Evaluating the Utility of General Additive Models for Tracking San Francisco Bay Water Quality Over Time

Ian Wren, Marcus Beck, Perry de Valpine, Rebecca Murphy, and David Senn

December 2018
1. **INTRODUCTION**

San Francisco Bay (SFB) has high nitrogen and phosphorus concentrations, but has historically not experienced the eutrophication problems typical of other nutrient-enriched estuaries. However, over the past ~20 years, analysis of long-term monitoring data identified interannual shifts in the Bay’s response to nutrients – specifically, substantial increases in phytoplankton biomass in South Bay (Cloern et al. 2007, 2010).

Concerns about SFB nutrient-related water quality prompted the SFB Regional Water Quality Control Board (SFBRWQCB) and stakeholders to establish the SFB Nutrient Management Strategy (NMS; SFBRWQCB, 2012), a collaborative applied science program. Over the past several years the NMS has been pursuing a diverse set of studies, motivated by the program’s management questions (Table 1), to inform potentially-costly nutrient management decisions. A key priority for the NMS is the development of a framework for assessing water quality condition (Management Question #1).

This report provides a progress update on a pilot project that is work exploring statistical approaches for characterizing water quality trends in San Francisco Bay (SFB). We evaluated the utility of Generalized Additive Models (GAMs) for detecting trends in chlorophyll-a concentrations – including seasonal patterns, long-term trends, and interannual variability – at long-term monitoring stations in Central Bay, South Bay, and Lower South Bay having biweekly to monthly data for the period of 1992-2017.

**Table 1.1** Overarching Nutrient Management Strategy Management Questions

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What water quality conditions would be considered healthy, or protective of beneficial uses?</td>
</tr>
<tr>
<td>2. How do observed water quality conditions compare to healthy/protective conditions?</td>
</tr>
<tr>
<td>3. What are the dose:response quantitative relationships between nutrient loads and key water quality indicators?</td>
</tr>
<tr>
<td>4. If nutrient loads are negatively impacting water quality, what nutrient load reductions are needed to achieve healthy/protective conditions?</td>
</tr>
</tbody>
</table>
2. BACKGROUND, GOALS, AND APPROACH

Cloern et al (2007, 2010) detected a sharp increase in summer-fall chl-a concentrations in South Bay, beginning around 1995 and continuing through 2006-2008 (Figure 2-1). This change in chl-a focused attention on SFB’s elevated nutrient loads, including by catalyzing the launch of the NMS. It also demonstrated the importance of tracking changes over time, or trends, in nutrient-related water quality indicators.

![Figure 2-1. Interquartile range of Chl-a concentrations in South Bay. Source: Cloern et al 2007](image)

We have continued exploring trends in South Bay summer-fall chlorophyll levels (SFEI 2015, 2016, 2017). Concentration increases were not driven by a single station or month, but rather by statistically-significant increases across multiple stations and months (Figure 2.2). Continuing to track subembayment-wide chl-a through 2016, concentrations plateaued between 2005-2010, and, in recent years appear to be returning to near mid-1990s levels (Figure 2.3).

These observations raise several relevant questions for the NMS:

1. Since the initial trend was both clear to the eye and statistically-significant, when did changes cease to be significant?
2. Is the trend now negative, and is that trend statistically significant?
3. How is chl-a varying in other regions of SFB?
4. How do other relevant nutrient-related indicators changing over time? e.g., dissolved oxygen, gross primary productivity, nutrient concentrations, suspended sediment concentrations (or light attenuation).
5. Given natural variability and realistic constraints on data collection, what trend magnitudes can be detected with statistical significance? Once conditions begin changing, how long will it take to detect a sustained change?
6. What physical or biological factors could be causing or contributing to observed changes in water quality indicators?

Cloern et al (2007) and SFEI (2015; Figure 2.2) used the well-established Seasonal Kendall test to quantify the chl-a trends. While the seasonal Kendall test is useful for detecting a consistent change in one direction (i.e., monotonic: positive or negative), and can certainly be used for questions 3 and 4, it is not well-suited for exploring trends that themselves vary in direction (or become zero) over time, and so cannot help with questions 1 or 2, and would have limited applications for question 5. Furthermore, the test is not designed to also test other explanatory variables (question 6).
Figure 2.2. Left: Monthly plots of chl-a interquartile ranges from 1975-2013, subdivided into 3 eras for stations in South and Lower South Bay. Right: Trend analysis (Theil-Sen slope) over the entire time period, with blue indicating statistical significant trend (p<0.05). See station locations in Figure 3.1. Source: SFEI 2015.

