Criteria Assessment Protocol Workgroup (CAP) Meeting

Monday, May 06, 2024
1:00-2:30PM

Meeting Materials

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

ACTION ITEMS

- ✓ August and Peter will schedule a virtual only June meeting CAP meeting, which will be discussion-focused, and a longer in-person/hybrid August CAP meeting which will feature several topics.
    - Future meeting topics:
    - More details on how the 4-D interpolator and GAMs work.
    - The viability of conditional attainment, and other options.
    - Technical documentation outline.
    - SAV and satellite based assessment needs.
- ✓ The CAP Workgroup will be invited to the next Bay Oxygen Research Group (BORG) public meeting which will include a detailed overview of data selection which the CAP is interested in.

MINUTES

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

Jacob Greene (MDE) introduced himself. Jacob started with MDE a year ago.

Peter shared his thoughts on his investigations into past documentation discussing short duration criteria assessment methods.

Peter: Since the publication of US EPA 2003 (the criteria and their assessment), we have operated without agreed upon short duration criteria assessment. But that is not for lack of trying. Appendix I in the document shows we had recommended methods we might pursue. Over the course of our 20 years, we've had recommendations for logistic regression models, spectral analysis using long-term and short-term patterns in data, conditional attainment, and direct assessment with enhanced monitoring. One that we took a lot of time with was spectral analysis. There was a 2010 STAC workshop on the topic which highlighted how we could make better use of long- and short-term data. However, we concluded it wasn't suitable for our purposes. Regarding conditional attainment, we found this is potentially applicable and was feasible with methods and techniques we had. However, it does require an agreement on acceptable uncertainty. We didn't bridge that in 2017. Low sample density translates to large uncertainty bounds on means requiring massive changes in water quality conditions to

overcome the boundaries of uncertainty if you want/need 90% or 95% confidence in a change in attainment status large uncertainty bounds in the means. Functional capability may challenge us unless we accept changes in confidence. We'll need to come back to that conversation.

We have other methods. There are multiple approaches. New approaches are effectively looking to address all those criteria assessment application periods. New monitoring approaches that effectively address all criteria assessment application periods are being developed. New monitoring support of high temporal density information in deep waters presents revisiting even basic questions – i.e., how do we calculate a mean? How do we handle missing data? How do we interpret the result against a criterion?

The 4-Dimensional (4-D) interpolator goes back to 2008 recommendations to develop this tool and use it for delisting and assessments. Findings from the panel have been reviewed before, we needed to do some things before we had such a tool. To the credit of the science community, that is happening, and we're overcoming the limitations viewed back in 2008.

Where we're headed – assessment of all Bay oxygen criteria in the Beyond 2025 world. This work has been going on since the formation of the team and we are mid-way in the process. We are at a point where sufficient development has progressed and there is more need to hear from the community as development is proceeding, to hear what the questions are and where we need co-development opportunities with states and jurisdictions in terms of its applications, outputs and calculations. In our discussions we will produce issues, ideas that need to be documented in the years ahead.

Some of the questions from workgroup members have been directing discussion include:

1. What can the 4D interpolator do, what can't it do?

2. We'd like to make sure that the 4D interpolator can handle and use all of the continuous monitoring data that our DNR partners are collecting and be adaptable enough to incorporate new data streams (from both government entities and NGOs) as they become available.

3. Be able to have it be recalibrated over time to improve its accuracy

4. Be fully understood by and maybe even run by the states.

We are looking for the co-development now with a need for continued community input. We will continue to work on a new technical document. Please see this as a discussion focused meeting, and we'll also come back together in June. Please use this opportunity for education, discussion, insight, and questions and answer time. We'll come back making sure we have the issues of concern documented and make sure to move forward in tandem with the BORG.

Upcoming Conferences, Meetings, Workshops and Webinars:

- Chesapeake Community Research Symposium – June 10-12, 2024, Annapolis, Maryland.

1:15 PM Presentation: Incorporating uncertainty into criteria assessment – Elgin Perry

Elgin: I've never been to a CAP meeting. Criteria assessment usually doesn't take uncertainty into account so there's usually not a job for a statistician. I also think there's a good reason it doesn't take uncertainty into account, and that's because it's challenging to do it in a way that is fair to all the parties involved. I don't present any solid answers, but hopefully will help get the discussion rolling.

