Science of the Total Environment

Multi-scale trend analysis of water quality using error propagation of generalized additive models

--Manuscript Draft--

Manuscript Number:	STOTEN-D-21-10662R1
Article Type:	Research Paper
Keywords:	chlorophyll; generalized additive models; Meta-analysis; San Francisco Estuary; Trend analysis
Corresponding Author:	Marcus W Beck Tampa Bay Estuary Program UNITED STATES
First Author:	Marcus W Beck
Order of Authors:	Marcus W Beck
	Perry de Valpine
	Rebecca Murphy
	Ian Wren
	Ariella Chelsky
	Melissa Foley
	David Senn
Abstract:	Effective stewardship of ecosystems to sustain current ecological status or mitigate impacts requires nuanced understanding of how conditions have changed over time in response to anthropogenic pressures and natural variability. Detecting and appropriately characterizing changes requires accurate and flexible trend assessment methods that can be readily applied to environmental monitoring datasets. A key requirement is complete propagation of uncertainty through the analysis. However, this is difficult when there are mismatches between sampling frequency, period of record, and trends of interest. Here, we propose a novel application of generalized additive models (GAMs) for characterizing multi-decadal changes in water quality indicators and demonstrate its utility by analyzing a 30-year record of biweekly-to-monthly chlorophyll-a concentrations in the San Francisco Estuary. GAMs have shown promise in water quality trend analysis to separate long-term (i.e., annual or decadal) trends from seasonal variation. Our proposed methods estimate seasonal averages in a response variable with GAMs, extract uncertainty measures for the seasonal estimates, and then use the uncertainty measures with mixed-effects meta-analysis regression to quantify inter-annual trends that account for full propagation of error across methods. We first demonstrate that nearly identical descriptions of temporal changes can be obtained using different smoothing spline formulations of the original time series. We then extract seasonal averages and their standard errors for an a priori time period within each year from the GAM results. Finally, we demonstrate how across-year trends in seasonal averages can be modeled with mixed-effects meta-analysis regression that propagates uncertainties from the GAM fits to the across-year analysis. Overall, this approach leverages GAMs to smooth data with missing observations or varying sample effort across years to estimate seasonal averages and meta-analysis to estimate trends across years. Methods are provided in the wqtrends R package.<o:p></o:p>
Response to Reviewers:	Please view the response to reviewers included in our submission.

Cover Letter

July 28th, 2021

Dr. Damià Barceló, Dr. Jay Gan, Dr. Philip Hopke
Co-Editors-in-Chief
Science of the Total Environment

We are pleased to resubmit our manuscript, “Multi-scale trend analysis of water quality using error propagation of generalized additive models” to be considered as an original research paper in Science of the Total Environment. We sincerely thank the two reviewers for providing thoughtful comments on the written text and proposed methods. Our response to reviewers is included with this resubmission as a point-by-point response to each comment provided. We are confident that these changes have improved our manuscript and are hopeful that is now appropriate for publication.

Sincerely,

Dr. Marcus W. Beck
Program Scientist
Tampa Bay Estuary Program

Multi-scale trend analysis of water quality using error propagation of generalized additive models

Marcus W. Beck* (mbeck@tbep.org), Tampa Bay Estuary Program, St. Petersburg, FL

Perry de Valpine (pdevalpine@berkeley.edu), University of California Berkeley, Berkeley, CA

Rebecca Murphy (rmurphy@chesapeakebay.net), University of Maryland Center for Environmental Science, Annapolis, MD

Ian Wren (ianw@sfei.org), San Francisco Estuary Institute, Richmond, Ca

Ariella Chelsky (ariellac@sfei.org), San Francisco Estuary Institute, Richmond, Ca

Melissa Foley (melissaf@sfei.org), San Francisco Estuary Institute, Richmond, Ca

David B. Senn (davids@sfei.org), San Francisco Estuary Institute, Richmond, Ca

*Corresponding author

We sincerely thank the reviewers for providing thoughtful comments on our manuscript. We have provided a point-by-point response to these comments below. We are confident that these additions and revisions improve our manuscript for readers of Science of the Total Environment.

In addition to making revisions that address points raised by the reviewers, we have also taken the opportunity to make minor improvements that we identified.

Reviewer 1

This manuscript is a welcome addition in the world of trend analyses. The authors propose combinations of GAM and mixed-effects meta-analysis regression to account for uncertainty. The methodology is sound and results are convincing. My recommendation would be even more enthusiastic if the authors had formatted their methods sections in a more traditional manner. I am of the opinion that models should be presented using clear mathematical equations and not in R code semantics. This comments and a few other can be found in the attached document.

- **Response**: We really appreciate these comments that affirm our opinion on the value of this work for trend analysis of long-term water quality data. We have changed our formulas and text to a conventional format, as opposed to R code notation. Briefly, here are the updated equations for the models described in the methods. All other instances in the main text referring to R code (except in a few locations) have been changed. Also please note that the attached pdf that was included with the reviewer's response did not include any markup comments. This is contrary to what the reviewer indicated.

$$\text{Model S: } y_i \sim \beta_0 + f_1(cont\_year_i) + \epsilon_i \qquad (1)$$

$$\text{Model SY: } y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + \epsilon_i \qquad (2)$$

$$\text{Model SYD: } y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + f_2(doy_i) + \epsilon_i \qquad (3)$$

$$\text{Model SYDI: } y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + f_2(doy_i) + f_3(cont\_year_i, doy_i) + \epsilon_i \qquad (4)$$

Reviewer 2

Overview

This manuscript presents a new application of GAMs approaches to fit irregular time series data (gaps, changing frequency of measurements) to estimate trends. The paper presents a case study using Chl a data from San Francisco Bay. The manuscript is presented as a general approach with applications to other similar environmental time series data. A comparison of methods shows that the GAMs plus meta-analysis method provides some different trend conclusions from ordinary least squares and plain GAMs. The paper is generally well written and well-suited for the audience for STOTEN. I have some clarifications that would help improve/clarify the utility and readability of the manuscript. These are presented below.

- **Response**: Thank you for your comments on our manuscript. We have addressed your concerns on the utility and readability of our manuscript below.

Utility (1):

We applied the wqtrends code in R to a time series data set of X2 (salinity intrusion measure) over approximately in San Francisco Bay. Some observations:

-the code did not work with the full time series of daily X2 data. But we were able to make it work using a random sampling of 5 points per month. Could the reason for this error be explored? As time series data are concerned, we provided an input of ~100 years of daily data. A sample result for the trends are shown in the figure below. I think it would be helpful for readers/users to understand what the practical limitations of the code/algorithm are.

- **Response**: Very rarely do reviewers actually apply the proposed methods on novel datasets as a proof of concept. For that, we are very appreciative of your efforts in exploring the R package. These applications are important for testing both the method and software. As such, you exposed an important limitation of our method that was not clearly articulated in the draft. We provide two responses to this limitation.

  First, the X2 dataset is at the daily scale, so one aspect of the GAM smoothing methods proposed in our manuscript for "irregularly spaced or missing" monitoring data is unnecessary. The GAMs provide two benefits: smoothing irregularly spaced data and estimating the non-independence of sequential data so that propagated uncertainty does not incorrectly assume sequential data are independent. The modelling approach is designed to provide both a continuous (e.g., daily) estimate of the long-term trend and a measure of uncertainty in the absence of more regularly collected values. So, application of our method to datasets with sampling intervals longer than the daily scale would be the most compelling use case, but they are nevertheless relevant for uncertainty propagation even with daily data. However, because some of the daily X2 values are in fact imputed, it would be appropriate to use the raw data only in our method, and possibly to enrich the GAM with whatever domain-specific model is used for smoothing X2 values, in order to fully propagate uncertainty.

  Second, and more important, the practical limitations of applying a GAM with flexible smoothing to nearly 100 years of daily data is beyond the computational power of most desktop computers. This is asking too much of the smoothing splines used in the mgcv package. Given this limitation, we do feel it is important to clearly state the intended use case for these models to avoid confusion in application. We have added clarification both in the manuscript and the package vignette to clearly indicate the type of data that is appropriate for the methods. In practice, if one wants to analyze such a long series, one could break it down into multiple (possibly overlapping) shorter series for the first two stages (GAM estimation and seasonal averaging with uncertainty propagation). The only difference would be that the estimated optimal degree of smoothness might change across different time windows, but that may be appropriate and supported by having so much data.

  In the manuscript, lines 194 - 196: "Approximate monthly or biweekly sampling with coverage of at least a decade is common for many long-term monitoring programs and is the motivating use case for the methods herein."

In the package vignette, first paragraph (https://tbep-tech.github.io/wqtrends/articles/introduction.html): "These models are appropriate for data typically from surface water quality monitoring programs at roughly monthly or biweekly collection intervals, covering at least a decade of observations (e.g., Cloern and Schraga 2016). Daily or continuous monitoring data covering many years are not appropriate for these methods, due to computational limitations and a goal of the analysis to estimate long-term, continuous trends from irregular or discontinuous sampling."

We have also added a paragraph in the discussion regarding limitations of the approach, one of which is related to the issues above:

Lines 532 - 541: "Several limitations of the proposed methods deserve mention. First, if sampling is so irregular that important fluctuations are missed entirely in some years, the GAM estimates and uncertainty propagation could become dubious in interpretation and usefulness. Second, estimation of GAMs for very long series can be computationally demanding. When this is an obstacle, one could do the first two analysis stages using temporal windows of the full data, with the only implication being that different degrees of smoothness may be estimated for different windows, which indeed might be justified by the data. Third, meta-analysis regression results for a very small number of years, particularly confidence intervals and associated p-values, may be inaccurate (e.g., in confidence interval coverage). In such cases, one could make alternative use of the GAM seasonal averages and standard errors, such as for pairwise comparisons among years."

Utility (2):

From figure 7 it seems that there are very different conclusions for Station 36 depending on the method used. From Figure 8, however, it seems that Station 36 is a bit of an anomaly in this regard, and that other 8 stations in the case study have findings on trends that are similar among the three methods—and that the results are not dramatically different. Perhaps this should be the basis for added discussion in the final section—when is the added complexity of the GAMS and meta analysis method is appropriate in a real-world setting? Alternatively, it would be helpful if the authors were to examine another data set from the region that they work with, to review the relative findings of trend significance and what this depends on.

- **Response**: You are correct to note that many of the comparisons had similar results and we agree that it is useful to clarify that our approach is warranted relative to more conventional methods. If a simpler approach is used, there is no way to know for sure that an invalid result is obtained without applying our proposed methods. As such, we argue that using our approach is still the wiser option in all cases. Otherwise, a user cannot be certain the correct conclusions are obtained using simpler methods. The paragraph on lines 498 to 513 was revised to address these points:

  "Incorrect conclusions on trends can have dramatic consequences for regulated parties under existing water quality compliance frameworks (Smith et al., 2001). Our examples in Figures 7 and 8 demonstrate these risks if propagation of uncertainty from raw observations across methods is unaccounted for in trend assessment. The 'naïve' method using OLS regression applied to seasonal averages from the raw observations fails to propagate uncertainty, similarly to averaging results within a year and applying a simple

Kendall test. In some cases the results may be similar to those from fully propagating uncertainty, but the loss of information can lead to increased Type I or II error rates depending on characteristics of the raw data and the method used for their evaluation (Shabman and Smith, 2003). Our examples demonstrated the increased potential for incorrect conclusions at specific monitoring locations and, at larger spatial scales, across all stations if simpler trend analyses are used. Even though simpler methods may produce similar results in some cases, particularly with frequent sampling and similar effort between years, the only way to confirm such an outcome would be to compare results, relying on the method with full propagation of uncertainty to be the more robust method. Use of methods that fully account for uncertainty is recommended to obtain statistically valid results in a wider range of conditions."

We also want to mention that we recently discovered a minor bug in the seasonal feature extraction of wqtrends that prevented an estimate for years on the tail ends of the time series with incomplete data. Because of this, we have updated Figure 7 and 8, now using stations 30 and 34 as examples in the manuscript. None of the conclusions change in the manuscript.

Readability/Clarity:

The term meta-analysis seems to be an important part of the paper and is referred to multiple times, but it is never clearly explained in the manuscript. For example, at line 164. See the following sentences.

"Third, we used a mixed-effects meta-analysis to estimate trends and test hypotheses about the change in seasonal averages across years. While 166 meta-analysis methods arose from analyses of results from multiple studies, their distinguishing characteristic is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). Meta-analysis uses response data that includes standard errors (uncertainties) as needed to address our questions."

It is not clear to me what exactly is meant by meta-analysis and what is being tested in the context of the time series data. An explanation that is understandable by a general audience of potential users of this method would be helpful. It does seem likely, as that sentence points out, that most readers would be more familiar with meta-analysis methods in the context of aggregating the results of multiple studies. There are other helpful citations afterwards, but a more explicit connection between the familiar use case and the use case here of combining different seasonal features from a single regression model would have been helpful for reading this work.

- **Response**: Thank you for pointing out this ambiguity. We agree that a distinction between the more common understanding of "meta-analysis" and the one applied here needs to be made early in the manuscript. We have added content in the introduction to clarify our methods.

  Lines 125 - 135: "Meta-analysis regression incorporates a known (or estimated) standard error for each response datum. Usually meta-analysis is used when each response summarizes a dataset from a separate study, along with its standard error, and the meta-analysis looks for across-study patterns in effect size (i.e., Lortie et al. 2014). Here each response summarizes one year or season of data, with standard error from the GAM, and

the meta-analysis looks for patterns across years. Thus, while meta-analysis methods are most commonly associated with combining results from multiple studies into a larger analysis, their key modeling step is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). To do this, meta-analysis makes use of a known (estimated) standard error for each response datum, which is the required priority here to propagate standard errors from the GAM into a regression of seasonal averages."

Abstract

~~Accurate and flexible trend assessment methods are valuable tools for describing historical changes in~~Effective stewardship of ecosystems to sustain current ecological status or mitigate impacts requires nuanced understanding of how conditions have changed over time in response to anthropogenic pressures and natural variability. Detecting and appropriately characterizing changes requires accurate and flexible trend assessment methods that can be readily applied to environmental monitoring datasets. A key requirement is complete propagation of uncertainty through the analysis. However, this is difficult when there are mismatches between ~~time scales~~sampling frequency, period of ~~monitoring data~~record, and trends of interest. Here, we propose a novel application of generalized additive models (GAMs) ~~to model seasonal and~~ for characterizing multi-decadal changes in water quality indicators and demonstrate its utility by analyzing a ~~long-term monitoring dataset~~30-year record of biweekly-to-monthly chlorophyll-a concentrations in the San Francisco Estuary. GAMs have shown promise in water quality trend analysis to separate long-term (i.e., annual or decadal) trends from seasonal variation. Our proposed methods estimate seasonal averages in a response variable with GAMs, extract uncertainty measures for the seasonal estimates, and then use the uncertainty measures with mixed-effects meta-analysis regression to quantify inter-annual trends that account for full propagation of error across methods. We first demonstrate that nearly identical descriptions of temporal changes can be obtained using different smoothing ~~splines for annual or seasonal components~~spline formulations of the original time series. We then extract seasonal averages and their standard errors for an *a priori* time period within each year from the GAM results. Finally, we demonstrate how across-year trends in seasonal averages can be modeled with mixed-effects

meta-analysis regression that propagates uncertainties from the GAM fits to the across-year analysis. Overall, this approach leverages GAMs to smooth data with missing observations or varying sample effort across years to estimate seasonal averages and meta-analysis to estimate trends across years. Methods are provided in the *wqtrends* R package.

Key words: chlorophyll, Generalized Additive Models, meta-analysis, San Francisco Estuary, Trend analysis

Introduction

Accurate quantification of environmental trends must consider variation at different temporal scales when ignoring variation at one scale could lead to incorrect conclusions about variation at another scale. Many environmental monitoring programs collect temporally resolved but irregular time series data to quantify trends for regulatory, management, or research purposes. The mismatch between the scales of monitoring versus analysis questions or management goals can present statistical challenges (Cumming et al., 2006; Forbes and Xie, 2018; Urquhart et al., 1998). At short temporal scales typically less than a year, environmental systems exhibit variability caused by multiple factors (e.g., weather events, management, or seasonal changes). Such fluctuations may not be related to inter-annual trends or may not be well-suited to multi-scale smoothing methods~~.~~, yet they can perturb the time-series such that the sub-annual fluctuations must be addressed within the trend analysis to allow accurate quantification of inter-annual trends. Many trend analysis methods lack the flexibility to evaluate ~~one to many~~multiple independent variables in an extendable structure that accommodates hypothesis testing at different time scales of interest.

In this paper, we develop methods to estimate across-year trends of within-year features, such as seasonal averages, while accounting for uncertainties across analysis steps. Our overarching goal for this work was to develop a flexible set of tools for accurately characterizing inter-annual changes in seasonally-averaged water quality metrics that can be robustly applied to diverse time-series data. At the outset, we identified several specific requirements and priorities for the trend analysis methods: 1) Complete propagation of uncertainty through the analysis; 2) Separation of trends on different time scales; 3) Ability to estimate linear and nonlinear responses; 4) Flexibility to evaluate multiple independent variables (and random effects), although the examples herein include time as the only independent variable; and 5) Robust to missing observations or varying sampling effort across years, which was considered a high priority to allow the methods to be applied to diverse time-series datasets and monitoring programs. Existing methods that begin to address some of the above requirements and priorities can be generalized into four groups: seasonal Kendall tests (and other non-parametric tests); seasonal trend decomposition using loess (STL); weighted regression on time, discharge, and season (WRTDS); and generalized additive models (GAMs).

~~Existing methods that begin to address our objectives in water quality trend analysis can be generalized into four basic approaches: seasonal Kendall tests, seasonal trend decomposition (STL), weighted regression on time, discharge, and season (WRTDS), and generalized additive models (GAMs).~~ Seasonal Kendall tests ~~or~~and related non-parametric approaches have been used for decades in water quality trend assessments to identify inter-annual, monotonic changes ~~over several years~~ while accounting for ~~the~~ predictable patterns among seasons (~~Helsel et al., 2020~~Cloern et al., 2007; Helsel et al., 2020; Hirsch et al., 1982). ~~Wan et al. (2017) showed that~~ While seasonal Kendall and other non-parametric approaches have been among the most

Formatted: First Paragraph

commonly used methods in long-term water quality trend analysis ~~despite critical limitations.~~(Wan et al., 2017), they do not satisfy several of our requirements. For descriptive decomposition of long-term monitoring data, ~~they~~ these approaches assume seasonal patterns within years do not change~~,~~ and require regularly spaced or balanced data~~,~~. In addition, seasonal Kendall tests do not ~~include~~allow for additional ~~predictors~~independent variables to explain variation, ~~and~~ do not estimate a model that could be useful for other purposes~~. Thus, while these non-parametric approaches have some degree of robustness to assess magnitude~~ (e.g., prediction), and ~~direction~~do not easily allow for propagation of ~~trends, they apply only to narrow goals~~uncertainty to other trend analysis methods.

~~The seasonal trend decomposition using loess (STL)~~STL decomposes a time series into additive components of a long-term trend, a seasonal pattern, and residuals (Cleveland et al., 1990; Cloern, 2018; Cloern and Jassby, 2010; Stow et al., 2015). While useful and widely applied, this method ~~also has important limitations.~~does not address all of our requirements. STL decomposition does not ~~incorporate~~allow for incorporating explanatory variables ~~besides~~other than time. In addition, it is ~~defined~~often characterized more as an algorithm of statistical steps than as a ~~coherent~~ statistical model with estimated parameters (e.g., Wan et al., 2017~~),~~) and it does not usually estimate standard errors to allow hypothesis testing (but see Hafen, 2010). STL methods may also over-simplify trends into ~~stationary~~fixed components that do not change over time, e.g., a seasonal estimate that is constant across years. This limitation presents challenges when addressing questions relevant to long-term water quality data, such as timing of seasonal peaks that can suggest system response to changing environmental conditions (Cloern and Jassby, 2010; Navarro et al., 2012).