Figure 2.3. Extension of the interquartile range calculations for summer-fall chl-a in South Bay through 2016, following the approach for spatially and temporally binning data described Cloern et al. 2007. Source: SFEI 2016, 2017
The Chesapeake Bay Program (CBP) has been wrestling with similar questions, in particular characterizing and testing the significance of non-monotonic changes in key water quality indicators, and exploring the importance of causal factors. Over the last several years, CBP’s Integrated Trends Analysis Team (ITAT) has been evaluating General Additive Models (GAMs) as model structures that are potentially well-suited for both of these applications. A GAM is a statistical model in which a response of interest is represented mathematically by the sum of multiple smooth functions of explanatory variables. Research applications of GAMs to water quality time series data are relatively new, but growing (Beck and Murphy 2017; Testa et al. 2018). GAMs use of smoothing splines is of particular use for seasonally-varying data that may also experience non-linear long-term changes. A smoothing spline is a method for estimating signal from noise by fitting a smooth curve through data based on relatively few assumptions. As described more below, the smoothing spline method determines the appropriate degree of smoothing based on optimizing approximate out-of-sample prediction error. For additional information regarding the application of GAMs for interpreting estuarine water quality the reader is referred to reports available on the CBP-ITAT’s website and other recent papers (Ellis et al, 2016; Beck and Murphy 2017; Murphy and Perry, 2018; Testa et al., 2018).

This project’s primary goal was to evaluate the ability of GAMs to characterize long-term water quality trends in San Francisco Bay. As an initial step, we applied GAMs to analyze the same chl-a data described above (Figures 2.1-2.3), building on approaches developed by the CBP ITAT.

3. METHODS

3.1. Data

Analysis focused on near-surface (depth = 1-3m) chl-a data from Central Bay, South Bay, and Lower South Bay (stations 18-36, Figure 1), collected biweekly to monthly through long-term monitoring by USGS (Cloern and Schraga, 2016; Schraga et al. 2017). For this initial work, we selected the time period 1990-2017 because it represented a suitable balance among three factors relevant to testing the statistical approaches: a. sufficient length of record; b. consistent biweekly-monthly sampling; and c. a diverse set of stations, across multiple subembayments, for which a and b were satisfied. While sampling frequency varied somewhat over time or by station, all data were treated as individual date/station/concentration records within the statistical models (no spatial or temporal binning or averaging).

Commented [DS2]: Need to work any pieces of this into paragraph above, which I pulled from the edits in the google drive copy document.
### Table 3.1 GAM Model Structures Evaluated for Chlorophyll-a Data

