Response to Reviewer

Marcus Beck, beck.marcus@epa.gov, Rebecca Murphy, rmurphy@chesapeakebay.net

24 March, 2016

The following is our response to reviewer comments on our manuscript "Comparison of weighted regression and additive models for trend evaluation of water quality in tidal waters" to be submitted to the Journal of the American Water Resources Association (JAWRA). The review was provied by Dr. Jeffrey Chanat (USGS). Comments are shown (edited for brevity) followed by our response in italics. All line numbers refer to the original manuscript.

To Marcus and Rebecca: Re-visiting the work after having set aside for a while, you would probably see ways to improve clarity and presentation without any review comments at all. Most of my annotations in the document are entered to call easily-fixable issues of wording/presentation to your attention, under the assumption that you intend to submit the ms in something close to its present form; I've tried to pull out and summarize some of the "must-dos" below. (Since I'm a non-tidal WRTDS user who is also interested in GAMs, you'll also find some "commentary" annotations not directly related to revision).

My only comment on the work that even approaches criticism is that it tends, at least in the document body, to present itself as both a methods paper and an interpretive paper in parallel. But I struggle to come up with a useful suggestion as to how to remedy that. Key features of the data for these sites are in some ways your motivation and basis for the model comparison. To that end my "knee-jerk" suggestion would be to talk about the data themselves first, point out the key features that drive the need for advanced modeling techniques, and then present and compare the models just as you have, focusing in particular at how they handle the most challenging features. Problem is, it's hard to describe, or even see the most problematic features without applying the models. Sheesh! And if you looked at another pair of sites after you finished the paper, you would likely see some different features on which you would like to run a comparison. I guess the only thing I can say is that I can personally identify with that plight, having brought WRTDS into an "operational" mode using data that's full of real-world problems. I guess in summary your choices are to gather comparative data across a wide range of conditions, and sometime "later" publish a model comparison on the most frequently-occurring "pathologies", much as Bob Hirsch did with his "flux bias" paper, or proceed with a submission such as you have here, qualifying it as the "first word, not the last". Only one practical suggestion on this matter: try taking out any interpretive content/hypotheses that isn't essential to the model comparison, just as an experiment, set it aside, and see if it improves readability/focus. You can always put it back in if it doesn't. If the paper seems too dry without that content, consider looking further into the areas where the models really differed (I'm thinking of the two areas I mentioned at the end of the first paragraph above), and try expanding your discussion of them – I found them to be fascinating.

Major points: from document body annotations, in roughly decreasing order of perceived need for revision:

- Throughout, there are instances where clarifying whether a result is reported, or a comparison performed in log or arithmetic space are needed.

We have clarified that all results or comparisons are reported in log-space (see comment below), except for a few instances in the results section

- Clarify whether and how you applied a bias correction factor when back-transforming GAM results from log to arithmetic space.

Data were not back-transformed to arithmetic space except for a few instances in the text that are explicitly noted. Although methods for back-transforming WRTDS results are available, they have not been implemented for GAMs. The following was added to line 153 for clarification: "Mean models require an estimation of the back-transformation bias parameter for response variables in log-space (Hirsch et al. 2010). Although

back-transformation is developed for WRTDS, a similar approach has not yet been implemented for GAMs. All units for chl-a are reported in log-space unless otherwise noted."

- For regression comparison, state which model result was the "predictor" and which was the "response." Consider re-doing this analysis with results in arithmetic space instead of log.

Text was added to clarify which model was the "predictor" and which was the "response" on line 259: "Results between models were also evaluated using regressions comparing WRTDS (as the response) and GAMs (as the predictor)." Text was also added to line 264 to better describe inerpretation of the results from the regression comparisons: "Although the signs of the slope and intercept estimates for the comparisons depended on which model was used as the predictor, we were primarily concerned with magnitude of the parameter estimates in the regression comparisons as evidence of systematic differences between each model."

For reasons noted above, we have not done the analysis in arithmetic space.

- Expand and clarify your description of methods for creating the synthetic data sets. See specific in-line comments/questions.

See comments below, we have also moved a portion of these methods to a supplementary appendix.

