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: Wed Jul 22 14:04:36 2015 -0500

Abstract

Long-term monitoring datasets that describe water quality in coastal systems can provide a means of interpreting the effects of environmental changes or management actions on ecosystem condition. The ability to link causal effects with potential changes is highly dependent on the chosen method of interpreting trends in the observed data. Two statistical approaches, weighted regression on time, discharge, and season (WRTDS) and generalized additive modelss (GAMs), have recently been used to evaluate long-term trends in chlorophyll time series in estuarine systems. Both models provide a similar approach to trend analysis by using context-dependent parameters or smoothing functions that vary continuously and are responsive to multiple drivers throughout the time series. However, the quantitative capabilities of each model, including descriptions of observed and flow-normalized trends, have not been rigorously compared to determine most appropriate use of each model. We evaluated WRTDS and GAMs using thirty years of data for a monthly time series of chlorophyll in the Patuxent River Estuary, a well-studied tributary to Chesapeake Bay. Each model was evaluated based on goodness of fit to the observed data and ability to reproduce flow-normalized trends with simulated data that had statistical properties that were comparable to the original dataset. Each model was also evaluated based on concordance between conclusions of water quality changes, and causes thereof, in different time periods. This information will provide researchers with valuable guidance for trend evaluation using statistical models, with particular relevance for computational requirements, desired products, and future data needs.

{acro:WRTI

{acro:GAM

Key words: chlorophyll, estuary, generalized additive models, Patuxent River Estuary, trend analysis, weighted regression

1 *Introduction*

The interpretation of enviromental trends can have far-reaching implications for the management of evironmental resources and developing an improved understanding of ecological factors that mediate system dynamics. An accurate interpretaion of observed data can depend on the chosen method of analysis, and more importantly, the ability of the analysis method to consider the effects of multiple drivers on response endpoints. For coastal environments.... Challenges for trend analysis in tidal waters

- Duarte neverland
- Challenges inherent in any system
- tidal flux
- freshwater inflow
- need to distinguish between environmental and management changes
- 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 applications of statistical models for water quality time series have shown promise in tidal waters, specifically weighted regression on time, discharge, and season (WRTDS) and generalized additive modelss (GAMs). {acro:WRTDS}{acro:GAM} The WRTDS method was initially developed to describe water quality trends in rivers (Hirsch et al. 2010, Hirsch and De Cicco 2014) and has recently been adapted to describe chlorophyll trends in estuaries (Beck and Hagy III 2015). A defining characteristic of the WRTDS model is a weighting scheme that fits a continuous set of parameters to the time series by considering the relative influence of location in the record and contextual flow inputs to the period of interest. To date, the WRTDS model has been used to model pollutant delivery from tributary sources to Chesapeake Bay (Hirsch et al. 2010, Moyer et al. 2012, Zhang et al. 2013), Lake Champlain (Medalie et al. 2012), and the Mississippi River (Sprague et al. 2011). A comparison to an alternative regression-based model for evaluating nutrient flux, ESTIMATOR, suggested that WRTDS can produce more accurate trend estimates (Moyer et al.

2012). GAMs are a more generic statistical model that can describe variation in a response variable as a sum of smoothing functions for different predictor variables (Hastie and Tibshirani 1990, Wood 2006). In application to water quality time series, GAMs are similar to WRTDS in that variable effects through time of different drivers can be described in the context of seasonal or annual changes. Application of GAMs to describe eutrophication endpoints in tidal waters have not been as extensive as WRTDS, although exploratory analyses have suggested that results are comparable. Moreoever, GAMs are particularly appealing because they are less computationally intense and provide more accessible estimates of model uncertainty than WRTDS. Despite the potential for both approaches to guide management applications, the relative merits of each have not been rigorously evaluated. Quantitative comparisons that describe the accuracy of the empirical description and the desired products could inform the use of each model to describe long-term changes in ecosystem characteristics.

The goal of this study is to provide an empirical description of the relative abilities of WRTDS and GAMs to describe long-term changes in time series of eutrophication response endpoints in tidal waters. A long-term time series of monthly chlorophyll observations from the Patuxent River Estuary is used as a common dataset for evaluating each model. The Patuxent Estuary is a well-studied tributary of the Chesapeake Bay system that has been monitored for thirty years with fixed stations along the longitudinal axis. Two stations were chosen as representative time series that differed in the relative contributions of watershed inputs and influences from the mainstem of the Chesapeake. This provided a unique opportunity to evaluate each model's ability to interpret effects of different drivers on primary production. The specific objectives of the analysis were to 1) provide a narrative comparison of the statistical foundation of each model, both as a general description and as a means to evaluate water quality time series, 2) use each model to develop an empirical description of water quality changes at each monitoring station in the context of known historical changes in water quality drivers, 3) apply the models to simulated data to evaluate descriptions of flow-normalized trends, and 4) compare each technique's ability to describe changes, as well as the differences in the information provided by each. We conclude with recommendations on the most appropriate context for using each method, with particular attention given to computational requirements, uncertainty assessment, and potential needs for additional monitoring data.

