Chesapeake Bay Program
A Watershed Partnership

410 Severn Avenue, Suite 109
Annapolis, MD 21403
Phone: 1 (800) YOUR BAY
Fax: (410) 267-5777
www.chesapeakebay.net

May 2, 2017

Dr. Lisa Wainger, Chair
Scientific and Technical Advisory Committee
Chesapeake Bay Program
645 Contees Wharf Road, P.O. Box 28
Edgewater, Maryland 21037

Dear Dr. Wainger,

Thank you for the opportunity to respond to the Scientific and Technical Advisory Committee's (STAC) report entitled "Scientific and Technical Advisory Committee Review of the Generalized Additive Model (GAM) Approach for Water Quality Trends in Tidal Waters." The STAC panel's review and approval of the use of GAMs for analyzing Chesapeake Bay water quality tidal trends is valuable at this critical time for understanding long term trends and water quality dynamics in Chesapeake Bay.

The following pages contain a detailed response to the STAC review report and suggestions. The primary users of this approach are the Bay Program's state agency partners. These team members are using the approach currently and, along with the review panel's suggestions, have guided our short- and near-term plans. In the short-term, we will be fully implementing the intervention approach for method changes and, as recommended by the panel, will make publically available our lists of major changes and results from analyzing them. We will also be continuing development of a flow-adjustment approach. Based on the review panel's report, this year we will also be compiling documentation of the tidal-trends GAM approach which will address most of the panel's comments, in particular:

- Identification of all defaults and selections to aid the user in applying the software either in an automated or hands-on way and what prior knowledge would be needed for a user;
- The percent change computation, which we agree requires more thorough description to explain the standard error calculation and justification;
- Description of uncertainty bounds and interpretation of component p-values.

Additional statistical work recommended by the panel such as residual diagnostics and continued improvement to the censored-date approach are important and slightly long-term tasks.

On behalf of the Management Board, thank you for the timely recommendations. Please extend the gratitude of the MB and STAR to STAC and the review panel for the time and effort involved in the review. We greatly appreciate the ongoing role of STAC in serving as an independent review body helping to improve the overall management of the Chesapeake Bay and watershed restoration efforts.

Sincerely,

Nicholas A. DiPasquale, Chair
Management Board

Attachment: Detailed Response to STAC Review of the GAM Approach
Ec: Management Board Members with Attachment

Detailed Response to: Scientific and Technical Advisory Committee Review of the Generalized Additive Model (GAM) Approach for Water Quality Trends in Tidal Waters April 2017

We appreciate the time and effort of the STAC review panel in their thorough review. Below we detail our response to the review report, in the order of the original charge questions. Our work on the tidal trends GAM method is moving forward, and we have used the comments from the panel to prioritize the development, as outlined with a summary list at the end of this response. Any questions on this response can be directed to Rebecca Murphy, UMCES at CBPO (rmurphy@chesapeakebay.net).

Response to Comments on Questions 1 to 10:

1. Please comment on whether the resource materials and references provided are adequate for conducting this review.

The reviewers agreed that the materials provided were sufficient for review. They suggest that further material be made available describing the Seasonal Kendall test previously used for tidal trend analysis. In the near-term, we will be expanding the report analyzed by the STAC panel to be documentation for this GAM-based approach, and many of the panel's recommendations, including this one, will shape the content of that documentation. The application of the Seasonal Kendall test for Chesapeake Bay tidal trends is described in a few places (Ebersole et al. 2002; Marshall et al. 2009). The test is a non-parametric test for a monotonic trend (Gilbert 1987). The strength of the approach for our data is that it does not require any particular distribution, but a weakness is that it is only designed to test for trends that do not change direction over time. With more than 30 years of data now, our team has observed many nonlinear trends and trends that change direction over time. Other reasons for exploration of alternative methods is that the Seasonal Kendall approach is not constructed for testing hypotheses for the factors affecting trends or flow-adjustment, two things that we want the flexibility to pursue in explaining tidal trends.

