TECHNICAL MEMORANDUM

TO: Ken Reimer, Chair of Independent Expert Advisory Committee (IEAC)

FROM: Iris Koch

SUBJECT: Comments on Memo Re "Clarification of the Effect of Measurement Error on Analyses of Methylmercury Concentrations in the Lower Churchill River"

DATE: February 8, 2018

CC: Marina Biasutti-Brown

The purpose of this technical memorandum is, at the request of Ken Reimer and the IEAC, to provide comments on a memorandum submitted to Peter Madden, Nalcor Energy, by Ryan Hill, Brian Pyper, and Randy Baker of Azimuth Consulting Group Partnership (December 20, 2017).

In an earlier memorandum (dated December 10, 2017), I estimated measurement uncertainty for water sampling results from the monitoring program being conducted in the Lower Churchill River and Lake Melville. The estimate was derived from a "Ten + Ten Study" reported in Azimuth 2017. The standard uncertainties were around 20% for both dissolved and total methylmercury for field sampling and laboratory subsampling and analysis, where the two sources of uncertainty were indistinguishable, indicating that the majority of the uncertainty was found at the laboratory subsampling and analysis stage. The expanded uncertainty was around 40%. The expanded uncertainty is the standard way of expressing measurement uncertainty and it provides a high confidence (95%) that the actual or true value is within the interval defined by it around the measured value.

The memorandum from Azimuth aimed to refute (at least in part) a concluding statement made by the IEAC from my earlier memorandum (Koch 2017). The (paraphrased) concluding statement made by IEAC was that differences between dissolved methylmercury at monitoring stations in the Lower Churchill River and the reference monitoring station were obscured by the measurement uncertainty, since *most* (my italics) of the values were within 40% of each other.

Azimuth argued that sample replication reduces the uncertainty and they mentioned a well-designed monitoring program. Presumably they are referring to a monthly sampling design where three or four time points per month were averaged (or the Bayesian equivalent) for assessment in Azimuth 2017. The IAEC will likely wish to examine exactly how the monitoring program is set up and the details of the power analysis (if any) that was used to produce the design. Azimuth also pointed out that superimposing

measurement uncertainty to monthly means was “misleading”. This superimposition was done in my December memorandum as an illustration whilst trying to keep the illustrative figure simple. This is why the comparison was also shown for individual values, for which *most* of the values were within the measurement uncertainty of each other.

To assume that all monitoring programs are always based on replication is somewhat of a generalization; for example drinking water checks and clinical tests are usually based on the collection and analysis of a single sample. Nevertheless, it is well recognized that collection of more than one sample is a crucial way of increasing the power of an experiment to discern differences in trends and population means. Ideally the estimate of variance that is used in a power analysis includes all sources of uncertainty, including measurement uncertainty. Others have seen that analytical measurement uncertainty was very small compared with uncertainty from spatial variability, which then dominated the variance value used in power analysis (Lovell et al. 2001).

In the present situation, we already know that the relative standard deviation (RSD) associated with sample collection – the best possible scenario of improving uncertainty in this way – is equivalent to approximately 20% when n = 10 (we know this from the “Ten + Ten” study). The 95% confidence interval (CI) around the mean, for 10 samples, is approximately 30% of the mean.

As an exercise in numbers, I randomly generated 20 standard deviations and RSDs for Station N1 field replicate data (from the “Ten + Ten” study), using n= 3, 5, 7 and of course 10. I calculated standard errors (SE = RSD/sqrt(n)) and confidence intervals (CI SE * Student’s t for n-1) from these values. (Note that the usual way of calculating SEs and CIs is to use the measured standard deviation; carrying out this calculation and then expressing the result relative to the average gives the same results as those shown below).

		n=3	n=3	n=5	n=5	n=7	n=7	n=10	n=10
		TMeHg	DMeHg	TMeHg	DMeHg	TMeHg	DMeHg	TMeHg	DMeHg
min	RSD	8.2	0.0	17.5	12.5	19.5	18.8		
max	RSD	41.6	33.8	32.3	29.5	27.4	24.6		
avg	RSD	23.5	19.7	24.0	21.0	25.2	22.1	24.0	20.6
min	SE	4.8	0.0	10.1	7.2	11.3	10.8		
max	SE	24.0	19.5	18.7	17.0	15.8	14.2		
avg	SE	13.6	11.4	13.9	12.1	14.5	12.8	13.9	11.9
t		4.3	4.3	2.78	2.78	2.45	2.45	2.26	2.26
min	CI	20.5	0.0	28.1	20.0	27.6	26.6		
max	CI	103.2	83.9	51.9	47.3	38.7	34.8		
avg	CI	58.3	48.9	38.5	33.7	35.6	31.3	31.3	26.9

TMeHg = total methylmercury; DMeHg = dissolved methylmercury; min = minimum RSD value; max = maximum RSD value; avg = mean RSD value; SE = standard error = standard deviation/sqrt(n); t = students t value 95% probability for degrees of freedom (n-1); CI = 95% confidence interval

It can be seen that the average 95% confidence interval (CI) generally decreases as n increases, but very small standard errors (SEs) and consequent reasonable CIs also occur when lower n's are use (e.g., the minimum values for n = 3 and n=5). If these were used for statistical analysis, it is possible that a discernible difference could be found because these % CIs and RSDs are small; their small size is probably because of accidental groupings, rather than their representativeness of the true mean and standard deviation (or a representative estimate of error around the mean). In other words, while it is understood that n>1 is required to ensure the ability to detect differences in population means with good power, it is possible that differences may be seen on the basis of measured standard deviations that are misleadingly small, and less than the true (or best estimate of true) standard deviation and error around the mean.