Elgin: Kudos to Gary on developing the diagram. Note the team has grown, we have Wes Slaughter working with us now.

Statement of goal: build a tool to allow us to assess criteria we have not assessed before like the weekly mean and daily minimum. We do NOT expect this tool to accurately predict DO in a specific hour on a specific date, but instead track the mean patterns of DO over time and reflect the variability around those means and give us a fairly accurate prediction of how frequently we're violating the criteria in an assessment period.

GAMs = generalized additive models. GAMs are the tool we use to get predicted means. GAM is a multiple linear regression tool that instead of being rigid and having straight lines, it allows the lines to bend. After we get those daily means, we develop a second model that tries to capture short term variability. Some of that variability will be deterministic, maybe a diel or tidal driven cycle, and some will be stochastic, random noise. But not completely random since we know if we're predicting observations at a high time frequency that ones that are closer together will be more alike than those that are farther apart. There will be a temporal dependency which we'll model with an autoregressive process. There will also be spatial dependency in this model. The third thing I will talk about is this water quality assessment box. This hasn't had any development at this point, so this is the kickoff discussion for that. I'll put some ideas on the table, but I find this to be a tricky thing to implement where it's totally fair. The GAM tool uses certain variables across this 4-D lattice of time, depth, longitude and latitude to create a prediction model. The variables we are using are Estuarine Longitude, Estuarine Latitude, Sample Depth, Bottom Depth, Long term trend, Seasonal Trend. You might think, there's other important things. You have to keep in mind in order to be able to use it to predict across a 4-D lattice, anything used as predictor has to be known in this 4-D space which limits what we can use (like depth or pycnocline). When we have this model in place it gives us predictions in time at a scale of one per day, predictions in depth at a scale of one per meter, longitude by 1 per km, latitude at 1 per km. So far these models are coming up with r squares around 0.85 so they have pretty good prediction power. This panel shows what predictions might look like.

Lew Linker (EPA): Is shallow water monitoring included in GAMs? Since it's continuous monitoring (con-mon) how is it handled since it looks like GAMs takes one observation of time per day?
Elgin: The GAMs doesn't require; it predicts one observation of time per day. This particular example was based on just the fixed station network data, but since this example was generated, we've updated our database for fitting the GAMs to include fixed station and con-mon data which gets us out to the edges of the Bay.
Lew: So the continuous monitoring calibration data are sampled once a month I think. I think I'm hearing that the continuous stations of dissolved oxygen (DO), chlorophyll are not being used but the calibration data is being used?
Elgin: For the GAM that's correct. For getting the daily predictions, we are going to use con-mon data and data flow data when we get to the second phase doing the small scale spatial and temporal variability.
Rebecca Murphy (UMCES): We're definitely using the continuous monitoring data to get the high frequency. As an addition to what Elgin said we've also recently been including the continuous data at a daily level. So selecting one daily value for any of the con-mons at the right period of time and using that in the GAM. That's new. Jon Harcum and I used that method and it seems to be working.
Lew: It sounds like there's an expanding wave of data inclusion. I'm interested to hear how drunken sailor plots will be included but that will be at some future time.
Matt Stover (MDE): For the continuous data sets, could you elaborate on the fact there's only one value per day used? Thinking about datasets we get values every 15 minutes, how do you pick one value for that?
Rebecca: Be assured the high frequency information will be used. Because we wanted to generate daily predictions and there's this rich high frequency data, we wanted to use that in the GAM also. I have a memo I can share with you on how we did the sub selection. To generate daily GAM predictions it wouldn't be beneficial to include 15 minute data because that's higher density than we need for daily predictions and would overwhelm the model fitting and shade out fixed station data. There are different ways to get one value. We picked data that was selected around the same time that the fixed station data was selected. I can share more information on that.
Matt: Of course fixed station data is usually collected at a time of day when it's not the lowest DO value. So if you're using this to estimate instantaneous minimum, it might not represent the do sag in the morning because folks are not out there sampling at 5 or 6am. I'm curious about this.
Amanda Shaver (VA DEQ): We'd be interested in those details in VA as well.
Rebecca: One thing we'll get at with the high frequency estimation is getting those daily minimums which isn't necessarily at the mid-day when the fixed station sampling is.