Regarding the reviewers' second set of comments about the description of GAMs in general and the name of the approach -- Although we agree that GAMs as a statistical technique is much broader than the single implementation we have selected, we do not think changing the term "GAMs" as a description of what we are doing to "Mixed GAM Computational Vehicle" would help clarify the issue. We found a response to a FAQ in which Simon Wood describes what 'mgcv' stands for. He writes that "Mixed GAM Computational Vehicle" is not ideal, but was selected to keep the acronym 'mgcv' once the functionality of the package changed substantially and he no longer could call it "Multiple GCV" (because it no longer only allowed GCV fitting) (https://rdrr.io/cran/mgcv/man/mgcv-FAQ.html). Furthermore, in the 'mgcv' package, there is a mixed model approach available, and we are not using that feature at this point. So including "mixed" in our name could lead to misunderstanding as to which functionality in mgcv we are using. We believe that keeping our name general, but then providing details of how we are implementing GAMs – both in the package documentation itself as well as any future publications will give us flexibility to continue developing package capability and be clear to our audience. We will clarify in future documentation the statement that implied to the reviewers that perhaps we think the GAM approach has converged on this method. Our audience is usually not familiar with the GAM technique, and we have used this statement to describe to members of groups such as

the WQGIT and management board that GAMs are not new technique that we just came up with, but have had a solid statistical foundation through years of research.

The reviewers also suggest results from some time-cost experiments to aid the first-time user with knowing how long an analysis should run. We appreciate this suggestion and will add it to our documentation.

2. Please comment on whether the choice to build the R package around the 'mgcv' package raises any concerns and on whether the functionality contained within the R package addresses the STAC MEOWQT workshop recommendation to implement a process for automating the GAM application.

The review team asked who the users of the package are and what level of knowledge is needed to run the analyses. The development of this approach has been an entirely collaborative effort between staff at the Chesapeake Bay Program Office and the MDDNR and VADEQ water quality analysts who annually conduct trend assessments for MD and VA tidal waters. The analysts who have been doing these analyses for years, plus their managers, have been part of monthly phone calls with the CBPO team (including UMCES, USGS, Tetra Tech consultant and our statistical consultant). The development has been a collaborative effort to meet the needs of the states in analyzing their water quality data, and the state agency analysts are our primary audience. To use the package, some statistical knowledge is needed, but not necessarily previous familiarity with GAMs.

The R package is automated to fit GAM models to multiple tidal monitoring stations in the same run. In the summer of 2016, it was run separately by MDDNR for 71 MD tidal stations for 10 parameters and ODU for VADEQ for 65 stations at a time. The need to analyze multiple stations and parameters is why we describe using the default 'mgcv' options in most cases. When the defaults have not been sufficient, we have conducted analyses to be able to provide default settings or model-selection within the analysis so that automation is possible. These choices and automations have all been made through multiple discussions, analyses, and example runs. In many cases these tests were done by the MD or VA team analysts, and so there is no "blind" automation of the program that has not been discussed with the primary users. The reviewers are correct that a thorough list of these choices should be available. This information is available with the help documentation for each function, but we will work on compiling a list of all model options and defaults in one place. In specific answer to the questions about lambda, thin plate regression splines, etc, all of these model options use the 'mgcv' defaults. Although we are exploring changing the basis dimensions (i.e., k-values) when the length of analysis goes beyond 15 years.

In terms of other users, we are very open to the scientific research community building on the approach. CBPO office staff have already used it in a few collaborative tidal trends synthesis projects, but we are also able to share the package with anyone who asks. The open-source nature of R allows for functions to be extracted and modified to a user's needs. This is the most likely avenue for the package to be used by research teams, at least in the near-term. We have plans to put the package on CRAN when budget allows.

3. Please comment on whether the three model options (i.e., gam0 – linear trend with seasonality; gam1 – nonlinear trend with seasonality; and gam2 – nonlinear trend with seasonality and interactions) for temporal analysis built into the R package are appropriate for representing the patterns over time and whether the conclusion to generally select the gam2 model is scientifically sound.

Regarding the three temporal models built into the current R package, all three models can easily be implemented in a single run. In fact, this is our default. The challenge is in interpretation of the results, and thus we are pleased the reviewers agree that gam2 would usually capture the overall trend in the data.