2 *Methods*

2.1 Study site

The Patuxent River Estuary... background, history

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 GAMs requires identifying the optimal degrees of freedom for the smoothing parameter. Overfitting a model with excessively small window widths or excessive 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 different datasets. From a statistical perspective, the optimal model parameters provide 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)...[insert GAMs methods]. Similarly, the tidal adaptation of WRTDS used k-fold cross-validation to identify the optimal half-window widths. For a given set of half-window widhts, 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 test datasets across k folds provided 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 implemented in R (Byrd et al. 1995, RDCT (R Development Core Team) 2015) 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. 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 or salinity-normalized trends in chlorophyll were evaluated using simulated datasets with known components. This approach was used because the flow-independent component of chlorophyll is typically not observed in raw data such that the true signal must be empirically estimated. Accordingly, the ability of each model to isolate the flow-normalized trend cannot be evaluated with reasonable certainty unless the true signal is known. Simulated time series of observed chlrophyll (Chl_{obs}) were created as additive components related to flow (Chl_{flo}, analogous to salinity) and a flow-independent biological component of chlorophyll (Chl_{bio}):

$$Chl_{obs} = Chl_{flo} + Chl_{bio} \qquad (1)$$ {chlobs}

A distinction between Chl_{flo} and Chl_{bio} is that the former describes variation in the observed time series with changes in discharge (e.g., concentration dilution with increased flow) and the latter describes a true, desired measure of chlorophyll in the water column that is directly linked to primary production. The biological component of chlorophyll is comparable to an observation in a closed system that is not affected by flow and is the time series that is estimated by flow-normalization with WRTDS and GAMs.

The simulated time series was based on a stochastic model derived from actual flow and water quality measurements to ensure the statistical properties were comparable to existing datasets. This approach allowed us to evaluate GAMs and WRTDS under different sampling regimes (e.g., monthly rather than daily), while ensuring the simulated datasets had statistical properties that were consistent with known time series. Daily flow observations were obtained from the US Geological Survey (USGS) stream gage station 01594440 near Bowie, Maryland {acro:USGS (38°57′21.3″N, 76°41′37.3″W) from 1985 to 2014. Daily chlorophyll 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 concentrations were estimated from fluorescence values that did not include blue-green algae blooms. Our primary concern was simulating chlorophyll concentrations at monthly or bimonthly timesteps such that taxa-specific concentrations on a daily time step were not relevant.

Four time series were estimated or simulated from the actual datasets to create the complete, simulated time series: 1) estimated discharge as a stationary seasonal component ($\hat{Q}_{seas}$), 2) simulated error structure from the residuals of the seasonal discharge model ($\varepsilon_{Q,\,sim}$), 3) estimated chlorophyll independent of discharge as a stationary seasonal component ($\hat{Chl}_{seas}$), and 4) simulated error structure from the residuals of the seasonal chlorophyll model ($\varepsilon_{Chl,\,sim}$). Unless otherwise noted, chlorophyll and discharge are in ln-transformed units. Each of the four components was used to simulate the components in eq. (1):

$$Chl_{flo} = I\left(\hat{Q}_{seas} + \sigma \cdot \varepsilon_{Q,\,sim}\right) \tag{2}$$ {chlflo}

$$Chl_{bio} = \hat{Chl}_{seas} + \sigma \cdot \varepsilon_{Chl,\,sim} \tag{3}$$ {chlbio}

The estimated flow time series within the parentheses, $\hat{Q}_{seas} + \sigma \cdot \varepsilon_{Q,\,sim}$, is floored at zero to simulate an additive effect of increasing flow on Chl_{obs}. Although the actual relationship of water quality measurements with flow is more complex, we assumed that a simple addition was sufficient for the simulations where the primary objective was to create an empirical and linear link between flow and chlorophyll. Moreover, the vector I (where $0 \leq I \leq 1$) can be manually changed to represent an independent effect of flow based on the desired simulation. For example, a flow effect that changes from non-existent to positive throughout the period of observation can be simulated by creating a vector ranging from zero to one. For the simulated Chl_{bio} time series, the seasonal and error components were characterized using the daily time series at Jug Bay that likely included an effect of flow in the observed data. For the simulated models, we assumed that the actual flow effect was part of the seasonal component to obtain an accurate estimate of the error component that was independent of both flow and season. Methods for estimating each of the components in eqs. (2) and (3) are described in detail below.

