is that you still have a model, but your model is that all stations have the same spatial importance, and the only thing is assigning greater spatial importance to those places which become important.

11:30 AM 10-minute Break

11:40 AM Session 1: Monitoring and assessment framework - continued

- 11:40 VA DEQ will share their ideas for how we can assess all the Bay DO criteria with existing datasets and assessment tools – Tish Robertson (VA DEQ)

Presentation:

There are 11 DO criteria applicable to the tidal waters of Chesapeake Bay, and have been in regulation since 2005. To date only 3 of the criteria have been implemented. It's important to assess all the designated uses for all the criteria to do a comprehensive reporting of progress towards the TMDL. The criteria currently assessed on a routine basis are 30 day mean for open water fish and shellfish use, 30 day mean for deep water fish and shellfish use, and instantaneous minimum for deep channel seasonal refuge use. The criteria not being assessed are often referred to as short duration criteria (except for the instantaneous minimum).

This workgroup has spent a lot of time working out assessment methodologies. The short duration criteria has been on our radar. The latest iteration of the workgroup's work on the short duration criteria assessment method is published in the 2017 technical addendum. There's a chapter devoted to the DO criteria assessment methods. There are two frameworks described:

Conditional probability analysis – making an inference of attainment of 7 day mean criteria from 30 day mean concentrations

When we are implementing the 30 day mean criteria, we're looking to see are the 30 day mean concentrations in this particular segment in the 3 year period are they tending towards meeting the goal. If consistently meeting it exactly, it's meeting the 30 day mean criteria. But if you want to make an assumption of its attainment of the 7 day mean criteria you don't have enough information or comfort level to make a conclusion. If you were to apply a higher threshold, though, you could say the 7 day mean criteria was likely attained in that segment. So that's how this works – using a higher threshold looking at the 30 day mean attainment and making an inference for the 7 day mean.

Why isn't DEQ using the conditional probability analysis? We're trying to assess all of the criteria, not just the 7 day mean criterion. We're interested in an approach that allows us to assess everything. It also allows us to see if we're meeting the 7 day mean criterion, but it doesn't allow us to say a segment is failing the 7 day mean criterion. Ideally we'd be able to determine both attainment and nonattainment. This method also requires using judgement calls/decision rules. Do we come up with a threshold for each segment, or a blanket threshold applied everywhere? However, the most prominent reason is it is an unusual method for 303(d)/305(b) assessments, and DEQ is looking for approaches that fit all of their state waters.

In addition to the conditional probability analysis approach, there is the sub-segmentation of segments approach, which is where you take segments and carve them out into different zones and take a piecemeal approach to assessment via enhanced monitoring. The CAP WG came up with different zones for a segment. The thinking if you can assess the criteria in different zones you can make a conclusion about the overall status of the segment. The CAP WG put a lot of work into this. Tish had some questions about how this would be rolled out.

Why shouldn't the con-mon data be used to assess all the criteria? The framework laid out limits assessment of the con-mon to the instantaneous minimum criteria. The instantaneous minimum procedure for con-mon is unlike anything else (no more than 2 consecutive days with 2.5+ hours worth of exceedances in a 3 year period.) This is both really hardcore and not hardcore enough. In a worst case scenario where a data set indicates a segment has 50-60 days of really low DO, say 5 hours of exceedances – as long as those days aren't consecutive, we'd say that segment is attaining the instantaneous minimum criterion. On the other hand, if you have wonderful DO except for 3 consecutive days, what is the basis for saying aquatic life is not able to recover from that? I think it's worth re-visiting so it's protective but not overly protective.

Assessing the instantaneous minimum criteria wherever data are available – discrete as well as con-mon. It's the lowest hanging fruit of the criteria, so we should always be assessing that one.

What about profiler and array datasets? They weren't on our radar in 2017, and we didn't lay out a procedure for analyzing the data.

This assessment approach hinges on us doing enhanced monitoring, which is only possible if we can get additional resources. At a minimum we'd probably only be able to do this type of hyper focused monitoring on one or two segments in every given reporting period, which wouldn't allow us to report on progress as much as we'd like to. We'd like to be able to report on all of our segments with respect to the TMDL at any given time.

VA DEQ's tentative assessment approach for Bay DO criteria that addresses how to assess all Bay DO criteria using all readily available data: Combine two tools used in risk assessment - screening values, and multiple lines of evidence.