<table>
<thead>
<tr>
<th>TITLE/DESCRIPTION</th>
<th>STRUCTURE</th>
<th>MODEL ASSUMPTIONS</th>
<th>SCHEMATIC OF MODEL ASSUMPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>gam0 / Linear trend with seasonality</td>
<td>$\log(\text{chl-a}) = \text{cyear} + s(\text{doy}, \text{bs=’cc’})$,</td>
<td>A linear trend in time over the course of the time series (cyear); and</td>
<td><img src="image1.png" alt="Diagram 1" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within-year seasonal fluctuations that follow the same pattern every year - e.g. large spring bloom and small fall bloom ($s(\text{doy}, \text{bs = ’cc’})$)</td>
<td><img src="image2.png" alt="Diagram 2" /></td>
</tr>
<tr>
<td>gam1 / Nonlinear trends with seasonality (constrained k)</td>
<td>$\log(\text{chl-a}) = \text{cyear} + s(\text{cyear}, k=\text{gamK1}) + s(\text{doy}, \text{bs=’cc’})$,</td>
<td>Treats cyear and doy same as gam0; and</td>
<td><img src="image3.png" alt="Diagram 3" /></td>
</tr>
<tr>
<td></td>
<td>where gamK1 = c(10,2/3) means that the maximum of 10 or 2/3 * number of years is selected</td>
<td>A smooth non-linear trend through time, such that the seasonal peaks are similar across years, but interannual change is evident ($s(\text{cyear}, k = \text{gamK1})$)</td>
<td><img src="image4.png" alt="Diagram 4" /></td>
</tr>
<tr>
<td>gam2 / Nonlinear trend with seasonality (plus interaction)</td>
<td>$\log(\text{chl-a}) = \text{cyear} + s(\text{cyear}, k=\text{gamK1}) + s(\text{doy}, \text{bs=’cc’}) + ti(\text{cyear, doy, bs=’cc’})$,</td>
<td>Treats cyear and doy same as gam1; and</td>
<td><img src="image5.png" alt="Diagram 5" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within-year seasonal fluctuations that vary across years but in a constrained pattern compared to gam6 ($ti(\text{cyear, doy, bs = (’tp’,’cc’)})$)</td>
<td><img src="image6.png" alt="Diagram 6" /></td>
</tr>
<tr>
<td>gam6 / Nonlinear trends with seasonality (large k)</td>
<td>$\log(\text{chl-a}) = \text{cyear} + s(\text{cyear, bs = ’tp’}, k = &lt;\text{large}&gt; ) + s(\text{doy, bs = ’cc’})$,</td>
<td>Variant of gam1 with much higher values of k, allowing the splines to follow greater fluctuations both within and across years.</td>
<td><img src="image7.png" alt="Diagram 7" /></td>
</tr>
</tbody>
</table>

An estimated spline has an “equivalent” (or “effective”) degrees of freedom (edf). This is conceptualized as a measure of model complexity on the same scale as the number of parameters in a model. A user provides a maximum allowed degrees of freedom (or accepts a default value set by software), labeled k in the mgcv package, and then the edf is determined by the optimal smoothness, which is chosen by GCV. The more technical term related to maximum degrees of freedom is “basis dimension”, which determines the number of knots in knot-based bases. In order to allow GCV to determine the optimal smoothness, one needs to provide more knots than necessary (Wood 2006).
3.2. Statistical Approaches

Data analysis was performed in R (R Core Team 2014), with GAM analyses performed using the package mgcv (Wood 2018). The SFB chlorophyll-a data are log-normally distributed, and were log-transformed prior to analysis.

To evaluate GAMs with SFB chl-a data, we followed an approach that built on past work by the CBP-ITAT, including 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).

Following standard practice when comparing model fit, the most representative model for the South Bay chl-a data set was identified by calculating the Akaike Information Criterion (AIC), the generalized cross-validation score (GCV), and the R2 values. Lower values for AIC and GCV and higher values for R2 indicate improved model fit.

As part of this project, we developed a publicly-accessible web tool to aid in visualizing and exploring results from the different GAM structures (http://162.243.131.102:3838/SFbaytrends/mods.Rmd), which is periodically updated as our methodologies evolve.

4. RESULTS AND DISCUSSION

GAM evaluation began by focusing on station s32, near the study area’s southern extreme. After refining the approach through tests with s32 data, we extended the set of analyses to other stations along the transect. Below, results from s32 are examined first in greater detail, followed by an overview of model output at remaining stations.

4.1. Model Comparison Example for USGS Station 32

Chlorophyll-a levels at s32 exhibit a strong seasonal cycle, with pronounced peaks observed in Feb-Apr (Figure 4.1), but substantial interannual variability in peak magnitude. Earlier work has also shown that summer-fall chl-a concentrations at South Bay stations, including s32, began increase in the mid-1990s (Cloern et al, 2007; Cloern and Jasby 2012; Figure A.1, SFEI 2014). The four GAM structures were fitted to s32 data, with model fits depicted in Figures 4-2 and 4-3 and model selection output summarized in Table 4-1. Key observations based on visual inspection of model fits include:

gam0: The linear trend and constant (across years) seasonal fails to capture the large interannual variability and the increase-then-decrease that emerges in the other models.

gam1: While the smooth non-linear fit over time time captures the change in trend, but the constant seasonal pattern is unable to represent both the smaller peaks at the end of the time period and the large peaks early in the record.
gam2: The interaction term (ti(year, doy)) - which allows the seasonal (within-year) trend to vary smoothly across years - allows the model to capture additional variability. However, gam2 also has insufficient flexibility to represent the extreme peaks and valleys.

gam6: Although gam6 does not include an interaction term, the large degree of freedom in the time spline permits the model to capture interannual fluctuations in the seasonal patterns. Comparisons of AIC across the GAM structures indicate that gam6 is the best model for chl-a at s32, and explains a large portion of the variance ($r^2 = 0.75$).

