Application of Generalized Additive Models for long-term and seasonal trend analysis of water quality: South San Francisco Bay case study

Marcus W. Beck ([mbeck@tbep.org](mailto:mbeck@tbep.org)), Perry de Valpine ([pdevalpine@berkeley.edu](mailto:pdevalpine@berkeley.edu)), Rebecca Murphy ([rmurphy@chesapeakebay.net](mailto:rmurphy@chesapeakebay.net)), Ian Wren ([ianw@sfei.org](mailto:ianw@sfei.org)), Ariella Chelsky ([ariellac@sfei.org](mailto:ariellac@sfei.org)), Melissa Foley ([melissaf@sfei.org](mailto:melissaf@sfei.org)), David B. Senn ([davids@sfei.org](mailto:davids@sfei.org))

Compiled 2020-04-16 11:11:55

# Abstract

# Introduction

# Methods

## Study area and data sources

The San Francisco Estuary (SFE) is the largest estuary on the Pacific Coast of North America and drains an area of approximately 200 thousand km in the US state of California. The estuary is subdivided into six sub-embayments having distinct geomorphological characteristics: Suisan Bay, San Pablo Bay, North Central Bay, Central Bay, South Bay,and Lower South Bay. Major freshwater inputs enter the system through the Sacramento-San Joaquin Delta complex upstream of Suisun Bay, where the combined inflow from both rivers is approximately 28 km per year. The northern subembayments are river-dominated (salinity ranging from 0 to 15 ppt) with seasonal freshwater pulses primarily occurring in the winter and spring during the rainy season and subsequent snowmelt in the upper watershed. The southern subembayments are marine-dominated with salinity ranging from 5 to 35 ppt depending on the tidal cycle, effluent discharge from wastewater treatment plants, and stormwater runoff. South Bay is heavily urbanized and includes thirty-seven wastewater treatments plants that serve 7.2 million people. Secondary treatment occurs at a majority of the treatment plants and the remaining effluent is discharged into the SFE. Agricultural runoff from the upper watershed also contributes to nutrient loading in the SFE with the annual nutrient export estimated as approximately 30 thousand kg dy of nitrogen from the Delta.

Nitrogen and phospohorus levels in SFE generally exceed concentrations that have been observed to promote excess primary production in other large estuarine systems. However, eutrophic conditions have not been regularly observed since routine monitoring began in the 1970s. The resistance of SFE to eutrophication has been attributed to several factors, including elevated suspended sediments that reduce light penetration in the water column, regular exchange and mixing with low-nutrient marine waters and export of estuarine nutrients to the Pacific Ocean, and benthic grazing by filter-feeding bivalves that reduce algal concentrations. Renewed interest in the potential for nutrient loading to negatively affect water quality has occurred recently, particularly in South Bay, where harmful algal blooms (HABs), increases in summer-fall chlorophyll concentrations, and low dissolved oxygen concentrations were observed beginning in 1999. These relatively recent occurrrences motivated the San Francisco Regional Water Quality Control Board and stakeholders to establish a Nutrient Management Strategy, with the purpose, among others, to develop a framework for assessing water quality condition. Critical questions that were to be addressed by the NMS included identifying water quality conditions to be considered healthy and protective of beneficial uses, how do current conditions compare to desirable water quality conditions, and what potential management actions are needed to achieve desired conditions if they currently do not persist. Although visual changes in observed data were apparent, statistical analyses to quantify current status and to provide estimates of annual and seasonal trends with appropriate bounds on uncertainty have not been sufficiently developed.

The SFE Research Program of the US Geological Survey has been collecting water quality data at fixed locations in the bay since 1969 (Cloern and Schraga, 2016; Schraga et al.2017). This dataset is one of the longest continuous monitoring datasets available in a North American estuary and includes at least monthly sampling at 37 fixed sampling locations along a 145 km longitudinal transect from the lower Sacramento River in the Delta to lower South Bay. Water quality constituent measurements available for each station include salinity, temperature, light attenuation, chlorophyll-a, dissolved oxygen, suspended particulate matter, and dissolved inorganic nutrients. The current analysis focused on near-surface chlorophyll-a data collected biweekly to monthly along the South Bay axis extending from Central Bay (stations 18-23), South Bay (stations 24 - 32), and Lower South Bay (stations 34-36). Discrete chlorophyll concentrations at each station were determined by fluoremetric analysis with 90% acetone pigment extraction on GFF filters. Data collected between 1990-2017 were selected for analysis because it represented a suitable balance among three factors relevant to testing the statistical approaches, including sufficient length of record, consistent biweekly-monthly sampling, and a diverse set of stations covering the salinity gradient across multiple subembayments. While sampling frequency varied somewhat over time or by station, all data were treated as unique time series within the statistical models (i.e., no spatial or temporal binning or averaging was done).