Screening values or screening thresholds are used to determine with high confidence where there's a low probability of adverse risks. They allow for rapid detection of sites where further investigation is needed. This is a benefit with limited resources, to allow for strategic use of resources to tackle problems. Screening values are applied to instantaneous data rather than spatially or temporally aggregated data. VA DEQ proposed ignoring the durations of the criteria and treating each one as though it's an instantaneous threshold. For each segment DU, they propose taking all the discrete data within a 3 year period, and for each Bay DO screening value, go through the data set and calculate exceedance rates. They would say an exceedance rate <=10% indicates a high likelihood the criterion was met, so we would say the criterion was attained; and an exceedance rate >10% would indicate lack of criteria attainment and/or additional information is needed.

The basis for the 10% rule is associated with USEPA 1997 documentation when states were doing their 305(b) reporting separately from the 303(d) list. That became integrated in the 2000s. Tish gave an example of a hypothetical segment assessed using the DEQ proposed method.

Some things that would need to be worked out include the minimum number of discrete samples (which helps guard against Type 2 errors), and whether or not the 10% threshold should always indicate nonattainment regardless of the criterion.

Multiple lines of evidence: There are different data types and data can be processed in different ways. Currently Bay DO assessments have been based on one data type (discrete) and one assessment procedure (interpolator/Cumulative Frequency Distribution/CFD). Weight of evidence decision making allows for the integration of multiple lines of evidence and make decisions in the face of uncertainty in evidence. Discrete data is weak in the temporal dimension while con-mon data is weak in the spatial dimension. The screening value approach is not good at distinguishing marginally attaining/non attaining segments, and the interpolator/CFD also has trouble with the grey zones, and is not being used for short duration criteria assessments. However, if we focus on the strengths rather than flaws of different methods, we can see how they can be combined collectively for decision making – this mitigates their individual weaknesses.

Continuous data analysis hasn't been explored, but can be used for 30 day, 7 day, 1 day means and daily minimums (using >10% exceedance rate as nonattainment). For screening values a greater than 10% exceedance rate would mean nonattainment or more information is needed. We would continue using the reference CFD. Tish then went over a hypothetical assessment using these methods.

Things that still have to be worked out include: what the minimum deployment duration of a con-mon dataset should be, how much of a particular period should be observed by a con-mon for an average to be calculated, how can array data be directly assessed in segments with Deep Water/Deep Channel uses. Finally, there is the question of how much weight should each line of evidence be given: should more weight be given to profiler results, etc, and if so, how.

VA DEQ's approach to assessing Bay DO would allow us to get the most out of our monitoring data, make use of all available data, report on incremental progress, identify segments where enhanced monitoring would be most beneficial, make sound decisions despite the uncertainty, have assessment results that can be readily understood and replicated by stakeholders and the general public, have a process that can be implemented by the jurisdictions, and have a process that is in line with established procedures in other 29 303(d)/305(b) assessments.

- 12:20 VA updates on CB7PH – Tish Robertson, VA DEQ
  - DEQ will talk about the new deep trench station they just added to CB7PH

VA DEQ identified a data gap in CB7. It's not immediately apparent because they do have a lot of stations in CB7, but in the deepest part of CB7 there aren't any stations. What the interpolator is doing is using measurements taken in the deep channel in CB5, and using those to stand in for the deep trench. We dropped a station in the deep trench next to CB7.3e. ODU has been monitoring there since July 2024. VA DEQ hopes to get a better assessment in the deep trench.

Discussion (on both presentations):

Matt Stover: Thank you. We're interested in pursuing those options, and maybe other options that involve looking at measured data instead of models.

Breck Sullivan: One of your requests was that the stakeholders and public can run the assessment. Has the public asked if they can run the assessment?

Amanda Shaver: Citizen science groups want to be able to check how the assessments are performing. We want to make sure it's a statewide approach that can be replicated.

Tish Robertson: Even people that aren't part of an organization can follow along with what we're doing. We get questions from retirees who used to be in the business.

Breck Sullivan: Hopefully in the next presentations we'll be able to show what the 4D interpolator can do. It will be using con-mon data.

