MINUTES
Data Integrity Work Group (DIWG)

Virtual Meeting

Wednesday, February 26, 2025
1:00PM-3:00PM

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

Participants:

Allison Welch (CRC), Amber DeMarr (UMCES), Jay Armstrong (DGS), Breck Sullivan (USGS), Tracee Cain (DNREC), Anessa Carter (ODU), Suzanne Doughten (ODU), Durga Ghosh (USGS), Elgin Perry, Ellyn Campbell (SRBC), Emily Young (ICPRB), Scott Hasinger (VA DEQ), Isabella Bertani (UMCES), Jaclyn Mantell (UMCES), Jacob Kilczwski (MDH), Jamie Shallenberg (SRBC), Jerry Frank (CBL), Cindy Johnson (VA DEQ), Kristen Heyer (MD DNR), Kelly Krock (EPA), Michael Lane (ODU), Lara Johnson (MDH), Clinton Leiby (PA DEP), Christopher Mason (USGS), Ian McMullen (DNREC), Meg Maddox (UMCES), Mike Mallonee (ICPRB), Kevin Minga (ODU), Nianhong Chen (UMCES), Rebecca Murphy (UMCES), Scott Schroeder (FIAlab), Sidney Anderson (UMCES), Mariah Smith (ODU), Tyler Shenk (SRBC), Jimmy Webber (USGS), Meighan Wisswell (DEQ), Heather Wright (ODU), Emma Jones (DEQ), Liz Chudoba (Alliance for the Chesapeake Bay), Becky Monahan (MDE).

1:00 PM Introductions, Announcements
Monitoring, Laboratory and Community Science updates
In-person Meeting Poll

Cindy Johnson (Chair) welcomed everyone to the meeting and introduced herself. All other meeting participants introduced themselves after Cindy.

Jay Armstrong introduced himself and let the group know that he has taken a new role with the Virginia Division of Consolidated Laboratory Services (DCLS). He is now managing the group that he was the technical lead for. He will be bringing his colleague, Ryan Lewis, into the DIWG meetings.

Liz Chudoba provided the community science updates from the Chesapeake Monitoring Cooperative (CMC). They are currently prepping the tier 3 groups for the 2025 sampling season. All groups are good to go. The Severn River Association had some staff turnover. They were able to get their Quality Assurance Project Plan (QAPP) updated with the new staff member and a field audit this summer is not going to be needed. So once the QAPP is approved they will still be able to be tier 3. Nanticoke, MDE Shellfish, Blue Water Baltimore, Anne Arundel Community College, Arundel Rivers, and Severn Rivers Association are the tier 3 groups for 2025. They submitted data to the Maryland Integrated Report at the beginning of the month. This was data through 2023, including the tier 3 groups, as well as some of the tier 2 groups, which were Shore Rivers, Black Water National Wildlife Refuge, Wicomico Creek Watchers, Elk and Northest Association, and Patapsco Heritage Greenway.

CMC has made some updates to their Data Explorer homepage. They would love it if everyone would check it out. They have made it more interactive with a map that will update based on filters and searches.

Over the past year, CMC has been working on a process to integrate bacteria data into Maryland's integrated reporting process. They didn't have a way to access labs for tier 3. Liz has been working with MDE, the Bay Program, and monitoring groups in Maryland. They are excited to be rolling out a process to have an EPA approved QAPP. They needed to figure out the lab certification process. One thing they found to check that box was to do proficiency testing through NSI laboratories, where each of the labs that will be running the i-dex or membrane filtration system for either E-coli or enterococcus can participate in the proficiency testing, run their samples, and send their results to the NSI labs and they will determine whether those samples are in the appropriate range. They attempted to get this rolling in the January study that NSI does. Unfortunately, there was an issue with shipping, so those samples were not actually obtained by the labs. They are going to try again in the March study. If they pass the proficiency testing, they will follow up with a field audit, similar to the tidal process, that will cover the bacteria samples, turbidity, dissolved oxygen, and pH. The groups that are piloting this in Maryland are the Anacostia River Keepers, Anne Arundel Community College (already a tidal tier 3 program), Arundel Rivers Federation (already a tidal tier 3 program), Blue Water Baltimore, Potomac River Keeper, and the Upper Potomac River Keeper's lab at Hood College. The groups are excited to have a process for this data to be used as tier 3 in the integrated reports. Liz will keep the group posted on whether they can get the sample to the groups in March.