The weighted regression on time, discharge, and season (WRTDS) method addresses the problem of inflexibility in STL by using a more general local regression scheme (Beck et al., 2018; Beck and Hagy, 2015; Hirsch et al., 2010~~);~~.; Hirsch et al., 2015). Designed for evaluating water quality in rivers where separating the effect of discharge on constituent concentration is important, WRTDS estimates a moving window regression model with components that allow parameters to vary smoothly in relation to both time and discharge. This yields parameters that are specific to season, year, and flow regime. The WRTDS approach is conceptually similar to local kernel smoothing methods, with specific application to explanatory variables relevant for water quality constituents (i.e., season, year, and discharge). Standard error estimates of predictions from WRTDS are available through a block bootstrap approach applied to the model results (Hirsch et al., 2015). Although a useful addition to the original method (Hirsch et al., 2010), the approach requires extensive resampling using a previously fitted model. Alternative methods that include standard error estimates simultaneously with model output may be preferred for intensive or more iterative applications.

~~Generalized additive models (~~Finally, GAMs~~) are~~ can satisfy the requirements and priorities identified above and were adopted as a central ~~to this paper and form the basis of~~component for the ~~fourth method to separate fluctuations on different time scales.~~trend analyses herein. GAMs combine one or more smoothing splines to model patterns in data and ~~may~~can be ~~seen~~reasonably viewed as generalizing the concepts behind STL and WRTDS (Haraguchi et al., 2015; He et al., 2006; Morton and Henderson, 2008; Murphy et al., 2019; Pearce et al., 2011). The basis functions used to formulate GAMs can be customized based on expected patterns in the data. ~~One example includes~~Examples include cyclic splines, which can be used to model seasonal patterns, and low-dimensional interactions~~.~~ (Wood, 2017). GAMs have added flexibility because

they can include both parametric (e.g., linear or quadratic) components and non-parametric (spline) components. Multiple approaches have been developed to determine the optimal degree of smoothness~~.~~ (Wood, 2004; 2017). These approaches are based on optimization of out-of-sample prediction error, which addresses a key concern around methods like WRTDS that do not have analogs for choosing optimal degrees of smoothing. GAMs can also produce results comparable ~~results similar~~ to those provided by WRTDS (Beck and Murphy, 2017) and have readily obtainable uncertainty estimates. Further, GAMs have natural frequentist and Bayesian interpretations, are naturally extensible to include random effects (i.e., generalized additive mixed models or GAMMs), and have computationally efficient implementations (Wood, 2017).

GAMs have been applied previously to evaluate trends in water quality time series from long-term monitoring programs (Haraguchi et al., 2015; Murphy et al., 2019). For example, Murphy et al. (2019) used GAMs to decompose water quality time series from Chesapeake Bay into long-term and seasonal trends (Murphy et al., 2019) and test trend hypotheses between two points in time. Other studies of environmental time series with GAMs have addressed the use of transformed response data (Yang and Moyer, 2020), serial correlation in high resolution data (Morton and Henderson, 2008; Yang and Moyer, 2020), and quantifying time lags in relationships between response and predictor variables (Lefcheck et al., 2017). The ~~study~~method development and analyses described herein ~~generalizes~~generalize the approach to analyzing trends of seasonal spline features, describes the relationships among alternative spline formulations when spline flexibility is allowed to vary (Wood, 2017, 2003) rather than being constrained *a priori* for different time scales, and prioritizes full incorporation of uncertainty.

~~Our motivating problem has several characteristics that are only partially addressed by previous methods and can further build on GAMs as a starting point. Our general goal is to understand~~

~~interannual changes in seasonally averaged water quality metrics, such as chlorophyll. However, the seasonal average within each year must be robust to inconsistent sampling times and intervals, and any trend analysis must consider the uncertainties in seasonal averages. The critical need is the ability to obtain an accurate estimate of uncertainty (e.g., a standard error) of seasonal averages, even with irregular sampling and serial correlation, which is common in time series data. This paper develops the use of GAMs with mixed-effects meta-analysis (~~To incorporate the uncertainty of seasonal estimates into trend analysis, we integrated GAMs with mixed-effects meta-analysis (Gasparrini et al., 2012; Sera et al., 2019). In this integration, the GAMs framework addressed a critical need by providing an estimate of uncertainty (e.g., a standard error) of seasonal averages, even in situations with irregular sampling and serial correlation, which are common in time series data. Meta-analysis regression incorporates a known (or estimated) standard error for each response datum. Usually meta-analysis is used when each response summarizes a dataset from a separate study, along with its standard error, and the meta-analysis looks for across-study patterns in effect size (i.e., Lortie, 2014). In this study, each response summarizes one year or season of data, with standard error from the GAM, and the meta-analysis looks for patterns across years. Thus, while meta-analysis methods are most commonly associated with combining results from multiple studies into a larger analysis, their key modeling step is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). ~~)~~ to address multi-scale trend analysis questions for which seasonal Kendall tests and the more complex STL and WRTDS methods are not well-suited.~~

~~We~~To do this, meta-analysis makes use of a known (estimated) standard error for each response datum, which is the required priority here to propagate standard errors from the GAM into a regression of seasonal averages.

To describe the approach and demonstrate ~~the proposed methods by analyzing water quality monitoring data from~~its utility, we analyze a 30 year record (1990-2019) of biweekly-to-monthly chlorophyll-a concentration data, collected at 9 stations in the southern portion of the San Francisco Estuary, California, USA. Approximately twice-monthly monitoring has been conducted for several decades at fixed locations (stations) ~~on~~along the longitudinal axis of the Bay. Analysis of these data is complicated by irregularities in timing and consistency of data collection~~, which can generate artifacts affecting simple seasonal averages of the data.~~. We were interested in questions such as: Are there significant trends in spring mean chlorophyll at multi-year time-scales? At what across-year ~~scale do~~window does summer-fall mean chlorophyll levels change? Is there a spatial difference in chlorophyll trends? We provide examples illustrating how these questions can be addressed using GAMs to estimate seasonal ~~trends~~patterns and ~~evaluated between years using~~use meta-analysis ~~methods. This approach is new~~ to ~~environmental trend-detection problems~~evaluate trends between years. The techniques are incorporated into an open-source and ~~is provided in the *wqtrends*~~publicly available R package, *wqtrends*, developed by the authors (Beck et al., 2021, available at https://tbep-tech.github.io/wqtrends, including an online dashboard for viewing results at ~~https://nutrient-data.sfei.org/apps/SFbaytrends/~~https://nutrient-data.sfei.org/apps/SFbaytrends/).

Formatted: Default Paragraph Font

Methods

Study area and data sources

The San Francisco Estuary (SFE) is the largest estuary on the Pacific Coast of North America~~.~~ ~~Its~~, and its watershed covers 200 thousand km^2 in the US state of California. ~~Major freshwater~~

~~inputs enter the system through~~ Flows from the Sacramento-San Joaquin Delta ~~complex upstream of Suisun~~, entering from the northeast, account for the vast majority of SFE-wide annual-average freshwater inputs (Cloern and Jassby, 2012). Freshwater contributions to southern SFE (South Bay~~.~~, Lower South Bay) come primarily from local tributaries during the wet season (Nov-Apr), and from wastewater treatment plant discharges during the dry season (May-Oct). Salinity ~~ranges from 0 to 15 ppt~~values in ~~the northern~~southern SFE subembayments ~~and~~ (the focus of this work; Central Bay, South Bay, Lower South Bay) range from 5 to 35 ppt ~~in southern subembayments closer to the Pacific Ocean, depending~~and depend strongly on ~~the~~ season (stormwater runoff), tidal cycle, and effluent discharge from wastewater treatment plants~~,~~ ~~and stormwater runoff (~~ (Cloern and Jassby, 2012). ~~An estimated 73.8 metric tons dy^{-1}~~ SFE receives 70,000 kg per day of dissolved inorganic nitrogen ~~are discharged into the Bay, primarily from wastewater (Novick and Senn, 2014). Agricultural runoff from~~ (DIN; annual average), with the ~~upper watershed contributes 30 metric tons dy^{-1} of nitrogen to the SFE via~~ majority of that DIN coming from wastewater treatment plant discharges. Flows from the Delta~~.~~ ~~Nitrogen and phosphorus levels in~~ deliver 30,000 kg per day of DIN to the SFE (annual average), with the ~~SFE usually exceed concentrations that cause eutrophication in other~~ Delta's DIN load varying over a 5-fold range annually (SFEI, 2014a). Based on its areal DIN loads, SFE ranks among the most nutrient-enriched estuaries~~. However~~ worldwide (SFEI, 2014a, b; Cloern et al., 2020). Despite its nutrient-enriched status, the SFE has ~~demonstrated resistance to eutrophication, which has been~~generally not experienced some of the water quality impacts common to other nutrient-enriched estuaries (e.g., excessive phytoplankton blooms, low dissolved oxygen), with SFE's muted response attributed to ~~high concentrations of suspended sediment that reduce~~its highly turbid waters (reduced light penetration ~~in~~within the water

column~~, low residence time caused by vigorous river flushing, and removal of primary producers by~~); strong tidal mixing (limiting duration of water column stratification to less than several days); and strong phytoplankton grazing pressure from abundant suspension feeding bivalves in some regions (Alpine and Cloern, 1988; Cole and Cloern, 1984; Jassby, 2008; Kimmerer and Thompson, 2014~~; Lehman et al., 2017).~~).

Studies over the past decade have identified changes in responses or sensitivity within SFE to nutrients in deep subtidal habitats via increased phytoplankton biomass (chl-a) and gross primary production (GPP) in South Bay (Cloern et al., 2007, 2010); recently documented occurrences of harmful algae and their associated toxins (Sutula et al., 2017; Peacock et al., 2018); and low dissolved oxygen in some tidal slough habitats (SFEI, 2021). These observations have raised concerns that SFE's resistance to its high nutrient inputs could be waning (SFEI, 2014b), prompting regulators to initiate the SFB Nutrient Management Strategy (SFBRWQCB, 2017).

The ~~Regional Water Quality Control Board has showed renewed interest in~~ early increases in South Bay chl-a (1995-2005) were quantitatively tested using seasonal Kendall, and the signal was sufficiently large and coherent it was also visually apparent in raw data. Given SFE's nutrient-enriched status, there is a critical need for on-going and comprehensive characterization of trends in chl-a, including the ability to examine variability at multiple time-scales and non-monotonic trends, that can also be readily applied to other nutrient-related indicators (e.g., dissolved oxygen, GPP). Beyond the role these tools can play in supporting improved understanding ~~the potential for nutrient loading to negatively affect water quality for more southern areas of the SFE where harmful algal blooms, elevated summer fall~~of system dynamics in SFE, water quality managers have emphasized the importance of robust trend detection for informing future nutrient management decisions.

Formatted: Font color: Custom Color(RGB(34,34,34)), Pattern: Clear (White)

Formatted: Font color: Custom Color(RGB(34,34,34)), Pattern: Clear (White)

For the trend analyses discussed below, we used near-surface (0-2 m) chlorophyll concentrations~~,~~ ~~and low dissolved oxygen concentrations began around 1999~~ (Figure 1) ~~(Cloern et al., 2020). Although changes in the data are visually apparent, statistical analyses to quantify these changes have been insufficient particularly with respect to seasonal differences between years.~~

~~We evaluated near-surface chlorophyll (chl-a) data~~ measured biweekly to monthly from 1990 to 2019 along the longitudinal axis of the SFE extending from Central Bay (stations 18-23), South Bay (stations 24-32), and Lower South Bay (stations 34-36) (Table 1, Figure 2). Monitoring data were obtained from the SFE Research Program of the US Geological Survey (Cloern and Schraga, 2016; Schraga et al., 2020). Sampling frequency varied somewhat over time and by station. Approximate monthly or biweekly sampling with coverage of at least a decade is common for many long-term monitoring programs and is the motivating use case for the methods herein. Every observation was included directly in the statistical models without spatial or temporal binning or averaging. Log_{10}-transformed chl-a was used for all analyses to meet assumptions of normally-distributed residuals. Methods for back-transformation of model results are provided in the supplement.

GAMs with uncertainty propagation

We implemented our analysis in three stages. First, we used a GAM to estimate a smooth temporal pattern in the raw data, along with ~~its~~the uncertainty of the smoother. Second, we calculated a feature of interest from the estimated GAM, along with its propagated uncertainty. For ~~this example,~~ the examples described here, we focused on extracting seasonal averages ~~were extracted, whereas other~~of chl-a values. Other features ~~could be~~can also be extracted using the same tools, including the timing or magnitude of a seasonal peak, but those are not ~~developed~~

~~here.~~presented here (see the *wqtrends* R package, https://tbep-tech.github.io/wqtrends). Third, we used a mixed-effects meta-analysis to estimate trends and test hypotheses about the change in seasonal averages across years. ~~While meta-analysis methods arose from analyses of results from multiple studies, their distinguishing characteristic is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). Meta-analysis uses response data that includes standard errors (uncertainties) as needed to address our questions.~~

First-stage analysis: GAM estimation

~~We considered four different GAMs to~~To smooth the raw data across time~~. Although they~~, we considered and tested four different GAM structures. While all four GAM structures can achieve similar fits, they ~~do so by partitioning~~differ in how they partition variation in the time series ~~differently~~ (Table 2~~).~~), which may unnecessarily influence understanding of temporal patterns. We discuss all four to clarify their relationships and interpretations. ~~Models are shown in the notation of~~ All models were created using the `mgcv` R package ~~as formulas for the `gam` function~~ (R Core Team, 2020; Wood, 2017), with utility functions included in the *wqtrends* package created by the authors (Beck et al., 2021).

The simplest GAM for this purpose is expressed as:

~~Model S: `y ~ s(cont_year, k = num_knots_Y)`~~

$$\text{Model S: } y_i \sim \beta_0 + f_1(cont\_year_i) + \epsilon_i \qquad (1)$$

where ~~*y*~~*y* is ~~the time series of interest, such as~~measured chl-a, ~~`cont_year`~~β_0 is an intercept, and *cont_year* is "continuous year," a continuous numerical date (e.g., July 1st 2019 would be 2019.5~~), `y ~ s(...)` indicates that `y` will be explained by~~). The $f_1()$ function is a smoothing

Formatted: First Paragraph

spline ~~(in this case as a function~~ composed of the sum of ~~cont_year), and num_knots_Y~~ ~~is~~multiple "basis functions" multiplied by coefficients. $f_1(cont\_year)$ describes the ~~number~~relationship of ~~knot or "connections" along~~y with $cont\_year$ in a way that smoothly follows the ~~spline~~data (Wood, 2017). The basis functions involve user-specified knots, a grid of values on the $cont\_year$ axis that is discussed more below. The ϵ term represents residuals following a normal distribution with mean zero and constant variance.

Smoothing was determined using generalized cross-validation (GCV, as implemented in `mgcv`), which ~~approximately~~ minimizes out-of-sample prediction error. GCV works by penalizing the net curvature of a spline (Wood, 2004). To allow GCV (or other alternatives) to work as intended, the number of knots ~~that~~chosen by the analyst, which determine the maximum degrees of freedom ~~chosen by the analyst,~~ must be sufficiently large so that the curvature penalty, rather than the number of knots, determines smoothness. Results should not be sensitive to the number of knots; if they are, the number of knots should be increased. In the examples below, we chose the number of knots~~, num_knots_Y,~~ for $f_1()$ as 12 times the number of years in the time series, i.e., one knot per month. If the data were too sparse to fit 12 knots per year, the number of knots was reduced by one knot per year until the model could be estimated (i.e., 12 * years, 11 * years, etc.).

The next three spline formulations (Model SY, SYD, and SYDI) provide progressively increasing complexity in how spline terms compose a model to smooth the raw data. Model SY describes the time series using a linear trend plus a spline for ~~cont_year~~$cont\_year$:

~~Model SY: y ~ cont_year + s(cont_year, k = num_knots_Y)~~

~~This model is~~Model SY: $y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + \epsilon_i \quad (2)$

Formatted: Comment Reference

where equation (2) is the same as equation (1) with the addition of a linear term for $cont\_year$ related to y_i by the β_1 slope parameter.

While Model SY contains the explicit linear trend term, $\beta_1 cont\_year_i$, it is in fact mathematically equivalent to model S (Table 3). The $f_1()$ spline for *cont_year* includes an unpenalized linear trend, so a trend will be estimated in model S. When *cont_year* is included explicitly as a linear term in model SY, `mgcv` adjusts the basis functions for the spline to exclude the linear term, thereby not over-parameterizing the model. Whereas an estimated linear trend in *cont_year* and its uncertainty can be extracted from the fitted spline in model S, model SY provides this trend directly, giving the equivalent result. Further, package `mgcv` ~~can~~offers the option to penalize linear trends in splines to provide a method for variable selection (option `select = TRUE`), such as when numerous splines are included in the model formulation for variables that may or may not be important. ~~For~~In the implementation of our approach described here, this option is not used and all models specify `select = FALSE`. Details in the supplement explain this justification.

Model SYD adds an average within-year cyclic pattern as a separate spline:

~~Model SYD: `y ~ cont_year + s(cont_year, k = num_knots_Y) + s(doy, bs = 'cc', k = num_knots_D)`~~

$$\text{Model SYD: } y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + f_2(doy_i) + \epsilon_i \quad (3)$$

where ~~doy~~equation (3) is the same as equation (2) with the addition of a smoothing spline for "day-of-year" (doy, i.e., Julian date, a count starting January 1 for each year~~), bs = 'cc' indicates that the spline will be~~). For $f_2()$, a cyclic ~~(constrained to~~ spline is specified (using `bs = 'cc'` in `mgcv`) to constrain the start and end at the same value~~), and num_knots_D is the~~. A user-

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: First Paragraph

specified number of knots ~~for the doy spline.~~is also included in $f_2()$. While model SYD is not mathematically equivalent to models S and SY, it should produce nearly identical results. The *doy* spline in model SYD gives the average within-year pattern and changes the interpretation of the *cont_year* spline to represent smoothed deviations from that pattern.

Models S, SY, and SYD can all potentially extract a similar signal from the raw data (Table 3). What differs between the models is the allocation of penalties for curvature used to determine smoothness for each spline. In model SYD, there are separate penalties for the two splines, as compared to S and SY that include penalties only for the *cont_year* spline. This is important because variation in the response variable can be differently attributed to each spline depending on the model, even while the sum of components for each model produces similar results between models. Our goal is to extract seasonal averages from the fitted time series~~, which is~~ that are not sensitive to different allocation of penalties among the splines in each model.

If the fits were to differ substantially between model SYD and models S or SY, an interpretation could be difficult because the penalties for smoothing splines based on curvature are heuristic (Wood, 2017). For example, if a lower AIC is achieved in one model compared to another, assuming both use sufficient knots, this may just reflect the outcome of alternative penalization heuristics implied by the different formulations and does not imply one model fit is better. In the examples here, model SYD achieves nearly identical fits to model S or SY, where the latter by definition also achieve identical fits.