Gary Shenk: Thank you. Another way to think about what you've done, is the simplest possible interpolation. You're running the same analysis we do with that. When we're putting together the TMDL, and realizing how much leverage in the TMDL the criteria assessment has, I was frightened. We ask do we do space first or time first. What does the 10% reference curve look like. Every decision was impactful. TMDL was based on deep water, which had a bioreference curve. We had observations, and an assessment method whose output was based on real biology.

Matt Stover: At the state level we agonize over assessment methodology developments every day when it comes to the IR. We look at it from a resource standpoint. If we call a segment wrongfully impaired, we have to develop a local TMDL, and they're not cheap. I share Gary's anxiety.

Clifton Bell: Thank you for the presentation. Very intriguing ideas. I think I know what you're doing with the screening values. In the examples, it seemed almost like they were being used more like pass/fail. That seems extremely conservative to ignore the time averaging if they end up being used on pass/fail criteria like that. There seems to be a link between this and conditional probability concepts. I wonder if the data could tell us about what is the exceedance rate, ignoring time averaging, that's associated with real exceedance. That might change the use of 10%.

Tish Robertson: I have thought about that, I don't know if we have the dataset to do that analysis. We would be mining the vertical profiler dataset. I don't think we have that kind of data at the moment.

Leah Ettema: As a technical note, TetraTech worked for Region 8 trying to answer the question of minimum data size – how many days are needed to accurately determine a 30 day, 7 day mean etc. I think that methodology might be applicable here and I'll share that with this group. Thank you, Tish, for outlining a potential framework for using these lines of evidence. I think it will be key to focus in on the temporal and spatial uncertainty in justifying which lines of evidence are used or not. I've been trying to sift through documentation. I think there was a recommendation to examine how monitoring frequency impacts uncertainty in 3D interpolator output. I wonder if something like that could be incorporated if the 4D interpolator is continued to be built. Can that be incorporated into the framework of what data is used and not used?

Richard Tian: The technology for data collection has evolved a lot over the last decade. When the 3D interpolation and the CFD method developed 20 years ago, con-mon data and profiler data basically didn't exist. Now we have a challenge to use this data. We need to find a way. When MDE suggested can we run the data through the interpolator, I would think it is a bigger challenge to do that. Additionally, up to now, the 3D interpolation is overstated, because in some segments there is only one station. The whole segment is only one value. It depends on the data availability. I support the approach that Tish laid out.

Bryant Thomas: I mentioned earlier the idea of consistency in methodology. On that idea of consistency, and timing as well, the 4D interpolator is under development. The 3D interpolator we have now is a great tool, with shortcomings of its own as Richard talked about. We have a rich dataset available which we all agree can be utilized more than it has been. As we look ahead, we have a couple stops along the way: the IRs coming up every 2 years with the next in 2026. For us, what we'd like to do is to move forward and not be beholden to the release of the 4D interpolator. The idea of using the weight of evidence and different data does help to support the decisions being made. We don't have to be stuck in any one decision as things evolve and upgrade. As 2026 IR comes up, we want to be able to incorporate use of the data that's available more than has been done to date. We want to continue conversations with the CAP on procedures and methodologies, but we don't want to be waiting on the release of the 4D interpolator before some of these decisions are carried out and implemented.

12:30 PM Lunch

1:15 PM Session 2: Interpolator 101

Interpolator 101 Presentation – Peter Tango, USGS

Questions that will be addressed include (but are not limited to):

- Why do we interpolate?
- Revisiting the initial challenges that led to choosing the 4D Interpolator as the solution.
- What is the 4D Interpolator being designed to do? And what it will not be able to do?
- How does the interpolator make full use of our monitoring data?
- How does the 4D Interpolator benefit our assessment practices?
- What is the value added by the 4D Interpolator to our program?

Presentation:
Peter Tango thanked Angie Wei for working on visualizations of 4D vs 3D interpolation.

We're moving towards blending the gathering of information with actionable knowledge and creating environmental intelligence for our assessments. We have no segments with full monitoring data accounting to support all criteria assessments needed to evaluate criteria underpinning WQ Standards. Unassessed criteria remain a hurdle for delisting decisions of State-adopted water quality standards with our existing framework. Historically there has been a mismatch between criteria and the monitoring and assessment to support their evaluation.