Leah Ettema (EPA): The focus on preserving the discrete stations, that's for spatial coverage, right?
Rebecca: Yes. We want to cover spatially as well as temporally.
Leah: To Matt's point, more detail is needed on why there is a large focus on keeping discrete stations.
Matt (in the chat): What about the instantaneous minimum?
Peter (in the chat): I believe that is the daily minimum reference but we can be sure to clarify that.
Matt (in the chat): There is also a 1-day mean for the Deep Water (DW) use.
Rebecca (in the chat): Yes, that is what is meant. We intend to cover the instantaneous, daily and weekly for each Designated Use (DU).
Matt (in the chat): Just wanted to make sure we're covering all of them. Great thanks Rebecca!
Elgin: It's a broader station area but much of the con-mon data is only spring through fall, so fixed station fills in winter and fall data. The fixed station network also gives us vertical profiles where we get observations once per meter from surface to bottom, while most con-mon and data flow data are mostly shallow. I think all these datasets have things to contribute.
Elgin: Here is how we plan to use con-mon data. In the surface we can see there is typically a peak in the middle of the day. With fixed station network we are collecting during the day when DO is high. Once of the reasons we wanted to pick one observation per day from con-mon data is so the fixed station data and con-mon data would all be representative of day time observations. So when we feed them to the model they would be estimating the same thing. When we get to the part of the model when we start feeding in daily cycles, that will be part of daytime observations. From that intercept point the DO will go down and swing up to the next day. Something Interesting about this data is at the bottom we see two peaks per day suggesting there might be a tidal cycle. Jim Hagy suggested we look for it and when we did we found it.
Lew: One thing to consider is there are times when the DO is absolutely at the bottom. When we'd have DO less than 1 mg/l, the biochemistry is that nutrients really get bumped out of the sediments, so it's generally an amplifier of nutrients in shallow water. Data on the time below 1mg/l DO would be of interest to many.
Elgin: This is true. When we get some nutrient criteria, I'm sure we will be interested in that. To capture the diel signal, we are proposing to use a Fourier model. It captures the peaks of the oxygen data. We're hoping by adding an autoregressive error term we can simulate what the behavior of DO would be on a daily basis. To the question about daytime sampling of the fixed station network, we'd consider that to be a prediction at the peak of the curve. When we composite this Fourier series with the daily prediction we'd be subtracting values at the peak down. We'd consider peak to be at 0 and other observations would be negative numbers we

add to the daily prediction. By subtracting off cycles we'll get a better picture of how far down the DO goes during the night.

Water Quality Assessment:
Elgin: This simulation process is something we can run over and over again with different sets of random numbers and come up with different multiple feasible hourly realizations. For every point in our 4-D lattice, we'll come up with multiple predictions. We'll have to look at this variability and somehow make a decision on whether we're meeting or failing the criterion. Next, I will talk about uncertainty in addressing whether or not we are meeting criterion.
You could just average all the points from different simulations. Or you could take each one of the simulations and compute a cumulative frequency diagram, and that would give you a cloud of frequency diagrams. From that cloud you'd make an inference on whether you were passing or failing the criteria. Moving forward you'll have to develop a philosophy of dealing with uncertainty. Here are some ways on how you might do that. As a note, I've never seen a great tool for incorporating uncertainty into criteria assessment, but here are some possibilities.

One approach is the even-handed approach. Assuming you have a score occurring between 1-10, and the criteria is at 5. The score we compute is based on a sample from the whole system. It's almost certain that a statistic computed from a sample is not going to match what the same statistic would be if you could compute it from the whole system. That gives you a whole uncertainty range which I've represented here with parentheses. If your sample endpoint is above the criterion, you pass, and if it's below the criterion, you fail. The problem of this approach is if your sample endpoint is close to the criterion, there's a good probability that your true system mean is above the criterion so it would be a pass but due to sampling error you might get a value that's below the criterion and indicates a fail. That's what I would call a false fail. That probability can get high, up to 50%, which I think most people would think is unacceptably large. And it can go the other way too with an endpoint statistic indicating a pass that is a false pass. You can have high probabilities of getting the wrong answer, but it's even handed. It imposes just as much burden on regulating parties as it does on the environment.