Model SYD has the appealing feature that, if some parts of some years have limited data, model SYD will impute an average seasonal pattern with the *doy* spline, thereby considering data from the same period in other years in the prediction of the period with missing data. However, an

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

interpretation of these imputations may be challenging. For example, the spring chl-a peak is a notable feature every year in the SFE. If the peak occurs at the same time every year but the magnitude varies, then the average within-year pattern can be interpreted as the average magnitude. However, if the magnitude is the same but the timing varies across years, then the magnitude of the average peak cannot be similarly interpreted ~~and instead underestimates~~; instead, the average peak extracted by $f_2(doy_i)$ will underestimate the magnitude that usually occurs. Moreover, the width or duration of the peak will be longer than typically occurs in a given year.

Finally, the raw data can be smoothed using a bivariate spline representing an interaction between *cont_year* and *doy*. This can be expressed as:

~~Model SYDI: y ~ cont_year + s(cont_year, k = num_knots_Y) + s(doy, bs = "cc", k = num_knots_D) + ti(cont_year, doy, bs = c("tp", "cc"), k = c(num_knots_Y_ti, num_knots_D_ti))~~

Model SYDI:

$$y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + f_2(doy_i) + f_3(cont\_year_i, doy_i) + \epsilon_i \quad (4)$$

where ~~ti() specifies~~ equation (4) is the same as equation (3) with the addition of a tensor-product smoothing spline ~~for a surface~~(`ti()` in `mgcv`) that varies smoothly as a function of both ~~cont_year~~$cont\_year$ and ~~doy. The~~ doy. Both $cont\_year$ and doy include their own number of knots, such that the total number of knots for the spline is the product of ~~num_knots_Y_ti in the cont_year axis and num_knots_D_ti in the doy axis. In SYDI, the~~ the two. The need for sufficient knots in SYDI can be satisfied either by allowing for sufficiently ~~large values for num_knots_Y_ti and num_knots_D_ti~~many knots for $f_3()$ or ~~a~~ sufficiently ~~large value for~~

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: First Paragraph

~~knots in num_knots_Y and num_knots_D,~~many knots for $f_1()$ or $f_2()$, but not both, given limits on the model degrees of freedom.

Following the rationale above, the relationship of model SYDI to model S is similar to that of model SYD to model S. Model SYDI differs from model S to a greater extent than model SYD, but all of the splines use the same inputs to smooth the same data. The univariate splines in *cont_year* and *doy* will combined likely not capture as much variation ~~in~~within model SYDI ~~compared to~~ as captured by model S, given the fewer knots that are available to ~~the former.~~$f_1()$ or $f_2()$ within SYDI. The ~~ti term~~tensor-product spline represents an interaction by allowing the pattern in ~~cont_year~~$cont\_year$ to vary by ~~doy~~doy and vice-versa. The interaction term ~~in model SYDI~~ provides an appearance that ~~this~~ model SYDI is fundamentally different from those provided by the other models. However, models S, SY, and SYD ~~all~~also allow within-year fluctuations to vary across years by allowing a spline to be fit through the entire time series. Although model SYDI is the only model that includes an explicit interaction term, all of the models support the interaction conceptually. By providing this term with sufficient knots, the raw data can be fully smoothed with model SYDI to a similar degree as for the other models. However, a very large number of knots in both the ~~cont_year~~$cont\_year$ spline *and* in both dimensions of the interaction spline is impossible to achieve. The distinct aspect of model SYDI is the anticipation that within-year fluctuations will vary smoothly from year to year~~, which~~. However, this is ~~unlikely~~an assumption about ecosystem dynamics that may not be appropriate to *a priori* parameterize into a statistical model, including for ~~the~~ SFE ~~data~~ and ~~chl-a dynamics in~~ many other estuaries ~~because~~where bloom ~~size typically~~magnitude often varies between years. Thus, the conceptual motivation for model SYDI and its practical application are not necessarily

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

supported in the more generalized application for which we are developing this ~~analysis~~set of analyses.

Murphy et al. (2019) used spline formulations for Chesapeake Bay water quality ~~related~~analyses that are comparable to those proposed here, but for different goals and with different handling of smoothness. They evaluated a "`gam0`" with only ~~s(~~a cyclic spline for *doy*~~)~~ and linear *cont_year* terms, a "`gam1`" like our SYD, and a "`gam2`" like our SYDI. In application, only "`gam2`" was used, including the addition of splines as functions of hydrologic variables to account for finer-scale variation. ~~Murphy et al.~~ Murphy et al. (2019) allowed a maximum number of knots ~~in~~for the ~~s(~~*cont_year*~~) term~~ spline ($f_1()$) of 2/3 times the number of years and do not explicitly consider the number of knots in the interaction spline, following an *ad hoc* allocation of variation in the data to different components based on previous interpretations of water quality dynamics in the system. Constraining splines with insufficient knots could inflate Type I error rates for temporal changes and we seek to lower this risk by increasing the upper limit for the knots for the ~~s(cont_year)~~$f_1()$ term. Finally, Murphy et al. (2019) present large AIC differences between their spline formulations. We instead emphasize that, given sufficient knots, the models represent alternative formulations of conceptually similar explanations for the data and yield similar fits (Table 3), resulting in near ties for AIC between models.

~~We~~To evaluate the theoretical and conceptual similarities and differences among the GAM structures discussed above, we visually ~~compare~~compared chl-a estimates from models ~~SY~~S, SYD, and SYDI ~~to emphasize that similar fits can be achieved by all of the presented models~~ (Figure 3~~,~~; note that SY is identical to S and is not shown). Models S, SYD, and SYDI were fit to chl-a data from station 34 using ~~large k values~~a sufficiently high number of knots for the ~~arguments num_knots_y, num_knots_D, num_knots_Y_ti, and num_knots_D_ti~~respective

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Times New Roman, 12 pt, Italic

splines for each model. ~~Predictions~~As expected, predictions by day of year from each model are visually similar (Figure 3a) and closely follow the 1:1 line (Figure 3b). However, ~~when contrasting~~ the ~~estimates using only~~component of predictions explained by the continuous year smoother (~~`s(cont_year))`~~,$f_1()$) differs between models that include additional smoothers (Figure 3c), offering a graphical impression of the ~~fits differ substantially because of how each model allocates variation~~degree to ~~the splines~~which data variability was partitioned to different GAM components. These results are also reflected in differences in the effective degrees of freedom among the additive components of each model (Table 3).~~ Accordingly, even though the models differ by which structural component describes variation in the chl-a time series, they provide similar predictions.~~

For all results, model S was used with enough knots in ~~`num_knots_y`~~$f_1()$ to evaluate chl-a trends across the monitoring stations in the SFE. This model was chosen because of the relatively faster processing time to fit the model, while providing nearly identical explanatory power as compared to the other models (Table 3).

Second-stage analysis: Seasonal features with uncertainties

In the second-stage analysis, we estimated a seasonal average, such as the mean spring chl-a concentrations, along with ~~the~~its associated uncertainty, in each year. We defined μ_t as the seasonal average in year t, $\hat{\mu}_t$ as an estimate of μ_t, and $\hat{\sigma}_{\hat{\mu},t}$ as the estimated standard error of $\hat{\mu}_t$. The season includes n days. For simplicity, the following text omits subscript t.

Point estimates of response values for the fitted GAM take the form $\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the vector of parameter estimates and $\mathbf{X}$ is a model matrix of explanatory variables, including spline basis function values. Vector $\widehat{\boldsymbol{\beta}}$ includes both fixed effect parameters and spline parameters, and

X contains columns corresponding to each. For example, using model SY, if a point estimate for chl-a is needed for a single day, given as *cont_year* = r, then **X** would have a row with 1 in the first column (for the intercept parameter), r (for the linear time trend) in the second column, and an evaluation of each spline basis function at r in the remaining columns. The number of spline basis functions is related to the number of knots. Note that r can be any time, not necessarily the time of an observation.

To obtain a vector, $\hat{\mathbf{y}}$, of fitted point estimates for every day in a season, **X** would have one row for each day. Here, the seasonal averages ~~used in our examples~~ were calculated at the resolution of days. The estimated spline yields both $\widehat{\boldsymbol{\beta}}$ and $\hat{\Sigma}_{\widehat{\beta}}$, an estimate of the covariance matrix of the sampling distribution of $\widehat{\boldsymbol{\beta}}$. The scalar standard errors of $\widehat{\boldsymbol{\beta}}$ are the square roots of the diagonal elements of $\hat{\Sigma}_{\widehat{\beta}}$, whereas the off-diagonal elements are the correlations among the elements of $\widehat{\boldsymbol{\beta}}$. Since parameter estimates are correlated, the covariance of $\hat{\mathbf{y}}$ is $\hat{\Sigma}_{\hat{y}} = \mathbf{X}\hat{\Sigma}_{\widehat{\beta}}\mathbf{X}^T$.

The estimated seasonal average was calculated from the vector of daily values for each of the n days in the season of interest with $\hat{\mu} = A^T\hat{\mathbf{y}}$, where A^T is a row vector with all values equal to $1/n$. The variance of $\hat{\mu}$ is $\hat{\sigma}_{\hat{\mu}}^2 = A^T\hat{\Sigma}_{\hat{\mathbf{y}}}A$, and its standard error is $\hat{\sigma}_{\hat{\mu}}$. Each of these estimates are from the approximate multivariate normality of the sampling distribution of $\widehat{\boldsymbol{\beta}}$.

Third-stage analysis: Trend analysis of seasonal features with uncertainties

In stage three of the analysis, we used a meta-analysis method to evaluate linear trends across years of seasonal-average water quality, characterized by the within-year means ($\hat{\mu}_t$) and their standard errors ($\hat{\sigma}_{\hat{\mu},t}$) that we estimated in stage two of the analysis. This analysis provided ~~a direct answer to~~the basis for directly answering the question~~:~~ *Is there a significant linear trend*

Formatted: Font: Times New Roman, 12 pt, Italic

Formatted: Font: Italic

Formatted: Font: Italic

across a group of years in a seasonal average~~?~~*, taking into account uncertainty*? For example, ~~is~~was there a trend in the spring chl-a average from 1990 to 2000? This question can also be posed in a moving-window manner across a time series (e.g., spring average trend from 1990-2000, 1991-2001, etc.). For all analyses, the response data of interest are $\hat{\mu}_t$, $t = 1, \dots, N$, with their associated standard errors, $\hat{\sigma}_{\hat{\mu},t}$. N is the number of years of the study.

A mixed-effects meta-analysis model can estimate linear trends when each observation has an associated standard error, which is the case with our estimates $\hat{\mu}_t$ and $\hat{\sigma}_{\hat{\mu},t}$. Differences in standard errors, which may result from different monitoring effort between years, are explicitly considered in the analysis. The model can be expressed using notation similar to Sera et al. (2019):

$$\hat{\mu}_t = \beta_0 + \beta_t t + b_t + \epsilon_t \qquad (~~1~~5)$$

where β_0 is the intercept, t is the year, β_t is the slope, b_t is the random effect for year t, and ϵ_t is the residual for year t. Accordingly, the seasonal average for year t is $\mu_t = \beta_0 + \beta_t t + b_t$. The "residual," ϵ_t, represents estimation error in $\hat{\mu}_t$, namely $\hat{\mu}_t - \mu_t$. The residuals are assumed to be independent and normally distributed with mean 0 and ~~variance~~variances $\hat{\sigma}^2_{\hat{\mu},t}$, where the latter is estimated from the calculations above. The random effect, b_t, is the difference between μ_t and $\beta_0 + \beta_t t$ and is considered the "residual" in the sense of unexplained variation not due to the estimation error. The random effect follows a normal distribution with mean 0 and variance, σ_b^2, to be estimated.

We estimated the model (equation (~~1~~5)) using the *mixmeta* package in R (Sera et al., 2019). Results from *mixmeta* have a similar interpretation as those from regression analysis, but parameter estimates and their standard errors incorporate the known standard errors of the

response values. The default estimation method for *mixmeta*, restricted maximum likelihood (REML), was used. The meta-analysis models were applied to a chosen sequence or "window" of years for estimating the linear trend.

Trend comparisons

The above methods were applied to each station by evaluating changes in seasonal averages ~~from January to June and July to December~~ for approximately ten year moving windows from 1991 to 2019. ~~The choice~~The overall technique and *wqtrends* R package allow for easy and flexible selection of ~~within-year~~seasonal windows. As an example focused on illustrating the approach (as opposed to extracting mechanistic interpretations), we selected two broad seasons, January-June and July-December, that are generally relevant to phytoplankton bloom phenology in the SFE (Cloern et al., 2020~~). The~~), and combined include conditions over the full year. A moving-window approach ~~applied~~encoded within *wqtrends* was used to apply the meta-analysis to each decadal window (e.g., 1991-2001, 1992-2002, etc.), allowing changes in slope and its significance to be interpreted as the window is shifted one year at a time. We interpret the slope as representative for the central year for each block, but a predictive trend for the final year of the window could also be interpreted. ~~For~~As a means for summarizing some results, we focus on the windows 1991-2000, 2000-2010, and 2010-2019.

Finally, trend results from the meta-analysis regression method for each season and different time periods were compared to "naive" across-year regressions that do not propagate uncertainty to demonstrate how different and potentially misleading conclusions can be obtained. Trend estimates were compared to 1) trends from ordinary least squares (OLS) regression applied to seasonal averages from the raw data and 2) trends from OLS regression applied to GAM

seasonal averages. Select examples ~~were used~~ where differences were pronounced ~~to illustrate false positive or negatives that may occur with alternative methods.~~ were used for illustration. This analysis was then applied to all stations. The method that formally propagates uncertainty should have more robust statistical properties, such as accurate confidence interval coverage, than naïve methods. For this reason, even when results are similar across methods, the more robust method provides the best support for those results.

Results

Model performance and predictions

Model predictions for chl-a trends across all stations had an average R-squared value of 71% (Table 4) and ~~ranging~~range from 59% (station 22) to 78% (station 18). GAM predictions from north to south showed more pronounced annual and seasonal changes in chl-a towards the more southern stations (Figures S1-S9). All the models suggested 1) increasing chl-a from 1990 until 2005 to 2010, followed by decreasing chl-a until the end of the record in 2019, 2) a spring chl-a peak, particularly at southern stations, and 3) a fall chl-a peak that was smaller than the spring peak. The magnitude of the fall peak did not vary noticeably by location (Figures S1-S9).

Inter-annual trend estimates

Estimates ~~of linear trends in seasonal averages~~from the seasonal trend analyses (mixed-effects meta-analysis regressions) across roughly ten-year windows for different seasons are shown for station 34 (Figure 4). Plots a-c show trends in January to June averages while plots d-f show trends in July to December averages. ~~The seasonal trend analyses showed that~~ January to June

chl-a increased ($\log_{10}$ chl-a slope 0.03 μg L^{-1} yr^{-1}, 0.01-0.~~06~~05 95% confidence interval) from 1991 to 2000, whereas July to December chl-a did not change significantly. Chl-a also increased from 2000 to 2010, but only for July to December ($\log_{10}$ slope 0.03, 0~~.01~~-0.05 95% confidence interval). Finally, chl-a decreased from 2010 to 2019 but again only for July to December ($\log_{10}$ chl-a slope -0.02, -0.04-0 95% confidence interval). Because the trends were confined to certain times of the year, the seasonal estimates provide additional information beyond coarser estimates that cover the entire year.

Temporal changes varied among regions of the Bay and were more pronounced at southern stations. Figure 5 shows results from similar analyses as those in Figure 4, but applied to all stations. ~~Mixed-effects meta-analysis regressions applied to~~The seasonal ~~averages~~trend analyses showed that increases (based on $p < 0.05$) for the January to June period were observed at stations 32~~, 34,~~ and ~~36~~34 from 1991 to 2000~~ and station 18 from 2000 to 2010~~; decreases were observed at stations 30 and 32 from 2010 to 2019. For the July to December period, increases were observed at stations 24, 27, 30, and 32 from 1991 to 2000 and stations 18, 21, 22, 24, and 34 from 2000 to 2010, whereas decreases were observed at stations 30, 32, and 34 from 2010 to 2019.

Results from a ten-year moving window comparison of seasonal trends provided additional context on when significant changes were occurring at each station (Figure 6). Trends were observed at all stations that followed a general pattern of increases early in the record followed by decreases later in the record. Increases and decreases were observed in both the January to June and July to December seasonal periods, with some notable exceptions. In particular, the most southern stations (32, 34, 36) had increasing trends prior to 2005 ~~only~~that were more often observed in the July to December period. Additionally, chl-a at the more northern stations has

not changed in recent years for ~~both~~either seasonal ~~periods~~period. For most stations and seasonal periods, a change from increasing to decreasing chl-a occurred around 2007.

Importance of uncertainty propagation

Results showing trend estimates from meta-analysis on GAM seasonal ~~estimates~~averages provided different conclusions than those from either OLS regression through seasonal averages from raw data (Figure 7 row 1) or OLS regression through GAM ~~estimates~~ seasonal averages without uncertainty propagation (Figure 7 row 2). Figure 7a shows trend estimates for station ~~36~~30 for January to ~~July~~June averages from ~~1991~~2000 to ~~2000~~2010. Only the meta-analysis regression results show a trend in this example (based on $p < 0.05$). The OLS regression on observed estimates (top plot) and OLS regression on GAM estimates (middle plot) did not identify trends. Figure 7b shows trend estimates ~~for the same~~at station 34 for July to December averages from 1991 to 2000~~ to 2010~~. Unlike the first example, only the ~~top~~middle figure shows a trend, whereas the top and bottom ~~two~~ plots do not show trends. In both cases, only the meta-analysis results give reliable conclusions because of full propagation of uncertainty across methods. Even in cases where the p-value threshold is not of interest, the confidence intervals from the alternative methods will be inaccurate.

Applying the same comparison to all stations showed that different trend analysis methods provided conflicting information on the magnitude and significance of the seasonal chl-a changes in each decade (Figure 8). In many cases, the slope estimates were similar in magnitude, with some exceptions at the more southern stations where the OLS estimates suggested a larger trend than the meta-analysis methods. More importantly, differences in the magnitude of the confidence intervals between the OLS models applied to the GAM averages and the meta-

553 analyses were also observed, reflecting the ability of the latter to more accurately assess
554 significance of trends by accounting for uncertainty in the average estimates.

555 Discussion

556 Propagation of uncertainty from within-year features of estimated GAMs to across-year trends
557 using mixed-effects meta-analysis is a new approach that can address different questions than
558 previous methods. Our approach has several advantages over more conventional approaches for
559 analysis of water quality data from long-term monitoring programs. GAMs are capable of
560 modelling time series with missing observations or irregular sampling which can complicate
561 trend assessment and comparison of trends between locations (Junninen et al., 2004; Racault et
562 al., 2014). As noted above, non-parametric approaches (i.e., seasonal Kendall tests) are by far the
563 most common trend analysis methods applied to long-term water quality data (Helsel et al.,
564 2020; Hirsch et al., 1982). These methods only assess the direction and significance of
565 comparisons ~~between year pairs~~across years, and importantly, do not account for full propagation
566 of uncertainty inherent in raw observations if the raw data are aggregated to meet test
567 requirements. Aggregation of raw data, e.g., averaging of observations within a year or season to
568 comply with the requirements of Kendall tests, risks loss of information by removing variation
569 between observations at smaller time scales. The logical outcome is increased risk of incorrect
570 conclusions from test results.

