# Comparison of weighted regression and additive models for trend evaluation of water quality in tidal waters

**Marcus W. Beck**[1], **Rebecca Murphy**[2]

[1]*ORISE Research Participation Program*
*USEPA National Health and Environmental Effects Research Laboratory*
*Gulf Ecology Division, 1 Sabine Island Drive, Gulf Breeze, FL 32561*
*Phone: 850-934-2480, Fax: 850-934-2401, Email: beck.marcus@epa.gov*

[2]*UMCES at Chesapeake Bay Program*
*410 Severn Avenue, Suite 112, Annapolis, MD 21403*
*Phone: 410-267-9837, Fax: 410-267-5777, Email: rmurphy@chesapeakebay.net*

Version Date: Fri Jun 26 16:58:54 2015 -0500

## *Abstract*

*Key words*:

# 1 Introduction

Needs

- Quantitative tools that describe trends in water quality time series are needed to identify factors that influence ecosystem condition and to evaluate the effects of management activities in the context of multiple drivers
- Recent adaptation of statistical models for evaluating water quality time series have shown promise for application in tidal waters, specifically generalized additive models (GAM) and weighted regression on time, discharge, and season (WRTDS) {acro:GAM} {acro:WRTI
- These similar techniques can be used to quantify relationships between response measures and different drivers that may vary over time, in addition to an evaluation of trends independent of variation in freshwater inputs
- The relative merits of each approach have not been evaluated, particularly related to accuracy of the empirical description and the desired products for trend evaluation
- Such a comparison could inform the use of each model for addressing management or restoration needs or for developing more robust descriptions of long-term changes in ecosystem characteristics

Goal: Provide a description of the relative abilities of GAMs and WRTDS to describe long-term changes in time series of response endpoints in tidal waters Objectives:

- Provide a narrative comparison of the statistical foundation of each technique, both as a general description and as a means to evaluate water quality time series
- Use each technique to develop an empirical description of water quality changes in a common dataset with known historical changes in water quality drivers
- Apply the models to simulated data to evaluate ability of the models to describe true changes
- Compare each technique's ability to describe changes, as well as the differences in the information provided by each

- Provide recommendations on the most appropriate context for using each method

# 2 *Methods*

## 2.1 Study site

The Patuxent River Estuary...

Observed trends over time

longitudinal gradient from watershed to mainstem influences, LE1.2, TF1.6

Show plots of trends over time in observed data

## 2.2 Model descriptions

How, Similarities, differences, optimal smoothing

The selection of optimal model parameters is a challenge that represents a tradeoff between model precision and ability to generalize to novel datasets. Weighted regression requires identifying optimal half-window widths, whereas GAM requires identifying the optimal degrees of freedom for the smoothing parameter. Overfitting a model with excessively small window widths or too many degrees of freedom will minimize prediction error but prevent extrapolation of results to different datasets. Similarly, underfitting a model with large window widths or very few degrees of freedom will reduce precision but will improve the ability to generalize results to a different dataset. From a statistical perspective, the optimal smoothing provides a balance between over- and under-fitting. Both models use a form of cross-validation to identify model parameters that maximize the precision of model predictions with a novel dataset.

The basic premise of cross-validation is to identify the optimal set of model parameters that minimize prediction error on a dataset that was not used to develop the model. For GAMs (Hastie and Tibshirani 1990, Zuur 2012)... Similarly, the tidal adaptation of WRTDS used k-fold cross-validation to identify the optimal model parameters. The dataset was separated into ten disjoint sets, such that ten models were evaluated for every combination of k - 1 training and remaining test datasets. That is, the training dataset for each fold was all k - 1 folds and the test dataset was the remaining fold, repeated k times. The average prediction error of the training datasets across k folds provides an indication of model performance for the given combination of half-window widths. The optimum window widths were those that provided minimum errors on the test data. Evaluating multiple combinations of window-widths can be computationally

intensive. An optimization function was used to more efficiently evaluate model parameters using a search algorithm. Window widths were searched using the limited-memory modification of the BFGS quasi-Newton method that imposes upper and lower bounds for each parameter (Byrd et al. 1995). The chosen parameters were based on a selected convergence tolerance for the error minimization of the search algorithm.

## 2.3 Comparison of modelled trends

Explanatory power of each method - explained variance/fit in the response, histograms of errors (see page 14 in Moyer) - we can test for significant differences in the errors using a two-sided t-test. Also see page 24/25 in Moyer for average difference comparisons between methods.