Allison Welch ran a poll to see if the group members would like to have the next DIWG meeting in-person and where they'd like to have it. For the first question, "Are you interested in attending the next DIWG meeting in-person?" There were 12 votes for "I'm interested," 6 votes for "I'm not interested," and 5 votes for "I'm not sure." Then, for the second question, "Where would you prefer the in-person meeting be?" There were 7 votes for Annapolis, 6 votes for Colonial Beach, VA, and 8 votes for Yorktown, VA (VIMS).

1:20 PM Nutrient and Sediment Loads and Trends – *Jimmy Webber (USGS) and Alex Soroka (USGS)*

Summary: Nutrient and sediment loads and trends were recently updated for the Non-tidal Monitoring Network (NTN) using observed data through the water-year 2023. This updated information provides the latest assessment of how amounts of nitrogen, phosphorus, and sediment are changing in rivers and streams throughout the Chesapeake Bay Watershed. This work is only possible because of partnerships across the watershed and among agencies.

Major Messages: The team believes that the quality of the NTN data has never been better. It is all from observed monitoring data. They continue to see that loads are highest in the agricultural and urban streams, which has been a long standing pattern. The good news is that the long term patterns from 1985 to 2023 have an improved level of nitrogen and phosphorus. The downstream locations have had a decrease in nitrogen, phosphorus, and sediment since 2014. More than half of the stations are detecting no change or increasing results from 2014 to 2023. USGS plans to continue to work with everyone to understand these results.

Monitoring Network and Methods: The Non-tidal Monitoring Network has 123 stations throughout the watershed which collect measurements of streamflow and water quality samples monthly and during storm targeted conditions. All of that information is used to compute loads and trends for five different parameters. These are total nitrogen, nitrate and nitrite, total phosphorus, orthophosphate, and suspended sediment. With the most recent update, they can report loads for 122 stations and short term trends for 120 of the stations, which is something to celebrate. The stations reporting first time data are circled in red on slide 6.

The team has now created a reproducible suite of water quality samples that were all assembled from publicly accessible databases. Previously, when they provided load and trend updates for the NTN, they pulled all of the water quality samples from a static database that the team maintained. This database had all of the historical data and new data would be added on each year. It worked well, but it was challenging to verify old samples. They decided that they should be able to find this data in a publicly accessible database to build confidence in the results. They went back and pulled all publicly available data in the water-quality portal and Chesapeake Data Hub, which was compiled into a suite of data for the NTN. They ended up with a lot of high quality data used to compute loads and trends. They are writing a report now to document these methods. This work was led by James Colgin from USGS in Pennsylvania. If the group would like to know more, James could present to the group. They also compared this new dataset with the historical information and found that it compares well. They uncovered some new samples in publicly accessible datasets. They feel good about what they had but feel even better about what they are working with now. If recomputed with the last load and trend data release, the results wouldn’t have changed too much. Only 2 of the 300 trend results from the previous NTN update would have changed direction.

Streamflow and Per-Acre Loads (Yields): In 2023, the watershed delivered about 17% less water than the long-term average making 2023 a dry year. Our water quality information is based on flow normalized loads which is a statistical approach to remove the influence of streamflow from the data. However, we know that even flow normalized loads can be affected by long-term increases or decreases in stream flow. So they wanted to ask whether there was an increase or decrease in streamflow over time. At the long-term stations, most of them had no meaningful change in streamflow over time. One station had a significant increase in flow relative to 1985, which was the Choptank

River on the Eastern Shore. That is an active area of research to understand why and what impact that may have on water quality trend results.

On slide 11, the map shows the total nitrogen loads normalized by watershed area as an average of the most recent five years. In the map, the lighter yellows are lower loading and the darker brown are watersheds that contribute a relatively high amount of nitrogen in respect to their monitoring stations. These areas of darker brown tend to be in the Susquehanna River watershed, which isn’t surprising due to the area’s intensive agricultural development. You can also see some darker tones in the urbanized areas near Washington DC. The scatter plot shows the same information as the map on the y-axis and the x-axis is percent of agricultural land. Moving from left to right, you can see that with more agricultural development there are higher nitrogen loads. There are exceptions to this. On the Eastern Shore, there are two exceptional stations, Chesterville Branch and Morgan Creek. Chesterville Branch is our highest yielding watershed of total nitrogen. It is a very agriculturally dominated watershed. It is neighbors with Morgan Creek which also has a lot of agricultural land, but lower total nitrogen levels. There has been some research done here suggesting that there’s more natural denitrification happening in Morgan Creek, which is a function of the soils and geology of the area. All of these stations have intricate patterns affecting these nitrogen loads. The line at 4.3 lb/ac shows the nitrogen planning target for the Chesapeake Bay Watershed, which is the amount of nitrogen the TMDL is trying to reach. This helps prioritize watershed areas that need to have focus.