We interpolate to estimate the values of unknown data points that fall in between existing, known data points, and to fill in missing data based on known data points. Spatial interpolation is the process of estimating values of spatially continuous variables for spatial locations where they have not been observed, based on observations. Interpolation can vary in complexity. Peter showed an example where if we were using something more simplistic we would not be capturing as much in the pass/fail threshold as when we try to smooth the system out and understand the relationship to the behavior of the system. Peter showed more examples of interpolation.

Why do interpolation with all this new data available? Highlighted in the EPA 2001 and 2003 documentation is that establishing magnitude, duration and frequency of a condition is crucial for successful development and application of state water quality standards. Equally important is the spatial extent of a condition. The spatial and temporal dimensions of attainment assessment must be defined. Methods currently used for criteria attainment were based on temporal variations measured at individual stations. Now we're thinking about ways to use that data differently. The use of the cumulative frequency distribution was a recommended method and has been consistent across the Bay for providing the DO results, hopefully the chlorophyll results. The documentation highlighted that we're looking for a grid based spatial framework. Spatial interpolation provides estimates of water quality measures for all locations within a spatial assessment unit.

The value of interpolation comes through in looking at our newest data. The mainstem vertical arrays in 2022 at East and West Goose Reef (near CB4.3E and CB4.3W) are a good case to see how high frequency data compares, and if using just one to represent DO conditions (without interpolation) is accurate. Across the entire summer, the distribution of DO at the same depth at these 2 stations differ. At every depth in common, the W station usually has lower DO concentrations than the E station. This could lead to very different insights if only one of these stations was considered in a criteria evaluation. However, our 4D interpolation will fit a statistical relationship between DO, bottom depth, and distance E&W of the main channel that easily explains these DO differences we see in the data. Ultimately, this information from the data + statistical relationship will help make good DO estimates non-monitored places using bathymetry and spatial coordinates.

Peter reviewed the history of load estimation for the Chesapeake Bay as a comparison of how methods have evolved over time. Partnership interest in moving from 3D to 4D water quality interpolation came from a 2008 STAC panel and report which stated 4D interpolation would improve the ability to evaluate water quality for the 303(d) listing process. The panel found that there was insufficient information available to evaluate the feasibility of a 4D interpolator, so they recommended a study to evaluate the different approaches available for developing a 4D interpolator. They recommended data analysis studies should be initiated to develop the statistical basis for a 4D interpolator. There was also consensus we didn't have the frequency and spatial resolution of our data sets. It took another decade to build up this data through increased monitoring.

In 2019, two Chesapeake Bay Program working groups co-created a proposal for an innovation grant opportunity, to develop a robust, efficient, effective water quality monitoring tool that would collect real time, full water column water quality data in the challenging conditions of the open waters of Chesapeake Bay. The winner of the proposal was portable vertical water quality monitoring arrays. In 2020, a pilot study of the arrays was conducted in 20m of water in the open Bay, measuring DO, temperature and salinity. In parallel methods for 4D interpolation were being developed by Dan Openour and applied to the Gulf of Mexico. Their work brought in multiple data sets. They were able to reduce uncertainty in hypoxia estimates and it matched up better with fisheries data.

Working under STAR guidance, a small team of analysts of the Chesapeake Bay Program's monitoring and modeling teams, with consultation and collaboration from independent statisticians and academicians, re-evaluated the state of the science on 4D interpolation during spring and summer of 2021. The team agreed 4D interpolation of Chesapeake Bay water quality was now feasible, and multiple options for approaches were available to address the issue. The Hypoxia Collaborative and Bay Oxygen Research Group came into existence. Coupled with the PSC Monitoring Review, they were able to leverage the results of the pilot study of the monitoring arrays into the network of instrumentation needed to meet adequate monitoring.

Peter then went over the current process of 3D interpolation. Observed data including tidal DO in CBP DataHub for "cruise" period (Long-term fixed stations + Calibration from ConMon and Dataflow + Citizen/riverkeeper monitoring). Inverse Distance Weighted (IDW) interpolation is done of approximately 2- week "cruise" periods, providing snapshots of the Bay in 3 dimensions for each cruise (compiled to month and for 3 year periods).