We appreciate the reviewers' comments suggesting to clarify the legend text. The abbreviations are needed to keep the graph compact in our output reports. However, we need a table defining each of the terms. This is:

Legend item	Description
Obs.	Observed data
GAM	Full GAM fit
Seas	Seasonally-adjusted long-term mean estimates
95% CI	95% confidence interval on the seasonally-adjusted mean estimates
1/1, 4/1, 7/1, 10/1	Mean model estimates for each of these 4 days of the year over time

The reviewers mention that our text about looking at the p-values on the interaction term in gam2 is contrary to our description that we would not rely on component-specific confidence bounds. That last statement in Section 2.2 is perhaps too strong and will not be stated like that in the future. A better way to phrase it would be that "we will acknowledge that component-specific confidence bounds are more uncertain." As mentioned earlier in that same paragraph, the p-values can be "useful for testing for inclusion or exclusion of model terms (Wood 2013)." We agree that a weight-of-evidence approach using the AIC of the models would be a wise approach as well.

The reviewers also commented that when cyear and s(cyear) are in the same model there will be difficulty distinguishing between the effects of the two. We have built this nested model structure in order to compare the whole models to each other building from gam0 to 1 to 2. Some additional detail that may be helpful is that the mgcv package resolves linear dependencies among groups of columns in the design matrix using a function called `fixDependence()` (https://rdrr.io/cran/mgcv/man/fixDependence.html). This function identifies and removes linear dependencies based on the QR algorithm. We are relying on this 'black-box' tool to resolve dependencies in the design matrix and make the model estimable. We appreciate that the nature of the constraints invoked to achieve this has an effect on the interpretation of the parameter estimates that are produced. Because of this, one might argue that we should endeavor to understand more completely what `fixDependence()` is doing. However, as is noted above, our goal is simply to compare the goodness of fit of gam0 to gam1 and, in cases where the simple linear model is adequate, rely on this simpler interpretive tool which results in more exact inference in that it does not use approximate degrees of freedom. As the package continues to be deployed bay-wide, if we consistently find that gam0 is rarely adequate, we will likely remove this hierarchical structure.

4. Please comment on whether the Percent Change calculation applied to the GAM results is sufficient for generating conclusions as to whether the long- or short-term trends are up or down.

The reviewers had concerns about our computation and/or the description of the percent change computation. We agree that clearer explanation is needed, and will take many of their suggestions on points to clarify. We do think that our computation of percent change is the best option available currently, and address the reviewers' suggestions below.

As suggested by the reviewers, we can flush out the description in the main text of documentation. Instead of our main text that is not specific on how the standard errors are computed, we will use the suggestion, modified slightly: "The estimate of percent change is a linear function of the parameter vector, which is in turn a linear function of the data. Thus its variance is readily available through a quadratic form defined by the covariance matrix of the GAM-estimates. This computation is documented in Appendix 2.4A, as well as in Wood (2006)."

It appears that further explanation is needed to clarify the computation and concerns that the degrees of freedom of the standard error calculations need to be reduced. Specifically, as mentioned above because our estimate of percent change is indeed a linear function of the data, our estimate of the SE of percent change also propagates from the raw data and not from the twelve monthly values which are actually an intermediate computation. The reviewers suggest an alternative to subtract out the seasonal trend from the observed data and perform a linear regression on the residuals. However, we would be concerned that with this approach the residuals are not independent and thus an important assumption of inference is not met.

Another concern raised in this section was that the interpretation of the uncertainty/confidence bounds of the output are not adequately addressed. The grey 95% CI shown in Figure 1 is computed from the standard errors of the mgcv predictions, which are described by Wood as being "based on the Bayesian posterior covariance matrix of the parameters." We describe the grey bounds as an estimate of the 95% confidence interval of the mean prediction. As mentioned by the reviewers, comparing confidence intervals for overlap is a frequently employed heuristic for assessing statistical significance. However, while non-overlapping confidence intervals does imply a statistically significant difference, overlapping confidence intervals does not imply a lack of statistical significance. It is generally considered more correct to construct a t-test for a difference based on the ratio of the difference estimate to its standard error. On the other hand, for comparing all possible predictions along a regression line, computing t-statistics is not practical. Thus confidence intervals are presented as a useful heuristic method of making comparisons of points along the time series.

A feature that has been observed for smoothing models such as GAMs and LOESS is that estimates of central tendency become unreliable at the edges of the independent variable space, a phenomenon known as "edge effects". Because of this feature, a direct comparison of the predictions at the beginning and end of the period of record, which represent the very edges of the independent variable 'cyear,' is not advisable. Other researchers (Robert Hirsch, USGS, personal communication for WRTDS) rely on similar stiffening the edge effect by averaging predictions beginning at the edge to continuing some distance toward the interior of the independent variable space.