## GAM application

Generalized Additive Models are an extension of linear models that can incorporate additive smoothed terms to model non-linear effects. The basic structural form is (Hastie and Tibshirani 1986):

where the expected value of a dependent variable conditional on predictors X through X is the sum of smoothing functions for each predictor plus an interecept term . The smoothing functons are standardized and have expected values equal to zero so that . Because GAMs are an extension of Generalized Linear Models, they can also include linear predictors in addition to the smoothed functions. All methods herein use functions provided by the *mgcv* R package to fit GAMs using multiple parameter smoothing estimation methods (Wood, 2017).

The flexibility of the smoothing functions is a powerful component of GAMs that uses localized, non-linear spline functions to approximate the relationship of with predictor . A spline is a piecewise function, such as a polynomial, that is connected at knots (or *k*-values) that smoothly join the separate pieces of the function. Increasing the number of knots creates a smoother fit of the response variable against the predictor at the risk of overfitting, whereas choosing a small number of knots creates a more rigid function that may underfit the data. There are several smoothing functions available in the *mgcv* package and the default method uses a thin-plate spline as a computationally efficient approach to estimating the smoother (Wood, 2003). The thin-plate spline can provide a balance between over- and under-fitting to the bivariate smoother by identifying the optimal number of knots through generalized cross validation (GCV). In addition to providing an overall summary statistic of model performance, the GCV optimization informs the number of knots for the spline by balancing the fit subject to a smoothing penalty. This frees the user from manually selecting knots, and therefore, the degree of smoothness modelled by the function. However, the *mgcv* package requires the user to specify an upper limit as a reasonable expection for the number of knots applied to each spline.

An additional smoother provided by *mgcv* is the cubic spline, which provides similar functionality as the thin-plate spline, but has an explicit form as a cubic polynomial function that is continuous to the second derivative at each knot location. Although visually, smooth fits for different spline types may appear similar if the number of knots are sufficiently high, the cubic spline is a useful form for modelling periodic variation, such as seasonal changes on the annual scale or diurnal variation at shorter scales. A slight modification of the cubic spline is the *cyclic* cubic spline that has the added constrainst of values being equal at the beginning and end of component being modelled. This allows the seasonal component to not only be modelled as a separate component of an overall time series, but the seasonal estimates are joined continuously across years. In others words, not using a cyclic cubic spline to model season would cause a non-continous function across the time series with step increases or decreases for the seasonal estimates between the years. This discrepancy can likely lead to inaccurate or ambiguous trend estimates when evaluating seasonal differences over time.

To evaluate GAMs with SFE chloropyll data, we followed an approach that built on past work by the Chesapeake Bay Program (Murphy et al. 2019), including incorporation and adaptation of their hierarchical model structure to evaluate long-term time series. These hierarchical models vary the functional form for time in relation to chlorophyll by adding smoothers and linear predictors with increasing flexibility in the GAM functions. This is a powerful approach for time series analysis because separate components can be explicitly modelled with different smoothing terms and the results can be idendependently evaluated and further tested from the model output. Varying degrees of model complexity also provides a balance between descriptive power and computational efficiency, where simpler models may be preferred in routine reporting or more complex models may be preferred for hypothesis testing. We present four GAM structures, following a similar format as Murphy et al. (2019). The notation for gam3, gam4, and gam5 are not used because describe models used in Murphy et al. (2019) that were developed for purposes specific to Chesapeake Bay.