![Figure 4.1 Location of Station 32 and time series of chl-a data (µg/L, not log-transformed)](image)

**Table 4-1 Model selection output for s32 chl-a**

<table>
<thead>
<tr>
<th>model</th>
<th>AIC</th>
<th>GCV</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>gam0</td>
<td>232.88</td>
<td>0.09</td>
<td>0.32</td>
</tr>
<tr>
<td>gam1</td>
<td>162.57</td>
<td>0.08</td>
<td>0.41</td>
</tr>
<tr>
<td>gam2</td>
<td>108.06</td>
<td>0.07</td>
<td>0.49</td>
</tr>
<tr>
<td>gam6</td>
<td>-148.96</td>
<td>0.06</td>
<td>0.75</td>
</tr>
</tbody>
</table>

### 4.1. Model Comparison at Central, South, and Lower South Bay sites

The same four GAM structures were applied to chl-a data at the remaining stations in Central, South, and Lower South Bay sites. Table 2 presents a summary of the differences between the AIC for gam2 and gam6 and gam1, which for all stations was the third best performing model. A negative difference means that the gam2 AIC was lower, therefore suggesting a “better” model. For all stations evaluated, gam2 outperformed gam1, as indicated by the negative value. However, for all stations except s18, gam6 was the best model based on AIC. Model performance statistics can be explored in greater detail for all stations using the project’s interactive website.
Figure 4.2: Time series of log(chl-a) at Station 32, 1992-2017, with model fits for gam0, gam1, gam2, and gam6.

Figure 4.3: Interannual variations in seasonal chl-a at Station 32.

Commented [DSS]: I think we will want to include versions of figure 4.2 (maybe 4.3?) in the appendix for the other stations. They can just be screenshots from the website, and I can do it if you don’t have time.
Figure 4: $\log (\text{chl-a})$ vs. time. Black dots = data; yellow curve = gam6 fits for station s32, s27, s24, s21, and s18.

Table 2. Differences in AIC-Scores Between Each Model and gam1. (Lower score indicates better predictive ability)

<table>
<thead>
<tr>
<th>MODEL</th>
<th>18</th>
<th>21</th>
<th>24</th>
<th>27</th>
<th>32</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>gam2</td>
<td>-2.97</td>
<td>-57.99</td>
<td>-79.85</td>
<td>-100.12</td>
<td>-54.51</td>
<td>-34.82</td>
</tr>
<tr>
<td>gam6</td>
<td>-0.66</td>
<td>-145.92</td>
<td>-158.12</td>
<td>-196.19</td>
<td>-311.53</td>
<td>-247.54</td>
</tr>
</tbody>
</table>
5. SUMMARY AND NEXT STEPS

Through this trend analysis pilot project, we found that a fairly straightforward GAM structure (gam6) was capable of explaining much of the variability in biweekly-monthly chl-a concentrations across multiple stations over a ~25 year period in Central Bay, South Bay, and Lower South Bay. While work to date has only scratched the surface, GAMs appear to be a promising approach for analyzing trends in SFB water quality monitoring data. In future work, we plan to continue refining GAM approaches to pursue some or all of the following:

- Develop and apply GAM-derived approaches for quantifying and testing the statistical significance of trends over time
- Extend the analyses to other regions of San Francisco Bay and to other relevant parameters (e.g., dissolved oxygen, nutrient concentrations, suspended particulate matter, gross primary production, etc.).
- Explore the feasibility of evaluating trends at the subembayment-level (aggregated or pooled)
- Incorporating other explanatory variables into GAM analyses for mechanistic analysis and hypothesis-testing/generating
References


CBP Integrated Analysis Team: www.chesapeakebay.net/who/group/integrated_trends_analysis_team


SFEI 2015 Lower South Bay Nutrient Synthesis. San Francisco Estuary Institute & Aquatic Science Center, Richmond, CA. Contribution # 732.