First, a model for simulating flow-related chlorophyll (eq. (2)) was estimated from the stream gage data as the additive combination of a stationary seasonal component and serially-correlated errors:

$$Q_{seas} = \beta_0 + \beta_1 \sin\left(2\pi T\right) + \beta_2 \cos\left(2\pi T\right) \tag{4}$$ {qseas}

$$\varepsilon_Q = Q_{seas} - \hat{Q}_{seas} \tag{5}$$ {qerr}

A seasonal model of flow was estimated using linear regression for time, T, on an annual sinusoidal period (eq. (4)). The residuals from this regression, ε_Q (eq. (5)), were used to estimate the structure of the error distribution for simulating the stochastic component of flow. The error distribution was characterized using an Autoregressive Moving Average (ARMA) model to {acro:ARMA identify appropriate p and q coefficients (Hyndman and Khandakar 2008). The parameters were chosen using stepwise estimation for nonseasonal univariate time series that minimized Akaike {acro:AIC Information Criterion (AIC). The resulting coefficients were used to generate random errors from a standard normal distribution for the length of the original time series, $\varepsilon_{Q,\,sim}$. These stochastic errors were multiplied by the standard deviation of the residuals in eq. (5) and added to the seasonal component in eq. (4) to create a simulated, daily time series of the flow-component for chlorophyll, Chl_{flo} (eq. (2)).

The chlorophyll time series was created using a similar approach. The first step estimated the stationary seasonal component of the chlorophyll time series by fitting a WRTDS model (Hirsch et al. 2010) that explicitly included discharge from the gaged station using one year of data from the whole time series:

$$Chl_{seas} = \beta_0 + \beta_1 T + \beta_2 Q + \beta_3 \sin{(2\pi T)} + \beta_4 \cos{(2\pi T)} \qquad (6)$$ {chlseas}

$$\varepsilon_{Chl} = Chl_{seas} - \hat{Chl}_{seas} \qquad (7)$$ {chlerr}

This approach was used to isolate an error structure for simulation that was independent of flow and biology, where the seasonal component (as time T on a sinusoidal annual period) was assumed to be related to biological processes. The error distribution was then estimated from the residuals (eq. (7)) as before using an ARMA estimate of the residual parameters, p and q. Standard error estimates from the regression used at each point in the one-year time series were also retained for each residual. Errors were simulated ($\varepsilon_{Chl,\,sim}$, eq. (3)) for the entire year using the estimated auto-regressive structure 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. Finally, the simulated flow-component, Chl_{flo}, was added to the simulated bilogical model, Chl_{bio}, to create the final chlorophyll-flow time series, Chl_{obs}, in eq. (1).

A daily time series for the entire period of record was simulated using the above methods and then used to compare the relative abilities of WRTDS and GAMs to characterize flow-normalized trends. Multiple time series with a monthly sampling frequencies and varying contributions of the flow component, (Chl_{flo} in eq. (1)) were created from the daily time series. One day in each month for each year was randomly sampled to create a monthly time series. Varying contributions of the flow component on observed chlorophyll were creating by multiplying Chl_{flo} by different indicator vectors (I in eq. (2)). The contribution of the flow component varied from constant, non-existent, steadily increasing, and steadily decreasing. Respectively, the vector of coefficients applied to each flow component was a constant vector of ones, a constant vector of zeroes, a linear increase starting at zero and ending at one, and a linear decrease starting at one and ending at zero. This created four monthly time series that were used to evaluate each model. The flow-normalized results of WRTDS and GAMs for each simulated time series were compared to each other and to the original biological chlorophyll component of each time series (Chl_{bio}, eqs. (1) and (3)).

3 Results

Predictions 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

Beck MW, Hagy III JD. 2015. Adaptation of a weighted regression approach to evaluate water quality trends in an estuary. Environmental Modelling and Assessment, pages 1–19.

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.

Hirsch RM, De Cicco L. 2014. User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data. Technical Report Techniques and Methods book 4, ch. A10, US Geological Survey, Reston, Virginia. http://pubs.usgs.gov/tm/04/a10/.

Hirsch RM, Moyer DL, Archfield SA. 2010. Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. Journal of the American Water Resources Association, 46(5):857–880.