For each GAM, chlorophyll is modelled using a link function that relates the response variable to the independent variables and day of year (). The variable is a continuous measure for date in decimal time, where date is expressed as a continuous numeric value with year on the left of the decimal and date within the year to the right of the decimal (e.g., July 1st 2019 would be 2019.5). The variable is an integer value for the Julian day in each year (from 1 to 366) that represents the seasonal component of the time series. The estimate is an intercept value and is a slope estimate for a linear year effect included in all models. The and functions are different smoothers for and where the former uses a thin-plate spline and the latter uses a cyclic cubic spline to capture the periodic seasonal component. The function is a tensor product interaction between and that allows the seasonal component to change between years. Within the function are the two separate smoothers for year as a thin-plate spline and as a cyclic cubic spline. As with conventional linear models, the remaining variance in the response not captured by the functional form of the model is expressed as an error term that is normally-distributed having mean equal to zero and variance as . The residual variance is used to back-transform model estimates for trend testing, described below.

The link functions for each model were chosen as either logarithmic transformations (base 10) or best estimates of Box-Cox power transformations for chlorophyll. A comparison of GAM perforamcne between transformation methods has not yet been explored in the application of GAMs to long-term water quality monitoring data. The logarithmic transformation is commonly used for chlorophyll to approximate a log-normal distribution of the response variable. However, alternative transformations could provide improved model fits and reduced uncertainty estimates in the fitted parameters by better satisying assumptions for GAMs. The Box-Cox method was also used to transform chlorophyll as a comparison to the logarithmic transformation (Box and Cox 1964). The Box-Cox method requires an estimate of the parameter , whereby the the optimal value is based on a minimization function of a log-likelihood profile vector. The profile vector for each chlorophyll time series was obtained using the *boxcox* funtion from the MASS package (Venables and Ripley 2002) to evaluate log-likelihood estimates of across the range -4 to 4. Once the optimal value was identified, chlorophyll was transformed using the following power transformation (Box and Cox 1964):

where is the transformation function for the appropriate GAM at the esitmated value. Note that in some cases the Box-Cox transformation is a logarithmic transformation if is estimated as zero. The *BoxCox* function from the forecast package (Hyndman et al. 2020) in R was used to transform the chlrophyll time series with the optimal value. Equation (6) is only appropriate for positive values of a response variable, as for chlorophyll.

A challenge to optimizing GAMs for time series data with significant intra- and interannual variation is choosing an appropriate functional form (i.e., gam0, gam1, gam2, or gam6) and determining how much variation could be explained by the smoothers in each function. The four models above provide tradeoffs in the functional forms that balance descriptive and computational efficiency. However, the individual smoothing functions (i.e, , , and ) require an explicit declaration of how much variation could be explained for the separate and components. This requires user input on the potential upper limit for the *k*-values that could be used during model fitting. Initial testing by Murphy et al. (2019) showed that the default *k* values in the mgcv package were insufficient for describing the interannual variation in the response variable and specified the upper limit for the number of knots for the smoother as the maximum between 10 or the number of years in the time series. We follow the same approach herein for gam1 and gam2. Further, the default *k*-values for and were considered appropriate and not tested further. For the application to SFE chlorophyll data, gam6 was also added as an extension of gam1, where the total number of knots for the smoother, , was increased to 12 times the total number of years in the time series. This approach potentially allows the smoother for to be completely unconstrained in how much variance is fit within and between years.

All models were compared using standard summary statistics describing overall model fit to the observed data. In addition to the GCV score, models were comapred using the Akaike Information Criterion (AIC) and summmary statistics. Comparisons were made between the different GAMs in equations (2) to (5) and the transformation used to define the function for chlorophyll at each station. Model predictions were based on the standard model output from the *predict()* function from the mgcv package. A long-term trend indepedent of seasonal variation was also estimated for each model by subtracting the seasonal terms (i.e., and ), where appropriate. Back-transformation of results were obtained by exponentiation if the log-transformed variable was used or using the *InvBoxCox()* function from the forecast package if the Box-Cox transformation was used.

## Long-term and seasonal trend analysis

* Secondary methods for trend analysis, estimate of seasonal values, hypothesis tests including mixed meta-analysis

## Sensitivity analysis

* Sensitivity analysis - do trend estimates or hypothesis test results change by model type?

# Results

# Discussion

# Conclusions

# Acknowledgments

# Figures

# Tables

# References