The reviewers mention that gam.por.diff is hard-coded. Further clarification is needed on that -- we included code in Appendix 2.4A for that stand-alone version of the function in the interest of helping to understand how the computation is done. In the actual R package being used by our team, a percent

change function is called with options and is very flexible to the length of record and temporal scale. The user can choose to extend or decrease the number of years at the beginning and end of the period of record for computing percent change. Two years is the default. But in fact, it is possible in to use multiple start and end periods (say, 1, 2, 3, 4 or more years) at the beginning and end to compute percent change with various baselines from the same model and see if the conclusion about a change over time varies with these options. However, in response to the comment of computing percent change with less than 4 years of data, we do want to note that computing a trend analysis on a Chesapeake Bay tidal dataset with four years or less would not be advised. Generally for both tidal trends in the Chesapeake Bay region, we recommend no less than 8 years of data due to climatic variability. This is consistent with the nontidal trend analysis conducted by the USGS.

To specifically respond to the final 3 bullets:

- Why was the window length of two years chosen? Two years is somewhat arbitrary. The first thought was to simply compare the first and last year. However, because uncertainty tends to increase at the edges of the independent variable space, it is desirable to consider more than one year for stability. This prevents situations where adding one year of data to the period of record results in dramatic changes in estimates of change. Initially we chose three years as the averaging period. After some testing, we concluded this could be reduced to two with little change in the results.
- What makes this approach superior to using more traditional methods for comparing means (based on raw observations)? Traditional methods based on raw observations would not remove the impact of occasional high/low events that increase the variability of the raw data. We are interested in answering the question of whether the mean has increased or decreased over time. Using the raw data would result in the standard errors for comparing means to be inflated. It is not completely clear what is meant by "traditional methods", but we assume for discussion that the reviewers are suggesting something simple such as a t-test or a Wilcoxon rank sum test for comparing the first two years of data to the last two years of data. Such an analysis assumes that all of the variability within the two periods is random. We believe that our analysis show that this assumption is not tenable. Most dependent responses exhibit a deterministic seasonal pattern. Thus failing to account for this seasonal pattern will result in overestimation of stochastic variability and hence inflate the false negative rate of the test. Having reached this conclusion, one might then consider extending "traditional methods" to include stratifying the data by month to compare the first and last two years. This equates to using a monthly means model to remove the deterministic seasonal variation and thus conceptually equates to what we are doing. We use a model to adjust for deterministic effects when comparing the two time periods. By using this model-based estimate approach, if the model is expanded to include terms for deterministic effects of flow, salinity, or wind, it will be easy to adjust the time comparison for these effects as well. And just to reiterate, the approach we use does compute the difference estimate for the two periods as the raw data vector for the dependent variable multiplied by a matrix and is therefore a linear function of the raw data.
- Why should means, standard errors, difference tests, etc., not be performed in log space for log-transformed data? All means, standard errors, and difference tests are performed in log-space for log-transformed data. For simplification in interpreting the percent change results, the table presented in Section 2.4 (below also) includes a simple conversion of the baseline and current means only: exp(-2.2885) = 0.1014

Table 1. Draft Estimates of Change from 1999-2014 for TF5.5A, Surface TP

Calculation	Estimate
Baseline log mean (geometric mean)	-2.2885 (0.1014)
Current log mean (geometric mean)	-2.7503 (0.0639)
Estimated log difference	-0.4618
Std. Err. log difference	0.0758
95% Confidence interval for log difference	(-0.6104 , -0.3132)
Difference p-value	<0.0001
Period of Record Percent Change Estimate (%)	-36.99%

5. Please comment on whether the decision to derive conclusions based on log-transformed results without conducting any back-transformation are problematic for any of the conclusions we are trying to draw from the GAM model results.

As mentioned above, to clarify, for parameters analyzed with log-transformed data, all analyses were conducted on the log-transformed results. As we say in Section 2.4, "When the GAM model is fit to the log of the data values, this computation is conducted on the model predictions without back-transforming them, thereby providing an estimate of the percent change of the geometric mean." The decision to log-transform certain parameters is indeed based on the statistical distribution of the population of the data. It is also important to note that the user can select whether or not to log-transform any data set, the defaults we describe are based on what the Chesapeake Bay state analysts normally do (log transform cha, nutrients, and TSS), which is based on the common distributions of these data sets. Other distributions of data are possible if an analysis calls for it, although it has not been standard to use the suggested transformations for processing Secchi or DO.