The goal for the 4D interpolator is to develop a spatial-and-temporal interpolation tool for water quality monitoring data collected in the tidal waters of the Chesapeake Bay, thus enabling the evaluation of both long- and short-duration water quality criteria. Specifically, the tool should be able to: Interpolate observed dissolved oxygen in space and time ("4D") provide statistical estimates of uncertainty, reproduce daily and hourly variability of the data, and allow for post-processing of the interpolation output into designated uses (DU). **It is not a process based model, but based on observed data, and building statistical relationships that depend on parameters like time, location and other DO observations (autocorrelation) in order to estimate (with uncertainty) the DO in times and places where no data was collected.**

Uses of data in the 4D interpolator include exploration of patterns to inform development (such as identifying sub-daily cycles in DO), development and testing (such as estimating daily DO using a Generalized Additive Model), validation, and application. **The tool is still in the development phase, and beginning documentation.**

(The presentation was condensed due to time constraints.)

2:00 PM Discussion

Bryant Thomas: What types of data will be used for conducting assessments in the interpolator and what data will not be used? Will continuous monitoring data be integrated into the assessment side of things? Is there a cut-off date for accepting the data for development of the 4D interpolator?

Peter Tango: Yes, continuous monitoring data will be used.

Elgin Perry: A type of data we don't have yet is continuous data along the bottom, but when we get it, we will use it. The advantage of this tool is we can use a lot of different kinds of data.

Participant: Is it used for development and validation? Or can we use the interpolator to assess the dataset?

Elgin Perry: The interpolator is not a fixed product. It will be recalibrated for every assessment period. There is no deadline for submitting data. All data that comes along will be used, if not in the current assessment period, in the next one.

Tish Robertson: Just like we use the discrete data in the 3D interpolator, we would use any data in the 4D interpolator?

Elgin Perry: Correct. The data are summarized at a daily level. For a short term variability, which is variability over the 24 hour and the 1 meter depth increments, they are verified hourly. We had to do some standardization since the continuous monitors are sometimes on a 10 minute schedule or sometimes on a 15 minute schedule.

Peter Tango: Is it necessary to collect the data at sub hourly scales to inform the hourly?

Elgin Perry: Just hourly is enough. Even every two hours is enough because it has that ability to interpolate the hour in between. We would worry if we got down to less than every 6 hours because of the tidal cycle in our model. We picked the data point nearest to the hour. Most of the modeling tools assume the identical distribution of the data.

Bryant Thomas: Did I hear correctly the intention is for every two years, for every assessment cycle, the 4D interpolator will be updated to reflect new information available?

Elgin Perry: Is it a tool that fits to data, so as new data becomes available, it is recalibrated. We estimate model coefficients where smoothed curves should go. Just like the current interpolator, it will be updated with each assessment cycle.

Bryant Thomas: I didn't know the algorithms were updated as much as the input data was updated. Are the algorithms in the current interpolator updated for each cycle or just the data that's plugged in?

Elgin Perry: It could be but probably not. With the current interpolator you could change from using nearest neighbor averaging to Krieging and come up with different answers, and you could do that with the interpolator we're developing as well, but at this point we don't anticipate the model construct to change from one assessment to the next.

Bryant Thomas: But you do intend for the initial construction that the continuous monitoring data will be included in the assessing of the information? Not just in the calibration and verification.

Elgin and Peter: Correct.

Leah Ettema: Is one way to think about this, each assessment cycle, is you have observed data where you can see absolute minimum, but the 4D interpolator is almost generating a confidence interval around what that minimum could be or the multiple realizations. Given the availability of the entire assessment unit, what that actual minimum could be. Each cycle you're giving it new data and you'll have new predictions because of the variability within the assessment unit.

Matt Stover: If you have an intense data set from one segment, run it, and the next 3 year period you don't have that data, does it take the lessons learned from when you had intense spatial and temporal monitoring?

Elgin Perry: It could or it could not, that would be a choice.

Matt Stover: How long does it take to run? Is it designed more for hypoxic volume than assessments?

Gary Shenk: The way I think about it is the basis of the assessment procedure with the CFD is an estimate of a fraction of space below a certain level. So we say it's good for estimating hypoxic volume. I see criteria assessment as different estimates of hypoxic volume at different thresholds. I agree it's good for estimating hypoxic volume. It's always good to have a bunch of plans. We're trying to do this, we don't know if we can, but if we can we can use it to assess all criteria. Multiple lines of evidence, there's nothing wrong with having a bunch of plans.

Matt Stover: My concern is that it will be too complicated to run ourselves. Will the program have the capability to run it for us on the time scale needed?

