Scientific and Technical Advisory Committee Peer Review of *Revised* James River Chlorophyll-*a* Criteria and Assessment

Lora Harris[1], Tom Fisher[2], Jim Hagy[3], Dong Liang[1], Martha Sutula[4]

[1]University of Maryland Center for Environmental Science – Chesapeake Biological Laboratory, [2]University of Maryland Center for Environmental Science – Horn Point Laboratory, [3]US EPA, [4]Southern California Coastal Water Research Project

STAC Review Report
August 2017

STAC Publication 17-006

About the Scientific and Technical Advisory Committee

The Scientific and Technical Advisory Committee (STAC) provides scientific and technical guidance to the Chesapeake Bay Program (CBP) on measures to restore and protect the Chesapeake Bay. Since its creation in December 1984, STAC has worked to enhance scientific communication and outreach throughout the Chesapeake Bay Watershed and beyond. STAC provides scientific and technical advice in various ways, including (1) technical reports and papers, (2) discussion groups, (3) assistance in organizing merit reviews of CBP programs and projects, (4) technical workshops, and (5) interaction between STAC members and the CBP. Through professional and academic contacts and organizational networks of its members, STAC ensures close cooperation among and between the various research institutions and management agencies represented in the Watershed. For additional information about STAC, please visit the STAC website at www.chesapeake.org/stac.

Publication Date: August 4, 2017

Publication Number: 17-006

Suggested Citation:

Harris, L., T. Fisher, J. Hagy, D. Liang, M. Sutula. 2017. Scientific and Technical Advisory Committee Peer Review of Revised James River Chlorophyll-a Criteria and Assessment. STAC Publication Number 17-006, Edgewater, MD. 19 pp.

The enclosed material represents the professional recommendations and expert opinion of individuals undertaking a workshop, review, forum, conference, or other activity on a topic or theme that STAC considered an important issue to the goals of the CBP. The content therefore reflects the views of the experts convened through the STAC-sponsored or co-sponsored activity.

STAC Administrative Support Provided by:

Chesapeake Research Consortium, Inc.
645 Contees Wharf Road
Edgewater, MD 21037
Telephone: 410-798-1283
Fax: 410-798-0816
http://www.chesapeake.org

Executive Summary

This independent peer review panel was originally convened by the Chesapeake Bay Program's Scientific and Technical Advisory Committee (STAC) in fall of 2016 to provide a scientific review of two primary documents outlining proposed assessment methodologies and chlorophyll-*a* criteria (Robertson et al. 2016 and VADEQ 2016). The review was previously approved by STAC and published (Harris et al. 2016). In response to that review, the Virginia Department of Environmental Quality (VADEQ) engaged in new analyses and developed a new approach which they documented in two completely new documents that are the subject of this second subsequent effort.

The panel feels encouraged that VADEQ earnestly considered the recommendations outlined in Harris et al. (2016), and agree that this new effort from VADEQ is on a path that will lead to scientifically defensible criteria and assessment methods for the James River. The panel had a consensus view, however, that there are still serious concerns that should be addressed. Some of these concerns are expressed in new comments specific to the new approaches taken by VADEQ; others are issues that were originally identified in Harris et al. (2016). Perhaps most importantly, the panel remains concerned that the assessment methodology and criteria derivation were treated in isolation and not engaged in as a complementary effort. Specifically, the assessment methodology should be determined first, and then used in a separate analysis to determine the chosen approach for data aggregation and calculation of criteria values as we suggested in Harris et al. (2016). The panel also agreed that a comparison of the existing assessment methodology with the new assessment methodology is warranted, and this type of comparative analysis was likewise suggested in the previous review.

As in the previous effort, the panel found the extent of co-mingling between policy decisions and the statistical analyses in the documentation to be inappropriate. The panel strongly encourages VADEQ to separate policy discussions into a new, separate section to clearly differentiate from the analytical work. For the assessment methodology, the panel had remaining concerns about interpolation and aggregation of the available data. Additionally, the panel had remaining questions regarding selection of a metric of central tendency, as thoroughly discussed in this report.

Highlights from Question Responses:

Question #1:

- VADEQ 2017 – Criteria intermingles analytical results and discussion with policy decisions. These should be clearly separated.
- In general, the VADEQ 2017 – Criteria document does not clearly provide rationale for the various approaches applied. This is true both generally (the panel suggest a conceptual model and flow-chart for decision making), as well as in more specific examples such as selection of the baseline approach and the chosen statistics.

- A disconnect remains between testing and using the new assessment approach in concert with the criteria development.