Magna Carter approach: This approach moves the criterion to the lower bound of the uncertainty range. You're giving the benefit of uncertainty to the regulated parties. You're saying that you want to be certain beyond a reasonable doubt that the system is failing before you call it a fail. It's kind of like the legal concept of being innocent until proven guilty without a shadow of a doubt. The problem of this approach is while it keeps the probability of imposing a fail when system is passing low, the probability of declaring a pass when system is failing is high. It poses an undue burden to environment, so it's not a very satisfactory solution.

The IRS approach gives the benefit of uncertainty to protecting the environment. We set the criterion at the higher bound. High probability that if it is passing, we'll call it a pass. It ensures that the environment is protected but it could impose a huge burden on regulated parties. Because there is a high probability of calling it a fail when it's not really failing, regulated parties could be required to spend a lot of money accomplishing nothing.

The Uncertainty approach abandons the binary pass fail and introduces a third category that we call an uncertain category. It has some advantages. We're pretty certain a fail is a fail and a pass is a pass. But there could be a large number of situations where we're calling it uncertain.

The conditional approach sticks with the binary approach where we have passes and fails, but introduces the idea that passes could be conditional on how far you are from the criterion. You get a conditional fail and conditional pass based on how far you are from the lower bound and the upper bound of the criterion. This has the advantage that you could only require mitigation if it's a fail, and make it optional for a conditional fail, and even a conditional pass. I think it doesn't provide undue burdens on one party or another and gets rid of the uncertainty zone where all the stakeholders are unhappy because we can't make any decisions. In terms of approaches this is the best one I've seen.

1:45 PM Q&A

Peter: We have a 10% allowable exceedance built into the process as a measure of uncertainty if it's not the bioreference curve option. That reflects some level of uncertainty in assessment. Are you familiar with other places that set some form of uncertainty on any criteria assessment?
Leah: There are some states that incorporate a binomial assessment approach into evaluation of the data. There may be a few slight variations on that but I'd have to research a bit more to answer. Elgin when you say the probability of attaining or failing - I'm having a hard time conceptualizing what that is. What exactly is being evaluated, the daily mean? The uncertainty around the probability of attainment? Is that directly a criteria to criteria comparison or is there some level of 10% exceedance baked into that?
Elgin: I was thinking more of the latter. In this example we're developing a simulation process that will give us multiple feasible realizations of how DO might behave over a certain time period. If we generated 100 of those and only 5 indicated failing some criterion, that could be how we got to failure. We could compute a cumulative frequency diagram (CFD) for every simulation we run and see if more than 5% fail. If fewer than 5% fail we'd be confident in calling it a pass. If 50% fail that would be uncertainty. If only 5% pass by this cumulative frequency diagram we'd call it a fail.

Leah: That CFD represents space and time simultaneously?
Elgin: It's like an area computation. We can get the percent of space over the percent of time from that 4-D lattice for every simulation we run.
Peter: This feels like something we talked about a couple times. We've just never put a confidence interval around the CFD line. In my mind this is how to approach that based on simulation results that reflect the distribution around the number of attainments we get from multiple realizations. That's how I'm seeing it. It's never as simple as just putting a boundary line on the CFD curve or bioreference curve.
Elgin: The calculations you go through to get the CFD make it difficult to get a standard error. You'd have to go through sampling strategy or simulation strategy to get that. Here we're going through simulation strategy. I do recall that Jeni Keisman presented cloud plots of CFD curves.
Peter: Yes. That was how the bioreference curves were developed for deep water, through that cloud of points for evaluating where the good and bad Bay Index of Biological Integrity (BIBI) scores where and boundary line for outlining that bioreference curve. But I didn't see it for evaluating a plus minus confidence interval kind of approach.
Elgin: That's correct, I think she did a resampling thing. Trying to go plus minus two standard error confidence interval is difficult because it's difficult to compute a standard error for that curve.
Tish Robertson (VA DEQ): I like your presentation. I still have some questions about GAMs but I think we'll have a follow up meeting. I do want to quibble on something you said about how you frame the Magna Carter approach on slide 15. The regulated community doesn't get any relief on discharge limits or anything when we say a segment is meeting a criterion. I find the framing to be cringy. And similarly, the IRS approach doesn't lead to extra protection of the environment. When we say something is failing, we don't throw extra resources. I don't agree with the framing. We're not trying to provide extra protection. The TMDL, allocations and permit limits are not going away. The uncertainty and conditional approach are going to be difficult to swallow. I don't know how we would explain to management and the tax payer why we should invest in a Cadillac monitoring program if an end result could be a handwringing that we're not sure if it's meeting or failing. We want to make a management decision with the monitoring data. We don't want to collect millions of data points and say the end result is untenable. It wouldn't be a good practice for us. I'm not against coming up with uncertainty measures; we have some. We're always hedging against incorrect decisions by not basing management decisions off a single assessment. We have data requirements in our assessment program. You laid out a nice philosophical framework for us and I would add to those questions, what are we hoping to get out of adopting an approach like this, what benefit do we get against having a safeguard against a wrong yes decision? Is it so we can sleep better at night or are we trying to prevent some harm? What are we trying to protect us from doing that we're not already doing?

