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 2020-04-14 18:59:57

# 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

To evaluate GAMs with SFE chloropyll data, we followed an approach that built on past work by the Chesapeake Bay Program, including incorporation and adaptation of their hierarchical model structure to evaluate long-term time series. These methods leverage functions provided by the *mgcv* R package to fit GAMs using multiple parameter smoothing estimation methods. In generals, GAMs are an extension of linear models that can incorporate additive smoothed tersm 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.

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 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 esitmating the smoother. 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. The default is ten, and as described below, this default value was increased to prevent under-fitting the data.

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, the cubic spline is a useful form for modelling period variation, such as seasonal changes on the annual scale or diurnal variation on the daily scale. 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 modeal season would cause a non-continous overall function with step increases or decreases for the seasonal estimates between the years.

Because GAM structures are flexible, the functional form for time can be described wih varying levels of complexity depending on the desired structural form. The method is useful for time series analysis because separate components can be explicitly modelled with differnet smoothing tersm.

using some of the model structures they have documented and incorporated into their R package, baytrends. Table 1 summarizes the four primary model structures tested. The first three models (gam0, gam1, gam2) have the same structure as CBP’s base models. We added a fourth model, gam6. For all models, date is expressed as cyear, or centered decimal date, meaning that a date is turned into a decimal (i.e., 2002.41), and then a time series is centered so that the middle date in a record becomes zero. doy is the day of year as a numeric value from 1 to 366. All models include year as a linear effect. The functions s() model either year or doy as a smoothed, non-linear variable and ti() models the interaction between the two in the gam2 model. A central challenge to optimizing GAMs for applications such as this, where time series data is rather noisy with significant inner- and interannual seasonal variation, is estimating how smooth (vs. ‘wiggly’) a spline should be to fit the data without over-fitting (i.e. following noise). In the mgcv package this is done by GCV. The fourth model, gam6, has the same structure as gam1 but permits a much higher degree in the s(year) by increasing the maximum allowable number of knots, k, and relies on mgcv’s internal model selection statistics to determine the appropriate degree of smoothness (using GCV).

* GAM general format and structures (gam0, gam1, gam2, gam6)
* Model application to SF time series and model evaluation, including transformation for response variables (why Box-Cox?)
* Model predictions (complete time series, annual trend) and back-transformation estimates

## 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