Question #2:

- Based on the state's description of potential aquatic life impacts, criteria should be designed to limit both the average condition and the frequency of episodes involving very high values as these relate to endpoints. The methods presented suggest that this has been recognized and is a goal of both the criteria development and assessment approach.
- Aggregation procedures involve interpolation, spatial medians, and seasonal geometric means. It is unclear how the well the frequency of extreme values that may cause aquatic life impacts is reflected in measures of central tendency after these steps. The sensitivity of the central tendency metric to changes in extreme values should be demonstrated.
- A measure of central tendency could be appropriate to use; but it is the view of the review panel that it is not sufficient. The log-normal distribution that is assumed by the assessment procedure is specified by two parameters: one for variability and one for central tendency. Thus, it is possible for variability in the assessment period to exceed variability of the baseline period while central tendency for the two periods remains the same. If variability increases, there will be an increase in the frequency of high chlorophyll events which are associated with designated use degradation. Therefore, we assert that assessing central tendency alone is not sufficient to ensure protection of aquatic life.
- A well-presented model of spatial and temporal variability (*e.g.*, hierarchical model outlines in response to Question #3) could be useful as a theoretical foundation for the criteria development and assessment procedures.

Question #3:

- The Assessment document (VADEQ 2017 – Assessment) presents an improved and more coherent assessment process compared to the 2016 document (Robertson 2016). This enables us to provide more specific comments.
- VADEQ 2017 – Assessment should consider both the central tendency and the spread of the chlorophyll-*a* distribution to maintain consistency with the Criteria document.
- Log-normality should be assumed throughout the document, which provides a theoretical background for using geometric mean. Any probability statement regarding the chlorophyll-*a* distribution is maintained under the log-normal distribution because logarithm is a monotonic transformation.
- A hierarchical analysis is suggested to estimate criteria exceedance probability while accounting for the spatiotemporal dependence and the changing seasonal distribution.

Introduction

The purpose of this report is to provide a scientific review of two new VADEQ documents that propose a) a new chlorophyll-*a* assessment protocol (VADEQ 2017**a**) and b) new chlorophyll-*a* criteria (VADEQ 2017**b**) for the James River, which are referred to in this report as "VADEQ 2017 – Assessment" and "VADEQ 2017 – Criteria", respectively. In developing the report, the panel responded to three charge questions issued by the U.S. EPA Chesapeake Bay Program Office:

1. Please comment on the criteria derivation approach that was used -- *i.e.*, on the approach based on first defining more recent chlorophyll-*a* concentrations by tidal river segment by spring and summer seasons and then applying an effects-based threshold to determine if a lower protective chlorophyll-*a* concentration was warranted.
2. Please comment on whether a measure of central tendency is appropriate to use when determining the protectiveness of chlorophyll-*a* concentrations from an aquatic life perspective. If the panel agrees that the measure is generally appropriate, please comment on the choice of a measure of central tendency used in the proposed criteria derivation methodology and criteria attainment assessment methodology. If the panel believes the measure is inappropriate, please comment on possibly more appropriate measures or approaches to use and any additional data needed to justify a final selection.
3. Please comment on the proposed criteria attainment assessment methodology and whether its application is consistent with how the proposed criteria were derived. Please also comment on whether application of this methodology would, in combination with the proposed criteria, provide a reasonable indication of whether or not conditions are protective of aquatic life. If you find fundamental flaws with the methodology, please comment on possible alternative assessment methodology(s) that could provide a more scientifically defensible indication of whether or not chlorophyll-*a* conditions are protective of aquatic life

A team of five reviewers previously tasked with reviewing two 2016 reports related to the state of Virginia's numeric water quality criteria for chlorophyll-*a* in the tidal portion of the James River (Harris et al. 2016), returned to take on this second review effort. The panel's process again centered on assigning a lead expert to each of the above three questions, with secondary reviewers contributing their responses before integrating the panel's findings together. Each reviewer contributed to at least two of the three questions based on their expertise; comments were also made on the content, organization, and structure of the new 2017 documentation. Responses to each question were collated in the final report, and consensus recommendations and conclusions were identified. The contents of this report have been reviewed and approved by the entire panel.

Question #1: Please comment on the criteria derivation approach that was used -- i.e., on the approach based on first defining more recent chlorophyll-a concentrations by tidal river segment by spring and summer seasons and then applying an effects-based threshold to determine if a lower protective chlorophyll-a concentration was warranted.

In general, the approach to link the determination of chlorophyll-*a* criteria to the risk of adverse outcomes is a reasonable approach that will allow VADEQ to provide a rationale for its importance in protecting designated uses. There were significant improvements in the application of this approach, such as determination of thresholds using statistically defensible methods; however, the panel found some of the choices in methodology questionable. Although the panel did not consider these issues to be fatal flaws, addressing them would make the document more defensible and justifiable. Specific comments with respect to document organization and the criteria derivation approach are provided below. More technical considerations are considered in the response to Question #3.