The next is the map on total phosphorus. There are some areas similar to the total nitrogen data, including the lower Susquehanna, Eastern Shore, and urban areas around DC. Notably, there are larger amounts of phosphorus in the Virginia streams, the Rappahannock and lower James River Watershed. That is an interesting difference considering the low amounts of nitrogen in the Virginia streams. Looking at the scatter plot, there is also a relationship between the amount of agricultural land and the phosphorus levels but there the highly urbanized watersheds are also areas with elevated phosphorus yields. Watersheds with more than 50% urban areas are colored with a dark black outline. Watts Branch and Hickey run are two very urbanized watersheds right outside of DC with high amounts of phosphorus being delivered to the streams. The planning target has been added as a reference line at 0.2 lb/ac to prioritize watersheds.

Suspended sediment is the last indicator graphed. The Rappahannock and James River watersheds are continuing with the high yield seen in the phosphorus map. The urban watersheds are contributing a relatively large amount of sediment, along with the headwater Susquehanna stations. Eastern Shore now has very low suspended sediment data because this is a low gradient, flat system that doesn’t have the energy to mobilize sediment into the streams. The same plot is shown again to show the relationship with agriculture, which seems to fall apart under this indicator. There is a more distinct relationship between high levels of suspended sediment load and highly urbanized watersheds, like the Northwest Branch, Anacostia Creek, Rock Branch, and

West Creek. A lot of these are around the western shore of Maryland where we have the greatest sediment loads. Then, a target load of 184 lb/ac is added onto the graph to show how to best prioritize the watersheds.

Nutrient and Sediment Trends: Long-term nutrient trends are those that have been monitored since 1985. For nitrogen, 58% of the 43 stations have improved. The graph shows the range of increase or decrease that has occurred. The median change is a -13% reduction. All of the stations throughout the Susquehanna are showing improvement. For phosphorus, 56% of the 16 stations have improved. There is one station in the headwaters of the Susquehanna that shows no change, but the rest of the Susquehanna stations are improving. The degrading trends are found on the Eastern Shore and some of the Virginia stations. For suspended sediment, there is about an even mixture of increasing and decreasing trends in the 15 stations. Most of the increasing trends are in Virginia with the decreasing trends throughout some of the Susquehanna, on the Eastern Shore, the Potomac River, and the Western Maryland sites.

Since 2014, the stations nearest the Bay have decreased in amount of nitrogen, phosphorus, and sediment. It's important to keep in mind that loads delivered to the Bay are heavily influenced by the changes in our downstream Susquehanna and Potomac stations. These are the largest rivers entering the Bay and deliver about ¾ of the load. This map highlights the 33 most downstream stations, so closest to the tidal waters. The average is about a 8% reduction which is heavily leveraged by the Susquehanna and Potomac which are highlighted and both improving. Elsewhere, it is interesting to note that there have been increasing trends throughout the watershed, particularly the Eastern Shore, which all show increasing phosphorus. The Conowingo is driving the sediment reduction to tidal waters, with other areas increasing or showing no statistical change. This is an area where we would be interested in marrying these non-tidal results with our tidal water quality monitoring results, which is probably something we will focus on in the near future.

Through the entire watershed, you can see that in the Susquehanna through the mainstem there's a lot of areas of improvement. The western branch of the watershed is showing some degradation. The Potomac River is showing that the headwater tributaries are showing increased levels of nitrogen but that becomes decreasing levels of nitrogen as you move downstream. 43% of the 120 stations have improved, slightly more than what has degraded. The plot shows the change in context of their yield. We want to see improvements in the highest yielding parts of the watershed. Virginia streams tend to have relatively low amounts of nitrogen load.

For phosphorus, the Susquehanna shows a mixture of conditions, which is also true elsewhere. The areas that have more consistent degradation are on the Western Shore of Maryland, the entire Eastern Shore, and many of these stations on the James, Appomattox, Rappahannock, and even some of the York River stations. About half of the 105 phosphorus stations have degraded with only 24% improving. There may be more of a concern for phosphorus than nitrogen. The Pequey River in the lower

Susquehanna is the highest yield station for phosphorus and its trend is increasing, which is not what we are hoping for. Hickey Run is the Potomac's highest yield but has been decreasing. About half of the stations that are exceeding the planning target have degraded which is similar to the overall patterns of the watershed.