For our analyses on log-transformed data, we are thus assuming that the untransformed data have a log-normal stochastic component so that the log-transformed data have a normal stochastic component. Owing to symmetry of the normal distribution, the mean and the median are the same parameter. Thus the mean in log-scale is also an estimate for the median in log-scale. Because the logarithm function is monotonic, the geometric mean which is obtained by exponentiating the log-mean is also an estimate of the median of the data in their observed scale. Thus by doing the analysis and hypothesis tests in the log-scale, we are testing hypotheses about the geometric mean or the median instead of tests on the arithmetic mean or the expected value of the observed data distribution. It is an important distinction to keep in mind. The geometric mean or median are frequently cited as preferred estimators in skewed distributions such as the log-normal because they de-emphasize the influence of outliers to the high side. On the other hand, when trying to estimate population totals, it can be important to use an estimate that reflects the expected value of an observation so that when this value is multiplied by population size, it reflects the estimate for the total population. This would be important for the example of estimating loads at the fall line. When mean concentration is multiplied by total flow to estimate total load, the mean estimate used needs to reflect the expected value of the population and thus the geometric mean and median would not be acceptable. Our goal is to assess trends. We believe the geometric mean/median estimator is acceptable for this purpose. In addition, transforming to a normal distribution, opens the door to a large arsenal of normal theory inference statistics that have a long established track record.

6. Please comment on whether the choice of the Maximum Likelihood method via the Expectation Maximization algorithm is a reasonable approach to account for censored data in the historical record as compared to other options, including the Monte Carlo sampling approach tested.

We appreciate the reviewers' thoughts on the implementation of the EM algorithm and additional citations. We have reviewed those references on the application of the EM algorithm in constrained maximum likelihood estimation. This approach is essentially what we are doing when we use EM with censored data to get maximum likelihood estimates and let 'mgcv' invoke constraints to insure smoothness. These references make it clear that the approach is acceptable in terms of getting the estimates, but there remains the problem of how to test hypotheses about these estimates. Right now we are essentially pretending that the expected values for censored data in the last iteration of the EM algorithm are actually observed values. Then we just let 'mgcv' give us the usual F-values and P-values. However, these are based on underestimates of variance because they assume specific values for the censored data when in fact there is additional uncertainty due to censored data being known only within an interval. For the purposes of identifying long-term patterns and trends in situations where censoring is modest, we think that this approximation is okay. In cases where censoring is extreme (>50% of the data), we already are truncating our data sets to avoid analyzing that highly censored part of the record.

We would like to incorporate an improvement to this approach in the future, although it appears that it will require some research-based guidance. The dissertation by Lu Wang (2010) is very promising in this regard, and we are hoping to reach out to researchers who may be interested in furthering study in this area and use any new developments in our application with GAMs in 'mgcv.'

In addition, in the response to this question there appears to be some misunderstanding of the censored data in our dataset, and further clarification will be given in future documentation. Specifically, censored data only exist in the CBP dataset before 1997. In 1997, the state agencies collecting the data agreed with CBP to provide below detection limit (BMDL) values which they had not provided previously. These BMDL values from 1997 onward are available to anyone who requests them, and are what the state agencies use in their databases for their trend analyses. Therefore, there will not be more left censored data in the future, although there may be more BMDL data. We appreciate the suggestions for additional analyses, but believe that our tests of artificially censoring complete data sets (Appendix 3.1C in original documentation) covered the major censoring conditions that we have in the CBP database.

7. Please comment on the importance of developing an intervention analysis approach to account for changes in lab and sampling methods, as opposed to implementing the adjustment factors approach used previously for these issues.

The panel suggests an extensive analysis comparing the intervention approach to adjustment factors, and a summary of all changes would be beneficial. We agree with the panel, and this work is underway currently in spring of 2017 by the state agency analysts on our team. They have developed extensive lists of all possible changes in the data sets, and are using the newest version of our R package to test these with the intervention approach. Once that work is complete, in the documentation of our findings, we will also look at some of the past adjustment factors and see how they compare to the intervention results. This work is underway currently and will impact our first set of 1985 to present trends generated in 2017.