Gary Shenk: Using the process model to say what attainment would be like. The process starts with taking an observed data set and modifying it based on what we think the relevant change would be based on the model. The process water quality model creates a new observed data set of what we think we would have observed had the watershed observed in a certain way.

Tish Robertson: How is it going to work with the 4D interpolator? I get how it works in assessment, but looking retrospectively back to the '91-'93 period when we didn't have high frequency observations, how does the temporal interpolation work? Are we just going to make the assumption that the kind of temporal variability we see currently was in play back in the simulation period?

Gary Shenk: That's related to Matt's question to Elgin about if we stopped monitoring, are we going to use the historical data we can.

Tish Robertson: So the assumption that the variability we see currently was the same back in the '90s.

Gary Shenk: Or if there is a predictor of variability it is reliable predictor of change over time.

Amanda Shaver: I assume the documentation phase won't be needed every time we recalibrate. We need to have it every other year updated and run, I don't know if we can tell if it will be feasible. For running it in 2027, it'll only be calibrated and using data from 2024?

Peter Tango: Yes.

Elgin Perry: I think there's a misunderstanding – what we're developing is a method. As data comes along, it's easy to pull data into the method, as with the current interpolator. When we do a new assessment, pull the new data or data from assessment period into the process. It's not fixed.

Tish Robertson: How does the uncertainty piece work with the 4D interpolator?

Elgin Perry: Through simulation. For example with the daily means part of the model, there's a GAM that gets fit over space and time. Then using the observed data we estimate what the autocorrelation and variability structure of those daily means are. Then we do a simulation process, an estimate of a parameter vector and a variance-covariance matrix, so we can use a multi-variate, normal simulation process to simulate a population of parameter vectors at the correct variance-covariance structure. Then we pump them into the predict function of R and get a set of numbers for each simulation. We hope to collect the kind of variability we would say if we were able to collect data between every one of our current monthly cruises.

Tish Robertson: So the simulation that runs simulations uses the same input with different iterations?

Elgin Perry: When we estimate a parameter with the GAM, that a fixed number based on a fixed set of data, but if we go out and collected data in between times we'd get a different answer. We are assuming that what we want to do is look at that parameter vector if it is a mean, it has a variance structure. That population described by that mean that we're calling the spine scale variability which looks at the hours within the day. We're estimating from our continuous monitoring data a population with diel cycles and tidal cycles. We'll randomly simulate diel cycles and tidal cycles as we plug into the spine scale variability and add that to the daily means model. The goal is not to come up with what the DO was at this location, at this depth, time and day. The goal is to say, over this segment at the end of the assessment period, what was the frequency of the violations?

Tish Robertson: How will the uncertainty be expressed for a manager?

Elgin Perry: I'm anticipating it will be expressed as a probability – like the probability that we're failing in this segment is 75%.

Guido Yayacto: There is a lot of questions in terms of how are the observations used and how are the water quality assessment model used in this assessment. What are the Bay Program model elements that are used, and what would be the instances when we have continuous information that informs it.

Gary Shenk: There's no plan at all to use the water quality model in the 4D interpolator. It is not used in the 3D either. How we use the water quality model for scenario analysis is we start on the left side of the flow chart. We modify the observed data set and run the same assessment.

Guido Yayacto: So you are using the model framework to modify the data? How does that modification take place? It won't be the case in the 4D?

Gary Shenk: We will do the same thing in the 4D.

Guido Yayacto: But the reason we were doing that was we didn't have a continuous data set, and now we do. Let's say we have a segment with all the information we need.

Gary Shenk: Effects of the WIP on oxygen.

Matt Stover: Do you see the next phase of the interpolator being developing those assessments? I think Virginia and Maryland should be the ones to agree on those. We don't want a black box situation. Those decision rules that will determine whether it's impaired or not I think we should weigh in on. MD wants to get to where VA is where they are running it on their own.

Peter: I think that's a positive goal. I can't tell you that there will be a nice GUI where you just put stuff in and say what you want out, but transparency and transferability are important.

Amanda Shaver: Trends doesn't have the regulatory repercussions that this does for us, which is why it's so important we know what's happening and which data is being used.

Matt Stover: We want to have a say in which data to verify. We've been spending a lot of effort, money and time collecting continuous monitoring data. Admittedly there's a lot of details we haven't worked out on that (how should we deal with drift?).