Document organization

Selection of numeric criteria is ultimately a policy decision, supported by scientific analyses. The VADEQ 2017 – Criteria document is confusing to read because it mixes policy decisions related to selecting criteria with language presenting the results of the scientific analyses that ultimately informs selection of the criteria. The panel strongly encourages reorganization of this document to separate presentation and discussion of the scientific analyses from selection of the criteria that are informed by those analyses. The outcomes of the scientific analyses should stand alone in the results section. The discussion should weigh the results of those analyses and related uncertainty to provide a strong rationale for criteria selection. To provide several examples of this co-mingling of science and policy, see page 34 of VADEQ 2017 – Criteria where the authors note that "the baseline summer means for … was deemed sufficiently protective", and at the bottom of the same paragraph that "no adjustments were made for enhanced protection against HABs" [harmful algal blooms]. These statements describe policy decisions that were informed by the scientific analysis. Conversely, an example of a scientific statement in the same paragraph is the comment that "the best model (Figure 3) was produced from the TF5.5A data, and it predicts that potentially harmful levels of microcystin … are associated with chlorophyll-*a* concentrations that are at or above 53 µg/l."

Criteria Derivation Approach

Although the methods provide a basic description of the approach, the panel found the description to be inadequate to thoroughly explain the rationale behind criteria derivations. Specific comments include:

1. The document assumes that the reader understands the rationale behind the "baseline approach." Although this is drawing on previous work, it is not clear how the baseline is protective of aquatic life, particularly if the ecosystem is already considered to be impaired.

One symptom of change is an elevated baseline, representing chronically elevated productivity. How does this baseline approach protect against that? The panel noted that the document does explain a general approach of using this baseline that acknowledges unknown risks, however, clarification and discussion of the baseline is warranted and the document will also benefit from greater separation of policy and analytical interpretation as noted above.

2. The document is lacking an overarching description of approach, that clarifies the various timescales used in relating toxic harmful algal blooms (HABs) and other adverse pathways to a seasonal chlorophyll-*a*. A conceptual figure would be ideal as a starting point for this. The document acknowledges the issues with aggregation of episodic phytoplankton blooms across time and space, but does not lay out the justification for how the chosen method bridges the gap from event-based phenomena to seasonally averaged criteria, in a manner that assures a low risk of impairment of uses. The document also does not provide a clear description or flow chart of decision-making of event-based threshold derivation to spatial or seasonally-aggregated means.
3. Within the methods, the document has a section on risk thresholds, but that section only discusses the frequency, not the magnitude of the risks. This section should clearly present a summary of thresholds (magnitude, duration, frequency) for each impairment pathway that was tested in the analyses, and the rationale for those thresholds (literature, etc.).
4. A very thorough description of data treatment should be provided as part of the methods. This should link the duration elements given in the risk thresholds section above to data treatment. A big picture issue remains in terms of the connection between the assessment methodology and the criteria assessment. The derivation of the criteria is really not feasible until an assessment methodology is fine-tuned and then used in the criteria development itself. They should be connected – especially in terms of how the values are aggregated, computed, and assessed, given the potential for aggregation to obscure low frequency, high values of chlorophyll-*a* .
5. The choice of statistical methods and aggregate measures used needs refinement and improved justification. Two particular issues were apparent:
    - While the panel agreed that a variety of regression approaches can be used, there is no justification as to 3why the selected model represents the best choice. Second, uncertainty in these event-based chlorophyll-*a* criteria should be quantified, and some explanation should be provided regarding how uncertainty is addressed in the calculation of seasonally averaged chlorophyll-*a*. Furthermore, it is not clear why a cumulative distribution function (CDF) is the best approach to translate event-based chlorophyll-*a* to seasonal or spatial chlorophyll-*a*.
    - The description of how the chosen statistic (*e.g.*, mean of the central tendency) is protective of designated uses should be improved. For example, quantile regression,

wherein selected 10th quantile could be justified as having low likelihood of occurrence. The text below from VADEQ 2017 - Criteria illustrates selection of confidence limits without providing justification for why they are appropriate. Regardless of the statistical approach eventually selected, noting what statistic is used and why it is valuable would be an improvement.

> "The upper 99% confidence limit was chosen to represent the baseline chlorophyll-*a* concentrations for JMSTFU, JMSTFL, and JMSOH to account for the greater measurement uncertainty for these segments, while a more restrictive statistic—the upper 95% confidence limit of the arithmetic mean of season-year estimates—was used to represent the baseline chlorophyll-*a* concentrations for JMSMH and JMSPH8. Downward adjustments were made to the baseline values whenever they were determined to provide inadequate protection of algal-related effects."

Question #2: Please comment on whether a measure of central tendency is appropriate to use when determining the protectiveness of chlorophyll-a concentrations from an aquatic life perspective. If the panel agrees that the measure is generally appropriate, please comment on the choice of a measure of central tendency used in the proposed criteria derivation methodology and criteria attainment assessment methodology. If the panel believes the measure is inappropriate, please comment on possibly more appropriate measures or approaches to use and any additional data needed to justify a final selection.