- I think you can pare down the number of figures; see particularly comments for 7, 8, and 9 in Results.

Figure 7 was moved to supplementary material, Figure 9 was removed. See response to comments below.

- Run a spell-checker (You did this whole document in R, no? At some point I want a lesson. . .)

In-line comments from the text:

Line 9: how about "prediction performance against"?

Line 11: consider "average between-model differences were small".

Line 14: consider ". . . both models predicting a roughly 65 percent increase in chl-a concentration over the period of record. . . " (from top panels of fig 3b: > (exp(2.5)-exp(2))/exp(2) [1] 0.6487213)

Line 15: consider: ". . . a more dynamic pattern, with a nearly-100 percent increase in chl-a over the most recent 10 years. . . " from bottom panel of 3b: > (exp(2.7)-exp(2))/exp(2) [1] 1.013753

Line 16: I am assuming you're talking about the discussion of fig 6 in lines 609-621? If so, consider turning this sentence around so that it flows more naturally from preceding discussion of model fit to observed data. Maybe "Comparison of flow-normalized trends estimated from observed data suggested that GAM results were less sensitive to periods with sparse observations, although both models had comparable abilities to remove flow effects from simulated time series of chl-a."

Line 75: Most of your graphs seem to span the range 1986-2014 inclusive, so wouldn't that be 2014-1986+1=29 years. In any case, why not be specific, e.g. "Two time series of monthly observations, spanning the years 1986-2014, from two stations in the Pax R. estuary are used as a common dataset. . . "?

Line 81: Good statement of objectives, and Introduction in general, but see cover comments about scope of report.

Line 144: I prefer "response variable"

Line 197: Presumably these were used as bias-correction factors in back-transformation from log to arithmetic space? Comparable treatment of this issue between the two models will be important for some of the comparisons you perform, so I suggest you establish that here.

Line 199: consider different grammar: ".. in a manner consistent with that used for WRTDS."

Line 222: typo

Line 233: a little vague...

Line 233: Point mostly for thought only: Did the parameters result in a global minimum? How did you know? I really like your idea of optimizing window widths, but for the kind of messy data we see in the NTN, I wouldn't be suprised to see eqifinal results, e.g. many different window-width choices yielding "near-optimal" results.

Line 253: consider: "were performed similarly, using the equation:"

Line 256: As you probably know, in the NTN we are especially interested in bias relative to the observed data, evaluated in arithmetic space, as an indication of a model's tendency to over- or under-predict aggregated e.g. annual values; see Hirsch (2014). What factors led you to choose this statistic?

Line 259: which was the predictor and which was the response?

Added in response to general comment above.

Line 262: Note: This is true if the calculation is performed in arithmetic space. If performed in log space, a non-zero intercept indicates the "difference that varies with relative magnitude of the predictions" after back-transformation. Correspondingly, a non-zero slope indicates (I think) a difference that grows exponentially with the relative magnitude of the differences. In any case, state here in Methods whether each of the statistics you describe are computed in log or arithmetic space.

The interpretation that 'a non-zero slope indicates a difference that grows exponentially with the relative magnitude of the differences' would be correct for a single variable (i.e., log to arithmetic is linear to exponential) but perhaps incorrect for a log-log regression of two variables, as was done for our analysis. The relationships of two variable in log-space versus the same variables in arithmetic space would both be linear so I think our interpretation of differences related to slope is still correct. Although the magnitude of results for each model change exponentially at higher values, the relative differences do not.

Line 288: Point mostly for thought: Does a "true" flow-normalized "signal" even really exist?

Line 296: Point mostly for thought, further to above comment: "primary production" needs nutrients and generates waste. Wouldn't the "closed system" to which you refer soon starve or poison itself? In the non-tidal Bay community, the concept of flow-normalization is widely misinterpreted, especially when it comes to trying to identify specific factors driving flow-normalized trends; we have even bickered with Bob over its meaning. My personal advice is that you can avoid a lot of uncomfortable discussions if you are careful (and circumspect) about stating its meaning in your publications and public presentations. As widely as the term is used, I find that highly specific interpretations tend to evaporate on close inspection.