Tish Robertson: In Matt's presentation he talked about rotating fixed monitoring is something we should consider. I think for the 4D interpolator that would help justify that approach. It could give us a robust prioritization scheme so we know we have to go back to the segment in x years because we know our results will be very different.

Matt Stover: We don't want to lose all the lessons learned from all of our sampling. Maybe the 4D can answer that.

Tish Robertson: Ideally we should have a longer evaluation period for assessment. There's no reason we wouldn't have a broad window.

Peter Tango: Did you read the paragraph after the 3 years that says or pick 3 years out of a 5 year window?

Tish Robertson: In the SAV water clarity part of our criteria, it says use the most recent 3 years if you don't have the last 3. We don't have that for DO though. I understand why regulators might have a problem with a broader time window, but it's not really using the raw data, it's using the relationships in the data.

Matt Stover: It's easier to justify using it if you have a scarcity of data whereas if you have an equal or better data set.

Joseph Morina: If the 4D interpolator is working and we have our data set for this year, and we get our assessment results, in 2 years ago, if we input the same data will we get the same results? It seems like if it has this memory of lessons learned in the past, that it could be potentially changing over the years. Could you snapshot the 2024 version so we could reproduce our assessment results?

Elgin Perry: In order to have that memory, it has to look at that data and use that information to interpolate in the current 3 year period. As long as you use the same data you get the same results. We can't exactly reproduce the simulation. They should be close, but not exact.

Tish Robertson: That's where the uncertainty comes in.

Becky Monahan: I like what Gary said about having lots of options. We have Fishing Bay data and I still don't know what instantaneous minimum means. The interpolator that will come out in 2027 isn't really helping me now figure out the definitions of what we all agree on that are going to go into these decisions and models. I still think we can talk about the 4D interpolator, decisions and uncertainty, but until we all agree on how many data points go into a 7 day mean, how do we calculate a 7 day mean, what does instantaneous minimum mean, are we allowed to use continuous data or not, I don't see how we can use a model with that either.

Tish Robertson: Out of all the criteria, the instantaneous minimum is most tricky. All the other criteria are relatively straightforward. You can define an instantaneous minimum as a 15 minute observation, but it would not be consistent with how other criteria were developed for a one hour duration. The first principal would be looking at if one hour makes sense. Then the decision rule would be how do you determine 303(d) data. Are you looking at every caliber in the 3 year assessment period or are you stratifying it based on day. There are a lot of ways to square that one. If we use that 10% distribution for the others it's pretty straightforward. Maybe that's a decision, do we continue using the 10% CFD.

Becky Monahan: Until we can answer all the questions of this morning it doesn't matter what model we use.

Tish Robertson: We don't have a minimum data requirement now for Bay DO. If you look at our assessment methodologies for inland waters, we explicitly describe how many samples you need. We don't do that for Bay DO. Should we fix that? Even with the 4D interpolator I think we should still have a minimum data requirement. I don't think it's been a problem for us because we've had this fabulous monitoring effort, but for the sake of being transparent and complete we should say what's the minimum number of observations we need.

Elgin Perry: The interpolator will provide more information about the certainty/uncertainty about the assessment given the amount of data.

2:15 PM 10-minute break

2:25 PM Session 3: Short duration criteria attainment decision

- 2:25 Updates on dissolved oxygen monitoring and assessment investments – Peter Tango, USGS
    - Discussion
- 2:50 Reflection on past work on short duration criteria attainment – Peter Tango, USGS
- 3:05 Binomial decision structure North Carolina – Clifton Bell, Brown & Caldwell

Presentation:

Clifton explained he was asked to give a presentation considering uncertainty on the assessment side, particularly on the North Carolina precedent, but he would not just be sticking to North Carolina. The broader topic is considering uncertainty on the assessment side, and it ties in a little bit with Elgin's presentation in terms of what might the decision rules be for considering uncertainty in an assessment decision. Clifton clarified he wasn't recommending the binomial approach or advocating for any approach in particular, but hoped to share this precedent as a useful part of the conversation.

The background for this is with the desire to limit type one and type two assessment errors, put numbers on them and in some cases recognize that those probabilities of type one or type two assessment areas are high enough that we might say that we don't really know, and stick with category three. EPA makes this encouragement to the states to consider both type one and type two assessment errors. Even if I didn't recognize this I would want to balance/control/minimize both types of errors. Some of the IR guidance of 2006 specifically suggest that the decision rules be defined and the actual significance levels of any statistical test should be defined in the assessment.

