Bay Oxygen Research Group

Monday, May 20th, 2024
12:00 PM – 1:30 PM
Meeting Materials Link

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

Action Items/Next Steps

- ✓ Rebecca Murphy (UMCES) will do some testing of the GAM in the CB7 deep trench in VA waters where there is a lack of bottom data, and share results/findings with Tish Robertson (VA DEQ).

Minutes

12:00 PM **Introductions/announcements –** *Peter Tango (USGS), August Goldfischer (CRC)*

12:10 PM **Data usage in 4-Dimensional (4-D) interpolator development and application –** *Rebecca Murphy (UMCES)*

Presentation:

Rebecca reminded the group the purpose of the 4-D interpolator tool: to serve as one part in the process of a more complete criteria assessment of the tidal waters of the Chesapeake Bay. Currently assessed are the 30-day mean criteria and instantaneous minimum in the deep channel. There are some criteria that we are not able to assess such as 7-day mean and instantaneous and one day means that we haven't had the tools and data sufficient to assess them yet. But we are aiming to do that.

As a reminder of the current process, data is pulled together into an Inverse Distance Weighting (IDW) interpolator in 2-week cruise periods snapshots, designed to interpolate in 3 dimensions. This gets us 3-D snapshots. These are compiled to monthly for 30 day means or individually for instantaneous ones, and used for 3-year periods of water quality assessments. Data includes the tidal dissolved oxygen in the Bay Program data hub. It includes the long-term fixed stations, the calibration data from continuous monitoring (con-mon) and DataFlow, and citizen and river keeper monitoring and other inputs that meet the EPA's data requirements.

Our goal is to develop a spatial and temporal interpolation tool for water quality monitoring data to ultimately assess all of the criteria. We also want to output statistical estimates of uncertainty, reproduce the daily and hourly variability in the data, and split up the oxygen into the Bay designated uses for the assessments. So we need to know what the pycnocline is. From hourly and daily outputs we'll aim to help to do the water quality assessments.

The data used for the current interpolator will still be used. We refer to the 4-D interpolator as a model, but when we talk about it to larger audiences we need to make the point that it is not a process based model. It is solely focused on observed data and not modeling any processes. The modeling we're talking about is building in statistical relationships between the DO data and location, time and other DO observations to get things like autocorrelations. These statistical relationships will allow us to estimate with uncertainty what the oxygen was in places and times where there was no data collected.

The data will be used in 4 categories:

1) Explore patterns to inform development
2) Development and testing
3) Validation
4) Application

Other possible data uses include:

- Shallow water explorations:
    - NOAA daily products for salinity and temperature based on satellite data (Wes Slaughter's work)
- Pycnocline:
    - Possibly use freshwater flow measured at USGS gages
- Maybe meteorological data as well
- Bathymetry:
    - We need to know the bottom depth everywhere, and want to use any updated information if possible.

<u>Discussion:</u>

Matt Stover in the chat: Will the 4-D interpolator include all of the con-mon data being collected?

Rebecca: Yes, I think so. Not necessarily in all the different parts of it. We did the wavelet analysis and there's a chance we'll add some in if there's a space and time we want to look at in more data. As we go forward any new con-mon data will be incorporated into the tool. We need to use the con-mon data.

Matt Stover: The main question is that the current interpolator doesn't include any con-mon data and we want to make sure this new version does. The less interpolation we have to do the better, whether spatial or temporal.

Tish Robertson in the chat: To build on Matt's question, will we be able to assess con-mon data (along with the discrete data) using the 4-D Interpolator? And the DataFlow as well?

Rebecca: Yes. If you have a specific place you think there's a gap we can talk about it.

Tish Robertson in the chat: Are the Fourier analysis coefficients specific to those segments with high frequency data?

Elgin Perry: I will be addressing this in my presentation.

Jon Harcum: The bottom right-hand panel shows the distribution of data in the Data Hub and that includes the fixed station network and any of the calibration data from con-mon and DataFlow. Because that distribution is not uniform across 24 hours like you'd expect, it made sense for us to integrate the con-mon data in two phases. The first phase is the daily DO value. To combine data hub and con-mon data, it's appropriate for us to down select the con-mon data down to a single value for the day that mimics this distribution. Then in the phase 2 part that is where all of the con-mon data be it 10-minute, 15-minute, hourly etc., is used to inform the work. It leads into what Elgin's going to be talking about next.

