*Response to reviewer comments, JAWRA-16-0152-P.R1, “Numerical and qualitative contrasts of two statistical models for water quality change in tidal waters”*

*Author responses are noted in italics below, original comments have been shortened for brevity.*

4] Delete the duplicate version of the SI figure: as it is found in the SI file [Appendix S2], it does not need to be [and should not be] submitted separately.   
  
ASSOCIATE EDITOR'S COMMENTS TO AUTHOR:   
  
Associate Editor   
Comments to the Author:   
Thank you for resubmitting "Numerical and qualitative contrasts of two statistical models for water quality change in tidal waters." The manuscript presents a useful comparison between two statistical techniques. The one detailed review suggests that two major issues need to be tackled before the manuscript can be re-considered for publication: (1) more rigorous assessment of predictive performance of the two modeling approaches and (2) clearer description of differences (and similarities) between the two statistical techniques - suitable for a JAWRA (i.e. scientific but not necessarily statistcial) audience. Reviewer 1 outlines these issues in detail and provides a thoughtful set of additional comments aimed at improving the manucript. Please note that if you chose to make these revisions, the manuscript will likely be sent for re-review by the original or by different reviewers. Thank you for your submission.   
  
*Thank you kindly for reviewing our manuscript. Out responses to the first reviewer below addresses the two main issues noted above.*   
  
REVIEWER(S)' COMMENTS TO AUTHOR:   
  
Reviewer: 1   
  
General comments:   
  
In the abstract, the authors state that the models were compared based on “predictive performance against the observed data” (page 2, line 20). By this, the authors mean the fit of the calibrated models to the observed data. However, I don’t think this method of determining “predictive performance” is appropriate, particularly for non-parametric models (or semi-parametric, in the case of WRTDS), which can be easily overfit to the observed data during calibration, if desired. The authors somewhat acknowledge this on page 25 (line 40), but overall the manuscript indicates that the RMSE of the calibrated model is a valid predictive performance metric. For example, the abstract discusses the “predictive abilities” of the models (page 2, line 34). In the study, overfitting is mitigated to some degree through cross-validation algorithms used during calibration, but these algorithms are presented as a black box, and it would be hard to prove equivalency between the two models. Therefore, the “predictive abilities” of these models can only be assessed through a thoughtful validation exercise (separate from calibration) reflecting the time scale over which predictions are desired and relevant. Validation is an important and expected component of any predictive modeling exercise, and this is particularly true for non-parametric models. If the authors wish to make statements about the predictive abilities of these models, then they need to test the models in a more rigorous way.   
  
The study also aims to compare the “statistical foundation of each model”, but I found this comparison to be somewhat lacking. The explanation of the GAM (page 10) relies on a lot of jargon that isn’t explained or referenced. I don’t think the intended audience of this article is familiar with “knots” or “spline basis”, for example, and more description would be useful. I also note that GAMs often make use of LOESS smoothing functions (as an alternative to splines); an advantage of GAMs is that they are neutral in terms of which smoothing function to apply. For example, see Faraway, J. J. (2016). Extending the linear model with R, among others. So the contrast between GAMs and WRTDS at page 12 (line 13) is not so compelling. Also, it seems both models are “additive” in that they are summing up the different smooth components. So, perhaps the differences in “statistical foundation” between GAMs and WRTDS are more subtle than the authors suggest? I expect there are important differences between GAMs and WRTDS, but the comparison may need to be revised, and should rely less on jargon, given the intended audience.   
  
Specific comments:   
  
Page 10, Line 44. GAM “parameters” are mentioned here, but the nature of these parameters needs to be clearly described. What parameters, besides the smoothing parameter, are included in a GAM model?

*This statement is somewhat misleading as ‘parameters’ in the GAM context are not as straightforward as compared to WRTDS (e.g.., GAM smooths are multi-parameter polynomials). We have edited the sentence for clarity (‘Predictions with GAMs are straightforward to obtain after the model is fit and can be obtained…’) and hope that some of the ambiguities of the GAM structure have been more clearly described in our response to the general comments above.*   
  
Page 11, Line 23. This section indicates that WRTDS is based on a “single set” of model parameters. But as described elsewhere, there is a unique set of “parameters” for each prediction point. Revise to clarify. 

