Response to reviewer comments, Environmental Modelling and Assessment ENMO-D-14-00158
Prepared January 23, 2015, M. Beck, J. Hagy

Response to review comments are in italics. All line numbers refer to the original text.

AE: The manuscript provides an interesting contribution to model and evaluate water quality but needs a profound revision along the valuable comments of the reviewers, especially reviewer no. 1.

Thank you for your review. Please see our detailed comments below.

Reviewer no. 1: The models evaluated in this manuscript extend the weighted regression on time, season, and discharge (WRTDS) method developed by Bob Hirsch by using salinity fraction in place of discharge and estimating a lower (0.10) and upper (0.90) quantile as well mean changes in chlorophyll (the response y). While I'm supportive in general of more sophisticated regression modeling of this sort of water quality data, especially by incorporating quantiles to complemented mean responses, there are several aspects of the modeling that deserve additional explanation and consideration. One aspect relates to the implementation and interpretation of the quantile regression estimates and the other aspect relates to the extreme localized smoothing used by the WRTDS method.

(1) Quantile regression estimates applied to this sort of water quality data certainly have great potential utility for highlighting heterogeneous trends over time or space. There are several aspects of its implementation that are not explained well here and that I could not find in any of the Hirsch 2014 publications related to WRTDS.

The R1 coefficients of determination suggested by Koenker and Machado (1999) for quantile regression goodness-of-fit should probably not be characterized as "pseudo R2" as they differ from the latter by fundamentally different scaling, the former based on a measure of variation based on absolute deviations and the latter on variance with squared deviations. Thus, the R2 for the mean regressions are not directly comparable to the R1 coefficients of determination for the quantile regression, and the latter will always be smaller than the former. They can be equated after the fact if so desired, e.g., R*1 = 1 - (1 - R2)0.5 to convert R2 to absolute deviations.

All instances of 'pseudo-R^2' were removed from the text to more carefully describe the measure of fit. Additionally, the mean models in the original manuscript were replaced with median quantile regression models (i.e., $\tau = 0.5$) to allow comparability between measures of fit. The following text starting at line 91 was modified as follows to describe these changes:

'Quantile regression models [2] were used to characterize trends at both the median and extreme conditional distributions of the data. Specifically, we adapted the weighted regression approach to model the conditional response at the 10th, 50th, and 90th quantiles ($\tau = 0.1, 0.5,$ and 0.9, respectively) of the chlorophyll distribution. Quantile regression is analogous to least-squares regression such that a set of β parameters that minimizes the error term is estimated. However, the minimization function is the sum of the weighted absolute deviations of the fitted values from

the observed quantile rather than the conditional mean response as in ordinary regression. A general interpretation of the fitted values is the distribution of chl-a conditional upon time and salinity for low ($\tau = 0.1$) or high ($\tau = 0.9$) biomass events. The median values can be considered a model estimation of the central tendency of chl-a over time, although this is quantitatively distinct from mean models that characterize average chl-a. Additionally, back-transformation bias of predicted values does not occur with quantile models because estimates are equivariant to non-linear, monotonic transformations [18].'

It is not clear how you calculated the R1 coefficients of determination for the quantiles in your WRTDS implementation. I know how this should be calculated for a single model, e.g., for the non-weighted columns in Table 3. But I don't know how you would calculate these for the WRTDS implementation where every observation has a separate model. Average them across all n models for n observations or some other manual accumulation of residuals from the n observations into a single result? This needs to be explained as it certainly is not standard for quantile regression.

Methods for evaluating goodness of fit were more carefully described, in addition to an explanation of how residuals were estimated from the WRTDS models. The following was added to the methods starting on line 119:

'Model fit was evaluated using the quantile regression goodness of fit described in Koenker and Machado [20]. This measure has a similar interpretation as the standard R^2 for mean regression models, although it differs fundamentally by describing the relative success of the model at a specific quantile using the weighted sum of absolute residuals. Using notation in Koenker and Machado [20]:

$$R^1(\tau) = 1 - \hat{V}(\tau) / \tilde{V}(\tau)$$

where $R^1(\tau)$ is the proportion of explained variance of the model at quantile τ. Values for $\hat{V}(\tau)$ and $\tilde{V}(\tau)$ describe the sum of residual variance of the fully parameterized model and a null model (i.e., the non-conditional quantile of the response). The residual variance for each WRTDS model was based on an accumulation of residuals for each regression model specific to each observation.'