Matt: We still have to use one daily value? It's rarifying data from con-mon to one daily value and using the diel pattern found in that con-mon site to inform interpolation over time from that site over an entire day.

Leah Ettema (EPA): Can you explain again why you had to down select to a daily value?

Jon: We looked at the pros and cons of throwing all the data in vs down selecting to a common frequency. The value of down selecting is it allows us to preserve some of the statistical characteristics to inform Phase 1 (the daily estimate). We have one data set, the Data Hub, which are generally single samples at a particular day at a particular depth; for the other dataset, ConMon, if we push all that data together to a daily average, we don't preserve the statistical properties. We talked about it and decided down selecting to one daily value was the best way to handle con-mon for the daily model. When we get to looking for the hourly deviations and look for the diel signal for the day, that's where we will look at every value that comes out of con-mon, would it be used to inform these Fourier coefficients. More detail will be provided. The phase 1 and phase 2 results get combined for a single DO answer.

Leah: Some locations only have discrete data, some only have continuous. If you were to not use the discrete data, you would lose spatial interpolation power?

Elgin: Correct. The fixed station network data has the best spatial coverage of any data set we have and has been the primary tool for calibrating the daily mean part of the model. When we started extracting data from con-mon data, we wanted that data to have similar statistical properties to the fixed station data because there is an underlying assumption of the model fitting technique we use for the daily mean model that all of the data are independent and identically distributed. A violation of this assumption would be used daily means from con-mons alongside single point observations from the fixed station network. So we're pulling out a single once a day observation from the con-mon so the fixed station network single point data and the con-mon single point data are more compatible with this assumption of independent and identical distribution.

Carl Friedrichs (VIMS): If some of the data you're trying to find is associated with VIMS maybe I could help get it for you.

Rebecca: Appreciate it! I will let you know if we need help.

12:30 PM Exploration Tree of Stochastic Components of Daily Mean and Small Scale Variability – *Elgin Perry*

Presentation:

The GAM is the tool that produces the daily mean estimates. I'll focus on statistical simulations: statistical simulations trying to reproduce small scale (within day) variability, and reproducing variability at a broader space time scale that has to do with adding variability to the daily means themselves. This second part has had some road blocks.

Three approaches to reproducing small scale variability (all based on using continuous monitoring data):

1. Resample mean adjusted residuals with Day as Experimental Unit.
2. Model cycles with Fourier terms and resample cycle detrended residuals.
3. Model cycles with Fourier terms and Simulate cycle detrended residuals.

For the first approach, Elgin envisions taking the con-mon data (which there is a lot of) and visualize it as daily units, and stratify the daily units by season and space. Then, we could compute a mean adjusted set of residuals. We could store those as a huge data set. Then when we're doing our simulation from the GAM to get a prediction for the day, and add to that a set of residuals (Ris) from 1-24, that should be a pretty good representation of the kind of variability we'd see for DO. It takes care of the deterministic cycles and autocorrelated structure of the data all at once. I realized instead of subtracting a daily mean to get these residuals it would probably be better to subtract a linear trend of each day so when you composite it with the daily mean data you could take the difference between two sequential daily means and draw a straight line between them which would represent the trend over the day, and composite the linearly detrended residuals.

Discussion/Questions for #1:

Jim Hagy: These deviations from the daily mean, you could add to the GAM prediction provided that the GAM was really predicting the daily mean. But the GAM is predicting the expected value from the survey data which may not be at the mean value because of the temporal distribution of the samples. How have you dealt with that?

Elgin: I haven't yet but I'm working on it. It's easier to deal with it with the Fourier analysis which is the tool I'm planning on using. Say the peak of oxygen occurs in the middle of the day and that's when we're sampling, you would set the inner set of your daily cycle to start at noon. If that was the peak, your Fourier model would get down from there and start back up. When I say partition the con-mon data into daily units, those daily units could start at noon. And that

would be the peak of the trend, the intercept of the model rather than the mean. It could be done based on averages.

Jim: What if you calculated the residuals for each hour not from the daily mean but rather from a mean you calculated based on the density of the sampled hour of the day you described earlier. If you randomly sample the continuous data to simulate the hour that the GAM is representing and then calculate residuals from that.