Lew: Building on your comments and taking a high-level view of where we are today, we have deep water deep channel that is far off from being attained. We should all probably know we're at least a decade out from even getting close to attainment. While we're getting closer and closer we'll be lifting all segments to some degree. To a certain extent this analysis is for a future hopeful day when we have a stable designated use assessment throughout. It's also right on the cusp. We all know, it's sort of a high-level assessment, we know bad years we won't attain, and good years we will. Is there a time element in terms of demonstrated over x number of years it's in attainment, we can declare attainment?
Matt (in the chat): Tish brings up some good points from the perspective of implementing Section 303d of the CWA. To some extent we have a draw a single line in the sand.
Clifton Bell: For whole assessment periods they have a certain time element for what years are chosen for the assessment period itself. Someone was asking about precedents and binomial was mentioned; some other states use the binomial as the current attainment status. For example, North Carolina's approach is they use a binomial and if they're already attaining the burden of proof is to demonstrate that you're not in attainment, and the other way around. If you're already listed as not attaining it's a little harder statistically to get it delisted. It is not completely symmetrical. If it's currently attaining you need a 90% confidence that at least 10% samples were out to get it listed, and at least 40% the other way around. I think that's based on once you get something listed there's a lot of burden economically to address that with TMDLs. I don't know if this group would consider that. The current attainment status affecting in some asymmetric way, what kind of categories you use.
Peter: I read about something of that regard to chlorophyll.
Gary: Essentially we've been using the balanced approach all along. We've taken some heat at CBPO for not being able to articulate uncertainty. The IRS and Magna Carter are interesting but I don't think they are things we could use. I agree with Tish, the TMDL is set but when we redefine planning targets for 2017 and 2025, we do run these over again to create these planning targets, our reduction goals are dependent on a decision we'd make about this. It's interesting that Tish's comment, and Matt backing up Tish's comment that states are good with balanced approach. I went right to this conditional approach. But if the partners don't think so then it seems like we're angling back to the balanced approach at least in this conversation.
Joe Wood (CBF): I appreciate the presentation. It sort of seems like semantics. I want to focus on the label of that gray area. I think it was described as the uncertainty range. I heard dissatisfaction from Tish. I think that the finding is uncertain is not the point. I think I'd rather label it as fringe/vulnerable/on the cusp. When I think about it like that, that finding is very meaningful. That finding would make me want to focus work there because you're on the cusp of attaining. Thinking about it like a doctor's appointment, if you're 1 point above or below the threshold for having a heart attack, it's a meaningful finding. I don't think uncertainty is a meaningful label for that area. You're transitioning between, rather. There's a lot of value in