In this case a common sense approach is recommended, where practitioners question statistical differences if they are based on standard deviations that appear to be smaller than the estimated standard deviation: if the estimated standard deviation is based on the n = 10 value in the present case, RSDs smaller than approximately 20% would be suspect. This common sense approach is not a new idea; a similar idea is used in the Canada-Ontario Decision-Making Framework for Assessment of Great Lakes Contaminated Sediment when distinguishing site sediment concentrations from reference sediment concentrations: "Note that in cases where there is little discriminatory power in statistical significance determinations due to very low variability in the reference areas (i.e., a very small difference from reference would be statistically significant but of arguable environmental significance), an additional comparison is possible, specifically: are concentrations of COPC less than 20% above those same COPC in reference areas?" (Environment Canada and OMOE 2008). The U.S. EPA also recommends the use of common sense in interpretation of statistical information especially when statistical differences are seen with large n values: "While statistics provide a strong and essential tool for environmental decision-making, the science of statistics is not a substitute for common sense and can lead to bad decisions if not tempered with practicality" (U.S. EPA 2006).

Recall also that the numbers exercise above is based on the best possible scenario, where variability is caused only by sampling (and analytical variability that is nested within it). The usual way of assessing variability is to add the different sources, and temporal variability would perhaps add to the total variability in a set of results.

The debatable point is then which common sense value should be applied. As seen above, the Canada-Ontario Framework uses a value of 20%, which they rationalize on the basis of this value being well within the "bounds of typical analytical variability" and that it "is highly unlikely to be of any environmental concern"; moreover, +20% "is the same criterion as used for sediment toxicity test results comparisons" (Environment Canada and OMOE 2008). Therefore 20% is considered to be a valid value from both an analytical chemistry and biological effects point of view and it seems reasonable to suggest that an applied value should not be *below* 20%.

The value of 20% happens to be the measurement uncertainty standard deviation estimated for the current project. In the previous memorandum, a typical application of this value was provided, where four times this value is often used to distinguish between results (Bettencourt da Silva and Williams 2015). Although this comparison is often used to compare single measurements to each other, the basis of the derivation is for a comparison of means, with enough measurements taken that the student's t value is close to 3; the example given was for the 99% probability level (Bettencourt da Silva and Williams 2015).

In other settings the expanded uncertainty (two times the measurement uncertainty standard deviation) has been used to provide 95% confidence in discerning a difference (Badrick and Stewart 2014, Gron et al. 2007).

The numbers exercise above shows that for n between 3 and 5, average CIs may be expected to be around 35 to 50% (with both smaller and larger CIs, as mentioned); this happens to be similar to the estimated expanded measurement uncertainty of 40%. For N1, when relative CIs were calculated from measured data between October 2016 and October 2017 (monthly means with n ranging from 1 to 9), a range of 8–125% (% CI of the mean) was seen for DMeHg and 18–520% for TMeHg, with average % CI around 50% for both types of measurements. It is understood that Azimuth did not use confidence intervals or other forms of frequentist statistics for their statistical analysis (instead they used Bayesian statistics) but the comparison is illustrative of the point that a 40% common sense check is not inconsistent with error estimates seen in the results.

Generally it seems prudent that if a statistical difference is identified on the basis of values (or means) that are within the expanded measurement uncertainty (40%), and certainly within 20% of each other, the conclusions should be carefully considered.

Respectfully,

Iris Koch, Ph.D.
Environmental Sciences Group, Royal Military College of Canada, Kingston, Ontario

References

Azimuth 2017. Technical Memorandum RE: Relationship Between Muskrat Falls Reservoir Elevation and Mercury Concentrations, Lower Churchill River October 2016 – September 2017. Prepared for Peter Madden, Nalcor Energy. Azimuth Consulting Group Partnership, Vancouver, BC. 50 pages (including appendices).

Badrick T, Stewart P 2014. Understanding Measurement Uncertainty. Common Sense Pathology. https://www.rcpa.edu.au/getattachment/463e837a-6674-41ca-9284-19b5d1096ad3/Measurement-Uncertainty.aspx

Bettencourt da Silva R and Williams A (Eds), Eurachem/CITAC Guide: Setting and Using Target Uncertainty in Chemical Measurement, (1st ed. 2015). www.eurachem.org. Accessed November 2017.

Environment Canada and Ontario Ministry of the Environment (Environment Canada and OMOE) 2008. Canada-Ontario Decision-making Framework for Assessment of Great Lakes Contaminated Sediment. Prepared by Peter Chapman (Golder Associates Ltd.) with the COA Sediment Task Group on behalf of Environment Canada and the Ontario Ministry of the Environment under the Canada-Ontario Agreement (COA 2002). March 2008.

Grøn C, Hansen JB,Magnusson B, Nordbotten A, Krysell M, Andersen KJ, Lund U 2007. Uncertainty from sampling: A Nordtest Handbook for Sampling Planners on Sampling Quality Assurance and Uncertainty Estimation, NT Technical Report 604, Approved 2007-06.

Koch I 2017. Technical Memorandum: Measurement uncertainty associated with methylmercury concentrations in the Lower Churchill River and Lake Melville from October 2016 to November 2017COA. Prepared for Ken Reimer, Chair of Independent Expert Advisory Committee (IEAC).

U.S. EPA 2006. Data Quality Assessment: A Reviewer's Guide. EPA QA/G-9R. United States Office of Environmental Information Washington, DC. EPA/240/B-06/002. February 2006