571 Incorrect conclusions on trends can have dramatic consequences for regulated parties under
572 existing water quality compliance frameworks (Smith et al., 2001). Our examples in Figures 7
573 and 8 demonstrate these risks if propagation of uncertainty from raw observations across
574 methods is unaccounted for in trend assessment. ~~Our assessment of trends~~ The "naïve" method

using OLS regression applied to seasonal averages from the raw observations ~~is effectively like~~fails to propagate uncertainty, similarly to averaging results within a year and applying a simple Kendall test. In ~~many~~some cases the results may be similar to those from fully propagating uncertainty, but the loss of information ~~with averaging~~ can lead to increased Type I or II error rates depending on characteristics of the raw data and the method used for their evaluation (Shabman and Smith, 2003). Our examples demonstrated the increased potential for incorrect conclusions at specific monitoring locations and, at ~~much~~ larger spatial scales, across all stations if simpler trend analyses are used. Even though simpler methods may produce similar results in some cases, particularly with frequent sampling and similar effort between years, the only way to confirm such an outcome would be to compare results, relying on the method with full propagation of uncertainty to be the more robust method. Use of methods that fully account for uncertainty is recommended to obtain statistically valid results in a wider range of conditions.

Results here also show that GAM structure (i.e., choice of smoothing terms) was less important than allowing the model sufficient freedom to fit the data. This is an important conclusion that provides guidance on how GAMs could be used to model time series from long-term environmental monitoring programs. Models with separate smoothers for continuous year and day of year can produce nearly identical results in the predicted trends if the knots are sufficiently high to allow the GAMs to be fit as intended by the methods in the mgcv package (Figure 3). The approach presented here leverages the ability of GAMs to objectively estimate smoothed trends across years by identifying an optimal level of smoothing using generalized cross-validation to extract an underlying signal in the observed data (Wood, 2017, 2004).

The underlying cross-validation methods used by GAMs in the mgcv package also reduce the decisions that may be necessary for the implementation of alternative trend assessment methods.

For example, WRTDS and similar smoothing approaches (e.g., LOESS) require decisions on appropriate window widths or bandwidths to define the neighborhood of observations for smoothing (Hirsch et al., 2010; Wan et al., 2017). This is especially problematic for policy analysis or regulatory decisions if the results change based on arbitrary decisions of the analyst. Because these decisions are not needed for GAMs, the results can be considered a more objective and potentially ~~"true"~~accurate signal of actual trends that are minimally influenced by process or observation error present in the raw data.

Several limitations of the proposed methods deserve mention. First, if sampling is so irregular that important fluctuations are missed entirely in some years, the GAM estimates and uncertainty propagation could become dubious in interpretation and usefulness. Second, estimation of GAMs for very long series can be computationally demanding. When this is an obstacle, one could do the first two analysis stages using temporal windows of the full data, with the only implication being that different degrees of smoothness may be estimated for different windows, which indeed might be justified by the data. Third, meta-analysis regression results for a very small number of years, particularly confidence intervals and associated p-values, may be inaccurate (e.g. in confidence interval coverage). In such cases, one could make alternative use of the GAM seasonal averages and standard errors, such as for pairwise comparisons among years.

Future work

Additional work could be conducted to further strengthen the conclusions based on trends from meta-analysis regression applied to the GAM seasonal averages. Our third stage analyses require *a priori* decisions on long-term time scales of interest and future work could generalize these choices. Although there are undoubtedly many scenarios where years of interest can be chosen

objectively by the needs of an analysis (e.g., regulatory compliance periods, time since management intervention), a more general question of ~~"~~when~~"~~ changes occur independent of user decisions is also important to address. Additional methods could be developed using objective criteria to identify inflection points or other important periods where changes occur independent of a user choice. Assessing water quality changes beyond an evaluation of seasonal averages could also be possible with our approach, such as assessing changes in the timing or magnitude of a seasonal peak across years.

Additional explanatory variables could be identified ~~that may be associated with~~to explain trends in either the ~~trend after~~GAM stage or the ~~trend has been adequately described.~~meta-analysis stage of analysis. This information ~~has~~would have obvious implications for management decisions on factors that influence water quality changes, e.g., wastewater treatment upgrades, large-scale climatic factors, or flow regulation practices. ~~An advantage of GAMs is their flexibility in including alternative predictors, such that the significance of a predictor or comparison of nested models with and without different predictors can provide evidence of which predictors are driving the observed trends~~Including alternative predictors in the GAMs (Wood and Augustin, 2002; Zuur et al., 2009~~). In such cases, considerations of model structure can have direct implications on conclusions given how GAMs could be used to assess different questions. Our goal was to describe chl-a changes relative to time, where the predictors were variations on a general theme (e.g., season vs. year). This is a different application from using GAMs with predictors selected to explain those changes over time. Therefore, using our approach to evaluate explanatory variables will require testing of different model structures~~) could reduce the uncertainties of its estimates and, if relevant, allow the influence of those variables to be removed. Including alternative predictors in the meta-analysis could help in

explaining long-term trends of seasonal averages or other metrics obtained from the GAMs. Our goal here was to describe chl-a changes relative to time, so the single predictor in both modeling stages was time.

Finally, the evaluation of trends for alternative water quality variables in addition to chl-a is a simple and logical extension of the methods proposed in this study. The long-term monitoring program maintained by USGS includes multiple parameters in addition to chl-a that can provide additional context into broader water quality trends in the SFE (Cloern and Schraga, 2016; Schraga et al., 2020). These parameters include salinity, temperature, light attenuation, dissolved oxygen, suspended particulate matter, and dissolved inorganic nutrients, which collectively can be used to provide a broader understanding of potential eutrophication patterns or ecosystem shifts at seasonal and multi-decadal scales. Chl-a measurements can also be used to estimate gross primary production to assess process rates that may be more indicative of system function (Cloern et al., 2007; Jassby et al., 2002). The open-source *wqtrends* R package (Beck et al., 2021) developed for this manuscript can be used for these analyses to provide additional insight into potential drivers of water quality change in the SFE and other estuarine systems.

Formatted: Font: Italic

Acknowledgments

This work was supported by funding from the San Francisco Bay Nutrient Management Strategy (NMS). We thank the staff of the US Geological Survey that collect and maintain long-term monitoring data in San Francisco Bay. This work benefited from discussions with the ~~San Francisco Bay~~NMS Nutrient Technical Workgroup and Steering Committee. We thank James D. Hagy III and two anonymous reviewers for ~~reviewing an earlier draft of~~providing helpful comments on this manuscript.

665

Formatted: First Paragraph, Line spacing: Double

References

Alpine, A.E., Cloern, J.E., 1988. Phytoplankton growth rates in a light-limited environment, San Francisco Bay. Marine Ecology Progress Series 44, 167–173.

Beck, M.W., de Valpine, P., Murpy, R., Wren, I., Chelsky, A., Foley, M., Senn, D., 2021. tbep-tech/wqtrends: v1.1.0 (Version v1.1.0). Zenodo. http://doi.org/10.5281/zenodo.4509638.

Beck, M.W., Hagy, J.D., III, 2015. Adaptation of a weighted regression approach to evaluate water quality trends in an estuary. Environmental Modelling and Assessment 20, 637–655. https://doi.org/10.1007/s10666-015-9452-8

Beck, M.W., Jabusch, T.W., Trowbridge, P.R., Senn, D.B., 2018. Four decades of water quality change in the upper San Francisco Estuary. Estuarine, Coastal and Shelf Science 212, 11–22. https://doi.org/10.1016/j.ecss.2018.06.021

Beck, M.W., Murphy, R.R., 2017. Numerical and qualitative contrasts of two statistical models for water quality change in tidal waters. Journal of the American Water Resources Association 53, 197–219. https://doi.org/10.1111/1752-1688.12489

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A seasonal-trend decomposition procedure based on Loess. Journal of Official Statistics 6, 3–73.

Cloern, J.E. 2018. Patterns, pace, and processes of water-quality variability in a long-studied estuary. Limnology and Oceanography 64, S192-S208. https://doi.org/10.1002/lno.10958

Cloern, J.E., Jassby, A.D., 2012. Drivers of change in estuarine-coastal ecosystems: Discoveries from four decades of study in San Francisco Bay. Reviews of Geophysics 50. https://doi.org/10.1029/2012RG000397

Cloern, J.E., Jassby, A.D., 2010. Patterns and scales of phytoplankton variability in estuarine-coastal ecosystems. Estuaries and Coasts 33, 230–241.

Cloern, J.E., Jassby, A.D., Thompson, J.K., Hieb, K.A., 2007. A cold phase of the East Pacific triggers new phytoplankton blooms in San Francisco Bay. Proceedings of the National Academy of Sciences of the United States of America 104, 18561–18565.

Cloern, J.E., Schraga, T.S., 2016. USGS measurements of water quality in San Francisco Bay (CA), 1969-2015: U.S. Geological Survey data release. ~~https://doi.org/10.5066/F7TQ5ZPR.~~https://doi.org/10.5066/F7TQ5ZPR.

Cloern, J.E., Shcraga, T.S., Nejad, E., Martin, C., 2020. Nutrient status of San Francisco Bay and its management implications. Estuaries & Coasts 43, 1299–1317. https://doi.org/10.1007/s12237-020-00737-w

Cole, B.E., Cloern, J.E., 1984. Significance of biomass and light availability to phytoplankton productivity in San Francisco Bay. Marine Ecology Progress Series 17, 15–24.

Cumming, G.S., Cumming, D.H.M., Redman, C.L., 2006. Scale mismatches in social-ecological systems: Causes, consequences, and solutions. Ecology and Society 11, 14.

Forbes, D.J., Xie, Z., 2018. Identifying process scales in the Indian River Lagoon, Florida using wavelet transform analysis of dissolved oxygen. Ecological Complexity 36, 149–167. https://doi.org/10.1016/j.ecocom.2018.07.005

Gasparrini, A., Armstrong, B., Kenward, M.G., 2012. Multivariate meta-analysis for non-linear and other multi-parameter associations. Statistics in Medicine 31, 3821–3839. https://doi.org/10.1002/sim.5471

Hafen, R.P., 2010. Local regression models: Advancements, applications, and new methods (PhD thesis). Purdue University, West Lafayette, Indiana.

Haraguchi, L., Carstensen, J., Abreu, P.C., Odebrecht, C., 2015. Long-term changes of the phytoplankton community and biomass in the subtropical shallow Patos Lagoon Estuary, Brazil. Estuarine, Coastal and Shelf Science 162, 76–87.

He, S., Mazumdar, S., Arena, V.C., 2006. A comparative study of the use of GAM and GLM in air pollution research. Environmetrics 17, 81–93. https://doi.org/10.1002/env.751

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., Gilroy, E.J., 2020. Statistical methods in water resources, 2nd ed. U.S. Geological Survey Techniques; Methods, book 4, chapter A3, version 1.1, Reston, Virginia.

Hirsch, R.M., Archfield, S.A., De Cicco, L.A., 2015. A bootstrap method for estimating uncertainty of water quality trends. Environmental Modelling and Software 73, 148–166. https://doi.org/10.1016/j.envsoft.2015.07.017

Hirsch, R.M., Moyer, D.L., Archfield, S.A., 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, 857–880.

Hirsch, R.M., Slack, J.R., Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data. Water Resources Research 18, 107–121.

Jassby, A.D., 2008. Phytoplankton in the Upper San Francisco Estuary: Recent biomass trends, their causes, and their trophic significance. San Francisco Estuary and Watershed Science 6, 1–24.

Jassby, A.D., Cloern, J.E., Cole, B.E., 2002. Annual primary production: Patterns and mechanisms of change in a nutrient-rich tidal ecosystem. Limnology and Oceanography 47, 698–712.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kimmerer, W.J., Thompson, J.K., 2014. Phytoplankton growth balanced by clam and zooplankton grazing and net transport into the low-salinity zone of the San Francisco Estuary. Estuaries and Coasts 37, 1202–1218.

Lefcheck, J.S., Wilcox, D.J., Murphy, R.R., Marion, S.R., Orth, R.J., 2017. Multiple stressors threaten the imperiled coastal foundation species eelgrass (*zostera marina*) in Chesapeake Bay, USA. Global Change Biology 23, 3474–3483. https://doi.org/10.1111/gcb.13623

~~Lehman, P.W., Kurobe, T., Lesmeister, S., Baxa, D., Tung, A., Teh, S.J., 2017. Impacts of the 2014 severe drought on the Microcystis bloom in San Francisco Estuary. Harmful Algae 63, 94–108. https://doi.org/10.1016/j.hal.2017.01.011~~

Lortie, C.J., 2014. Formalized synthesis opportunities for ecology: Systematic reviews and meta-analyses. OIKOS 123, 897–902. https://doi.org/10.1111/j.1600-0706.2013.00970.x

Morton, R., Henderson, B.L., 2008. Estimation of nonlinear trends in water quality: An improved approach using generalized additive models. Water Resources Research 44, W07420. https://doi.org/10.1029/2007WR006191

Murphy, R.R., Perry, E., Harcum, J., Keisman, J., 2019. A Generalized Additive Model Approach to evaluating water quality: Chesapeake Bay case study. Environmental Modelling & Software 118, 1–13. https://doi.org/10.1016/j.envsoft.2019.03.027

Navarro, G., Caballero, I., Prieto, L., Vázquez, A., Flecha, S., Huertas, I.E., Ruiz, J., 2012. Seasonal-to-interannual variability of chlorophyll-*a* bloom timing associated with physical forcing in the Gulf of Cádiz. Advances in Space Research 50, 1164–1172. https://doi.org/10.1016/j.asr.2011.11.034

~~Novick, E., Senn, D., 2014. External nutrient loads to San Francisco Bay (No. Contribution Number 704).~~Peacock, M.B., Gibble, C.M., Senn, D.B., Cloern, J.E., Kudela, R.M. 2018. Blurred lines: Multiple freshwater and marine algal toxins at the land-sea interface of San Francisco Bay, California. Harmful Algae 73, 138–147. https://doi.org/10.1016/j.hal.2018.02.005

~~San Francisco Estuary Institute, Richmond, CA.~~

Pearce, J.L., Beringer, J., Nicholls, N., Hyndman, R.J., Tapper, N.J., 2011. Quantifying the influence of local meteorology on air quality using generalized additive models. Atmospheric Environment 45, 1328–1336. https://doi.org/10.1016/j.atmosenv.2010.11.051

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, R v4.0.2, Vienna, Austria.

Racault, M.F., Sathyendranath, S., Platt, T., 2014. Impact of missing data on the estimation of ecological indicators from satellite ocean-colour time series. Remote Sensing of Environment 152, 15–28. https://doi.org/10.1016/j.rse.2014.05.016

Schraga, T.S., Nejad, E.S., Martin, C.A., Cloern, J.E., 2020. USGS measurements of water quality in San Francisco (CA), beginning in 2016 (ver. 3.0, March 2020): U.S. Geological Survey data release. https://doi.org/10.5066/F7D21WGF.

Sera, F., Armstrong, B., Blangiardo, M., Gasparrini, A., 2019. An extended mixed-effects framework for meta-analysis. Statistics in Medicine 38, 5429–5444. https://doi.org/10.1002/sim.8362

SFBRWQCB (San Francisco Bay Regional Water Quality Control Board), 2017. Water quality control plan (basin plan) for the San Francisco Bay basin. Prepared by California Regional Water Quality Control Board, San Francisco Bay Region, Oakland, CA. https://www.waterboards.ca.gov/sanfranciscobay/basin_planning.html

SFEI (San Francisco Estuary Institute), 2014a. External nutrient loads to San Francisco Bay. SFEI Contribution Number 704. San Francisco Estuary Institute, Richmond, CA.

SFEI (San Francisco Estuary Institute), 2014b. Scientific foundation for the San Francisco Bay Nutrient Management Strategy. SFEI Contribution Number 979. San Francisco Estuary Institute, Richmond, CA.

SFEI (San Francisco Estuary Institute), 2021. Connections to tidal marsh and restored salt ponds drive seasonal and spatial variability in ecosystem metabolic rates in Lower South San Francisco Bay. SFEI Contribution No. 992. San Francisco Estuary Institute, Richmond, CA.

Shabman, L., Smith, E., 2003. Implications of applying statistically based procedures for water quality assessment. Journal of Water Resources Planning and Management 129, 330–336. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:4(330)

Smith, E.P., Ye, K., Hughes, C., Shabman, L., 2001. Statistical assessment of violations of water quality standards under section 303 (d) of the Clean Water Act. Environmental science & technology 35, 606–612. https://doi.org/10.1021/es001159e

Stow, C.A., Cha, Y., Johnson, L.T., Confesor, R., Richards, R.P., 2015. Long-term and seasonal trend decomposition of Maumee River nutrient inputs to western Lake Erie. Environmental Science and Technology 49, 3392–3400. https://doi.org/10.1021/es5062648

Sutula, M.A., Kudela, R.M., Hagy, J.D., III, Harding, L.W., Jr., Senn, D.B., Cloern, J.E., Bricker, S., Berg, G.M., Beck, M.W. 2017. Novel analyses of long-term data provide a scientific basis for chlorophyll-a thresholds in San Francisco Bay. Estuarine, Coastal and Shelf Science. 197, 107-118. https://doi.org/10.1016/j.ecss.2017.07.009

Urquhart, N.S., Paulsen, S.G., Larsen, D.P., 1998. Monitoring for policy-relevant regional trends over time. Ecological Applications 8, 246–257. https://doi.org/10.1890/1051-0761(1998)008[0246:MFPRRO]2.0.CO;2

Wan, Y., Wan, L., Li, Y., Doering, P., 2017. Decadal and seasonal trends of nutrient concentration and export from highly managed coastal catchments. Water Research 115, 180–194.

Wood, S.N., 2017. Generalized additive models: An introduction with r, 2nd ed. Chapman; Hall, CRC Press, London, United Kingdom.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686. https://doi.org/10.1198/016214504000000980

Wood, S.N., 2003. Thin-plate regression splines. Journal of the Royal Statistical Society (B) 65, 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157, 157–177. https://doi.org/10.1016/S0304-3800(02)00193-X

Yang, G., Moyer, D.L., 2020. Estimation of nonlinear water-quality trends in high-frequency monitoring data. Science of The Total Environment 715, ~~136686.~~ 10.1016/j.scitotenv.2020.136686.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed effects models and extensions in ecology with r. Springer-Verlag, New York, New York.

Figures

Formatted: Normal, Space After: 0 pt

Figure 1: Observed chl-a concentrations for all stations in central and south San Francisco Estuary (18-36, Figure 2), with (a) annual summer/fall concentrations (Aug - Dec) and (b) monthly concentrations by decade.

Figure 2: Station locations in the central and south San Francisco Estuary used for analysis. See Table 1 for station descriptions. Full dataset described in Schraga et al. (2020).

Figure 3: GAM output of estimated chl-a at station 32 for models S, SYD, and SYDI. Model SY is identical to S and is not shown. Plots in (a) show model predictions by day of year with separate lines for each year. Plots in (b) show pairwise comparisons of predicted chl-a between the models and plots in (c) show the same comparisons as in (b) but only for results from the estimated smoother for the cont_year variable. The plots demonstrate that results between the models are similar except for a few observations at extreme values (a, b), but they vary in how they allocate contributions to the predictions among different additive splines (c). The 1:1 lines are in red to facilitate comparisons.

*Figure 4: Examples of seasonal averages and trend estimates in ten year blocks from meta-analyses using results of GAM predictions for station 34. Plots (a), (b), and (c) show trend estimates for January to June averages and (d), (e), and (f) show trend estimates for July to December averages. The trend lines estimate the rate of change of chl-a per year, reported as the log_{10}-slope (+/- 95 % confidence interval) in the sub-plot titles. ns: not significant at α = 0.05, * $p < 0.05$*

Figure 5: Interannual trend estimates of seasonal averages by decade for chl-a at each station. Point type and color represent the direction and magnitude of an estimated trend as the log_{10} slope for chl-a concentration per year. Trends with $p < 0.05$ are marked with an asterisk. All results are from Model S.