Similarity of predictions - observed data, simple scatterplots, similarity coefficients, similarity by time periods, etc.

Indications of change - direction/magnitude of trends by different time periods

## 2.4 Comparison of flow-normalized trends

The relative abilities of each model to characterize flow-normalized trends in chlorophyll were evaluated using simulated datasets with known components. This approach was adopted because the flow-independent component of chlorophyll in estuaries cannot be known with absolute certainty. Accordingly, the ability of each model to isolate the flow-normalized trend cannot be faithfully evaluted unless the true signal is known. Simulated time series of observed chlrophyll ($Chl_{obs}$) were created as additive components related to flow ($Chl_{flo}$), a flow-indepenent biological component of chlorophyll ($Chl_{bio}$), and residual error ($\varepsilon$):

$$Chl_{obs} = Chl_{flo} + Chl_{bio} + \varepsilon \quad (1) \quad \{\texttt{chlsim}\}$$

The simulated time series were based on stochastic models derived from actual water quality measurements to ensure the statistical properties were similar to observed data. Daily flow observations were obtained from the US Geological Survey (USGS) stream gage station 01594440 near Bowie, Maryland (38°57′21.3″N, 76°41′37.3″W) from 1985 to 2014. Similarly, daily chloropyll records were obtained from the Jug Bay station (38°46′50.6″N, 76°42′29.1″W) of the Chesapeake Bay Maryland National Estuarine Research Reserve. Daily chlorophyll {acro:USGS

concentrations were estimated from fluorescence values that did not include blue-green algae blooms. However, our primarily concern was simulating chlorophyll concentrations at monthly or bimonthly timesteps such that taxa-specific concentrations on a daily time step were not relevant.

The statistical properties of both the flow and chlorophyll time series were characterized to create a stochastic model of water quality, where chlorophyll was directly related to the effects of flow. This approach allowed us to evalute the ability of GAMs and WRTDS under different sampling regimes while ensuring the simulated datasets were consistent with the statistical properties of known time series at much finer temporal resolution. First, daily flow data were simulated as the additive combination of a stationary seasonal component and serially-correlated errors:

$$\ln(Q_i) = \beta_0 + \beta_1 \sin(2\pi T) + \beta_2 \cos(2\pi T) + \sigma\dot{\varepsilon} \tag{2}$$

creating a stational seasonal regression of flow over time. The residuals from this regression were used to estimate the error distribution using an auto-regressive, moving average model. Results of this model were used to generate random errors from a standard normal distribution. The random, serially-correlated errors were multiplied by the standard deviation of the residuals, then added to the seasonal component of original model to create a simulated, daily log-flow time series.

The chlorophyll time series was created using a similar approach with minor differences. The first step was to estimate the error component of the chlorophyll time series by fitting a WRTDS model using one year of data from the whole time series. This approach was used rather than a simple seasonal model to remove any confounding effect of flow on the error structure. The error distribution was estimated from the residuals as before. Standard error estimates from the regression used at each point in the one-year time series were also retained for each residual. Random errors using the estimated auto-regressive structures were simulated for the entire year and multiplied by the corresponding standard error estimate from the regression. The entire year was repeated for every year in the observed time series. All simulated errors were rescaled to the range of the original residuals that were used to estimate the distribution.

The simulated chlorophyll time series was then created by estimating the seasonal component from the observed time series, with the assumption that this component represented a flow-independent time series. The chlorophyll error time series was then added to the seasonal

model. Finally, the simulated flow-component was added to the simulated chlorophyll time series to create a combined chlorophyll-flow time series. The flow-component was first centered at zero and multiplied by a vector of coefficient that represented the relative effect of flow through the time series.

Similarity of flow-normalized results - simulated data, simple scatterplots, similarity coefficients, similarity by time periods, etc.

# *3 Results*

Predicions with actual data  Simulations

# *4 Discussion*

Qualitative comparison

- Computational requirements and potential limitations
- Data needs or transferability of each technique to novel datasets
- Products, e.g., conditional quantiles of WRTDS, confidence intervals for GAMs, handling censored data, hypothesis testing vs description
- Appropriate context for using each approach

## 4.1 Conclusions

## *References*

Byrd RH, Lu P, adn C. Zhu JN. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing, 16(5):1190–1208.

Hastie T, Tibshirani R. 1990. Generalized Additive Models. Chapman and Hall, London, New York.

Zuur AF. 2012. A Beginner's Guide to Generalized Additive Models in R. Highland Statistics Ltd., Newburgh, United Kingdom.