Line 308: I wonder if you could just eliminate this statement completely? I'm not an estuarine scientist, but it seems like a fairly nuanced point.

Line 319: I think you are trying to define the terms in these equations in the surrounding paragraph, which is OK. But you seem to be relaxing the rigor, applied above, in time series notation e.g., defining subscripts "i" and stating their range. Also, I take it from the following text that the sigma values in the two equations are different? If correct, they need subscripts. Suggest you state units throughout the document.

Line 324: consider: "The vector I (where $0 < I < 1$) is a weighting and unit-conversion vector that a) translates the terms enclosed in parentheses from flow to chl-a concentration units, and b) allows for the effect of flow to be defined as time-varying. For example..."

Line 329: I think I got what this means after repeated readings, but it needs to be clarified. As stated, it implies that you did not include a flow term in your model, since you assumed it was subsumed by the seasonal term. But I think you did include a flow term (Eq. 10).

Line 341: From this I assume that the errors did not have a seasonal component?

Line 359: Clarify. Did you repeat the exact error series every year, or did you use the estimated ARMA model with a new white noise series to generate unique error series for each year?

Line 366: Clarify - you chose a random day-of-month (e.g., 16) and used it as the basis for a systematic random sample (3/16/2000, 4/16/2000, 5/16/2000, . . .)? Was that day-of-month used for each year or the entire POR? Or, was a new random day-of-month generated for each monthly sample?

Line 405: I know what you're saying, but "observed predictions" is a little confusing. I have the terms "absolute predictions" and "non-flow-normalized" predictions applied in this context - maybe consider one of those.

Line 406: Can you be more specific? "More variable than"? "Out of phase with"?

Line 410: See comments in tables.

Line 419, 420: RMSE is not a rate.

Line 433: "log-transformed chl-a concentration"

Line 441: January-February-March (JFM)

Line 444: quantile

Line 457 - 469: See comments in Methods section about significance of log transformation. Overall I think this comparison would be more useful in done in arithmetic space. But you would need to specify more clearly how you bias-adjusted the back-transformed GAM predictions. Either way, specify which model results were the "predictor" and which were the "response" in this comparison.

Line 481: this seems a little vague. Do you mean "simulations conducted with co-variates defined to vary over a regular grid"?

Line 493: Looks like the models differ in April for that site, too, but this is obscured a little by the plotting scheme.

Line 495: Overall, this section says that both models performed well on all three test cases. Like all your figures, I like the ones you used for this section, but am thinking maybe you don't need so many figures to make this point. Figure 7 essentially says your chl-a simulations correctly did what they were created to do, which I think we could maybe take on faith given you are reporting comparison results. I like Figure 8 the best. If you included a corresponding set of panels for the GAM simulation, and left in Table 7, I think you could get away without figs. 7 and 9.

Figures 7 was added to supplementary material, Figure 9 was removed. GAM panels were added to Figure 8

Line 536: very true.

Line 543: RMSE is not a rate.

Line 550: I think you need to explain what this term means in this context or find another choice of words.

Line 552: Cool!

Line 565: I don't know of anyone who "expected" this novel insight (although feel free to provide a citation). It may simply be that GAMs provide a very similar type of information as WRTDS with less computational burden. I don't know. It might be more helpful to specifically discuss what each approach can do that the other can't, if their capabilities really are that different.

Additionally, since computational time seems to be among the largest differences, how about presenting some data? As a steward of the NTN trend products, I would be delighted to see a model that can do everything WRTDS can in a lot less time.

Line 567: how about "predefined parameterization and fixed parameters"?

I think Bob was saying that conventional models are limited by their parameterization and the reliance on constant parameters. To me the term "parameter space" specifies the numeric ranges the parameters can take on, which really isn't a model limitation.

Line 598: log-chl-a

Line 637: typo

Table 2: RMSE = sqrt(SSE/(12*(2014-1986+1)(, no? I can't get these numbers to balance.

Table 3: Clarify in document body - there are trends over the period-of-record if only the selected set of months are considered, correct?

Table 5: State which is "x" and "y" in these regressions.