VADEQ 2017 – Criteria notes that the criteria "are expressed as seasonal means because they are intended to protect aquatic life from the negative effects of eutrophication that tend to occur over the scale of months". There is ample basis for the view that biotic responses to eutrophication are in many cases linked to seasonal means, rather than short term exposures. Hypoxia is possibly the best example, wherein average organic matter production tends to drive average rates of heterotrophy, which leads to hypoxia on a variety of temporal or spatial scales, possibly mediated on short time times by physical processes more than biological processes. Relationships between seagrass and water clarity may also reflect average exposures, since seagrasses can store and access carbon reserves as they respond to periods poor water clarity (the statement that "SAV only requires optimal water clarity 50% of the time over the course of the growing season" may be a generous oversimplification which requires a citation). These ideas are also explained in the Introduction on pages 28 and 29. VADEQ also notes that seasonal means are useful endpoints for water quality simulation modeling, which is often used to determine protective nutrient loading levels.

VADEQ 2017 – Criteria goes on, however, to argue that the main concerns due to chlorophyll-*a* in the tidal fresh James River are elevated risks associated with periods of high algal biomass and productivity, as indicated by chlorophyll-*a* values in the upper tail of the distribution. Such effects include increased risk of harmful algae (*e.g.*, *Microcystis aeruginosa* or *Cochlodinium* sp.), algal biotic integrity, and high pH resulting from high primary production. As is noted, some HABs may "cause extensive mortality after a single 96-hour exposure." This implies that key endpoints of concern are linked to episodes of particularly high biomass and associated production, as opposed to high average biomass and production. This logic largely guides the fundamental approach to developing the proposed chlorophyll-*a* criteria: (1) the criterion should have the effect of limiting the frequency at which chlorophyll-*a* reaches levels associated with impacts to aquatic life, (2) a measure of central tendency is associated with a known frequency of occurrence for higher values and therefore, (3) criteria may be set for the statistic that defines central tendency.

A further feature of the approach is the suggestion that unknown future risks can be limited by ensuring that chlorophyll-*a* does not increase from the present baseline ("While it is impossible to protect aquatic life from all potentialities, it is possible to hedge against some unknowns by

simply maintaining current conditions", VADEQ 2017 – Criteria). The baseline chlorophyll-*a* threshold, which defines the current condition, is based on data from 2005-2015 and is specific to season and segment. The need to further limit chlorophyll-*a* in any particular season/segment to prevent harmful biotic effects is evaluated independently via "dose-response" relationships and adjustments made as needed.

The panel was charged to address whether the approach of setting criteria for a measure of central tendency is "appropriate." In responding, the panel recognizes that there is a wealth of statistical theory and practical experience to suggest that measures of central tendency are a reliable indicator of the location of the majority of measurements. The same theory and practical experience also suggests that central tendency can be quantified with greater precision than an extreme quantile given any particular dataset because extreme values, by definition, occur infrequently. A greater number of observations are needed to define the tails of the distribution compared to the middle. If the distribution of observations follows a common probability distribution, such as the normal distribution, it may be possible to estimate the frequency of extreme values based on the known distribution. For normally or log-normally distributed values, the distribution is characterized by mean and standard deviation. Two additional parameters or "moments", skewness and kurtosis can be computed to quantify departures from these ideal distributions. Skewness is a measure of "symmetry," and kurtosis quantifies whether the distribution has more or less than the expected number of values in the tails. Often these are assumed to be zero, and the distribution is characterized by only the first two moments (*i.e.*, mean and standard deviation). In this sense, it is appropriate to define a chlorophyll-*a* threshold as a measure of central tendency. However, such an approach may not be entirely sufficient. The most important assumptions are (1) that the relationship between the measure of central tendency and the frequency of extreme observations is quantified adequately, and (2) that this relationship does not change from when the criteria are developed to when new data are assessed against it. With respect to (2), the state should verify that future distributions are unchanged from the assumed distributions. While the baseline data may suggest that the log-normal distribution is a satisfactory approximation, one cannot know what distribution the assessment period will bring. For example, the chlorophyll distribution could become bimodal, with an excess of high values. An analysis of the kurtosis could make this apparent, but a measure of central tendency alone could be insufficient.

VADEQ 2017 – Criteria acknowledges that it is important to quantify the statistical distribution of chlorophyll-*a* concentrations in addition to the central tendency, which they suggest was done by "exploiting the cumulative distribution function." This gives the impression that an empirical cumulative distribution function (or CDF) was used, which could potentially illustrate any possible relationship between measures of central tendency and the frequency of higher values. However, the distribution of the data appears instead to have been assumed or shown to follow a log-normal distribution and then any probabilities or high percentiles of interest were computed from the distribution. In particular, VADEQ 2017 – Criteria notes that spatial and temporal

variability was characterized "using standard deviations derived from interpolated Dataflow chlorophyll-*a* and daily-averaged continuous measurements of chlorophyll-*a*" and that these were used to construct CDFs. Although it is mentioned that log-transformation was used because the data were log-normally distributed, it is unclear which data were log transformed (*i.e.*, raw data, spatial or daily means) or "CDF parameters" (the specifics of which are unclear to the panel).