For suspended sediment, the map shows an interesting pattern where the headwater stations of the Susquehanna are increasing in their amounts of suspended sediments. That is interesting to understand what changes are happening in that portion of the watershed. As we get lower into the Susquehanna, we see more improvement and no change. Through the Potomac, there appears to be a lot more degrading than improving streams. The Eastern Shore of Maryland has a lot more increasing or no trend conditions. Virginia tends to be pretty mixed. Throughout the 105 stations, they are evenly split between no change, improving, and degrading.

The team maintains a project website with this data and information. The results will be uploaded on this website in the next few weeks. They also will be updating an interactive geonarrative where you can click on and explore individual stations to see their trend results. They are hoping to have all of that updated in the next few weeks.

Discussion

Mike Lane: Is there any connection between total phosphorus (TP) and suspended sediment (SS) loads and deforestation or habitat fragmentation?

- Jimmy Webber: Yes, total phosphorus and suspended sediment can follow similar transport pathways. When you deliver more sediment to the streams, that can often carry phosphorus attached to those sediment particles. We do see areas where when sediment goes up so does phosphorus. The question is what's driving the sediment increase? For the ideas that you posed there, deforestation and habitat fragmentation, there is research documenting that those land use changes can mobilize more sediment to the streams. That would be interesting for the group to think about. In those areas in the Susquehanna, are we seeing those types of land use changes that might deliver more sediment to the streams? Are there other challenges? I think there are interesting stories about the change of forest cover throughout some of those parts of the Susquehanna that might be worth exploring.
- Jamie Shallenberger: We see areas of the upper Susquehanna, particularly in New York state, that have glacial history generally provide more sediment, in some cases regardless of land use. I am not suggesting that the two factors might not be at work there. The glacial history up there does tend to present more mobile sediments in the watersheds.
- Jimmy Webber: That's great, Jamie. The local insights about what's potentially underlying the patterns, what's changed in the last ten years, and what people are seeing on the ground in these watersheds is so valuable.

1:50 PM Blind Audits Update – *Jerry Frank (CBL)*

Summary: After a bit of a hiatus, they're back online. Samples went out in December for the first half of the fiscal year (FY). They have 17 labs back participating, which is outstanding. Data is back from all but one. They have been in communication with them to use a delayed submission of their data. They should be getting that in soon. The second round for this FY should be going out by the end of March.

2:00 PM Coordinated Split Sample Program – *Mike Mallonee (ICPRB)*

Summary: Mike showed his presentation on the mainstem split sample data, which is from November 2023 to December 2024, and the tidal split sample data, which is from December 2023 to December 2024.

Discussion:

Jerry Frank: We are missing our TSS FSS data. The last two rounds of the mainstem splits data diverged quite a bit from the cohorts. We will get to the bottom of that. Just to put that on your radar. If I've contacted you in the past with these sorts of questions, expect an email from me this week.

- Mike Mallonee: Jerry, I'll get that five year summary to you and we can go from there.

Nianhong Chen: I have a few updates about the new instrument we have, the COAA 500 system. I started to submit the mainstem data last year and we recently had a problem with the phosphate channel. I had email communication about this problem and Jay has some experience with the same instrument. He recommended we use FFD6 instead of SDS for the phosphate channel. Recently this channel is almost recovered. Another update is about the new ammonium analysis. We figured out that there's some confusion in the manual. They were confused with the concentration of the ammonium standard with the molarity of the ammonium sulfate. I didn't really check into that and I used their number to make the standard and it turns out the actual concentration should be double. With the numbers they listed on the manual, I submitted the ammonium data to the mainstem, but that should be doubled. For the last two months, we have had a problem with the phosphate channel so I didn't submit the data. So that's my update.

2:10 PM Paperless Data Entry – *Scott Hasinger (VA DEQ) and Emma Jones (VA DEQ)*

Summary: The team works out of the Blue Ridge Regional office and are responsible for one non-tidal station. They are presenting on changing their data collection methods.

The traditional workflow for collecting field data involves taking out the sampling equipment and copying the field parameters to pre-generated field sheets, which come out of their centralized data repository. These are auto-generated when field teams schedule a run which include the sites and field parameters they need to collect. Unfortunately, things happen on the field. The data that is copied onto the field sheets is not always legible. You can also easily lose or damage your field sheet while sampling. By nature of copying things down from the probe screen, to the paper, and then later to the centralized database there are two chances for making human errors. They also have handheld displays that are difficult to read with a very small font and high glare screen.