Presumably the adjustment for bias of back transforming estimates (lines 126-139) was only applied to the mean regression estimates and not the quantile estimates as they are equivariant to nonlinear transformations like the logarithmic used here. This should be made clear. Indeed, this is nice advantage of the quantile regression estimates for these sorts of models and would suggest that the median (0.50 quantile) estimate might be a useful, simpler quantity to estimate to describe the center of the distribution rather than the mean as you can avoid estimating this back-transformation bias.

The correction for back-transformation bias was removed from the analysis because all models now use quantile regression, see the response to the first comment above.

The quantile regression estimates are readily extended to estimating left-censored data associated with below-detection limit responses. See Portnoy (2003. Censored regression quantiles. J. American Statistical Association 98(464): 1001-1012) and Koenker (2008. Censored quantile regression redux. J. Statistical Software 27(6): 1-25). This feature has been refined in the implementation of quantile regression in the quantreg package for R, partly based on some of my input regarding use with water quality data.

We modified the regression methods to more adequately handle censored data. The following paragraph was added to the methods beginning on line 119 prior to the description of evaluating model performance:

'A common issue with water quality data is the presence of observations that occur beyond the detection limit of the method used to measure the variable of interest. The most recent version of Weighted Regressions on Time, Discharge, and Season (WRTDS) method accounts for censored data by using a 'survival analysis' technique [25, 15], which is an adaptation of the weighted Tobit model for left-censored data [40]. Chlorophyll data for Tampa Bay are also left censored with the most commmon lower detection limit being 2.4 $\mu g\ L^{-1}$ *for individual survey years. A censored quantile regression approach was used based on methods described in Portnoy [28] and Koenker [18]. The method builds on the Kaplan-Meier approximation for a single-sample survival function by generalizing to conditional regression quantiles. The* `quantreg` *package in R [19] employs this method using recursive estimation of linear conditional quantile functions. Censored quantile regression models were used withe adapted weighted scheme to model observed chlorophyll-a (chl-a) in each segment. Data were based on median values for all stations within a segment such that the lower detection limits that applied to observations at individual stations were preserved in the combined data. A segment observation at a given time step was considered censored if it was equal to the known detection limit for a given year. The lower detection limits were identified by parsing the station data by year to identify values that were flagged accordingly in the original data [38].'*

(2) The WRTDS approach represents an extreme form of local smoothing of regression estimates that while perhaps desirable from the standpoint of maximizing fit to an observed temporal sequence of water quality measures has the undesirable features of (a) not readily providing measures of uncertainty (e.g., standard errors or confidence intervals) to characterize responses, (b) not readily nested within less complex models with less local smoothing to assess whether more or less local smoothing is adequate, and (c) can't readily incorporate other covariates (e.g., seagrass coverage) into the model. The WRTDS approach where every observation gets its own regression model is on the extreme opposite of a continuum of smoothing where a single regression relationship across all observations (your non-weighted estimates) is the other extreme. Your Table 3 results suggest about a 50% gain in variation explained by going from the single non-weighted regression model to the other extreme of an n-observations locally weighted WRTDS model, which while substantial is still only providing modest improvements in absolute variation explained (e.g., increasing R1 of quantile estimates from 0.30 to 0.45). It certainly is possible that less extreme locally weighted regression procedures that fit into conventional linear modeling approaches might provide some substantial improvements in variation explained but require far less than the n p estimated parameters of the WRTDS approach (where p is the

number of parameters estimated in each model). It is relatively simple to make the single regression model with parameters applicable across all observations (your non-weighted model) into a model that is piecewise linear in various regions of the predictor space (e.g, by time of salinity), either by using indicator variables (e.g., Neter et al. 1996. Applied linear statistical models: 474-478) for continuous or discontinuous pieces or with spline basis functions (e.g. b-splines). Decisions on how many piecewise linear (or quadratic or cubic) regions and where they should occur can be based on information criterion (e.g,, AIC or BIC) or cross-validation. For example, it might just require that the time component is broken into 3 regions such that 3 p parameter estimates provide adequate fit. Most importantly goodness-of-fit and information criteria can be compared with the simple no locally smoothing model to assess where substantial improvement has been obtained. These piecewise linear models may never fit quite as well as the extreme locally weighted WRTDS method, but coupled with quantile estimates they may provide a more than adequate characterization of the changing water quality that requires a less extremely parameterized model, that has conventional statistical summary information to characterize uncertainty in estimates and goodness-of-fit, and provides a more parsimonious interpretation of relationships that has greater potential to generalize to other places and times. It would be interesting in this manuscript if you compared one of these less extreme local smoothing approaches rather than just going to the extreme WRTDS approach.