Hyndman RJ, Khandakar Y. 2008. Automatic time series forecasting: The forecast package for r. JOurnal of Statistical Software, 26(3):1–22.

Medalie L, Hirsch RM, Archfield SA. 2012. Use of flow-normalization to evaluate nutrient concentration and flux changes in Lake Champlain tributaries, 1990-2009. Journal of Great Lakes Research, 38(SI):58–67.

Moyer DL, Hirsch RM, Hyer KE. 2012. Comparison of two regression-based approaches for determining nutrient and sediment fluxes and trends in the Chesapeake Bay watershed. Technical Report Scientific Investigations Report 2012-544, US Geological Survey, US Department of the Interior, Reston, Virginia.

RDCT (R Development Core Team). 2015. R: A language and environment for statistical computing, v3.2.0. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Sprague LA, Hirsch RM, Aulenbach BT. 2011. Nitrate in the Mississippi River and its tributaries, 1980 to 2008: Are we making progress? Environmental Science and Technology, 45(17):7209–7216.

Wood SN. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall, CRC Press, London, United Kingdom.

Zhang Q, Brady DC, Ball WP. 2013. Long-term seasonal trends of nitrogen, phosphorus, and suspended sediment load from the non-tidal Susquehanna River Basin to Chesapeake Bay. Science of the Total Environment, 452-453:208–221.