Elgin: You could do that. I also think it would work if you typically sample at 11, draw a straight line between 11 one day and 11 next day and calculate residuals from that straight line.

Continuation of presentation - #2

Option #2 is a hybrid of 1 and 3. It assumes that we're going to establish cycles in the data that we can model with a Fourier type analysis (using sin and cosin terms to reproduce the periodicity that we observe in the data). We could fit each bit of Fourier model to each individual day of data and remove the cycle and save the residuals. Then when we put together a daily DO prediction we would take the daily mean of one day and the daily mean of the next day and fill in between those daily means by first adding on a smoothed term to reproduce the harmonic cycles. Then we could grab a set of residuals calculated based on the harmonic cycles. That would take care of the deterministic and stochastic parts. The stochastic part would have the proper autocorrelation for the proper dependence through time. That would be reproduced because it's coming from the observed data. It has the same issue that Jim just raised; what we consider to be the intercept of that Fourier model maybe should reflect more what a day time observation is than just the mean observation over the 24 hours.

Discussion/Questions on #2:

Carl: With the stochastic part, I would guess that the stochastic part isn't stationary so that different stations would have different statistical stochastic properties. And maybe one station's statistical stochastic properties would change with time. Have you thought about it?

Elgin: I have thought about it. You'll see this problem is mathematically complex. Getting into things like non-stationarity is getting more complex. To me that would be an advantage of this resampling approach. If there's non-stationarity in the data and you resample that data to composite with your estimates of daily means, you would capture some of that non-stationarity. This idea of resampling is not something I put on the table until this presentation today.

Continuation of presentation - #3

Like in #2 we deal with the deterministic part by fitting a Fourier model to all the data. I introduce the idea that we might do a second GAM. Once we get all of these Fourier coefficients, we could model those as a function of time and space and come up with a GAM that predicts what the Fourier coefficients would be. For the random part of within day variability we would simulate these mathematically. We'd assume something like an

Autoregressive One (AR1) model. Using our vast supply of con-mon data we could estimate at the same time we do the Fourier analysis we could use the residuals to estimate an autocorrelation coefficient. Then we could generate a set of errors that have the type of time dependence we have in the observed data using this recursion relationship we have. That has the advantage of not having vast amounts of data to do resampling from and just generate a set of residuals that have a property similar to what we'd expect to see in the observed DO. This is the tool we've been thinking about using up to this point, but it doesn't deal with non-stationarity at all, it assumes the autocorrelation coefficient is constant across the board.

Another way to generate those temporally corelated residuals is to have a variance covariance matrix over time (AR-1 correlation matrix). When you want to generate a set of residuals that come from an AR-1 process with this variance-covariance matrix, you do what's called a Cholesky decomposition of the matrix which is kind of like taking the square root of the matrix and multiplying it by a set of independent residuals. You end up with a set of residuals that have the covariance you have specified. For an AR-1, you don't have to do a complicated mathematical procedure to get the Cholesky decomposition. It's very easy to generate a set of residuals that have a variance covariance structure as specified by an AR-1 process. This is the method I'd been planning to use before I thought about using resampling strategies.

No questions on #3.

<u>GAM tool for daily means:</u>

The GAM predictor variables are Estuarine Latitude, Estuarine Longitude, Sample Depth, Bottom Depth, Long term trend, Seasonal Trend. Currently we're planning to do this on a 4-D lattice where we do one prediction per day per meter per depth. In latitude, longitude it's typically 1 kilometer by 1 kilometer although may be smaller in some tributaries. When we talk about introducing a reasonable degree of variability to these daily means, we have to think about dependence over time on a scale of one per day, dependence over depth on scale of one per meter, and dependence in latitude and longitude on a scale of a kilometer. In an ideal world you would just go out and collect data on a lattice. We don't have that luxury, so we're trying to take the data we do have and figure out how to put those together to reasonably reproduce the spatial and temporal variability of daily means.

We're going to talk about two components of variability: uncertainty in the GAM daily mean model itself, and some kind of space time correlated errors.