VADEQ 2017 – Criteria notes that variability occurs on a variety of spatial and temporal scales. How spatial and temporal variability are addressed has a significant bearing on the "appropriateness" of a threshold defined as central tendency, particularly if, as has been argued, the frequency of observations in the upper tails of the distribution is most associated with aquatic life impacts. Aggregation is a common approach to addressing spatial and temporal variability in the context of management thresholds, such as criteria, in significant part because the objective is often to make a binary determination (meets/does not meet) at the level of a "segment" or other management unit. Given the likely need for aggregation, a threshold based on central tendency is appropriate only if the aggregation methods maintain a relationship between the central tendency and the frequency of higher values. Conversely, a measure of central tendency will not be sufficient or appropriate as an indicator of upper chlorophyll-*a* percentiles if the means of aggregation obscure the existence of higher values or weaken the relationship between upper percentiles of the distribution and measures of central tendency. As has been noted previously, these relationships also must not change between development of the criteria and future assessment.

Evaluation of whether a measure of central tendency is appropriate is inhibited to some degree by a lack of clarity regarding methods of aggregation, including what was actually done and the underlying statistical model and rationale. The section of VADEQ 2017 – Criteria addressing "Baseline characterization" (page 31) provides an example of how the explanation of the approach lacks clarity. The section intends to describe how the distribution of chlorophyll-*a* during 2005-2015 was characterized. It is noted that weekly spring and summer Dataflow cruises were conducted in JMSMH and JMSPH and that this resulted in "much less year-to-year variation … in seasonal estimates compared to other segments." Since variance is not expected to depend on sample size, while standard errors of the mean are expected to decrease with more observations, one is left to guess at how to properly understand this statement. As the document continues, it appears that the lower variance estimate is used to justify selecting an extreme statistic (the "upper 99% confidence limit) or an unspecified value to represent the baseline for some segments. Given less data in others, "the upper 95% confidence limit of the arithmetic mean of season-year estimates" was used to represent the baseline. No rationale is presented for selecting an upper confidence limit, for selecting a particular significance level for the confidence limit, or for selecting arithmetic means for this particular step of aggregation. Although it is not specified, it may be relevant that aggregation by averaging (*i.e.*, taking sample

means) should result in normally distributed values, regardless of the distribution of the original data (*i.e.*, Central Limit Theorem).

Other sections of the document suggest that an inconsistent combination of statistics has been used in the course of aggregation. In the second paragraph on page 33, it is implied that spatial central tendency is determined via the median. Similarly, on page 3 of the proposed assessment procedure it is noted that samples from a Dataflow run are "averaged (median)." (Note: some define the term average to mean any measure of central tendency, including arithmetic and geometric means, median, and mode. Due to this ambiguity, it would be most useful to specify which statistic is used, while avoiding the ambiguity of the term "averaging"). The median is often used as a measure of central tendency in environmental data when the objective is to eliminate the influence of extreme values. For example, use of the median eliminates the effect on central tendency of data below detection limits (left-censoring), off-scale (right censoring), or extreme values. A geometric mean is then used to aggregate spatial medians over season-year. As has been noted, it seems possible that spatial medians may not have a log-normal distribution as do the original observations.

A clear and systematic description of aggregation procedures would aid considerably in understanding the relationship between extreme values and measures of central tendency. As has been described in our response to charge Question #3, a hierarchical approach to describing spatial and temporal variability could be useful. Alternatively, a simplified aggregation procedure with a demonstrated capacity for high chlorophyll-*a* events to be reflected in the ultimately assessed metric would be desirable. If the raw data were independently and identically distributed, a statistical test of whether the observed frequency of raw observations exceeded a threshold would be appropriate. This could be a simple binomial test with a null hypothesis that the exceedance frequency is < 10%. Such a simple procedure would not be rigorous because of the noted spatial and temporal dependence, as well as changing variability among seasons. However, the panel agreed that there is not necessarily great value in replacing a simple and transparent, but erroneous, procedure with a much more complex, conceptually opaque, procedure that may also hide variability in unknown and untested ways. The ideal procedure would provide a more transparent assurance that the frequency of high biomass events is limited sufficiently to prevent unacceptable risks to aquatic life.

Question 3: Please comment on the proposed criteria attainment assessment methodology and whether its application is consistent with how the proposed criteria were derived. Please also comment on whether application of this methodology would, in combination with the proposed criteria, provide a reasonable indication of whether or not conditions are protective of aquatic life. If you find fundamental flaws with the methodology, please comment on possible alternative assessment methodology(s) that could provide a more scientifically defensible indication of whether or not chlorophyll-a conditions are protective of aquatic life.