We agree that WRTDS represents an extreme example of locally weighted smoothing and are aware that alternative, less-intensive approaches could provide comparable estimates. In fact, we have had recent discussions with researchers working on similar techniques on the relative merits of different approaches for trend evaluation, including WRTDS, generalized additive models, and loess regression. These discussions are on-going and we agree that more comprehensive comparisons of the different methods are warranted, particularly with respect to differences in parameterization techniques and level of information provided. We have added content to the discussion to address some of these concerns rather than additional examples as we believe that a simple comparison in this manuscript would incompleteley describe the validity of each technique. The use of goodness of fit as a sole metric of performance has also been discussed more thoroughly to emphasize that improved fit is not the sole advantage of WRTDS. Rather, the real value of WRTDS and our current adaptation is the ability to generate hypotheses of system drivers of change that are consistent with the parameterizations (i.e., figure 8), in addition to describings trends in the extreme distributions.

The following was added to the discussion:

'The WRTDS and the adaptation to modelling quantile distributions provides an approach for generating hypotheses describing factors that act as system drivers of change. Application of WRTDS to the Tampa Bay dataset allowed a temporally consistent description of water quality leading to the generation of such hypotheses. Alternative analysis methods may be more appropriate for different research or management objectives, particularly given sample size constraints or the need to predict future water quality trends under hypothetical scenarios. For example, generalized additive models or locally estimated (loess) regression are similar methods for describing conditional response within discrete periods of time. Such methods may provide similar improvements in the fit of the predicted values to the observed data, although predictive

performance as indicated in Table 3 is a relatively simple metric that inadequately describes the true value of a model. The true value of an analysis technique depends more on its ability to address a specific question of interest. For example, WRTDS may be most appropriate for describing historical patterns to better understand drivers of change, given the ability to remove flow effects and describe relationships that change over time. Quantitative comparisons of WRTDS with other techniques in the context of providing historical descriptions could clarify the relative merits of different approaches, although this falls beyond the scope of the current analysis.'

The following sentence was also added to the methods in the paragraph describing measure of fit (line 125).

'In this context, 'performance' describes the measure of fit to the observed data and is considered a relatively narrow definition of overall model value.'

Other comments by line numbers:

Lines 140-151: This explanation of normalized predictions doesn't make it real clear what the purpose of this normalization is and why it is required. Some elaboration is in order.

The following text was added/modified on line 140:

'Normalization is used to remove the variance in the response that is attributed to a predictor variable, allowing interpretation of trends that are independent of confounding sources of variation. For example, water quality trends that are potentially related to management actions can be more precisely evaluated if changes in pollutant concentrations due to natural variation in discharge are removed.'

Lines 239-242: Given that this residual pattern is axiomatic from the quantile regression estimation, it is not clear that it is worth mentioning.

The lines were removed.

Table 6: I think it is important to have sample sizes be meshed with these correlation values rather than just providing their range in the footnote.

The following text was added to the table caption:

'Samples sizes for correlations were 11 for seagrass area, 156 for ENSO index by season, 156 for ENSO index by year, 275 for nitrogen load, and 303 for nitogren concentration with slight variation by segment depending on data availability.'

Reviewer no. 2: Comments on Adaptation of a weighted regression approach to evaluate water quality trends in an estuary.

This is an excellent paper and it makes a great contribution to the complex problem of evaluating environmental trends in an estuary. Finding the means to account for the variability that is due to random variations in freshwater inputs to the estuary is crucial to such analyses, but the analyses must be designed in a way that accounts for the fact that the relationships of water quality variables to inflow and season can change substantially over a period of several decades. The authors' adaptation of the WRTDS method, designed for rivers, to a method that is appropriate to estuaries is highly creative and thoughtful. Their inclusion of quantile regression is a useful enhancement to the WRTDS method. Creation of new exploratory graphics such as Figure 8 is also a very worthwhile addition. I have a few specific comments.