Figure 6: Estimates of log_{10} chl-a change per year (+/- 95% confidence interval) from applying the meta-analysis across the seasonal averages for each station. Stations are arranged top to bottom from north to south. Plots in (a) show estimates for seasonal averages from January to June and plots in (b) show estimates for seasonal averages from July to December. Results are from a ten-year, centered moving window where each point shows a linear trend estimate from five years prior to five years after each year. Estimates prior to 1996 and after 2014 are not available because of an incomplete ten year record for estimating the trend. Significant estimates are shown in red.

Figure 7: Trend estimate comparisons (arithmetic scale) for three models applied to seasonal averages of chl-a in different annual periods at ~~station 36.~~stations (a) 30 and (b) 34. The first row shows OLS (ordinary least squares) regression applied to seasonal averages of chl-a from the raw data, the second row shows OLS regression applied to seasonal averages of chl-a from the GAM (without error propagation), and the third row shows meta-analysis regression applied to the seasonal averages of chl-a from the GAM. Regressions in each plot are fit through the seasonal estimates indicated in the plot titles for a specified year range.~~ ~~These examples on selected periods of time show that slope estimates can be similar, but the confidence intervals vary.~~

Figure 8: Trend estimate comparisons for three models applied to seasonal averages of chl-a in different annual periods at each station. The "OLS raw" trend model is based on an ordinary least squares (OLS) regression fit to the seasonal averages of chl-a from the raw data, the "OLS GAM" trend model is based on an OLS regression fit to the seasonal averages of chl-a from the GAM model (without error propagation), and the "Meta-analysis GAM" trend model is based on a meta-analysis regression fit to the seasonal averages of chl-a from the GAM model. Values for each model are the log_{10}-slope estimates (+/- 95% confidence interval) as annual change per year within each season, with line style denoting trend significance.

Formatted: Image Caption

Formatted: Font: 12 pt, Not Bold, Font color: Auto

Tables

Table 1: Station locations, sample sizes (from 1991 to 2019), and summary values (median, minimum, maximum) for chl-a ($\mu g\ L^{-1}$). Stations are arranged from north to south.

Station	Latitude	Longitude	n	Med.	Min.	Max.
18	37.836	-122.418	414	3.6	0.2	16.6
21	37.784	-122.351	576	4.4	0.6	40.0
22	37.752	-122.351	569	4.0	0.7	53.1
24	37.686	-122.334	595	4.2	0.7	47.3
27	37.617	-122.285	596	4.5	0.5	50.9
30	37.551	-122.184	608	5.1	0.8	112.2
32	37.517	-122.133	591	5.9	0.7	282.1
34	37.485	-122.086	544	6.5	0.6	158.3
36	37.468	-122.067	476	6.2	1.1	328.4

Formatted Table

Table 2: Summary and details for each of the GAM structures. In practice, a sufficiently large number of knots provided to the additive terms will produce identical or comparable estimates for a response variable. The models differ in the allocation of penalties for the smoothness of each spline ~~(s())~~.

GAM	Additive components	Details
S	~~s(cont_year)~~ $f_1(cont\_year)$	A single smoother over a continuous year variable
SY	~~cont_year + s(cont_year)~~ $\beta_1 cont\_year + f_1(cont\_year)$	A linear continuous year variable and a single smoother over a continuous year variable
SYD	~~cont_year + s(cont_year) + s(doy)~~ $\beta_1 cont\_year + f_1(cont\_year) + f_2(doy)$	A linear continuous year variable, a smoother over a continuous year variable, and a smoother over a day of year variable
SYDI	~~cont_year + s(cont_year) + s(doy) + ti(cont_year, doy)~~ $\beta_1 cont\_year + f_1(cont\_year) + f_2(doy) + f_3(cont\_year, doy)$	A linear continuous year variable, a smoother over a continuous year variable, a smoother over a day of year variable, and an interaction smoother across continuous year and day of year variables

Formatted Table

*Table 3: Comparison of the four model structures (S, SY, SYD, SYDI) described in the first stage analysis of GAM estimation. The four models provide either identical or comparable ability to describe chl-a trends at an example station (32) in the southern end of the San Francisco Estuary. The models differ in additive smoothers and the amount of effective degrees of freedom (edf) in the smoothers (measure of wiggliness in each component), but the overall model predictions are similar. AIC: Akaike Information Criterion, GCV: generalized cross-validation score, R2: r-squared values for predictions, edf: effective degrees of freedom, F: F-statistic, p-val: probability value, ** $p < 0.001$*

model	AIC	GCV	R2	smoother	edf	F	p-val
S	-138.2	0.06	0.74	s(cont_year)	242.23	6.27	**
SY	-138.2	0.06	0.74	s(cont_year)	242.23	6.27	**
SYD	-135.66	0.06	0.74	s(cont_year)	229.33	3.88	**
				s(doy)	8.07	0.11	**
SYDI	-123.57	0.06	0.73	s(cont_year)	136.88	3.13	**
				s(doy)	9.54	0.79	**
				ti(cont_year,doy)	69.31	0.74	**

Formatted Table

Formatted: Font: Not Italic

Table 4: Model performance statistics for each station as Akaike Information Criterion scores (AIC), generalized cross-validation scores (GCV), and r-squared values.

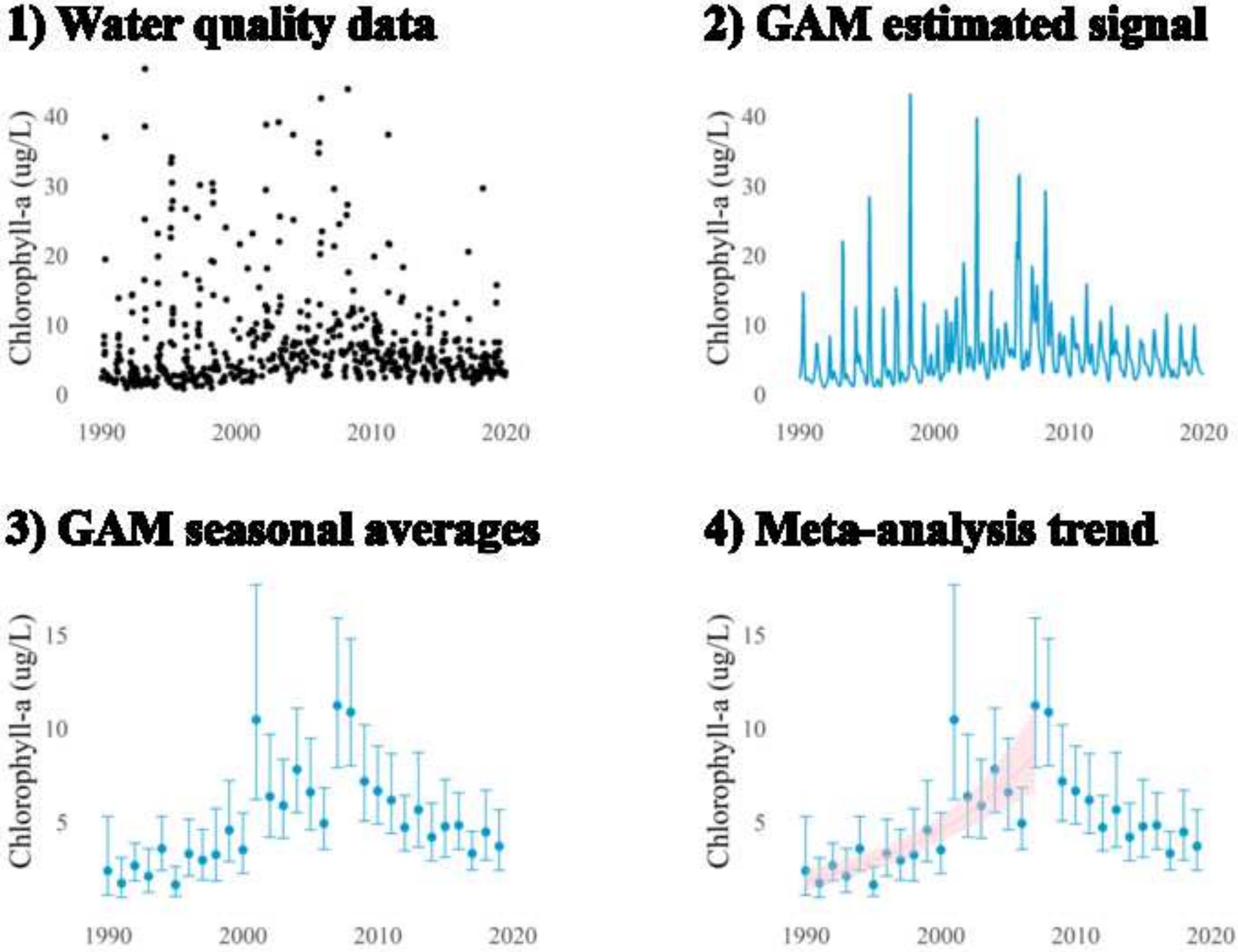
station	AIC	GCV	R-squared
18	-430.94	0.04	0.78
21	-305.93	0.04	0.70
22	-160.07	0.05	0.59
24	-250.77	0.04	0.69
27	-188.72	0.05	0.72
30	-172.79	0.05	0.74
32	-138.20	0.06	0.74
34	-3.17	0.07	0.68
36	-22.05	0.07	0.73

Formatted Table

Formatted: Font: 12 pt, Not Bold, Font color: Auto

Formatted: Body Text

Graphical Abstract



Highlights (for review : 3 to 5 bullet points (maximum 85 characters including spaces per bullet point)

- Trend analyses of water quality data must consider full propagation of uncertainty
- GAMs with different smoothing splines can extract nearly identical trends
- GAMs can estimate seasonal averages with uncertainty from monitoring data
- GAMs coupled with mixed-effects meta-analysis can accurately assess trends

Abstract

Effective stewardship of ecosystems to sustain current ecological status or mitigate impacts requires nuanced understanding of how conditions have changed over time in response to anthropogenic pressures and natural variability. Detecting and appropriately characterizing changes requires accurate and flexible trend assessment methods that can be readily applied to environmental monitoring datasets. A key requirement is complete propagation of uncertainty through the analysis. However, this is difficult when there are mismatches between sampling frequency, period of record, and trends of interest. Here, we propose a novel application of generalized additive models (GAMs) for characterizing multi-decadal changes in water quality indicators and demonstrate its utility by analyzing a 30-year record of biweekly-to-monthly chlorophyll-a concentrations in the San Francisco Estuary. GAMs have shown promise in water quality trend analysis to separate long-term (i.e., annual or decadal) trends from seasonal variation. Our proposed methods estimate seasonal averages in a response variable with GAMs, extract uncertainty measures for the seasonal estimates, and then use the uncertainty measures with mixed-effects meta-analysis regression to quantify inter-annual trends that account for full propagation of error across methods. We first demonstrate that nearly identical descriptions of temporal changes can be obtained using different smoothing spline formulations of the original time series. We then extract seasonal averages and their standard errors for an *a priori* time period within each year from the GAM results. Finally, we demonstrate how across-year trends in seasonal averages can be modeled with mixed-effects meta-analysis regression that propagates uncertainties from the GAM fits to the across-year analysis. Overall, this approach leverages GAMs to smooth data with missing observations or varying sample effort across years to

estimate seasonal averages and meta-analysis to estimate trends across years. Methods are provided in the *wqtrends* R package.

Key words: chlorophyll, Generalized Additive Models, meta-analysis, San Francisco Estuary, Trend analysis

Introduction

Accurate quantification of environmental trends must consider variation at different temporal scales when ignoring variation at one scale could lead to incorrect conclusions about variation at another scale. Many environmental monitoring programs collect temporally resolved but irregular time series data to quantify trends for regulatory, management, or research purposes. The mismatch between the scales of monitoring versus analysis questions or management goals can present statistical challenges (Cumming et al., 2006; Forbes and Xie, 2018; Urquhart et al., 1998). At short temporal scales typically less than a year, environmental systems exhibit variability caused by multiple factors (e.g., weather events, management, or seasonal changes). Such fluctuations may not be related to inter-annual trends or may not be well-suited to multi-scale smoothing methods, yet they can perturb the time-series such that the sub-annual fluctuations must be addressed within the trend analysis to allow accurate quantification of inter-annual trends. Many trend analysis methods lack the flexibility to evaluate multiple independent variables in an extendable structure that accommodates hypothesis testing at different time scales of interest.

In this paper, we develop methods to estimate across-year trends of within-year features, such as seasonal averages, while accounting for uncertainties across analysis steps. Our overarching goal

for this work was to develop a flexible set of tools for accurately characterizing inter-annual changes in seasonally-averaged water quality metrics that can be robustly applied to diverse time-series data. At the outset, we identified several specific requirements and priorities for the trend analysis methods: 1) Complete propagation of uncertainty through the analysis; 2) Separation of trends on different time scales; 3) Ability to estimate linear and nonlinear responses; 4) Flexibility to evaluate multiple independent variables (and random effects), although the examples herein include time as the only independent variable; and 5) Robust to missing observations or varying sampling effort across years, which was considered a high priority to allow the methods to be applied to diverse time-series datasets and monitoring programs. Existing methods that begin to address some of the above requirements and priorities can be generalized into four groups: seasonal Kendall tests (and other non-parametric tests); seasonal trend decomposition using loess (STL); weighted regression on time, discharge, and season (WRTDS); and generalized additive models (GAMs).

Seasonal Kendall tests and related non-parametric approaches have been used for decades in water quality trend assessments to identify inter-annual, monotonic changes while accounting for predictable patterns among seasons (Cloern et al., 2007; Helsel et al., 2020; Hirsch et al., 1982). While seasonal Kendall and other non-parametric approaches have been among the most commonly used methods in long-term water quality trend analysis (Wan et al., 2017), they do not satisfy several of our requirements. For descriptive decomposition of long-term monitoring data, these approaches assume seasonal patterns within years do not change and require regularly spaced or balanced data. In addition, seasonal Kendall tests do not allow for additional independent variables to explain variation, do not estimate a model that could be useful for other

66 purposes (e.g., prediction), and do not easily allow for propagation of uncertainty to other trend
67 analysis methods.

68 STL decomposes a time series into additive components of a long-term trend, a seasonal pattern,
69 and residuals (Cleveland et al., 1990; Cloern, 2018; Cloern and Jassby, 2010; Stow et al., 2015).
70 While useful and widely applied, this method does not address all of our requirements. STL
71 decomposition does not allow for incorporating explanatory variables other than time. In
72 addition, it is often characterized more as an algorithm of statistical steps than as a statistical
73 model with estimated parameters (e.g., Wan et al., 2017) and it does not usually estimate
74 standard errors to allow hypothesis testing (but see Hafen, 2010). STL methods may also over-
75 simplify trends into fixed components that do not change over time, e.g., a seasonal estimate that
76 is constant across years. This limitation presents challenges when addressing questions relevant
77 to long-term water quality data, such as timing of seasonal peaks that can suggest system
78 response to changing environmental conditions (Cloern and Jassby, 2010; Navarro et al., 2012).

79 The weighted regression on time, discharge, and season (WRTDS) method addresses the
80 problem of inflexibility in STL by using a more general local regression scheme (Beck et al.,
81 2018; Beck and Hagy, 2015; Hirsch et al., 2010; Hirsch et al., 2015). Designed for evaluating
82 water quality in rivers where separating the effect of discharge on constituent concentration is
83 important, WRTDS estimates a moving window regression model with components that allow
84 parameters to vary smoothly in relation to both time and discharge. This yields parameters that
85 are specific to season, year, and flow regime. The WRTDS approach is conceptually similar to
86 local kernel smoothing methods, with specific application to explanatory variables relevant for
87 water quality constituents (i.e., season, year, and discharge). Standard error estimates of
88 predictions from WRTDS are available through a block bootstrap approach applied to the model

89 results (Hirsch et al., 2015). Although a useful addition to the original method (Hirsch et al.,
90 2010), the approach requires extensive resampling using a previously fitted model. Alternative
91 methods that include standard error estimates simultaneously with model output may be
92 preferred for intensive or more iterative applications.

93 Finally, GAMs can satisfy the requirements and priorities identified above and were adopted as a
94 central component for the trend analyses herein. GAMs combine one or more smoothing splines
95 to model patterns in data and can be reasonably viewed as generalizing the concepts behind STL
96 and WRTDS (Haraguchi et al., 2015; He et al., 2006; Morton and Henderson, 2008; Murphy et
97 al., 2019; Pearce et al., 2011). The basis functions used to formulate GAMs can be customized
98 based on expected patterns in the data. Examples include cyclic splines, which can be used to
99 model seasonal patterns, and low-dimensional interactions (Wood, 2017). GAMs have added
100 flexibility because they can include both parametric (e.g., linear or quadratic) components and
101 non-parametric (spline) components. Multiple approaches have been developed to determine the
102 optimal degree of smoothness (Wood, 2004; 2017). These approaches are based on optimization
103 of out-of-sample prediction error, which addresses a key concern around methods like WRTDS
104 that do not have analogs for choosing optimal degrees of smoothing. GAMs can also produce
105 results comparable to those provided by WRTDS (Beck and Murphy, 2017) and have readily
106 obtainable uncertainty estimates. Further, GAMs have natural frequentist and Bayesian
107 interpretations, are naturally extensible to include random effects (i.e., generalized additive
108 mixed models or GAMMs), and have computationally efficient implementations (Wood, 2017).

109 GAMs have been applied previously to evaluate trends in water quality time series from long-
110 term monitoring programs (Haraguchi et al., 2015; Murphy et al., 2019). For example, Murphy et
111 al. (2019) used GAMs to decompose water quality time series from Chesapeake Bay into long-

term and seasonal trends (Murphy et al., 2019) and test trend hypotheses between two points in time. Other studies of environmental time series with GAMs have addressed the use of transformed response data (Yang and Moyer, 2020), serial correlation in high resolution data (Morton and Henderson, 2008; Yang and Moyer, 2020), and quantifying time lags in relationships between response and predictor variables (Lefcheck et al., 2017). The method development and analyses described herein generalize the approach to analyzing trends of seasonal spline features, describes the relationships among alternative spline formulations when spline flexibility is allowed to vary (Wood, 2017, 2003) rather than being constrained *a priori* for different time scales, and prioritizes full incorporation of uncertainty.

To incorporate the uncertainty of seasonal estimates into trend analysis, we integrated GAMs with mixed-effects meta-analysis (Gasparrini et al., 2012; Sera et al., 2019). In this integration, the GAMs framework addressed a critical need by providing an estimate of uncertainty (e.g., a standard error) of seasonal averages, even in situations with irregular sampling and serial correlation, which are common in time series data. Meta-analysis regression incorporates a known (or estimated) standard error for each response datum. Usually meta-analysis is used when each response summarizes a dataset from a separate study, along with its standard error, and the meta-analysis looks for across-study patterns in effect size (i.e., Lortie, 2014). In this study, each response summarizes one year or season of data, with standard error from the GAM, and the meta-analysis looks for patterns across years. Thus, while meta-analysis methods are most commonly associated with combining results from multiple studies into a larger analysis, their key modeling step is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). To do this, meta-analysis makes use of a known (estimated) standard error for each response

datum, which is the required priority here to propagate standard errors from the GAM into a regression of seasonal averages.