Model uncertainties in GAM itself: When we fit this GAM it gives us a parameter vector. That vector has an associated variance covariance matrix. One thing we can do to reproduce uncertainty in the model fit itself, when we do our predictions, we don't use beta hat every time, we create a new parameter vector called beta star. This will be a random draw from a multivariate normal distribution which has mean beta hat and variance covariance matrix beta sigma, hat to beta. The way we get that goes back to using this Cholesky decomposition. We start with a set of multivariate normal, give those a mean of beta hat, and add to it a set of

mean zero variance covariances beta hat sigma hat beta, and come up with a beta star with is a random draw from a multivariate normal. In this case there's no trick for getting the Cholesky decomposition. But beta is not that big so it won't be a problem. Then we'll get a set of daily mean predictions. For the error term we would like it to have some kind of realistic space time dependence.

For the time component, assuming everything has 5 levels, if you take a prediction vector and organized it and sorted it so it was in the order where the days are together and as you go from block of days to block of days you move along depth then latitude and longitude. The variance covariance among days I would expect to behave like an AR-1. Then when you start putting multiple blocks of days together (because you have different depths, latitudes and longitudes) you'd end up with a matrix composed of blocks of AR-1 variance covariance matrices. That is pretty easy to deal with; you can just write down the Cholesky decomposition and because it's diagonal, composite those into a big Cholesky and get a reasonable variance covariance matrix that reflects the autoregressive structure of the time.

Then I tried to incorporate some spatial dependence in addition to the time dependence. If you just had two depths side by side, I started trying to solve this matrix to see if it would yield a pattern like AR-1. The reason I'm doing this is when you blow out this matrix to a full segment of latitude, longitude, depth and time, is going to be absolutely huge. I don't think, even if we could postulate what it looks like, computers could deal with getting a Cholesky decomposition. So I'm trying to get a shortcut route. So far I've not gotten anywhere with that. There are some Cholesky methods that deal with blocks of matrices, so I'll explore that path. I'm also thinking about another way to get there by postulating a Cholesky factor and squaring it to get the variance covariance matrix. I'm looking at using nested random effects models.

Another possibility is to assume that space behaves like time so you just construct a variance covariance matrix as a big AR-1, but I don't think that's particularly realistic. A last resort would be to assume independence of space. The resampling option I talked about earlier doesn't work here because we don't have any data on a kilometer by kilometer grid that we could resample from to construct the level of covariance you'd expect to see at that distance. Our fixed station network tends to be tens of kilometers apart.

1:00 PM Discussion, Q&A

Carl: When you talk about AR-1 in space, I'm confused because AR-1 is a time series concept that doesn't translate to space unless you're talking about a 1-dimensional line where for some reason the space is only autocorrelated in 1 dimension.

Elgin: Mathematically the derivation of AR-1 assumes it's a forward moving process, you can't move backwards in time. The correlation structure says that 2 neighbors are correlated with a correlation of rho, and if you're two units apart it is rho squared. You're correct that the ideas we used to originally formulate autoregression models in economics are one direction but what I'm really worried about is how do I get a tool for reproducing certain variance covariance

structure. Doesn't seem totally ridiculous to say that the correlation over space degenerates in the same way that it does over time.

Carl: Yes. I thought those were called spatial autoregressive models and they have a 2-D structure. It's the same idea, the terminology confused me. Another comment, you said what you would use to look at spatial correlation. The Data Flow are available for that and they're typically many kilometers long on a particular day.

Elgin: Yes and I should've mentioned if I had to fill in a tau I would get it from the Data Flow data. And we would have to go back to the con-mon data to get estimates of rho. Instead of hour by hour we'd look at it day by day to get estimates of rho. We're doing that because we don't have any data that is kilometer by kilometer, meter by meter and day by day to estimate these all together. So we're trying to get it from different pieces of data we have and fill in this matrix and then see if we can figure out a way to factor that matrix so we can reproduce error terms that would have that variance covariance structure.

Carl: If only satellites could measure surface oxygen really well. We're getting daily satellite data. We can apply it to clarity and hopefully chlorophyll in the future. My other comment was using the con-mons for information that will then go into spatial grid that's a kilometer, I wonder if there's spatial conundrums there because the con-mons are mainly in shallow areas close to shore. If you moved a kilometer away from shore the environment would be completely different. I wonder if the kilometer spacing, if that's what used consistently through the 4-Dimensional (4-D) interpolator, does that mean the 4-D interpolator isn't really valuable for shallow water because you want to think at scales less than a kilometer if you're moving perpendicularly to shore.