having that designation. Focusing efforts on waterways like that you might have near term responses of achieving attainment.
Peter: Scientifically we use it in the GAMs and trends report we like to see p values of 0-0.5, 0.25-0.5 and beyond as these different levels of uncertainty on what we're saying is a strong trend or not. I hear Tish saying that may not necessarily be the way, even though it's great scientifically, that the regulatory community can work with that information. Maybe the conditional approach isn't something we can do and the balanced approach is something we have to work with.
Tish: I don't want to be the voice of no; I think we should think of other options and approaches. One way we've been talking about at DEQ is taking a weight of evidence approach to Bay DO. We have the tool the 4-D interpolator, which has a lot of promise, but it's not the only way to analyze monitoring data. What we could consider is when we do have uncertainty about the 4-D interpolator we would use another line of evidence. So that we can still use the monitoring data, but we need another line of evidence to make the decision. We use this approach for estuarine benthic assessments. It's a way of addressing uncertainty; having multiple tools mitigates the uncertainty of each one.
Peter: Good point. I've heard that about the benthic data.
Matt: The thing driving a lot of this is the regulatory process and the fact that under section 303 d we have to make a decision on whether something is impaired or not. It's a little tough the way section 303d is written, with the guidance we have from EPA for integrated reports, we kind of have to make a binary decision. That said I think uncertainty is valuable, but it may be better placed when talking about trends and when we're trying to tell a more nuanced story. Whether we meet our standards or not we're still putting resources in. We don't vacate the waters. I am also sensitive to the point where we're spending a lot of money on monitoring and assessment and it's a tough message to sell to people who fund us that we can't come up with an answer despite all the monitoring. I'm also still having a little trouble understanding GAMs in the 4-Dimensional interpolator. I was wondering if we could go through a fictitious segment and see how the data is used. You can use me as practice for pitching this to the public. Also, we have a rich dataset from the Fishing Bay mesohaline segment, data from shallow water and tributaries. We also have a profiler there in the deeper channel. We've got fixed station monitoring and con-mons there. It's about as densely sampled as you can get. I think VA did something similar in the James. Granted, you can have errors in monitoring, but it's easier for people to grasp when it comes to basing your assessment on these tools.
Leah: As we go forward, I'm curious to think about if there's a scenario where your observations from continuous data would give you a different result than your probability of attainment? Being aware where that may happen I think will be important for this workgroup as we assess the data. Maybe that's where what Tish suggested comes in. That would be a challenging thing to explain.

Peter: That reminds me of what MD and VA did with benthic results that were given a mean of 1 but variance pushed it into consideration for another category, so needed different thinking to make decisions about classification. We will keep that in mind. And Matt, we can talk about that for a next BORG meeting. Knowing we are still developing the part where the short duration criteria result of the daily mean function has a solid background and now the team is working on building in the short duration behavior. It may take a little time but that represents a target in terms of how we apply the tool and where a good case study might be.
Elgin: I think that's correct. I glossed over details of daily means and short-term variability part of the model. We could come back to this group and show more details and plots of how the model tracks the data and show that this is a model that is empirically driven. It's the data that tells the model how to behave. The model is just a reflection of the data.
Matt: I wonder if an in-person meeting might be more fruitful. Like Tish is doing in VA, we're also looking at alternative methods of assessment to make sure what comes out of the 4-D interpolator is matching what comes out of our data. In places like fishing bay where we have enough data where we can say conclusively whether we are meeting or not meeting water quality standards, if the two agree or don't agree it's helpful ground truthing. We want to add another method that's easy to explain to the public.
Peter: I'm up for organizing an in-person meeting.

2:00 PM Documentation outline, next steps – Peter Tango

2:30 PM Adjourn

Participants:
Amanda Shaver (VA DEQ), August Goldfischer (CRC), Becky Monahan (MDE), Breck Sullivan (USGS), Carl Friedrichs (VIMS), Carol Cain (MD DNR), Cindy Johnson (VA DEQ), Claire Buchanan (ICPRB), Clifton Bell (Brown and Caldwell), Elgin Perry (independent statistician), Gary Shenk (USGS), Jacob Greene (MDE), Joe Morina (VA DEQ), Joe Wood (CBF), Juan Vicenty-Gonzalez (EPA), Kaylyn Gootman (EPA), Leah Ettema (EPA), Lew Linker (EPA), Mark Trice (MD DNR), Matthew Stover (MDE), Melinda Cutler (MDE), Peter Tango (USGS), Rebecca Murphy (UMCES), Renee Karrh (MD DNR), Richard Tian (UMCES), Sandy Mueller (VA DEQ), Suzanne Trevena (EPA), Tish Robertson (VA DEQ)