The current criteria and assessment documents present a more coherent process than that of earlier documents. Since the underlying science and models are clearer, the panel was able to offer more specific comments and address details that were too vague for comment in the earlier version. As a result, the volume of comments on this version is large, but the nature on the comments are specific and the issues can be addressed. With additional work, these documents provide a pathway to a scientifically credible James River Chorophyll-*a* Criteria and Assessment Procedure.

The Assessment document proposed an approach based on the geometric mean to examine the central tendency of the chlorophyll-*a* distribution; however, the assessment approach does not consider the spread of the chlorophyll-*a* distribution. As has been described above in the response to charge Question #2, it is important to quantify the statistical distribution of chlorophyll-*a* concentrations in addition to the central tendency. The assessment approach should consider the frequency of chlorophyll values in the assessment data that are above the harmful effect threshold. The rigor of statistical analysis can be improved by analyzing raw observations on a natural logarithmic scale, rather than interpolated or aggregated observations at the original scale. Care should be taken to align the Assessment and the Criteria documents with respect to methodology and spatial scales. A hierarchical spatial analysis based on the log-normal distribution is given in this section as a potential alternative Assessment tool.

Consistency between criteria and assessment methodology

The Criteria document acknowledges that it is important to consider the spread of the chlorophyll distribution in addition to the central tendency when setting the chlorophyll criteria. This is clear from the sections that propose using the CDF to check if the 10% quantile in the upper tail for the observed mean and standard deviation is less than the threshold for detrimental effects. However, VADEQ 2017 – Assessment does not reflect this understanding. If both central tendency and spread are important for setting criteria, then both central tendency and spread should be assessed to ensure that designated use is not impaired. VADEQ 2017 – Assessment makes the strong assumption that spread in the assessment period will be equal to spread in the baseline period. An approach is needed to assess whether this assumption holds.

VADEQ 2017 - Criteria considers the chlorophyll distribution at an instantaneous scale. This is clear from the section that develops spatial CDF at bi-weekly temporal scale and temporal CDF at continuous monitoring (*i.e.*, ConMon) stations. However, VADEQ 2017 - Assessment focuses on the scale defined by the fixed station sampling design. The misalignment between the scales of criteria development and assessment makes the probability statement in the criteria document difficult to interpret. The power analysis in VADEQ 2017 - Assessment suggests a potential bias in the central tendency estimate based on the fixed station data. Specifically, large error rates were observed in the tidal fresh region. This suggests that a multiscale approach is important for both criteria and assessment. The panel acknowledges the challenges of maintaining an intensive monitoring program, but an effort should be invested to make the best use of any available data in assessment.

General comments on the assessment methodology

The Assessment document utilizes several statistical tools such as geometric mean, standard deviation and power analysis to evaluate the robustness of the procedure. These statistical tools rely on assumptions, which should be documented and checked based on empirical data. Such model diagnostics would improve the rigor of the assessment methods and the defensibility of the approach. Recommendations for such diagnostic work are provided below.

In using the fixed station design, it is unclear if VADEQ intends to consider these stations as sentinel sites wherein measurements taken at these sites are considered random samples of the chlorophyll-*a* process. The power analysis partially checks this assumption in estimating central tendency and provides a surprising result. In the mesohaline region, the data from two fixed stations were as good as those from three more stations for estimating the central tendency. However, biased estimates were observed in tidal fresh segments based on the current fixed station design. Thus, fixed stations may serve as sentinel sites in some segments, but are not suitable in others. Additional effort should be invested to correct the possible bias through incorporating the intensive monitoring data. A hierarchical analysis would incorporate dynamic spatial trends and time through structured random effects. It also calibrates Data Flow, ConMon, and the fixed station data. Thus, inference can be made regarding the possible bias in using only the fixed station data. The panel details this approach in the section describing an alternative assessment methodology on page 17.

For a sample $x_i, i = 1, \ldots, n$ following a log-normal distribution, let $\hat{\mu} = n^{-1} \sum_{i=1}^{n} \ln(x_i)$ and $\hat{\sigma} = \sqrt{(n-1)^{-1} \sum_{i=1}^{n} (\ln(x_i) - \hat{\mu})^2}$. The sample standard deviation is $\sqrt{(\exp(2\hat{\mu} + \hat{\sigma}^2)[\exp(\hat{\sigma}^2) - 1]}$, and the geometric standard deviation is $\exp(\hat{\sigma})$. Due to the convexity of the exponential function and Jensen's inequality, the geometric standard deviation is always an underestimate of the standard deviation if the random variable follows a log-normal distribution. Given this negative bias of the geometric standard deviation, power analysis based on it would lead to over-estimation of accuracy. Instead, the power analysis should be conducted

at the natural log scale, since any probability statement is maintained under monotonic transformation.