Out of the Blue Ridge Regional office, they often partner with EPA for the National Aquatic Resource Surveys (NARS). NARS has integrated digital data collection to replace paper forms since the mid 2010s. Using digital forms that are managed on the iPads and sent back to the centralized EPA servers has hugely changed the nature of the program. It has eased the data collection in the field, improved the quality assurance, and has helped with the reporting timelines of these databases, since pages of field forms don't have to be transposed.

The VA DEQ Blue Ridge Regional office wanted to take some of those ideas and apply them to their statewide survey programs. They started by using iPhones with excel sheets and OneDrive. Now they are using toughbooks from Dell and sending data directly to an internal GIS portal which is able to write to their centralized data repository, the same one that generates the field sheets. It took a lot of changes and getting through challenges like waterproofing iPads, long-term battery life, withstanding the sun and conditions, figuring out a VPN that could connect the devices to the warehouse, teaching staff to work through technical issues on the water, and learning the new hardware. They have been able to improve screen visibility and have made great advancements in efficiency and quality control. By using a survey 123 product, they could build in a lot of quality assurance parameters, like thresholds that shouldn't be exceeded with warnings that indicate the wrong number may have been typed in. They have reduced transcription errors by copying numbers from a digitized multi-probe and pasting them onto the computer screen of the data collection. They have moved from the tablet units to a more heat tolerant laptop (one of the slides shows this method).

This is the first season of using Survey123. They collected almost 70,000 samples at 51 sites. Using this system, staff saved 97 hours of time that would have been spent in the initial strategy.

They would like their summer intern to convert the lessons learned from using Survey 123 for lakes into a digital data collection form for their one non-tidal network station. They think that leveraging the lessons learned has a lot of potential to adapt from lakes to non-tidal and tidal stations. There is also interest from other offices throughout the state.

2:35 PM **Instrument Comparison Study Analysis Results –** *Mike Lane (ODU) and Elgin Perry*

Summary: Elgin shared that the instrument and lab comparison studies have been happening for a long time in the Bay Program. They would do the study and often find some small statistically significant difference between one instrument and another. It would be so small compared to environmental variability that they would conclude it wasn't a problem, but in Elgin's mind there was a problem because it is statistically significant. The study that will be discussed today will be introducing a new method of analyzing this method comparison data using a random effects model.

Suzanne Doughten and Heather Wright collected the data for this study. Mike Lane did the data analysis and sent it to Elgin. Elgin has put together a little set of results.

Heather gave a description of how they collected the data and the study design. Before the comparison study was done, Heather got the method working on the SEAL. The initial tests were done with the reporting limit, the initial precision recovery, LOD, LOQ, MDL, carry over test, and DOC. Once that was taken care of, she went into the comparison study. They used an auto diluter to create the curve and run check standards. On the SEAL, it's all manual, so she made up the curve. The high standard from that went on the Latchet and the whole curve went on the SEAL. Samples were saved from archived cruises and splits. She tried to choose a variety of samples with variation in the value, location, water column level, and season, to get a well-represented suite of samples. Then they were run on both the Latchet and the SEAL on the same day.

Elgin continued explaining the statistical process. The traditional model for doing a methods comparison statistically is to say your concentration (Y_{ij}) is equal to some overall mean and then there would be an effect (β_i) that would represent what the instrument contributes to the concentration. In this case, typically we would have two instruments that were in comparison and all the other variability from sample to sample would be a random error. Often, that model is simplified. The data are usually collected in pairs, as Heather was describing. Both instruments measure the same sample on the same day. You can take the difference of those two concentrations, call that D, and model that as an intercept and an error term. The intercept of this simplified model is going to be the difference of the two instrument effects. The error of the simplified model is the difference of the error terms of the previous model. That's the traditional way of looking at this data from a statistical point of view. They run statistics using either the paired T test or a Wilcoxon sign rank test, which is a nonparametric test. It would give you a p-value which would give an idea to whether there was a significant instrument effect. When they do these tests, they frequently find significant differences that don't make much sense. Elgin was concerned that there is possibly another random effect here that is giving correlated data or pseudo replicated data. One random effect that came to Elgin's mind was the act of doing the calibrations of the instruments.