To describe the approach and demonstrate its utility, we analyze a 30 year record (1990-2019) of biweekly-to-monthly chlorophyll-a concentration data, collected at 9 stations in the southern portion of the San Francisco Estuary, California, USA. Approximately twice-monthly monitoring has been conducted for several decades at fixed locations (stations) along the longitudinal axis of the Bay. Analysis of these data is complicated by irregularities in timing and consistency of data collection. We were interested in questions such as: Are there significant trends in spring mean chlorophyll at multi-year time-scales? At what across-year window does summer-fall mean chlorophyll levels change? Is there a spatial difference in chlorophyll trends? We provide examples illustrating how these questions can be addressed using GAMs to estimate seasonal patterns and use meta-analysis to evaluate trends between years. The techniques are incorporated into an open-source and publicly available R package, *wqtrends*, developed by the authors (Beck et al., 2021, available at https://tbep-tech.github.io/wqtrends, including an online dashboard for viewing results at https://nutrient-data.sfei.org/apps/SFbaytrends/).

Methods

Study area and data sources

The San Francisco Estuary (SFE) is the largest estuary on the Pacific Coast of North America, and its watershed covers 200 thousand km^2 in the US state of California. Flows from the Sacramento-San Joaquin Delta, entering from the northeast, account for the vast majority of SFE-wide annual-average freshwater inputs (Cloern and Jassby, 2012). Freshwater contributions

to southern SFE (South Bay, Lower South Bay) come primarily from local tributaries during the wet season (Nov-Apr), and from wastewater treatment plant discharges during the dry season (May-Oct). Salinity values in southern SFE subembayments (the focus of this work; Central Bay, South Bay, Lower South Bay) range from 5 to 35 ppt and depend strongly on season (stormwater runoff), tidal cycle, and effluent discharge from wastewater treatment plants (Cloern and Jassby, 2012).

SFE receives 70,000 kg per day of dissolved inorganic nitrogen (DIN; annual average), with the majority of that DIN coming from wastewater treatment plant discharges. Flows from the Delta deliver 30,000 kg per day of DIN to the SFE (annual average), with the Delta's DIN load varying over a 5-fold range annually (SFEI, 2014a). Based on its areal DIN loads, SFE ranks among the most nutrient-enriched estuaries worldwide (SFEI, 2014a, b; Cloern et al., 2020). Despite its nutrient-enriched status, the SFE has generally not experienced some of the water quality impacts common to other nutrient-enriched estuaries (e.g., excessive phytoplankton blooms, low dissolved oxygen), with SFE's muted response attributed to its highly turbid waters (reduced light penetration within the water column); strong tidal mixing (limiting duration of water column stratification to less than several days); and strong phytoplankton grazing pressure from abundant suspension feeding bivalves in some regions (Alpine and Cloern, 1988; Cole and Cloern, 1984; Jassby, 2008; Kimmerer and Thompson, 2014).

Studies over the past decade have identified changes in responses or sensitivity within SFE to nutrients in deep subtidal habitats via increased phytoplankton biomass (chl-a) and gross primary production (GPP) in South Bay (Cloern et al., 2007, 2010); recently documented occurrences of harmful algae and their associated toxins (Sutula et al., 2017; Peacock et al., 2018); and low dissolved oxygen in some tidal slough habitats (SFEI, 2021). These observations have raised

concerns that SFE's resistance to its high nutrient inputs could be waning (SFEI, 2014b), prompting regulators to initiate the SFB Nutrient Management Strategy (SFBRWQCB, 2017). The early increases in South Bay chl-a (1995-2005) were quantitatively tested using seasonal Kendall, and the signal was sufficiently large and coherent it was also visually apparent in raw data. Given SFE's nutrient-enriched status, there is a critical need for on-going and comprehensive characterization of trends in chl-a, including the ability to examine variability at multiple time-scales and non-monotonic trends, that can also be readily applied to other nutrient-related indicators (e.g., dissolved oxygen, GPP). Beyond the role these tools can play in supporting improved understanding of system dynamics in SFE, water quality managers have emphasized the importance of robust trend detection for informing future nutrient management decisions.

For the trend analyses discussed below, we used near-surface (0-2 m) chlorophyll concentrations (Figure 1) measured biweekly to monthly from 1990 to 2019 along the longitudinal axis of the SFE extending from Central Bay (stations 18-23), South Bay (stations 24-32), and Lower South Bay (stations 34-36) (Table 1, Figure 2). Monitoring data were obtained from the SFE Research Program of the US Geological Survey (Cloern and Schraga, 2016; Schraga et al., 2020). Sampling frequency varied somewhat over time and by station. Approximate monthly or biweekly sampling with coverage of at least a decade is common for many long-term monitoring programs and is the motivating use case for the methods herein. Every observation was included directly in the statistical models without spatial or temporal binning or averaging. Log_{10}-transformed chl-a was used for all analyses to meet assumptions of normally-distributed residuals. Methods for back-transformation of model results are provided in the supplement.

GAMs with uncertainty propagation

We implemented our analysis in three stages. First, we used a GAM to estimate a smooth temporal pattern in the raw data, along with the uncertainty of the smoother. Second, we calculated a feature of interest from the estimated GAM, along with its propagated uncertainty. For the examples described here, we focused on extracting seasonal averages of chl-a values. Other features can also be extracted using the same tools, including the timing or magnitude of a seasonal peak, but those are not presented here (see the *wqtrends* R package, https://tbep-tech.github.io/wqtrends). Third, we used a mixed-effects meta-analysis to estimate trends and test hypotheses about the change in seasonal averages across years.

First-stage analysis: GAM estimation

To smooth the raw data across time, we considered and tested four different GAM structures. While all four GAM structures can achieve similar fits, they differ in how they partition variation in the time series (Table 2), which may unnecessarily influence understanding of temporal patterns. We discuss all four to clarify their relationships and interpretations. All models were created using the `mgcv` R package (R Core Team, 2020; Wood, 2017), with utility functions included in the *wqtrends* package created by the authors (Beck et al., 2021).

The simplest GAM for this purpose is expressed as:

$$\text{Model S: } y_i \sim \beta_0 + f_1(cont\_year_i) + \epsilon_i \qquad (1)$$

where y is measured chl-a, β_0 is an intercept, and $cont\_year$ is "continuous year," a continuous numerical date (e.g., July 1st 2019 would be 2019.5). The $f_1()$ function is a smoothing spline composed of the sum of multiple "basis functions" multiplied by coefficients. $f_1(cont\_year)$

221 describes the relationship of *y* with $cont\_year$ in a way that smoothly follows the data (Wood,
222 2017). The basis functions involve user-specified knots, a grid of values on the $cont\_year$ axis
223 that is discussed more below. The ϵ term represents residuals following a normal distribution
224 with mean zero and constant variance.

225 Smoothing was determined using generalized cross-validation (GCV, as implemented in `mgcv`),
226 which minimizes out-of-sample prediction error. GCV works by penalizing the net curvature of a
227 spline (Wood, 2004). To allow GCV (or other alternatives) to work as intended, the number of
228 knots chosen by the analyst, which determine the maximum degrees of freedom, must be
229 sufficiently large so that the curvature penalty, rather than the number of knots, determines
230 smoothness. Results should not be sensitive to the number of knots; if they are, the number of
231 knots should be increased. In the examples below, we chose the number of knots for $f_1()$ as 12
232 times the number of years in the time series, i.e., one knot per month. If the data were too sparse
233 to fit 12 knots per year, the number of knots was reduced by one knot per year until the model
234 could be estimated (i.e., 12 * years, 11 * years, etc.).

235 The next three spline formulations (Model SY, SYD, and SYDI) provide progressively
236 increasing complexity in how spline terms compose a model to smooth the raw data. Model SY
237 describes the time series using a linear trend plus a spline for $cont\_year$:

238 $$\text{Model SY: } y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + \epsilon_i \qquad (2)$$

239 where equation (2) is the same as equation (1) with the addition of a linear term for $cont\_year$
240 related to y_i by the β_1 slope parameter.

241 While Model SY contains the explicit linear trend term, $\beta_1 cont\_year_i$, it is in fact
242 mathematically equivalent to model S (Table 3). The $f_1()$ spline for *cont_year* includes an

unpenalized linear trend, so a trend will be estimated in model S. When *cont_year* is included explicitly as a linear term in model SY, `mgcv` adjusts the basis functions for the spline to exclude the linear term, thereby not over-parameterizing the model. Whereas an estimated linear trend in *cont_year* and its uncertainty can be extracted from the fitted spline in model S, model SY provides this trend directly, giving the equivalent result. Further, package `mgcv` offers the option to penalize linear trends in splines to provide a method for variable selection (option `select = TRUE`), such as when numerous splines are included in the model formulation for variables that may or may not be important. In the implementation of our approach described here, this option is not used and all models specify `select = FALSE`. Details in the supplement explain this justification.

Model SYD adds an average within-year cyclic pattern as a separate spline:

$$\text{Model SYD: } y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + f_2(doy_i) + \epsilon_i \qquad (3)$$

where equation (3) is the same as equation (2) with the addition of a smoothing spline for "day-of-year" (doy, i.e., Julian date, a count starting January 1 for each year). For $f_2()$, a cyclic spline is specified (using `bs = 'cc'` in `mgcv`) to constrain the start and end at the same value. A user-specified number of knots is also included in $f_2()$. While model SYD is not mathematically equivalent to models S and SY, it should produce nearly identical results. The *doy* spline in model SYD gives the average within-year pattern and changes the interpretation of the *cont_year* spline to represent smoothed deviations from that pattern.

Models S, SY, and SYD can all potentially extract a similar signal from the raw data (Table 3). What differs between the models is the allocation of penalties for curvature used to determine smoothness for each spline. In model SYD, there are separate penalties for the two splines, as

compared to S and SY that include penalties only for the *cont_year* spline. This is important because variation in the response variable can be differently attributed to each spline depending on the model, even while the sum of components for each model produces similar results between models. Our goal is to extract seasonal averages from the fitted time series that are not sensitive to different allocation of penalties among the splines in each model.

If the fits were to differ substantially between model SYD and models S or SY, an interpretation could be difficult because the penalties for smoothing splines based on curvature are heuristic (Wood, 2017). For example, if a lower AIC is achieved in one model compared to another, assuming both use sufficient knots, this may just reflect the outcome of alternative penalization heuristics implied by the different formulations and does not imply one model fit is better. In the examples here, model SYD achieves nearly identical fits to model S or SY, where the latter by definition also achieve identical fits.

Model SYD has the appealing feature that, if some parts of some years have limited data, model SYD will impute an average seasonal pattern with the *doy* spline, thereby considering data from the same period in other years in the prediction of the period with missing data. However, an interpretation of these imputations may be challenging. For example, the spring chl-a peak is a notable feature every year in the SFE. If the peak occurs at the same time every year but the magnitude varies, then the average within-year pattern can be interpreted as the average magnitude. However, if the magnitude is the same but the timing varies across years, then the magnitude of the average peak cannot be similarly interpreted; instead, the average peak extracted by $f_2(doy_i)$ will underestimate the magnitude that usually occurs. Moreover, the width or duration of the peak will be longer than typically occurs in a given year.

Finally, the raw data can be smoothed using a bivariate spline representing an interaction between *cont_year* and *doy*. This can be expressed as:

Model SYDI:

$$y_i \sim \beta_0 + \beta_1 cont\_year_i + f_1(cont\_year_i) + f_2(doy_i) + f_3(cont\_year_i, doy_i) + \epsilon_i \qquad (4)$$

where equation (4) is the same as equation (3) with the addition of a tensor-product smoothing spline (`ti()` in `mgcv`) that varies smoothly as a function of both $cont\_year$ and doy. Both $cont\_year$ and doy include their own number of knots, such that the total number of knots for the spline is the product of the two. The need for sufficient knots in SYDI can be satisfied either by allowing for sufficiently many knots for $f_3()$ or sufficiently many knots for $f_1()$ or $f_2()$, but not both, given limits on the model degrees of freedom.

Following the rationale above, the relationship of model SYDI to model S is similar to that of model SYD to model S. Model SYDI differs from model S to a greater extent than model SYD, but all of the splines use the same inputs to smooth the same data. The univariate splines in *cont_year* and *doy* will combined likely not capture as much variation within model SYDI as captured by model S, given the fewer knots that are available to $f_1()$ or $f_2()$ within SYDI. The tensor-product spline represents an interaction by allowing the pattern in $cont\_year$ to vary by doy and vice-versa. The interaction term provides an appearance that model SYDI is fundamentally different from those provided by the other models. However, models S, SY, and SYD also allow within-year fluctuations to vary across years by allowing a spline to be fit through the entire time series. Although model SYDI is the only model that includes an explicit interaction term, all of the models support the interaction conceptually. By providing this term with sufficient knots, the raw data can be fully smoothed with model SYDI to a similar degree as

for the other models. However, a very large number of knots in both the $cont\_year$ spline *and* in both dimensions of the interaction spline is impossible to achieve. The distinct aspect of model SYDI is the anticipation that within-year fluctuations will vary smoothly from year to year. However, this is an assumption about ecosystem dynamics that may not be appropriate to *a priori* parameterize into a statistical model, including for SFE and many other estuaries where bloom magnitude often varies between years. Thus, the conceptual motivation for model SYDI and its practical application are not necessarily supported in the more generalized application for which we are developing this set of analyses.

Murphy et al. (2019) used spline formulations for Chesapeake Bay water quality analyses that are comparable to those proposed here, but for different goals and with different handling of smoothness. They evaluated a "`gam0`" with only a cyclic spline for *doy* and linear *cont_year* terms, a "`gam1`" like our SYD, and a "`gam2`" like our SYDI. In application, only "`gam2`" was used, including the addition of splines as functions of hydrologic variables to account for finer-scale variation. Murphy et al. (2019) allowed a maximum number of knots for the *cont_year* spline ($f_1()$) of 2/3 times the number of years and do not explicitly consider the number of knots in the interaction spline, following an *ad hoc* allocation of variation in the data to different components based on previous interpretations of water quality dynamics in the system. Constraining splines with insufficient knots could inflate Type I error rates for temporal changes and we seek to lower this risk by increasing the upper limit for the knots for the $f_1()$ term. Finally, Murphy et al. (2019) present large AIC differences between their spline formulations. We instead emphasize that, given sufficient knots, the models represent alternative formulations of conceptually similar explanations for the data and yield similar fits (Table 3), resulting in near ties for AIC between models.

To evaluate the theoretical and conceptual similarities and differences among the GAM structures discussed above, we visually compared chl-a estimates from models S, SYD, and SYDI (Figure 3; note that SY is identical to S and is not shown). Models S, SYD, and SYDI were fit to chl-a data from station 34 using a sufficiently high number of knots for the respective splines for each model. As expected, predictions by day of year from each model are visually similar (Figure 3a) and closely follow the 1:1 line (Figure 3b). However, the component of predictions explained by the continuous year smoother ($f_1()$) differs between models that include additional smoothers (Figure 3c), offering a graphical impression of the degree to which data variability was partitioned to different GAM components. These results are also reflected in differences in the effective degrees of freedom among the additive components of each model (Table 3).

For all results, model S was used with enough knots in $f_1()$ to evaluate chl-a trends across the monitoring stations in the SFE. This model was chosen because of the relatively faster processing time to fit the model, while providing nearly identical explanatory power as compared to the other models (Table 3).

Second-stage analysis: Seasonal features with uncertainties

In the second-stage analysis, we estimated a seasonal average, such as the mean spring chl-a concentrations, along with its associated uncertainty, in each year. We defined μ_t as the seasonal average in year t, $\hat{\mu}_t$ as an estimate of μ_t, and $\hat{\sigma}_{\hat{\mu},t}$ as the estimated standard error of $\hat{\mu}_t$. The season includes n days. For simplicity, the following text omits subscript t.

Point estimates of response values for the fitted GAM take the form $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimates and $\mathbf{X}$ is a model matrix of explanatory variables, including spline

basis function values. Vector $\widehat{\boldsymbol{\beta}}$ includes both fixed effect parameters and spline parameters, and $\mathbf{X}$ contains columns corresponding to each. For example, using model SY, if a point estimate for chl-a is needed for a single day, given as *cont_year* = r, then $\mathbf{X}$ would have a row with 1 in the first column (for the intercept parameter), r (for the linear time trend) in the second column, and an evaluation of each spline basis function at r in the remaining columns. The number of spline basis functions is related to the number of knots. Note that r can be any time, not necessarily the time of an observation.

To obtain a vector, $\hat{\mathbf{y}}$, of fitted point estimates for every day in a season, $\mathbf{X}$ would have one row for each day. Here, the seasonal averages were calculated at the resolution of days. The estimated spline yields both $\widehat{\boldsymbol{\beta}}$ and $\hat{\Sigma}_{\widehat{\beta}}$, an estimate of the covariance matrix of the sampling distribution of $\widehat{\boldsymbol{\beta}}$. The scalar standard errors of $\widehat{\boldsymbol{\beta}}$ are the square roots of the diagonal elements of $\hat{\Sigma}_{\widehat{\beta}}$, whereas the off-diagonal elements are the correlations among the elements of $\widehat{\boldsymbol{\beta}}$. Since parameter estimates are correlated, the covariance of $\hat{\mathbf{y}}$ is $\hat{\Sigma}_{\hat{\mathbf{y}}} = \mathbf{X}\hat{\Sigma}_{\widehat{\beta}}\mathbf{X}^T$.

The estimated seasonal average was calculated from the vector of daily values for each of the n days in the season of interest with $\hat{\mu} = A^T\hat{\mathbf{y}}$, where A^T is a row vector with all values equal to $1/n$. The variance of $\hat{\mu}$ is $\hat{\sigma}_{\hat{\mu}}^2 = A^T\hat{\Sigma}_{\hat{\mathbf{y}}}A$, and its standard error is $\hat{\sigma}_{\hat{\mu}}$. Each of these estimates are from the approximate multivariate normality of the sampling distribution of $\widehat{\boldsymbol{\beta}}$.

Third-stage analysis: Trend analysis of seasonal features with uncertainties

In stage three of the analysis, we used a meta-analysis method to evaluate linear trends across years of seasonal-average water quality, characterized by the within-year means ($\hat{\mu}_t$) and their standard errors ($\hat{\sigma}_{\hat{\mu},t}$) that we estimated in stage two of the analysis. This analysis provided the

basis for directly answering the question *Is there a significant linear trend across a group of years in a seasonal average, taking into account uncertainty*? For example, was there a trend in the spring chl-a average from 1990 to 2000? This question can also be posed in a moving-window manner across a time series (e.g., spring average trend from 1990-2000, 1991-2001, etc.). For all analyses, the response data of interest are $\hat{\mu}_t$, $t = 1, \ldots, N$, with their associated standard errors, $\hat{\sigma}_{\hat{\mu},t}$. N is the number of years of the study.

A mixed-effects meta-analysis model can estimate linear trends when each observation has an associated standard error, which is the case with our estimates $\hat{\mu}_t$ and $\hat{\sigma}_{\hat{\mu},t}$. Differences in standard errors, which may result from different monitoring effort between years, are explicitly considered in the analysis. The model can be expressed using notation similar to Sera et al. (2019):

$$\hat{\mu}_t = \beta_0 + \beta_t t + b_t + \epsilon_t \qquad (5)$$

where β_0 is the intercept, t is the year, β_t is the slope, b_t is the random effect for year t, and ϵ_t is the residual for year t. Accordingly, the seasonal average for year t is $\mu_t = \beta_0 + \beta_t t + b_t$. The "residual," ϵ_t, represents estimation error in $\hat{\mu}_t$, namely $\hat{\mu}_t - \mu_t$. The residuals are assumed to be independent and normally distributed with mean 0 and variances $\hat{\sigma}^2_{\hat{\mu},t}$, where the latter is estimated from the calculations above. The random effect, b_t, is the difference between μ_t and $\beta_0 + \beta_t t$ and is considered the "residual" in the sense of unexplained variation not due to the estimation error. The random effect follows a normal distribution with mean 0 and variance, σ_b^2, to be estimated.