VADEQ 2017 - Assessment acknowledges the spatial structure in the dataflow measurements and addresses the spatial pseudo-replication by interpolation the raw observations to grid cells within segments using Kriging. While Kriging is optimal for local estimation, it is not appropriate for estimating global central tendency, and is especially problematic for estimating the spread of the distribution because it results in excessive smoothing of the raw data. This is particularly true for highly skewed and non-normally distributed data. Instead of Kriging, variogram estimation should be performed by segment to quantify the strength of spatial autocorrelation. A generalized least squares approach should be used to estimate the central tendency and spread given the variogram estimates.

Specific comments on the Assessment method

Listed below are some specific consensus comments from the panel about the Assessment method, and proposals for means of addressing these in order to strengthen the statistical rigor of the approach and its documentation.

- From Page 2/2

"Distributions were compared using a Kruskal-Wallis test followed by a pairwise Wilcoxon Rank Sum test ($p<0.05$)."

The Kruskal-Wallis test and Wilcoxon Rank Sum Test are tests for central tendency, not tests for comparing distributions.

- From Page 3/3

"First, the clustering was deemed consistent if the average pseudo-R^2 of all the analyzed cruise data was at least 0.60. … Secondly, for the majority of the "high R^2" cruises, the percent difference between group medians was at least 100%."

The post-hoc analyses of the grouping results may require some justification. How were the thresholds of 0.60 and 100% determined?

- From Page 4/4

"When rolling up into a single "segment-wide" seasonal mean, each "zone-specific" seasonal mean would be weighted based on the aerial size of its respective zone."

Is it required by the water quality standard that assessments be done on a segment basis? If not, then would it not make more sense to do assessments at the level of zones? There is a real possibility here that a subzone with high chlorophyll is experiencing detrimental effects and this will get masked by estimating a segment wide weighted mean.

- From Page 5/5

"Interpolation estimates, rather than individual Dataflow observations, are to be averaged since interpolation smooths out the effect of any biased monitoring that may occur while the vessel is underway, such as when the vessel slows down to bring bloom samples shipboard."

While it is true that interpolation reduces bias due to oversampling in a region, but interpolation also reduces variability because it smooths the observed data.

"The proposed procedure would require that all observations taken within a grid cell (see Figure 7) be averaged (median)"

Local estimation usually requires estimates of variogram to ensure optimal interpolation. The proposed averaging approach assumes a specific type of variogram model. Justification should be made for this choice based on empirical variogram estimates from the observed data.

"The proposed procedure would limit the generation of estimates to only those cells containing at least one observation"

Is this not tantamount to assuming that the unmeasured cells have a violation rate similar to the measured cells and that measured cells are a random sample of the data frame containing all cells? Such restriction also leaves empty grid cells in the segments; how were these empty cells treated in the simulation study?

- From Page 6/6

"Using the mean standard deviation, the minimum sample sizes needed to generate central tendency estimates within a margin of error of 1 µg/l and 2 µg/l were determined for each segment."

Power analysis is usually based on Normal theory. Thus data should be analyzed in natural logarithm transformed unit given the positive skewed chlorophyll-*a* distribution. It is not clear whether the original scale or natural log scale was used in the power analysis. The geometric standard deviation reported suggests that log scale was used, but the margin of error reported suggests the original scale (µg/l). Consequently, the results in Table 2 and Table 3 are hard to reproduce and interpret.

"A statistical power of 80% was selected as the "tolerable" false negative error rate (e.g., for twenty out of every 100 assessments, a decision of "attainment" will be made when the segment is in nonattainment). The tolerable false positive error rate (alpha) was assumed to be 5% (e.g., for five out of every 100 assessments, a decision of "non-attainment" will be made when the segment is in attainment)."

It is a conventional "rule of thumb" in power studies to set the type II error rate at 0.20 and the type I error rate at 0.05. This rule of thumb is based on an assumption that a type I error

is more serious (*i.e.*, results in greater loss) than a type II error. In this case, a type II error results in harm to the environment while a type I error results in loss to stakeholders whose job it is to protect water quality. A rationale should be presented on why it is acceptable to have greater risk of harming the environment (20%) and lesser risk of being overprotective of the environment (5%). Finding non-attainment triggers "further study" and development of a plan to remedy the non-attainment, which may dilute management resources and focus away from real non-attainments. These costs should be more fairly considered and discussed. Furthermore, there is not a justification for a large difference in type I and type II error rates. Power analysis starts with the null hypothesis of attainment; would it be more defensible to assume non-attainment and find evidence against that?

- From Page 7/7 and Page 8/8

"For the former examination, it is assumed that the segment is always in non-compliance of a criterion that is equal to whatever the Dataflow seasonal mean is minus 2 µg/l, … For the latter examination, it is assumed that the segment is always in compliance of a criterion that is equal to whatever the Dataflow seasonal mean is plus 2 µg/l."