The random effects model takes the error term and splits it into two parts. There is still the μ, overall mean, and β_i, instrument effect, but it introduces a new error term, δ_{ij}, which would represent the effect of deviations that may be created when calibrating the instruments. When you create your calibration curve, there's a bunch of samples that you run after you create that curve and all of those will have deviations as well. Those deviations are represented as the ε_{ijk}. With this new model, the random effect is now represented by two terms, one will have a lot of replication, while the other, the calibration term, does not have a lot of replication.

A difference version of that model can be created as well. It still gives the difference between a pair of samples, d_{jk}, and the γ which represents β_2 - β_1, or the difference between the two instruments. The ρ_j term represents the deviations between the two calibration curves, if you do those separately for the two instruments. The ε_{jk} term represents the sample to sample variability in the differences. In the results that follow, they are focusing on this difference model because it's simpler. When Mike did the analysis, he always computed the concentration measured by the Latchet instrument minus the concentration measured by the SEAL instrument.

The big question is, is this calibration effect, or new random effect, a real thing? Elgin confesses he chose an example that makes it look real. There are other examples we will see that aren't as dramatic as this. On the Y axis, they have plotted the difference in the measurements by the two instruments and on the X axis, they have three different calibrations at Time 1, Time 2 and Time 3. For the box plots, the dark bar across the middle represents the median of the data, the box represents the middle 50% of the data, the lower end of the box is the 25th percentile, the upper end of the box is the 75th percentile, the vertical bars represent the range from the minimum to the maximum of the data, and the points beyond the vertical bars are outliers. The first two calibration curves give a median that is slightly positive, so that would say that the Latchet is giving slightly larger values than the SEAL, but the boxes or interquartile range overlap zero, which is a good indication that the differences are not significantly different from zero. For the first two calibrations everything is looking good, but for the third it has gone from having a positive difference to a negative difference. With this third calibration, the SEAL is giving larger concentrations than the Latchet and the interquartile range no longer overlaps zero. This third calibration makes Elgin think of those studies that result in those small differences that look statistically significant but don't have a good explanation. If there was only data on NH3 from the third calibration, they would probably reach that conclusion, but they can see that if they start averaging across the different calibrations, then you would reach a different conclusion that they're probably not statistically different. The figure on the left (slide 6) shows the differences in the raw data in the scale in which the data were observed. Typically concentration data follows a log normal distribution. When Mike did his analysis, he log transformed the data and took the differences of the log. He created the figure on the right hand side. The conclusions that you would reach looking at the log data are pretty much the same as the conclusions from the raw data.

The next example looks different (slide 7). Looking at the three calibrations, it looks as if you would reach the same conclusion on each one. The median is below zero. The interquartile range is overlapping 0. It looks like you have a statistical significance between these two instruments on this parameter too. Elgin argues that from a statistical point of view that even if the statistics came out saying that this is significant, it is not very convincing. Think about a situation where you have 3 coins in your hand.

You toss those in the air and let them land on the table. It would not be an unusual event to get 3 heads in a row. It should happen about 12.5% of the time. So even if this is statistically significant, there is not enough replication at the calibration level to reach a solid conclusion, because it's just as high of a probability that those three coin tosses might go the same way. It's widely recommended that if you have a random factor, like this calibration factor, that you should aim for at least five levels of that random factor. That will be discussed later. That is how Elgin would explain random effects and how he is looking at this data now, after years and years of looking at it through a simple model, which he now thinks is a flawed model.

Elgin moves into individual parameter results. On slide 8, Elgin shows graphs that were already previously shown. This is ammonium again. In the same table below the graphs, Elgin shows two effects he has looked at. One of them is the calibration effect, which is measured by the standard deviation among the calibrations, and then there is the p-value for that, that decides whether it's statistically significant, and then there is the instrument effect, which is the mean difference of the log transform data averaged across all of the calibrations. In this example, the conclusion is that it is not statistically significant. The error term that's used for this instrument effect size is compared to a standard error. That standard error is a weighted average of the standard deviation across the calibrations and the standard deviation of the sample variation within each calibration. That's part of this mixed effects analysis.

On the next slide, NO23 is graphed. Again, there is a significant calibration effect and a marginal significance on the instrument effect. This is the graph that was looked at earlier that was very significant across calibrations. In fact, it was so consistent that the analysis failed. It says this is so small that it can't be measured. This is one that should maybe be reconsidered and instead, use the original model without the mixed effects term. Again, with only three calibrations, maybe there isn't enough information to make a judgement that this calibration effect is too small to be measured.