*Yes, WRTDS includes multiple parameter sets for each point in the time series. This statement was meant to contrast the functional model (i.e. simple regression) used at each time step with the functional model used by GAMs. A follow-up statement was added for clarity: ‘This simple regression is used at each time step with different weights such that a combined parameter set equal in length to the time series is created.’*

Page 14, Line 23. A comparison between two model outputs is not really an error. I suggest calling this something else, like root mean square difference (RMSD). 

*Equation 3 was changed, all text updated accordingly.*

Page 14, Line 37. I’d recommend dividing by [½ the sum of GAM predictions plus ½ the sum of WRTDS predictions]. This would avoid any irregularities associated with arbitrarily choosing one or the other model to average over.

*Good suggestion, we modified the equation and updated the table 5 with the new results. Note that the new values were only slightly different from the original and the conclusions have not changed.*   
  
Pages 16-17. The description of the pseudo data generation is hard to follow. I recommend adding a flow chart or outline to help guide the reader through it. 

*An additional figure was added to Appendix B (Figure B1) that describes this workflow.*

Page 19, Line 13. Clarify what variables these half-window widths apply to.

*Sentence was revised: ‘…0.25 as a proportion of each year (seasonal component, sinusoidal terms in eq. (1)), 13.59 years (T in eq. (1)), and 0.25 as a proportion of the total range of salinity (Sal in eq. (1)) for LE1.2…’*  
  
Page 19, Line 20. “seasonal (annual proportion)” is unclear.

*Changed the sentence as follows: ‘…minimizing the seasonal and flow components’. An earlier sentence was also modified for clarity (line 13: ‘…as proportion of each year (seasonal component, sinusoidal terms in eq. (1))…’).*  
Page 19, line 27-30. This is jargon-heavy. And again, I’m not sure what is meant by “parameters” in the context of a spline-based GAM model. Do the authors mean “variables”? 

*This sentence was revised: ‘The smoothing method used for the GAMs does not split the degrees of freedom among the three interacting variables’.*

Page 24, lines 42-45. The “suggestion that GAMS are not separating the effect of flow and time” may not be obvious to readers. Explain. 

*The following sentence was added as a follow-up for qualification: ‘Specifically, results for WRTDS with no influence and a constant influence of flow showed less variation than GAMs in the relationship between chlorophyll and flow over time, consistent with the empirical relationships used to create the simulated time series’*

Page 26, line 30. I understand the authors’ point here, but I think it’s a bit extreme to say that conventional modeling approaches “mold the data to the model”. In conventional regression, the model is still fit to the data. Suggest rewording. 

*This statement was revised: ‘Conventional modelling techniques have been described as ‘statistical straightjackets’ that can inadequately characterize variation in the data with a limited parameter space and structural constraints.’*

Page 26, line 34. I don’t understand how GAMs could be considered “over-constrained”. Splines can be very flexible. More explanation is required to justify this assertion. 

*This statement was meant to help the reader understand potential differences between the models based on different structural components. We attempted this distinction by contrasting the multi-parameter space of WRTDS with the one smooth/one variable approach used by GAMs. The text was revised to make this clearer: ‘WRTDS is meant to provide a contrasting approach where the data mold the results using multiple parameter sets. In contrast, one might expect GAMs to be over-constrained by following a potentially less flexible model composed of one smoothing function per explanatory variable. However, the results do not provide a compelling numeric contrast between GAMs and WRTDS, despite the alternative statistical foundations. Both models are extremely flexible through fine control of window widths for WRTDS and degree of smoothing in GAMS, although at the cost of losing generality with increased precision.’*

Page 26, line 44. I don’t think “theories” is the right word here, as if statistical theories were developed specifically to describe water quality in the Patuxent River Estuary. Suggest revising. 

*This sentence was simplified for clarity, ‘Similarity in results for WRTDS and GAMs may suggest that relationships between time, season, and flow in the Patuxent were adequately described by each approach…’*

Figure 5: This figure could probably go in supporting information (at least most of it). 

*Agreed, figure 5 was placed in the supplementary material in Appendix C.*

Figure 7: “no flow” category name is confusing and inconsistent with text. Also, x-axis numbers are wrong in either the top or bottom panels, I think.

*The facet label was changed to ‘no influence’ and the x-axis numbers were corrected.*  
  
Reviewer: 2 

*We thank the second reviewer for reviewing our manuscript. We have shortened some of the figure and table captions in response to the comment.*