Rebecca: We will likely use a finer grid in the tributaries. In the moment any testing we're doing is on the current interpolator grid, which does get very fine in the smaller regions of the Bay. The question was whether we want to do less fine because it gets really small, like 50 meters.

Elgin: Most of our con-mon data is located along a shoreline, and even if you move into just the mid-channel of a sub-estuary, you can get a fairly different behavior. Once I worked on a study with Walter Boynton looking at a center channel buoy vs a con-mon which were less than a kilometer apart but the behavior was very different between those two locations (though we were looking at chlorophyll, not DO). I don't know if we would expect DO to have similar differences. I think that's one of the shortcomings of some of the resampling plans I presented earlier. They will be heavily weighted towards con-mon data collected in shallow water. That might make it a better strategy to use the Fourier models and try to model those coefficients as a function of space, so we'll have some continuous data from mid-channel. If we see that has a very different signal we could use what I'm calling the Harmonic Coefficients GAM to interpolate that structure from shoreline to mid-channel and give us a continuum of predictions in between. Then there's the issue that we don't have anything at this point that is comparable to

Data Flow that is not surface data. VIMS might have some. You had that scan fish that collected continuous DO as it moved up and down in water column?

Carl: Yes we did. I bet there's a lot of unanalyzed data. Iris Anderson and Mark Brush were doing that in the York.

Tish: It was called an acrobat.

Peter: And Larry Haas.

Mark Trice (MD DNR): There was something at Horn Point that was similar.

Tish: I was going to throw something out on the table regarding the weakness of the current interpolator. For CB7PH we have a really deep trench. The deepest station we have in CB7 doesn't come close to get to the bottom of it. There's like 20 meters we're not monitoring in CB7PH. So what happens in the 3-D interpolator is it looks for the nearest neighbors, and the nearest neighbors are almost into MD waters. We know the MD waters have true hypoxia. What happens is the assessment of CB7 really reflects what we observed in MD than what we see in VA because we don't have any data in the bottom 15 meters of CB7. It's a monitoring weakness, it's not a weakness of the interpolator. We don't have the data. I'm curious how the 4-D interpolator would handle a situation like that where we don't historically have any observations at certain depths.

Rebecca: I'm doing some testing of the GAM grouping segments. That'll be a good spot to look at and see what happens with my testing. I'll show you and we can talk about it

Tish: I know a lot of the arrays are in MD waters, for good reason. I am concerned that with the lower Bay, specifically CB7, that we'll be in the same situation with the 4-D interpolator as we are with the 3-D interpolator and what are some ways to heal that situation.

Rebecca: If there's no data, the model will extrapolate. But it's informed by the data. We'll take a look at it.

Peter: I think of how many locations the fisheries and benthic surveys do. I thought our benthic folks take bottom DOs when they're out there.

Tish: Occasionally they land in the deep trench but not as often as you think they would. That data could be useful for calibration. It's been a long time since I looked at the ChesMap data. Last time I looked at it was when Mary Ellen was the QA coordinator. It was impressive the scale of it so it would be good to roll that in.

Elgin: That issue came up with the GAM in some of my early testing when I would see in the profile of the channel you could see every now and then you could see a column that had visibly lighter color of DO than others. I found out that what was causing that was a cell in the prediction space that was deeper than any cell we had in the fixed station network. It was extrapolating to depths beyond where we had observed data. In particular I had a term in the

model that was the bottom depth. It was biasing that whole column. Any time we get to doing predictions in a place with observed data we have to keep any eye on it, and make sure we're not reporting obviously bogus results. It happens with GAMs, not just with the IDW interpolator.

1:30 PM Adjourn

Participants:

August Goldfischer (CRC), Jon Harcum (Tetra Tech), Matt Stover (MDE),Elgin Perry, Rebecca Murphy (UMCES), Peter Tango (USGS), Mark Trice (MD DNR), Tish Robertson (VA DEQ), Isabella Bertani (UMCES), Marjy Friedrichs (VIMS), Leah Ettema (EPA), Tom Parham (MD DNR), Andrew Keppel (MD DNR), Jim Hagy (EPA), Amanda Shaver (VA DEQ), Carl Friedrichs (VIMS), Zhaoying (Angie) Wei (UMCES)