The panel listed some disadvantages to the intervention analysis, listed below with our responses:

- An intervention analysis cannot be applied soon after a lab or method change. We understand this limitation, and luckily do not have many recent interventions. If major changes happen in the future, we will suggest the sampling program to conduct spilt sampling to help alleviate this problem.
- The possibility of method change co-occurring with regime shifts in the time series. We agree, this is a weakness. Our entire team is aware of this and we have been analyzing the intervention results with this reality in mind.
- P-values could be underestimated when there is autocorrelation in the residuals. We agree, and will also keep this in mind.
- Overfitting a short time series. We do not analyze time series less than 10 years long, but in addition, we will not keep possible step changes in the model that are shown not to be significant. Even so, we have identified cases where a series of lab changes with short intervals between the changes clearly results in overfitting even though the time series is long. In addition to the overfitting, there is the problem that both flow and lab administration tend to effect the response on annual intervals. As long as there are several flow cycles within a lab cycle, the two are identifiable. However when the lab cycles get short, the two are very much confounded. While we have identified this issue that was anticipated by the review committee, we have not resolved it. We do find that our graphical tools are helpful in identifying cases where this issue exists.
- Adding an intervention term is problematic if it is correlated with long-term changes in other environmental conditions. We agree, as with second bullet. This is a concern, and will be considered as we interpret results.

The Perry (2008) reference refers to a summary document submitted to MDDNR and CBP in 2008. This document is not available online, but we will include a reference that says "available upon request," and we can certainly share it with the review team if they are interested.

8. Please comment on the continuing research and development toward a comprehensive flow-adjustment procedure.

We appreciate the reviewers' positive comments about our work on flow-adjustment and agreement on the challenges. We are currently working on developing a flow/salinity adjustment technique addressing these challenges. The goal is to make it fairly automated, with defaults agreed upon by our team, but fully-adjustable for any individual analysis. This involves using salinity-adjustment in areas where there is a lot of mixing from multiple freshwater sources (e.g., mainstem) or where we have no freshwater monitoring (e.g., eastern shore tributaries), but to use river flow in the major tributaries where we do have that monitoring. When river flow is used, an informed automatic selection algorithm will be used that identifies an optimal number of preceding days of flow to average, within certain upper and lower bounds set based on published information about freshwater travel time within the tributaries.

9. At this time, are there any issues that you recommend the CBP investigate over the longer term (i.e. post-2017) regarding the application of GAMs for water quality trend analysis? The importance of these analyses for reporting changes in Chesapeake Bay water quality will continue to increase as we approach 2025 and beyond.

The reviewers' suggestions to allow seasonality to be turned on and off and analysis of annual averages are good suggestions. It is actually very easy to change the model structure in the package, and there would be no problem with removing s(doy) from a model and still using all the other functionality. We had worked in an earlier version on analyzing data that is collected only once per year, and agree that functionality should be included. Although it is technically possible to run data with one point per year currently in the package, we may want to check on some of the computation and graphics since they were all designed using data sets with multiple values per year. This is something we will work on soon.

Additional predictor variables are also good suggestions for using GAMs for hypothesis testing. Removing climatic variability with these variables should help to see the response to nutrient load changes. These suggestions all require research behind them to select the most appropriate explanatory variables (e.g., which wind direction and speed, from which monitoring stations). Hence in the near-term, any exploratory analysis with variables like these will be done as a more research-oriented effort than the operational-effort of using GAMs to quantify the temporal changes in water quality. In the future, we would hope to build additional functions into the package that would allow for this work. Until then, we will gladly share the package and our insights for any research efforts incorporating climatic variables beyond river flow.

The reviewers described some of the issues surrounding uncertainty of model components as "unresolved." We mentioned the uncertainty in the p-values for the components, and the 'mgcv' documentation says: "In general the p-values behave well, but neglecting smoothing parameter uncertainty means that they may be somewhat too low when smoothing parameters are highly uncertain. High uncertainty happens in particular when smoothing parameters are poorly identified, which can occur with nested smooths or highly correlated covariates (high concurvity)." Based on this, in evaluation of our model results, as long as we do not precisely think about a cutoff of something like $p<0.05$, but instead use the p-values on the individual components to understand the relative variability of the components, we should be fine. For this and other technical issues of applying 'mgcv,' we are relying heavily on the methods developed by the researchers that contribute to the 'mgcv'. Because this is a widely used package, if there are issues with their methods, we believe they will be discovered quickly and we will aim to stay on top of any developments.