For PO4, there was a similar result where the results across calibrations were very significant and the calibration term was not successfully estimated. When you look at the raw data plot, they look very consistent but in the log differences they look pretty different. This result is a little surprising. Even though it was a very small effect size, it looks statistically significant.

For SI, there were only two calibrations, but that was enough for the methodology to conclude that the calibration effect is significant and the instrument effect is pretty close to that 0.5 cut off.

Elgin's conclusions are that the calibration effects are real and it's something that should be made an account of when we do calibration studies, you need to be averaging over a number of calibrations (rule of thumb is five) to get an unbiased instrument comparison, and that these five calibrations seem like a lot of work. Elgin would like to know what the group thinks because he feels like this is a much bigger commitment than made in the past. From emails between Elgin, Suzanne, and Michael, a point from Suzanne was one thing that's getting lost in the DIWG is that these comparison studies are for CBP data analysts, so that if there are trends in the data, they can decipher whether they are due to an environmental change. That's why they do these studies. If there is an instrument effect, then there is a statistical methodology called intervention analysis where that can be adjusted. That's the value of those studies. Elgin thinks that Suzanne's point is valid. The people who build these instruments have already certified that they're giving accurate measurements. If that is trusted, are these studies necessary?

2:50 PM Discussion on Instrumentation Comparison Study Analysis Results

Jay Armstrong: I have thought about this in the past in a different way concerning these studies and the data that gleaned from them. On almost every occasion, we are comparing a newer technology. A lot of times we are comparing instruments that are 11-12 years old to newer technology with cleaner electronics, newer flow cell technology, newer ways to make glass, and improvements in fluidics. Sometimes I wonder if some of these differences are because we are comparing something old to something new.

- Scott Schroeder: I will comment on that because I've run the Latchet and there isn't clarity to me on that. The newer technology is actually flow injection and you're comparing segmented flow to flow injection. Yes, they're both colorimetric. I would have you think about how they are truly different technology, with the newer technology being flow injection.
- Jay Armstrong: Maybe I should rephrase, I don't disagree with what you're saying and the difference of technology. I might not have presented my thoughts correctly. Referring more to when we're comparing instruments, we're comparing something that we've bought that is new, all of its components are new, to an instrument that we have been utilizing over a long period of time. I wonder if that plays into anything as far as electrical noise. Just throwing ideas out there.
- Jerry Frank: To make that statement a little simpler maybe. A 2010 Honda Accord is not going to perform as well as a 2025 Honda Accord, how about that? Assuming that everything about the two cars is the same.
- Scott Schroeder: I absolutely agree on that, yes.

Elgin Perry: In my experience of doing these laboratory comparisons, I think I've only had one instance where I think we learned something really valuable. It turned out to have nothing to do with instruments. It was a lab comparison where it was a switch from one laboratory to another. The parameter that was problematic was suspended sediment, and I went through the data, looked at the effect size between the laboratories and plotted them on a map. When I plotted them I could see that all of the downstream stations had this effect and all of the upstream stations didn't. So I called Carolyn Keith at CBL and asked what would cause that. She says "oh, somebody's not rinsing all of the salt out of their samples." It was just a methodological issue, but the effect was big enough to be measurable. We were able to do some statistics and try to adjust for it, but I've been through a number of these studies and even if it's statistically significant, it's so small compared to spatial variability and environmental temporal variability that it won't cause a problem in our statistical assessments. There may be a legal requirement to do this, but where did it come from that these studies should be done?

- Durga Ghosh: From what I remember, this was built into the program when we started because, I think, the sentiment was that there needs to be a comparison. At the time, we didn't think about what else we could use to compare results, other than having side by side analysis. You bring up an important point about how valid these are or how useful they are, since in your experience of looking at the Bay Program data for the last 20 or so years, you have not found anything super significant when switches were made or instruments were phased out. This is an important point because moving forward, I think we need to reconsider what we have been including in our methodology. If we were to take it out, I'm still trying to grapple with what would be our median. Where are we going to draw the line, if we were to make significant changes, in order for us to see any trends? In your opinion, is there anything that precludes us from doing any of this and just making the switch? Or are there some rudimentary tests that we should do? I don't mean number wise because we barely ruled out having to run a certain number of samples, but do you think there are certain aspects of this we should consider when we make changes to our instrumentation?