We estimated the model (equation (5)) using the *mixmeta* package in R (Sera et al., 2019). Results from *mixmeta* have a similar interpretation as those from regression analysis, but

parameter estimates and their standard errors incorporate the known standard errors of the response values. The default estimation method for *mixmeta*, restricted maximum likelihood (REML), was used. The meta-analysis models were applied to a chosen sequence or "window" of years for estimating the linear trend.

Trend comparisons

The above methods were applied to each station by evaluating changes in seasonal averages for approximately ten year moving windows from 1991 to 2019. The overall technique and *wqtrends* R package allow for easy and flexible selection of seasonal windows. As an example focused on illustrating the approach (as opposed to extracting mechanistic interpretations), we selected two broad seasons, January-June and July-December, that are generally relevant to phytoplankton bloom phenology in the SFE (Cloern et al., 2020), and combined include conditions over the full year. A moving-window approach encoded within *wqtrends* was used to apply the meta-analysis to each decadal window (e.g., 1991-2001, 1992-2002, etc.), allowing changes in slope and its significance to be interpreted as the window is shifted one year at a time. We interpret the slope as representative for the central year for each block, but a predictive trend for the final year of the window could also be interpreted. As a means for summarizing some results, we focus on the windows 1991-2000, 2000-2010, and 2010-2019.

Finally, trend results from the meta-analysis regression method for each season and different time periods were compared to "naive" across-year regressions that do not propagate uncertainty to demonstrate how different and potentially misleading conclusions can be obtained. Trend estimates were compared to 1) trends from ordinary least squares (OLS) regression applied to seasonal averages from the raw data and 2) trends from OLS regression applied to GAM

seasonal averages. Select examples where differences were pronounced were used for illustration. This analysis was then applied to all stations. The method that formally propagates uncertainty should have more robust statistical properties, such as accurate confidence interval coverage, than naïve methods. For this reason, even when results are similar across methods, the more robust method provides the best support for those results.

Results

Model performance and predictions

Model predictions for chl-a trends across all stations had an average R-squared value of 71% (Table 4) and range from 59% (station 22) to 78% (station 18). GAM predictions from north to south showed more pronounced annual and seasonal changes in chl-a towards the more southern stations (Figures S1-S9). All the models suggested 1) increasing chl-a from 1990 until 2005 to 2010, followed by decreasing chl-a until the end of the record in 2019, 2) a spring chl-a peak, particularly at southern stations, and 3) a fall chl-a peak that was smaller than the spring peak. The magnitude of the fall peak did not vary noticeably by location (Figures S1-S9).

Inter-annual trend estimates

Estimates from the seasonal trend analyses (mixed-effects meta-analysis regressions) across roughly ten-year windows for different seasons are shown for station 34 (Figure 4). Plots a-c show trends in January to June averages while plots d-f show trends in July to December averages. January to June chl-a increased ($\log_{10}$ chl-a slope 0.03 μg L^{-1} yr^{-1}, 0.01-0.05 95% confidence interval) from 1991 to 2000, whereas July to December chl-a did not change

439 significantly. Chl-a also increased from 2000 to 2010, but only for July to December ($\log_{10}$ slope
440 0.03, 0-0.05 95% confidence interval). Finally, chl-a decreased from 2010 to 2019 but again only
441 for July to December ($\log_{10}$ chl-a slope -0.02, -0.04-0 95% confidence interval). Because the
442 trends were confined to certain times of the year, the seasonal estimates provide additional
443 information beyond coarser estimates that cover the entire year.

444 Temporal changes varied among regions of the Bay and were more pronounced at southern
445 stations. Figure 5 shows results from similar analyses as those in Figure 4, but applied to all
446 stations. The seasonal trend analyses showed that increases (based on $p < 0.05$) for the January
447 to June period were observed at stations 32 and 34 from 1991 to 2000; decreases were observed
448 at stations 30 and 32 from 2010 to 2019. For the July to December period, increases were
449 observed at stations 24, 27, 30, and 32 from 1991 to 2000 and stations 18, 21, 22, 24, and 34
450 from 2000 to 2010, whereas decreases were observed at stations 30, 32, and 34 from 2010 to
451 2019.

452 Results from a ten-year moving window comparison of seasonal trends provided additional
453 context on when significant changes were occurring at each station (Figure 6). Trends were
454 observed at all stations that followed a general pattern of increases early in the record followed
455 by decreases later in the record. Increases and decreases were observed in both the January to
456 June and July to December seasonal periods, with some notable exceptions. In particular, the
457 most southern stations (32, 34, 36) had increasing trends prior to 2005 that were more often
458 observed in the July to December period. Additionally, chl-a at the more northern stations has
459 not changed in recent years for either seasonal period. For most stations and seasonal periods, a
460 change from increasing to decreasing chl-a occurred around 2007.

Importance of uncertainty propagation

Results showing trend estimates from meta-analysis on GAM seasonal averages provided different conclusions than those from either OLS regression through seasonal averages from raw data (Figure 7 row 1) or OLS regression through GAM seasonal averages without uncertainty propagation (Figure 7 row 2). Figure 7a shows trend estimates for station 30 for January to June averages from 2000 to 2010. Only the meta-analysis regression results show a trend in this example (based on $p < 0.05$). The OLS regression on observed estimates (top plot) and OLS regression on GAM estimates (middle plot) did not identify trends. Figure 7b shows trend estimates at station 34 for July to December averages from 1991 to 2000. Unlike the first example, only the middle figure shows a trend, whereas the top and bottom plots do not show trends. In both cases, only the meta-analysis results give reliable conclusions because of full propagation of uncertainty across methods. Even in cases where the p-value threshold is not of interest, the confidence intervals from the alternative methods will be inaccurate.

Applying the same comparison to all stations showed that different trend analysis methods provided conflicting information on the magnitude and significance of the seasonal chl-a changes in each decade (Figure 8). In many cases, the slope estimates were similar in magnitude, with some exceptions at the more southern stations where the OLS estimates suggested a larger trend than the meta-analysis methods. More importantly, differences in the magnitude of the confidence intervals between the OLS models applied to the GAM averages and the meta-analyses were also observed, reflecting the ability of the latter to more accurately assess significance of trends by accounting for uncertainty in the average estimates.

Discussion

Propagation of uncertainty from within-year features of estimated GAMs to across-year trends using mixed-effects meta-analysis is a new approach that can address different questions than previous methods. Our approach has several advantages over more conventional approaches for analysis of water quality data from long-term monitoring programs. GAMs are capable of modelling time series with missing observations or irregular sampling which can complicate trend assessment and comparison of trends between locations (Junninen et al., 2004; Racault et al., 2014). As noted above, non-parametric approaches (i.e., seasonal Kendall tests) are by far the most common trend analysis methods applied to long-term water quality data (Helsel et al., 2020; Hirsch et al., 1982). These methods only assess the direction and significance of comparisons across years, and importantly, do not account for full propagation of uncertainty inherent in raw observations if the raw data are aggregated to meet test requirements. Aggregation of raw data, e.g., averaging of observations within a year or season to comply with the requirements of Kendall tests, risks loss of information by removing variation between observations at smaller time scales. The logical outcome is increased risk of incorrect conclusions from test results.

Incorrect conclusions on trends can have dramatic consequences for regulated parties under existing water quality compliance frameworks (Smith et al., 2001). Our examples in Figures 7 and 8 demonstrate these risks if propagation of uncertainty from raw observations across methods is unaccounted for in trend assessment. The “naïve” method using OLS regression applied to seasonal averages from the raw observations fails to propagate uncertainty, similarly to averaging results within a year and applying a simple Kendall test. In some cases the results

504 may be similar to those from fully propagating uncertainty, but the loss of information can lead
505 to increased Type I or II error rates depending on characteristics of the raw data and the method
506 used for their evaluation (Shabman and Smith, 2003). Our examples demonstrated the increased
507 potential for incorrect conclusions at specific monitoring locations and, at larger spatial scales,
508 across all stations if simpler trend analyses are used. Even though simpler methods may produce
509 similar results in some cases, particularly with frequent sampling and similar effort between
510 years, the only way to confirm such an outcome would be to compare results, relying on the
511 method with full propagation of uncertainty to be the more robust method. Use of methods that
512 fully account for uncertainty is recommended to obtain statistically valid results in a wider range
513 of conditions.

514 Results here also show that GAM structure (i.e., choice of smoothing terms) was less important
515 than allowing the model sufficient freedom to fit the data. This is an important conclusion that
516 provides guidance on how GAMs could be used to model time series from long-term
517 environmental monitoring programs. Models with separate smoothers for continuous year and
518 day of year can produce nearly identical results in the predicted trends if the knots are
519 sufficiently high to allow the GAMs to be fit as intended by the methods in the mgcv package
520 (Figure 3). The approach presented here leverages the ability of GAMs to objectively estimate
521 smoothed trends across years by identifying an optimal level of smoothing using generalized
522 cross-validation to extract an underlying signal in the observed data (Wood, 2017, 2004).

523 The underlying cross-validation methods used by GAMs in the mgcv package also reduce the
524 decisions that may be necessary for the implementation of alternative trend assessment methods.
525 For example, WRTDS and similar smoothing approaches (e.g., LOESS) require decisions on
526 appropriate window widths or bandwidths to define the neighborhood of observations for

smoothing (Hirsch et al., 2010; Wan et al., 2017). This is especially problematic for policy analysis or regulatory decisions if the results change based on arbitrary decisions of the analyst. Because these decisions are not needed for GAMs, the results can be considered a more objective and potentially accurate signal of actual trends that are minimally influenced by process or observation error present in the raw data.

Several limitations of the proposed methods deserve mention. First, if sampling is so irregular that important fluctuations are missed entirely in some years, the GAM estimates and uncertainty propagation could become dubious in interpretation and usefulness. Second, estimation of GAMs for very long series can be computationally demanding. When this is an obstacle, one could do the first two analysis stages using temporal windows of the full data, with the only implication being that different degrees of smoothness may be estimated for different windows, which indeed might be justified by the data. Third, meta-analysis regression results for a very small number of years, particularly confidence intervals and associated p-values, may be inaccurate (e.g. in confidence interval coverage). In such cases, one could make alternative use of the GAM seasonal averages and standard errors, such as for pairwise comparisons among years.

Future work

Additional work could be conducted to further strengthen the conclusions based on trends from meta-analysis regression applied to the GAM seasonal averages. Our third stage analyses require *a priori* decisions on long-term time scales of interest and future work could generalize these choices. Although there are undoubtedly many scenarios where years of interest can be chosen objectively by the needs of an analysis (e.g., regulatory compliance periods, time since management intervention), a more general question of when changes occur independent of user

549 decisions is also important to address. Additional methods could be developed using objective
550 criteria to identify inflection points or other important periods where changes occur independent
551 of a user choice. Assessing water quality changes beyond an evaluation of seasonal averages
552 could also be possible with our approach, such as assessing changes in the timing or magnitude
553 of a seasonal peak across years.

554 Additional explanatory variables could be identified to explain trends in either the GAM stage or
555 the meta-analysis stage of analysis. This information would have obvious implications for
556 management decisions on factors that influence water quality changes, e.g., wastewater treatment
557 upgrades, large-scale climatic factors, or flow regulation practices. Including alternative
558 predictors in the GAMs (Wood and Augustin, 2002; Zuur et al., 2009) could reduce the
559 uncertainties of its estimates and, if relevant, allow the influence of those variables to be
560 removed. Including alternative predictors in the meta-analysis could help in explaining long-term
561 trends of seasonal averages or other metrics obtained from the GAMs. Our goal here was to
562 describe chl-a changes relative to time, so the single predictor in both modeling stages was time.

563 Finally, the evaluation of trends for alternative water quality variables in addition to chl-a is a
564 simple and logical extension of the methods proposed in this study. The long-term monitoring
565 program maintained by USGS includes multiple parameters in addition to chl-a that can provide
566 additional context into broader water quality trends in the SFE (Cloern and Schraga, 2016;
567 Schraga et al., 2020). These parameters include salinity, temperature, light attenuation, dissolved
568 oxygen, suspended particulate matter, and dissolved inorganic nutrients, which collectively can
569 be used to provide a broader understanding of potential eutrophication patterns or ecosystem
570 shifts at seasonal and multi-decadal scales. Chl-a measurements can also be used to estimate
571 gross primary production to assess process rates that may be more indicative of system function

572 (Cloern et al., 2007; Jassby et al., 2002). The open-source *wqtrends* R package (Beck et al.,
573 2021) developed for this manuscript can be used for these analyses to provide additional insight
574 into potential drivers of water quality change in the SFE and other estuarine systems.

575 Acknowledgments

576 This work was supported by funding from the San Francisco Bay Nutrient Management Strategy
577 (NMS). We thank the staff of the US Geological Survey that collect and maintain long-term
578 monitoring data in San Francisco Bay. This work benefited from discussions with the NMS
579 Nutrient Technical Workgroup and Steering Committee. We thank James D. Hagy III and two
580 anonymous reviewers for providing helpful comments on this manuscript.

References

Alpine, A.E., Cloern, J.E., 1988. Phytoplankton growth rates in a light-limited environment, San Francisco Bay. Marine Ecology Progress Series 44, 167–173.

Beck, M.W., de Valpine, P., Murpy, R., Wren, I., Chelsky, A., Foley, M., Senn, D., 2021. tbep-tech/wqtrends: v1.1.0 (Version v1.1.0). Zenodo. http://doi.org/10.5281/zenodo.4509638.

Beck, M.W., Hagy, J.D., III, 2015. Adaptation of a weighted regression approach to evaluate water quality trends in an estuary. Environmental Modelling and Assessment 20, 637–655. https://doi.org/10.1007/s10666-015-9452-8

Beck, M.W., Jabusch, T.W., Trowbridge, P.R., Senn, D.B., 2018. Four decades of water quality change in the upper San Francisco Estuary. Estuarine, Coastal and Shelf Science 212, 11–22. https://doi.org/10.1016/j.ecss.2018.06.021

Beck, M.W., Murphy, R.R., 2017. Numerical and qualitative contrasts of two statistical models for water quality change in tidal waters. Journal of the American Water Resources Association 53, 197–219. https://doi.org/10.1111/1752-1688.12489

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A seasonal-trend decomposition procedure based on Loess. Journal of Official Statistics 6, 3–73.

Cloern, J.E. 2018. Patterns, pace, and processes of water-quality variability in a long-studied estuary. Limnology and Oceanography 64, S192-S208. https://doi.org/10.1002/lno.10958

Cloern, J.E., Jassby, A.D., 2012. Drivers of change in estuarine-coastal ecosystems: Discoveries from four decades of study in San Francisco Bay. Reviews of Geophysics 50. https://doi.org/10.1029/2012RG000397

Cloern, J.E., Jassby, A.D., 2010. Patterns and scales of phytoplankton variability in estuarine-coastal ecosystems. Estuaries and Coasts 33, 230–241.

Cloern, J.E., Jassby, A.D., Thompson, J.K., Hieb, K.A., 2007. A cold phase of the East Pacific triggers new phytoplankton blooms in San Francisco Bay. Proceedings of the National Academy of Sciences of the United States of America 104, 18561–18565.

Cloern, J.E., Schraga, T.S., 2016. USGS measurements of water quality in San Francisco Bay (CA), 1969-2015: U.S. Geological Survey data release. https://doi.org/10.5066/F7TQ5ZPR.

Cloern, J.E., Shcraga, T.S., Nejad, E., Martin, C., 2020. Nutrient status of San Francisco Bay and its management implications. Estuaries & Coasts 43, 1299–1317. https://doi.org/10.1007/s12237-020-00737-w

Cole, B.E., Cloern, J.E., 1984. Significance of biomass and light availability to phytoplankton productivity in San Francisco Bay. Marine Ecology Progress Series 17, 15–24.

Cumming, G.S., Cumming, D.H.M., Redman, C.L., 2006. Scale mismatches in social-ecological systems: Causes, consequences, and solutions. Ecology and Society 11, 14.

Forbes, D.J., Xie, Z., 2018. Identifying process scales in the Indian River Lagoon, Florida using wavelet transform analysis of dissolved oxygen. Ecological Complexity 36, 149–167. https://doi.org/10.1016/j.ecocom.2018.07.005

Gasparrini, A., Armstrong, B., Kenward, M.G., 2012. Multivariate meta-analysis for non-linear and other multi-parameter associations. Statistics in Medicine 31, 3821–3839. https://doi.org/10.1002/sim.5471

Hafen, R.P., 2010. Local regression models: Advancements, applications, and new methods (PhD thesis). Purdue University, West Lafayette, Indiana.

Haraguchi, L., Carstensen, J., Abreu, P.C., Odebrecht, C., 2015. Long-term changes of the phytoplankton community and biomass in the subtropical shallow Patos Lagoon Estuary, Brazil. Estuarine, Coastal and Shelf Science 162, 76–87.

He, S., Mazumdar, S., Arena, V.C., 2006. A comparative study of the use of GAM and GLM in air pollution research. Environmetrics 17, 81–93. https://doi.org/10.1002/env.751

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., Gilroy, E.J., 2020. Statistical methods in water resources, 2nd ed. U.S. Geological Survey Techniques; Methods, book 4, chapter A3, version 1.1, Reston, Virginia.

Hirsch, R.M., Archfield, S.A., De Cicco, L.A., 2015. A bootstrap method for estimating uncertainty of water quality trends. Environmental Modelling and Software 73, 148–166. https://doi.org/10.1016/j.envsoft.2015.07.017

Hirsch, R.M., Moyer, D.L., Archfield, S.A., 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, 857–880.

Hirsch, R.M., Slack, J.R., Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data. Water Resources Research 18, 107–121.

Jassby, A.D., 2008. Phytoplankton in the Upper San Francisco Estuary: Recent biomass trends, their causes, and their trophic significance. San Francisco Estuary and Watershed Science 6, 1–24.

Jassby, A.D., Cloern, J.E., Cole, B.E., 2002. Annual primary production: Patterns and mechanisms of change in a nutrient-rich tidal ecosystem. Limnology and Oceanography 47, 698–712.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kimmerer, W.J., Thompson, J.K., 2014. Phytoplankton growth balanced by clam and zooplankton grazing and net transport into the low-salinity zone of the San Francisco Estuary. Estuaries and Coasts 37, 1202–1218.