"A fixed station design enabling balanced false negative/false positive rates and a total error rate less than or equal to 30% is considered acceptable."

The 2 µg/l margin of error may have been derived from the power analysis, but no justification was given for this choice. Will the margin of error change between segments? How was the 30% choice determined? A Monte Carlo sample size of 50 is relatively small. Usually the Monte Carlo sample size is chosen so that numeric error is 5% of the standard deviation.

Alternative assessment methodology – Hierarchical analysis

VADEQ 2017 – Criteria recognizes both spatial and temporal variance of chlorophyll-*a* and presents separate CDF curves for these two variance components. However, chlorophyll-*a* is realized in a four dimensional domain with three dimensions for space and one for time. Since chlorophyll-*a* is realized in a temporal-spatial domain, it should be represented by a single model with temporal and spatial variance components. In the simple example of a unified model proposed here (see eq. 1 below), space is considered nested within time. In this model the variability over the three dimensions of space is assumed uniform so that space can be simplified to one dimension and observations in space and time are assumed independent. While these are very simplistic assumptions, this model does illustrate the concept of representing both spatial and temporal variance of chlorophyll-*a* so that the probability of exceeding the harmful effects threshold can be estimated from a single CDF curve.

The aggregation procedure through interpolation, geometric mean or standard deviation smoothed the individual measurements. The resulting probability statement would be optimistic.

The stepwise approach of treating spatial and temporal variation separately would also reduce the variance estimate. Consequently, the variance estimate does not provide a realistic representation of the chlorophyll-*a* distribution. A model-based approach assessment could be used to develop the spatiotemporal CDF in support of the single reference distribution.

There needs to be a conceptual model supporting this temporal-spatial structure. For example, a hierarchical structure where space is nested within month, which is then nested within season. With proper experimental design where samples are collected from the spatial-temporal lattice, it would be possible to obtain estimates for spatial variance components and temporal variance components from a single analysis using this hierarchical model. In practice, there are not sufficient data to do this combined analysis producing the spatial and temporal variance components in all segments. As a result, it may be necessary to approach estimation of the spatial and temporal variance components with separate analysis as was done for this report. None-the-less, it will be helpful to spell out the full model even if different parts are estimated by separate analyses.

Let y_{it} denote the observation at location i and time t. Throughout this section, log-normality is assumed and y represents the natural logarithm transformed chlorophyll-*a* concentration. This provides a theoretical foundation for the choice of statistics and avoids the issues with underestimation of the variability in geometric standard deviation. The proposed model could accommodate multiscale monitoring data. Data flow, ConMon, as well as the fixed station data could be fused in this model after suitable calibration. An example of such a model is as follows:

$$y_{it} = \mu + \theta_t + \phi_i + \delta_{it} + \epsilon_{it} \qquad (1)$$

In the equation above, μ denotes the overall mean, $\mu + \theta_t$ the season-year mean within each season-year segment, ϕ_i denote the static spatial mean, δ_{it} denote the dynamic spatial mean nested in time t, and ϵ_{it} denote the micro-scale variability. Under model (1), the random terms are assumed individual variance components where $\mathrm{Var}(\theta_t) = \sigma_\theta^2$, $\mathrm{Var}(\phi_i) = \sigma_\phi^2$, $\mathrm{Var}(\delta_{it}) = \sigma_\delta^2$ and $\mathrm{Var}(\epsilon_{it}) = \sigma_\epsilon^2$. The total variance of the process is $\mathrm{Var}(y_{it}) = \sigma_\theta^2 + \sigma_\phi^2 + \sigma_\delta^2 + \sigma_\epsilon^2$. This is the variance that should be used to construct the CDF that determines if the segment is failing the criteria and not protective from negative effects. It should be noted that the model does not contain an explicit specification of the spatial and temporal dependence, and model fitting based on the data-flow observation would be computationally challenging given the high dimensionality of the data. It is sufficient, however, to illustrate that temporal and spatial uncertainties are additive in determining the spread of the chlorophyll-*a* distribution.

Literature Cited

Harris, L., T. Fisher, J. Hagy, D. Liang, and M. Sutula. 2016. Scientific and Technical Advisory Committee Peer Review for the James River Chlorophyll-*a* Criteria Re-evaluation. STAC Publication Number 16-007, Edgewater, MD. 41 p.

Robertson, T. 2016. "Proposed Assessment Methodology for James River Chlorophyll Criteria". Virginia Department of Environmental Quality. 22 p.

VADEQ 2016. "Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary". A Report from the Science Advisory Panel for the James River Chlorophyll Criteria Study. April 14, 2016. 44 p.

VADEQ 2017**a**. "Proposed Assessment Methodology for James River Chlorophyll-a Criteria". Virginia Department of Environmental Quality. 35 p.

VADEQ 2017**b**. "Proposed Revisions to Numeric James River Chlorophyll-a Criteria". Virginia Department of Environmental Quality. 60 p.