There are a number of states that use some version of the of the binomial method at least for some of their criteria, if not for all. It's also described in EPA technical documents, so it's a pretty well used and well known method across the CWA programs in the States.

A binomial distribution is the distribution by binary outcome. The math you plug in gives the probabilities of success on a trial. It'll tell you the probability of getting exactly a certain number or a cumulative version getting at least or as many or more. This is used in water quality standards in a couple of different ways. Number one, it can be used to determine a frequency component – such as allowable frequency of exceedance on the water quality standards development side. Here's an example with Florida DEP based on annual geometric means criteria on some reference conditions where they had a pretty good idea of what the exceedance rate was for that reference condition. They could plug it through the binomial and say if they're really at the reference conditions they would expect in a three year assessment period having more than one exceedance only 10 % of the time. And they considered that to be an acceptable type one error rate. Type two error rate they felt they had covered without the way they developed the criteria.

Another example is North Carolina, which is in the process of adopting a water quality criteria for SAV protection. It's based on a single best year, but with no allowable exceedance rate, so it has to be achieved every year. You can use the binomial to look at reference sites. We've been in an advocacy role there saying these are some of your best sites. They exceed this ambitious criteria even though 20 % of the time on every five year assessment of period there is about a two thirds probability of a type one error rate. We can put this into an R shiny app where you could download data from a reference site and play around with the frequency component with your sliders and determine what's a reasonable type one error rate based on a combination of the criteria.

We're more interested in the assessment as opposed to criteria development in this conversation, so we'll use the North Carolina example to show how they've used it. And if you want more details, here's the link to their listing methodology. North Carolina will use a non parametric hypothesis testing approach using the distribution and basically they're trying to limit the type one error rate to 10 % here, and the null hypothesis in this test is that the overall exceedance probability is less than or equal to 10%. That's how you can quantify that type one error rate, but there's more to the assessment than simply plugging through the binominal exact task. There's a little asymmetry in the confidence level required for listing or delisting. It's a little easier to delist than it is to list. I infer it's a certain reluctance to commit TMDL resources unless you have a certain level of confidence. Also, what is considered is the current listing status influences the decision rules and the number of excursions in newer data. So you can't really understand how this works until you look at the flowchart.

You can see an example here where if you got a data set with more than 10% exceedance, you've got more than 90 % confidence that using the binomial you'd have a very low probability of getting that many exceedances if you're really only at 10% exceedance. Otherwise you consider what is your current listing status? And then you might end up listed even if you don't have that 90%, if you're already listed and you have some recent exceedances. Otherwise you'll end up in data inconclusive (Category three). The flip side is if your data shows less than 10% exceedance, but you have more than 70% confidence that you're meeting the criteria, you could be delisted. Otherwise you still might end up in category five if you're already in category five and you don't have a certain level of confidence in meeting the criteria and you have recent excursions. Otherwise there are different other ways that you can follow those error arrows where you might end up either data inconclusive or meeting the criteria.

The binomial is not the only test that that is used or could be used for this. There are certainly other ways to express and calculate uncertainty. The hyper geometric test is one that Washington Department of Ecology uses and that's a kind of binomial. The reason in that case is that their metric of interest is days in a year with an exceedance as opposed to this percent exceedance. The other things have been proposed are the sequential probability ratio test. You can get the same type one error rates with fewer data.

There is the concept of hypothesis testing of probability impairment, by quantifying not just a 10% rule or a 1% rule, whatever it is, but the confidence required to make a listing or delisting decision. The idea of considering the existing status, and then really this method will be biased towards the existing status you have a certain requirement of confidence to change. If it's currently listed as impaired and there's a certain level of confidence to change that, which I can imagine people would have different perspectives on, but in this case the North Carolina example, a higher confidence is needed to commit TMDL resources. The idea of having a quantitative basis, even though from an administrative standpoint and from a communication standpoint we may feel like we've failed if we're stuck in category three, but sometimes that might be the right category. And it provides a quantitative basis for saying we really don't have enough information and maybe that's where we need to target our next enhanced monitoring with all of those.

4:00 PM Adjourn