On line 91 it states that concentrations below the detection limit were set to half the detection limit. The paper does not mention how common such censored data are in the data set. Doing this kind of substitution can seriously bias the results of a trend analysis if censored data are common. The perils of such substitution methods are discussed in Helsel's "Statistics for Censored Environmental Data Using Minitab and R" (Second Edition, Wiley and Company, 2012). From looking at the graphics in their paper I expect that such censored data were rather rare and if that is the case, then the simple approach that they took to the censored data issue should be no problem. In my mind something less than about 3% of the data being censored may be considered low enough to make this a non-issue. However, I think it would be worthwhile if the readers were informed of the frequency of censored data in their data set. It should also be noted that since the original paper on WRTDS in 2010 the method has been modified to account for censoring using survival regression.

Based on the comments from both reviewers, we have modified the analysis to more carefully handle censored observations. Please see the comments to reviewer 1 for details.

The percentage of observations for each segment was also added to the paragraph describing censored data:

'The percentage of observations that were censored by segment was 1.8 for Hillsborough Bay, 3.6 for Old Tampa Bay, 4.3 for Middle Tampa Bay, and 24.8 for Lower Tampa Bay.'

I had an issue with the choice of words on line 178. The first sentence of the paragraph is: "Observed seasonality of chl-a was consistent with expected trends." The use of the word "trends" here is confusing, since "trends" is the central topic of their paper. I think they had a different meaning in mind for the use of this word in this sentence. Perhaps they could say "consistent with the behavior observed in many water bodies [or many estuaries]."

Sentence modified: 'Observed seasonality of chl-a was consistent with the behavior observed in many estuaries.'

I had some concern with the material in section 3.3, particularly the relationship to ENSO. They found no significant relationship between their model residuals and ENSO, but I think their statement is a bit misleading. The WRTDS model already considers the impact of rainfall on chl-a, because salinity is driven by temporally averaged streamflow, which is, in turn driven by

rainfall, and some portion of the variability in rainfall is a result of ENSO. A better way to express their finding here is to say that although salinity may be significantly influenced by ENSO, they found no significant relationship in the unexplained variability in chl-a and ENSO. They could also elaborate on this point a bit more and consider directly the possibility of a statistically significant relationship between salinity and ENSO, using a model that also accounts for the impact of seasonality. I think that some clarification of this would improve the paper, although I think the topic is rather tangential to the major points of the paper. The subsequent discussion of this topic around lines 379-383 does address this topic rather well.

The following text was added to line 157 in the methods to provide a better context for the results:

'ENSO effects were included to evaluate the potential effects of this climate cycle other than effects caused by river discharge, which is directly addressed in the model. El Niño/La Niña events have been associated with extreme variation in rainfall that influences freshwater discharge into Tampa Bay [33]. Although salinity as a model predictor may account for this variation, ENSO effects were evaluated to identify potentially unexplained changes in chl-a related to extreme climate events as compared to seasonal changes in freshwater inputs.'

Further clarification in the discussion was also added to line 382.

'Additionally, the normalized chl-a estimates in Fig. 6 differ substantially from predicted values in 1998 for all segments. High rainfall associated with an El Niño event in the winter of 1997-1998 contributed to increased discharge and nutrient inputs into the Bay, as indicated by higher predicted chl-a values. Normalized values that removed the effects of freshwater inputs show chl-a values independent of discharge, suggesting the model performs as expected by quantifying and removing this variation through normalization. However, the variation in chl-a response attributed to unique seasonal changes in discharge cannot be distinguished from extreme climate events using salinity as a proxy for freshwater inputs.'

There is an additional reference that the authors might want to consider including, which only became available around the time they completed this manuscript. It is a more comprehensive description of the WRTDS model and thus would be of value to readers wishing to understand more about this method. The reference is: Hirsch, R. M., and De Cicco, Laura, 2014, User Guide to Exploration and Graphic for RivEr Trends (EGRET) and dataRetrieval: R Packages for Hydrologic Data, USGS Techniques and Methods 4-A10, 95p. http://pubs.usgs.gov/tm/04/a10/

The reference for Hirsch and De Cicco 2014 was added.

I strongly encourage the journal to publish this manuscript. It should have broad applicability to estuaries worldwide, where river inputs of freshwater are a significant driver of water quality conditions in the estuary.