- Elgin Perry: One thing you could do, is to consider having the number of calibrations be five or greater in your methods comparison studies. I don't know anything about how much work is involved in doing that but maybe somebody should speak to that. I assume you do a separate calibration for each parameter, right? I see you shook your head. It multiplies the amount of work that needs to be done to go up to five calibrations, but if you wanted to make a valid statistical comparison, that would be required. Maybe there are other ways to get there. I was thinking about it when Mike was presenting the inter-laboratory comparison data that each laboratory has done its own calibration right? That would be a measurement of this inter-calibration error. It's mixed in with procedural things that might change from one lab to another, but it might be that rather than doing a formal statistical analysis, like this mixed model analysis, that we would do, like we've done in the past, a limited number of calibrations and a large number of samples. If the differences are small compared to the inter-laboratory variants measured in other studies, then you'd say it's not a problem and not try to do a formal statistical test on the methods comparison study. When you make a change from the Latchet to the SEAL, for example, is it customary that you try to make that change in all laboratories that are participating in the Bay Program?

- Durga Ghosh: Not necessarily, because of the logistics involved with making the switch. There are times, such as this, when the instrument is being phased out and there is no support for the older instrument and you are forced to make the switch. However, there are times when the switches are made to get better efficiency and that is up to the lab. The Bay Program would just want to know what the switch was, if appropriate controls were run, and comparison studies were done, but it would not be mandated on the entire watershed. So yes, the changes are not mandated by us.

- Elgin Perry: Ok. If you had the situation where three laboratories were going to switch from the Latchet to the SEAL, if each laboratory were to do three calibrations, then we would end up with six degrees of freedom, two from each laboratory. That would give us a fairly legitimate test for that random effect, without everybody needing to do five. That could be employed to spread the burden of measuring this calibration effect. Otherwise, I don't know. This is something that needs to be taken into account and I wish I could go back now and look at the calibration effects for some of these studies where I've observed significance in the past. I think I even saw one study where there were two laboratories making the switch and at one laboratory the old to new difference was positive and the other was negative. They both were statistically significant. I am certain that it was just two different calibrations giving us two different results and had almost nothing to do with the instruments.
- Durga Ghosh: In my mind, I feel happier seeing an upward and downward trend. I fear when we see a similar trend everywhere. That worries me. Are we overlooking something because everyone is seeing a downward trend, for example? It is a lot easier if all of the labs are making the switch because you have a lot more data sets then that you can look at. We should reconsider some of these things as we move forward. I appreciate your input, Elgin. I can revisit our chapters where we talk about method modifications, see if we can pick your brain, and make changes to the document. I might reach out to you.

Elgin Perry: I appreciate you all for doing this study because this is something that has been concerning me for quite a while. It was nice to do and get some resolution to what was going on.

Cindy Johnson: Thank you, Suzanne and Heather, for supplying the data. I was thinking that we normally do our comparison studies over multiple seasons. It would seem to me that we would have three or four calibrations easily and if we included our split samples, that would be five.

- Suzanne Doughten: How we are doing it, Cindy, is when we know we are going to buy an instrument, we archive samples. These aren't current samples. Heather will go on two or three different days to get the paired results. We're not doing it on two instruments each time the samples come in.
- Cindy Johnson: You still did it over three days? You did calibrations for each of those three days, right?

- Suzanne Doughten: Yes, two or three. That’s what Elgin was talking about, but he wanted five, then it’s getting longer. One of the reasons we had three was that when we originally did it, we had too many below detection samples. Then we added a third day to get some higher samples. When you have below detection limits, what are you comparing?
- Cindy Johnson: So when you do your comparison studies, you do a range of low, medium, and high range samples?
- Suzanne Doughten: Most of our samples are low, so we try to get some of them in there so it’s hard to compare. We have high samples too. If we know we’re going to be buying, like this time we knew the Latchet was being discontinued, we’d hold onto high samples.
- Cindy Johnson: So would you normally run them on one day?
- Suzanne Doughten: No, at least two days because the protocol called for 100 paired samples over at least two days.
- Cindy Johnson: Well, Durga will have to kick it around, right?
- Suzanne Doughten: I think this was beyond my understanding.

Cindy Johnson: Thank you to everyone who has presented today. If you didn’t vote on where the next meeting is and whether it is in-person, please send your information to Allison (awelch@chesapeakebay.net). I imagine there is some travel involved so we could send out a doodle poll to see when the best dates are for travel. We really appreciate your time today. Thank you all.

3:00 PM Adjourn