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

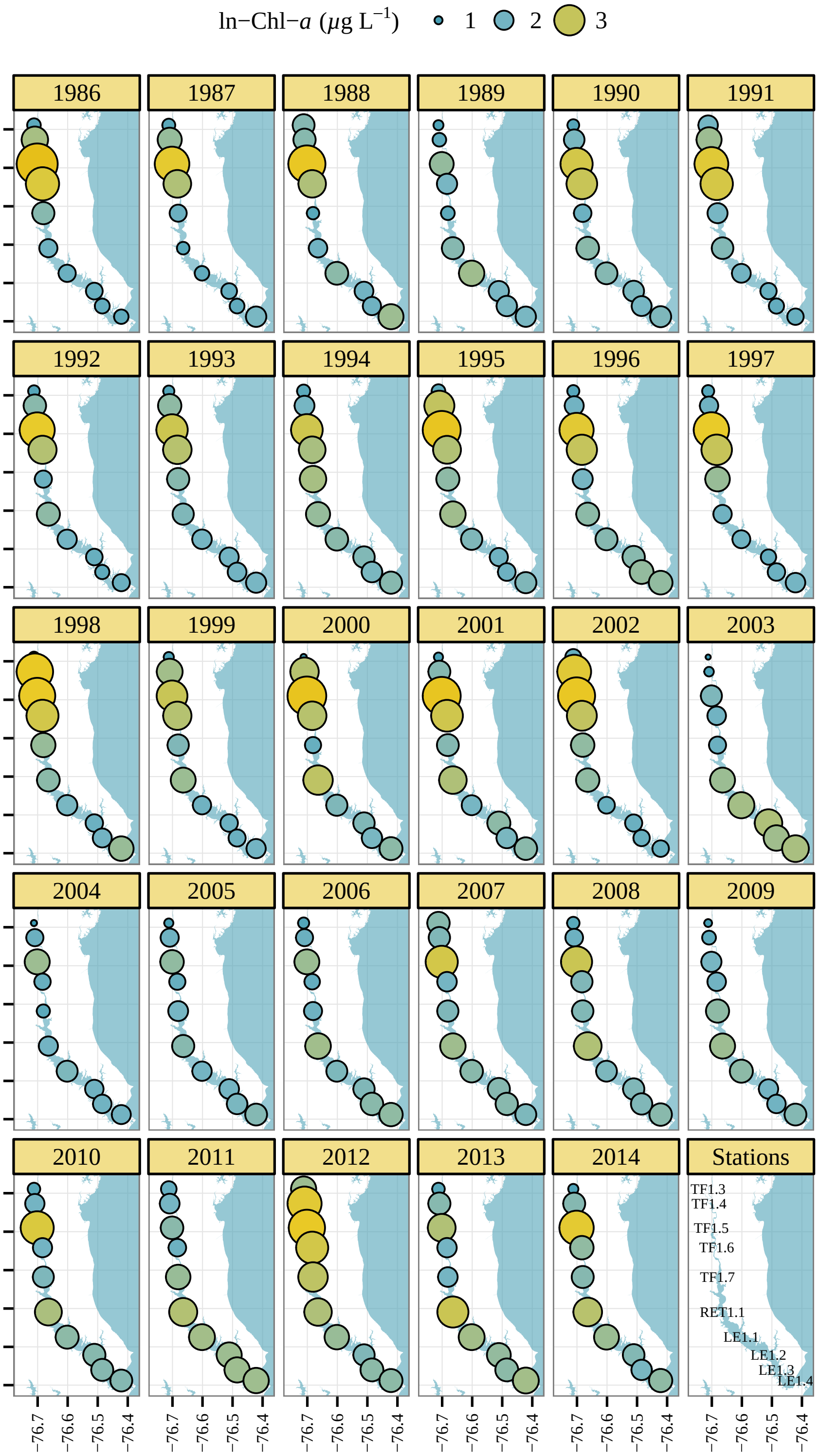

Fig. 1: Annual chlorophyll trends at each monitoring station in the Patuxent River Estuary. Values are annual medians of ln-chlorophyll-a with size and color proportional between years.

{fig:chly

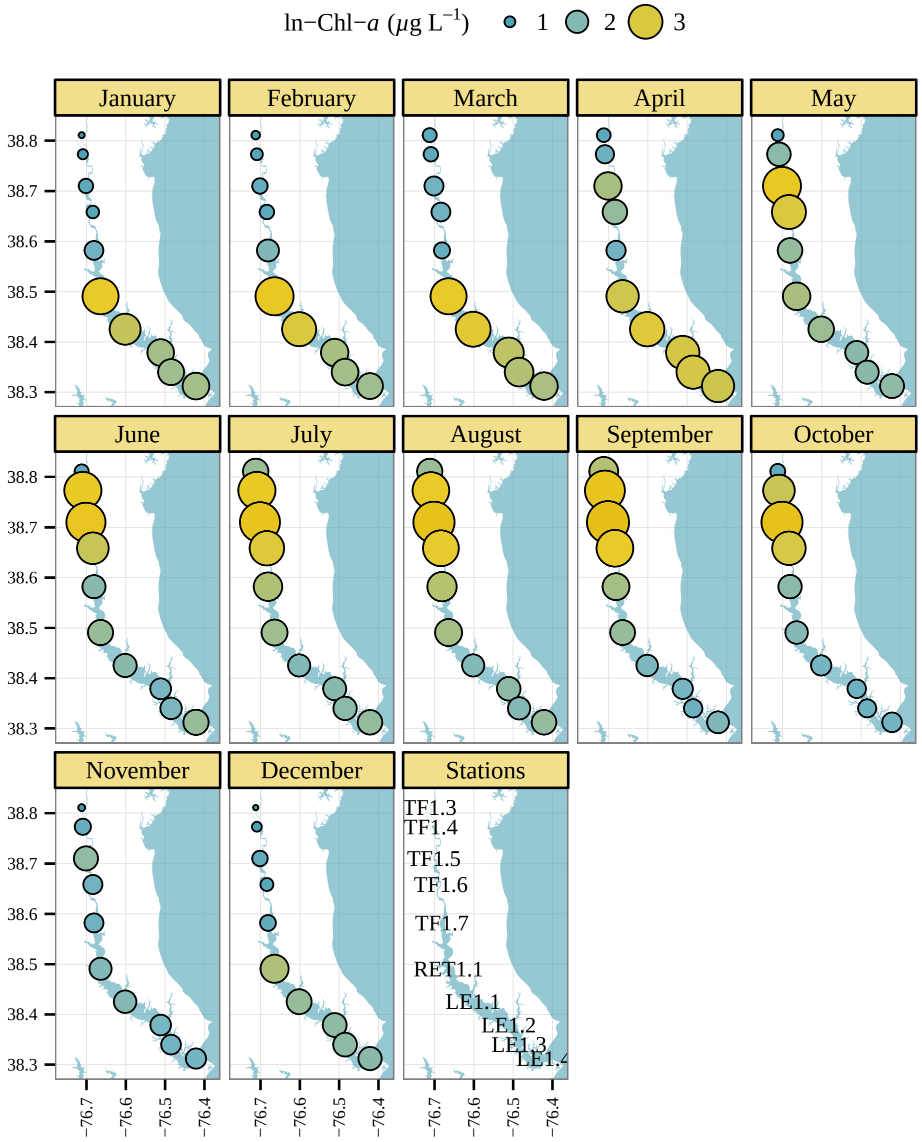

Fig. 2: Monthly chlorophyll trends at each monitoring station in the Patuxent River Estuary. Values are monthly medians of ln-chlorophyll-a with size and color proportional between months.

{fig:chlmo