*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. 

*We agree that this is a critical component of the analysis that was missing. However, WRTDS is meant more for description of historical records and not for prediction or extrapolation, so we present the addition below with these limitations in mind. The added analyses evaluate prediction performance on validation datasets to determine the effects of changing the amount and nature of missing data relative to different training and validation datasets. We think this information is useful to better appreciate differences between the models.*

*The following was added to the methods:*

*‘The final analysis provided a complementary comparison to those described above for model performance by evaluating RMSE on independent datasets. Prediction performance was evaluated for validation datasets to provide a measure that was completely independent of the data used to train the models. Although WRTDS is not meant for prediction or extrapolation, this analysis provided an approximation of the ability of both models to predict missing data in different contexts.*

*This analysis used the daily simulated time series with a constant flow effect that was described in the previous section. Weekly samples at a fixed interval from the daily time series were used to ensure sufficient data (i.e., not monthly) were used to train the models while also minimizing processing time (i.e., not daily). The weekly time series was split into different training and validation datasets to evaluate effects of 1) different ratios of training to validation (1:1, 2:1, etc.), and 2) characteristics of the missing data. We developed a sampling algorithm to separate the weekly time series into different ratios and block-sampling schemes, which was accomplished using repeated sampling of the complete time series with a given block size until the desired split ratio was achieved. RMSE of model predictions for GAMs and WRTDS were evaluated for split ratios of 5-50% at 5% increments (e.g., validation was 5% and training was 95% of total, validation 10% and training 90%, etc.) and sampling from completely random to blocks of increasing size up to a block equal in size to the desired split ratio. Block sampling, in addition to completely random sampling, was necessary to account for temporal correlation, i.e., missing data are more likely to occur in blocks of time rather than single observations missing at random. Because the data splits and blocks were stochastic, 1000 replicates were created for each split ratio and block sampling level to place a range on model performance.’*

*The following was added to the results, including Figure 7:*

*‘Both models performed similarly for the training datasets based on different splits of the weekly simulated data. Overall, median RMSE values decreased slightly as the ratio between training and validation datasets increased (5% to 50% validation), although the range of RMSE values increased. Sampling characteristics for the validation datasets (random and block samples) did not have a noticeable effect on training RMSE for either model. Overall, WRTDS had a slightly lower RMSE for all training datasets compared to GAMs (median RMSE 0.51 for WRTDS, 0.52 for GAMs). For the validation datasets, GAMs generally had similar RMSE values for all datasets (median RMSE 0.55 for all), whereas WRTDS prediction performance varied considerably. RMSE values for WRTDS with randomly sampled validation datasets were similar to all validation datasets for GAMs and did not vary as a function of the split ratio (median RMSE ~ 0.55 for both). However, RMSE values increased dramatically for WRTDS as the sampling block size and split ratio increased (median RMSE 0.62, 0.88, and 0.88 for blocks sizes of 10%, 50%, and 100%).’*

*The following was added to the discussion:*

*‘Although our results generally indicated that comparable information was provided by both models, our comparisons of prediction errors using validation data from simulated time series highlighted an important difference. The results suggested that GAMs predict observations in the independent with a much higher precision than WRTDS, particularly when the missing the data are in blocks as is common with time series data. Although these results are compelling, the differences must be considered relative to what each model is meant to provide. WRTDS was developed as a descriptive method for historical data such that prediction beyond the range of data used to fit the model, including forecasting, was never an intended use. This application relates directly to the statistical foundation of the model such that trend descriptions are driven by the data used to fit the model. WRTDS results are stored as a prediction surface that relates the response across seasonal, annual, and flow ranges in the data. This surface represents numerous regression models fit to the observed data such that no ‘universal’ model exists, as compared to GAMs that create a parameterized polynomial model in three-dimensional space. As such, it is expected that GAMs will predict novel data well if the validation dataset has the same characteristics as the training data, whereas WRTDS results become less precise the farther validation data are form the training data in either of the three dimensions (time, season, and flow). This does not represent a methodological flaw; rather it represent differences in results that are indirectly caused by differences in intended applications. Regardless, our initial results suggest that the use of GAMs for prediction or extrapolation in water quality time series could be promising, whereas WRTDS models should be used to describe historical trends for which they were intended, keeping in mind the characteristics of missing data in the time series.’*

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. 

*We agree that our previous descriptions were not sufficiently explained for an ecological or water quality audience. We have revised our description in the methods to provide a more general interpretation of the methods as relevant to trend analysis and interpretation:*

*Page 10, first paragraph was modified as follows (starting on line 18): ‘…Multiple types of smooth functions could be used in a GAM (Hastie 1990), and our implementation relies on thin plate regression splines (Wood 2006a). A spline is a piece-wise function (e.g., a polynomial) whose pieces are connected at knots, or breakpoints, where the functions are joined smoothly (Hastie 1990). The thin plate regression spline has the benefit that a user is not required to select knot locations for a spline explicitly, but only selects a reasonable upper limit on the flexibility of the function. Within that limit, the balance between model fit and smoothness is achieved by fitting a smoothness parameter that minimizes the generalized cross-validation score (Wood 2006a). To allow for interaction between the model covariates (e.g., seasonal differences in the long-term chl-a pattern), a tensor product basis (Wood 2006b) between all three covariates was constructed, which allows for the model covariates to interact (e.g., the long-term patterns to vary by season).’*

*Page 12, text in the final paragraph of the “Methodological Contrasts” was also modified (starting on line 13): ‘…By contrast, this implementation of GAMs estimates the smoothing functions for the explanatory variables using a spline-fitting process that results in single (although quite complicated) spline functions fit across the entire data set for each explanatory variable. Although parallels between GAM fits can be made with both LOESS and WRTDS, the relationship between response and explanatory variables described by the hyper-dimensional smoothing surface from WRTDS is a different theoretical approach than a set of spline functions fit across all the data with GAMs.’*

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.*