We have been investigating concurvity between s(cyear) and s(flow) when we are testing modeling with flow, and realize this is a challenge. Sorting out the contributions of different model components in the presence of concurvity is an active area of research in the academic statistical community. We will continue to track this research and test methods in our application as they come available. Finally, application of AIC and use of the uncertainty bounds in interpreting the results is a work-in-progress by our team, and we will be more explicit in how we are interpreting the model uncertainty in our future documentation.

Regarding building on this analysis tool to aggregate results spatially and examine spatial correlations between stations, we appreciate the suggestions of packages that might help with this approach. This also could be an extension on package, and we will examine these ideas in the future.

Also we appreciate the point about residual diagnostics and likely autocorrelation of the residuals. Some of our original investigation into using GAMs involved using a mixed-model approach and modeling the residual autocorrelation. As the example provided by the reviewers suggests, we did not see a major change in the conclusion about long-term trends or patterns, and the functionality for mixed models provided some challenges for our implementation. Therefore, we put the analysis of residual autocorrelation aside for the time. This also will continue to be something we will keep in mind and consider in future work.

10. Are there other technical approaches that we should investigate that can supplement the GAM approach in order to identify and analyze the effects of management actions on water quality in the estuary?

We agree that a multiple-models approach is important, hence our comparison with WRTDS and working comparisons with Seasonal Kendall. We will continue to conduct these on a case-by-case basis, and will encourage and work with any researchers who want to compare their methods to this one. We will investigate the use of MARS for cases where we may want to be using two multiple response variables.

Summary

We are pleased with the reviewers' positive response to our GAM application to tidal water quality analysis in Chesapeake Bay and helpful suggestions. Below is a summary of the suggestions provided by the reviewers and our timeline for implementing them.

1. Recommendations we are currently working on (approximate 6 month time-frame):

    - Develop a comprehensive list of method and lab changes by location and parameter for each sampling program.
    - Test the method/lab-changes as interventions and begin to deal with related challenges such as multiple method changes in a short time period.
    - Continue flow-adjustment method development.

2. Recommendations that will be incorporated in the near-term (approximate one-year time-frame):

    - Documentation additions: Our near-term plans are to compile documentation on the currently in-development version of the package which includes the intervention method and accounts for censored data. This documentation will include much of what the STAC review team evaluated, and their suggestions including:
        - More details on the Seasonal Kendall technique being replaced,
        - Documenting how the GAM approach built into 'mgcv' compares to the general category of GAM analyses,
        - Time cost experiments,
        - Documentation of defaults and other options,
        - Description of abbreviations in the legend,
        - More thorough description of uncertainty bounds and interpretation of component p-values, and

    - More details explaining the percent change computation and computation of standard errors.
- Build flow-adjustment functionality into the R package.
- Publically available documentation of method and lab changes that are relevant to long-term trend analysis.
- Functionality to turn seasonality off and analyze once-per-year data.
- Continue sharing the package with research teams, and using GAMs as part of scientific research projects aimed towards hypothesis testing.

3. Recommendations that will be considered in future years:

- Putting the package on CRAN.
- Based on further research on the EM approach applied to GAMs, incorporate modifications to the algorithm that more accurately account for the uncertainty in the model.
- Address concurvity between smooth model components, as the statistical research develops on this topic.
- Additional functions built into the package that allow for more explanatory variables and a hypothesis-testing approach for research-purposes.
- Consider spatial correlations between stations in grouped-station analyses.
- Consider mixed model approach and residual autocorrelation.
- Encourage and participate in multiple-models approaches to statistical tidal trends analyses.

References

Ebersole, E. M. Lane, M. Olson, E. Perry, and B. Romano. 2002. Assumptions and Procedures for Calculating Water Quality Status and Trends In Tidal Waters of the Chesapeake Bay and its Tributaries: A Cumulative History. Prepared for the Tidal Monitoring and Analysis Workgroup, June. Online at: http://eyesonthebay.dnr.maryland.gov/eyesonthebay/documents/stat_trend_hist.pdf

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold: New York. 253 p.

Marshall, H.G., M.F. Lane, K.K. Nesius, and L. Burchardt. 2009. Assessment and significance of phytoplankton species composition within Chesapeake Bay and Virginia tributaries through a long-term monitoring program. *Environ Monit Assess* 150:143–155.

Wang, L. (2010). Cure Rate Model with Spline Estimated Components. Doctoral dissertation, Virginia Polytechnic Institute and State University.