Lefcheck, J.S., Wilcox, D.J., Murphy, R.R., Marion, S.R., Orth, R.J., 2017. Multiple stressors threaten the imperiled coastal foundation species eelgrass (*zostera marina*) in Chesapeake Bay, USA. Global Change Biology 23, 3474–3483. https://doi.org/10.1111/gcb.13623

Lortie, C.J., 2014. Formalized synthesis opportunities for ecology: Systematic reviews and meta-analyses. OIKOS 123, 897–902. https://doi.org/10.1111/j.1600-0706.2013.00970.x

Morton, R., Henderson, B.L., 2008. Estimation of nonlinear trends in water quality: An improved approach using generalized additive models. Water Resources Research 44, W07420. https://doi.org/10.1029/2007WR006191

Murphy, R.R., Perry, E., Harcum, J., Keisman, J., 2019. A Generalized Additive Model Approach to evaluating water quality: Chesapeake Bay case study. Environmental Modelling & Software 118, 1–13. https://doi.org/10.1016/j.envsoft.2019.03.027

Navarro, G., Caballero, I., Prieto, L., Vázquez, A., Flecha, S., Huertas, I.E., Ruiz, J., 2012. Seasonal-to-interannual variability of chlorophyll-*a* bloom timing associated with physical forcing in the Gulf of Cádiz. Advances in Space Research 50, 1164–1172. https://doi.org/10.1016/j.asr.2011.11.034

Peacock, M.B., Gibble, C.M., Senn, D.B., Cloern, J.E., Kudela, R.M. 2018. Blurred lines: Multiple freshwater and marine algal toxins at the land-sea interface of San Francisco Bay, California. Harmful Algae 73, 138–147. https://doi.org/10.1016/j.hal.2018.02.005

Pearce, J.L., Beringer, J., Nicholls, N., Hyndman, R.J., Tapper, N.J., 2011. Quantifying the influence of local meteorology on air quality using generalized additive models. Atmospheric Environment 45, 1328–1336. https://doi.org/10.1016/j.atmosenv.2010.11.051

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, R v4.0.2, Vienna, Austria.

Racault, M.F., Sathyendranath, S., Platt, T., 2014. Impact of missing data on the estimation of ecological indicators from satellite ocean-colour time series. Remote Sensing of Environment 152, 15–28. https://doi.org/10.1016/j.rse.2014.05.016

Schraga, T.S., Nejad, E.S., Martin, C.A., Cloern, J.E., 2020. USGS measurements of water quality in San Francisco (CA), beginning in 2016 (ver. 3.0, March 2020): U.S. Geological Survey data release. https://doi.org/10.5066/F7D21WGF.

Sera, F., Armstrong, B., Blangiardo, M., Gasparrini, A., 2019. An extended mixed-effects framework for meta-analysis. Statistics in Medicine 38, 5429–5444. https://doi.org/10.1002/sim.8362

SFBRWQCB (San Francisco Bay Regional Water Quality Control Board), 2017. Water quality control plan (basin plan) for the San Francisco Bay basin. Prepared by California Regional Water Quality Control Board, San Francisco Bay Region, Oakland, CA. https://www.waterboards.ca.gov/sanfranciscobay/basin_planning.html

SFEI (San Francisco Estuary Institute), 2014a. External nutrient loads to San Francisco Bay. SFEI Contribution Number 704. San Francisco Estuary Institute, Richmond, CA.

SFEI (San Francisco Estuary Institute), 2014b. Scientific foundation for the San Francisco Bay Nutrient Management Strategy. SFEI Contribution Number 979. San Francisco Estuary Institute, Richmond, CA.

SFEI (San Francisco Estuary Institute), 2021. Connections to tidal marsh and restored salt ponds drive seasonal and spatial variability in ecosystem metabolic rates in Lower South San Francisco Bay. SFEI Contribution No. 992. San Francisco Estuary Institute, Richmond, CA.

Shabman, L., Smith, E., 2003. Implications of applying statistically based procedures for water quality assessment. Journal of Water Resources Planning and Management 129, 330–336. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:4(330)

Smith, E.P., Ye, K., Hughes, C., Shabman, L., 2001. Statistical assessment of violations of water quality standards under section 303 (d) of the Clean Water Act. Environmental science & technology 35, 606–612. https://doi.org/10.1021/es001159e

Stow, C.A., Cha, Y., Johnson, L.T., Confesor, R., Richards, R.P., 2015. Long-term and seasonal trend decomposition of Maumee River nutrient inputs to western Lake Erie. Environmental Science and Technology 49, 3392–3400. https://doi.org/10.1021/es5062648

Sutula, M.A., Kudela, R.M., Hagy, J.D., III, Harding, L.W., Jr., Senn, D.B., Cloern, J.E., Bricker, S., Berg, G.M., Beck, M.W. 2017. Novel analyses of long-term data provide a scientific basis for chlorophyll-a thresholds in San Francisco Bay. Estuarine, Coastal and Shelf Science. 197, 107-118. https://doi.org/10.1016/j.ecss.2017.07.009

Urquhart, N.S., Paulsen, S.G., Larsen, D.P., 1998. Monitoring for policy-relevant regional trends over time. Ecological Applications 8, 246–257. https://doi.org/10.1890/1051-0761(1998)008[0246:MFPRRO]2.0.CO;2

Wan, Y., Wan, L., Li, Y., Doering, P., 2017. Decadal and seasonal trends of nutrient concentration and export from highly managed coastal catchments. Water Research 115, 180–194.

Wood, S.N., 2017. Generalized additive models: An introduction with r, 2nd ed. Chapman; Hall, CRC Press, London, United Kingdom.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686. https://doi.org/10.1198/016214504000000980

Wood, S.N., 2003. Thin-plate regression splines. Journal of the Royal Statistical Society (B) 65, 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157, 157–177. https://doi.org/10.1016/S0304-3800(02)00193-X

Yang, G., Moyer, D.L., 2020. Estimation of nonlinear water-quality trends in high-frequency monitoring data. Science of The Total Environment 715, 10.1016/j.scitotenv.2020.136686.

727 Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed effects models
728 and extensions in ecology with r. Springer-Verlag, New York, New York.

729

Figures

Figure 1: Observed chl-a concentrations for all stations in central and south San Francisco Estuary (18-36, Figure 2), with (a) annual summer/fall concentrations (Aug - Dec) and (b) monthly concentrations by decade.

Figure 2: Station locations in the central and south San Francisco Estuary used for analysis. See Table 1 for station descriptions. Full dataset described in Schraga et al. (2020).

Figure 3: GAM output of estimated chl-a at station 32 for models S, SYD, and SYDI. Model SY is identical to S and is not shown. Plots in (a) show model predictions by day of year with separate lines for each year. Plots in (b) show pairwise comparisons of predicted chl-a between the models and plots in (c) show the same comparisons as in (b) but only for results from the estimated smoother for the cont_year variable. The plots demonstrate that results between the models are similar except for a few observations at extreme values (a, b), but they vary in how they allocate contributions to the predictions among different additive splines (c). The 1:1 lines are in red to facilitate comparisons.

*Figure 4: Examples of seasonal averages and trend estimates in ten year blocks from meta-analyses using results of GAM predictions for station 34. Plots (a), (b), and (c) show trend estimates for January to June averages and (d), (e), and (f) show trend estimates for July to December averages. The trend lines estimate the rate of change of chl-a per year, reported as the log_{10}-slope (+/- 95 % confidence interval) in the sub-plot titles. ns: not significant at $\alpha = 0.05$, * $p < 0.05$*

Figure 5: Interannual trend estimates of seasonal averages by decade for chl-a at each station. Point type and color represent the direction and magnitude of an estimated trend as the log_{10} slope for chl-a concentration per year. Trends with $p < 0.05$ are marked with an asterisk. All results are from Model S.

Figure 6: Estimates of log_{10} chl-a change per year (+/- 95% confidence interval) from applying the meta-analysis across the seasonal averages for each station. Stations are arranged top to bottom from north to south. Plots in (a) show estimates for seasonal averages from January to June and plots in (b) show estimates for seasonal averages from July to December. Results are from a ten-year, centered moving window where each point shows a linear trend estimate from five years prior to five years after each year. Estimates prior to 1996 and after 2014 are not available because of an incomplete ten year record for estimating the trend. Significant estimates are shown in red.

Figure 7: Trend estimate comparisons (arithmetic scale) for three models applied to seasonal averages of chl-a in different annual periods at stations (a) 30 and (b) 34. The first row shows OLS (ordinary least squares) regression applied to seasonal averages of chl-a from the raw data, the second row shows OLS regression applied to seasonal averages of chl-a from the GAM (without error propagation), and the third row shows meta-analysis regression applied to the seasonal averages of chl-a from the GAM. Regressions in each plot are fit through the seasonal estimates indicated in the plot titles for a specified year range.

Figure 8: Trend estimate comparisons for three models applied to seasonal averages of chl-a in different annual periods at each station. The "OLS raw" trend model is based on an ordinary

least squares (OLS) regression fit to the seasonal averages of chl-a from the raw data, the "OLS GAM" trend model is based on an OLS regression fit to the seasonal averages of chl-a from the GAM model (without error propagation), and the "Meta-analysis GAM" trend model is based on a meta-analysis regression fit to the seasonal averages of chl-a from the GAM model. Values for each model are the log_{10}-slope estimates (+/- 95% confidence interval) as annual change per year within each season, with line style denoting trend significance.

Tables

Table 1: Station locations, sample sizes (from 1991 to 2019), and summary values (median, minimum, maximum) for chl-a ($\mu g\ L^{-1}$). Stations are arranged from north to south.

Station	Latitude	Longitude	n	Med.	Min.	Max.
18	37.836	-122.418	414	3.6	0.2	16.6
21	37.784	-122.351	576	4.4	0.6	40.0
22	37.752	-122.351	569	4.0	0.7	53.1
24	37.686	-122.334	595	4.2	0.7	47.3
27	37.617	-122.285	596	4.5	0.5	50.9
30	37.551	-122.184	608	5.1	0.8	112.2
32	37.517	-122.133	591	5.9	0.7	282.1
34	37.485	-122.086	544	6.5	0.6	158.3
36	37.468	-122.067	476	6.2	1.1	328.4

Table 2: Summary and details for each of the GAM structures. In practice, a sufficiently large number of knots provided to the additive terms will produce identical or comparable estimates for a response variable. The models differ in the allocation of penalties for the smoothness of each spline.

GAM	Additive components	Details
S	$f_1(cont\_year)$	A single smoother over a continuous year variable
SY	$\beta_1 cont\_year + f_1(cont\_year)$	A linear continuous year variable and a single smoother over a continuous year variable
SYD	$\beta_1 cont\_year + f_1(cont\_year) + f_2(doy)$	A linear continuous year variable, a smoother over a continuous year variable, and a smoother over a day of year variable
SYDI	$\beta_1 cont\_year + f_1(cont\_year) + f_2(doy) + f_3(cont\_year, doy)$	A linear continuous year variable, a smoother over a continuous year variable, a smoother over a day of year variable, and an interaction smoother across continuous year and day of year variables

*Table 3: Comparison of the four model structures (S, SY, SYD, SYDI) described in the first stage analysis of GAM estimation. The four models provide either identical or comparable ability to describe chl-a trends at an example station (32) in the southern end of the San Francisco Estuary. The models differ in additive smoothers and the amount of effective degrees of freedom (edf) in the smoothers (measure of wiggliness in each component), but the overall model predictions are similar. AIC: Akaike Information Criterion, GCV: generalized cross-validation score, R2: r-squared values for predictions, edf: effective degrees of freedom, F: F-statistic, p-val: probability value, ** p < 0.001*

model	AIC	GCV	R2	smoother	edf	F	p-val
S	-138.2	0.06	0.74	s(cont_year)	242.23	6.27	**
SY	-138.2	0.06	0.74	s(cont_year)	242.23	6.27	**
SYD	-135.66	0.06	0.74	s(cont_year)	229.33	3.88	**
				s(doy)	8.07	0.11	**
SYDI	-123.57	0.06	0.73	s(cont_year)	136.88	3.13	**
				s(doy)	9.54	0.79	**
				ti(cont_year,doy)	69.31	0.74	**

Table 4: Model performance statistics for each station as Akaike Information Criterion scores (AIC), generalized cross-validation scores (GCV), and r-squared values.

station	AIC	GCV	R-squared
18	-430.94	0.04	0.78
21	-305.93	0.04	0.70
22	-160.07	0.05	0.59
24	-250.77	0.04	0.69
27	-188.72	0.05	0.72
30	-172.79	0.05	0.74
32	-138.20	0.06	0.74
34	-3.17	0.07	0.68
36	-22.05	0.07	0.73

Figure 1

Click here to access/download;Figure;Fig1.tiff

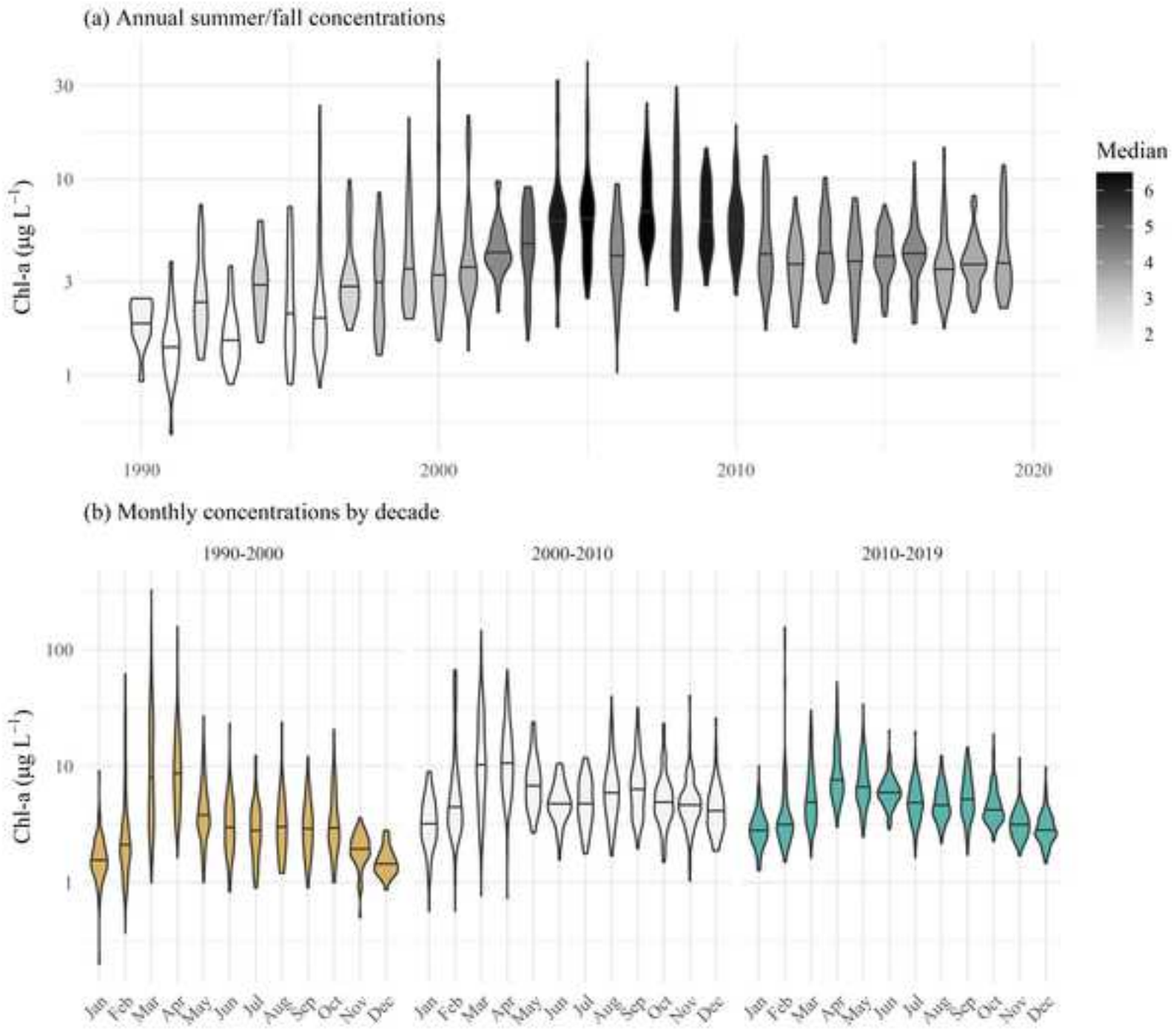


Figure 2

Click here to access/download;Figure;Fig2.tiff

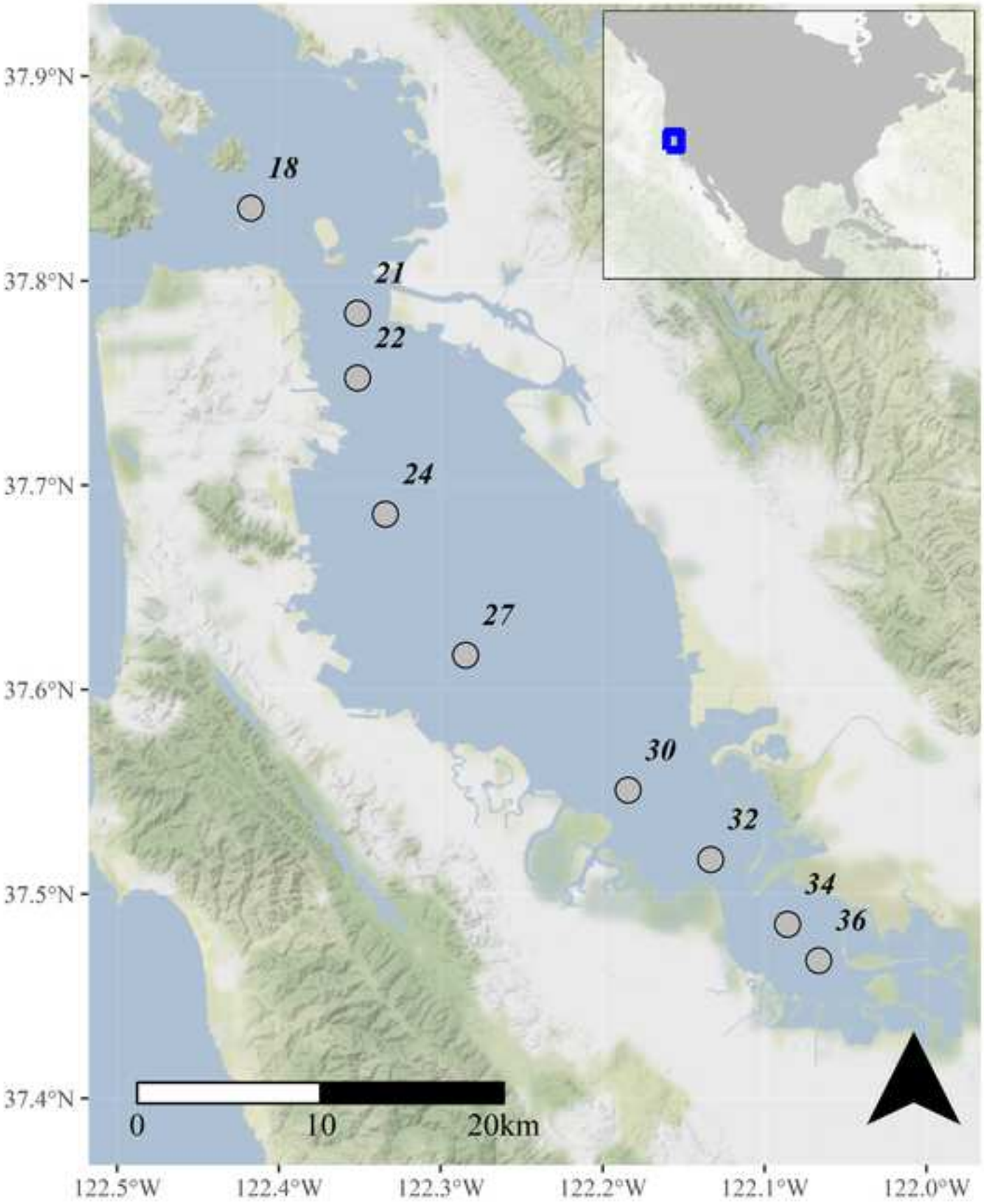


Figure 3

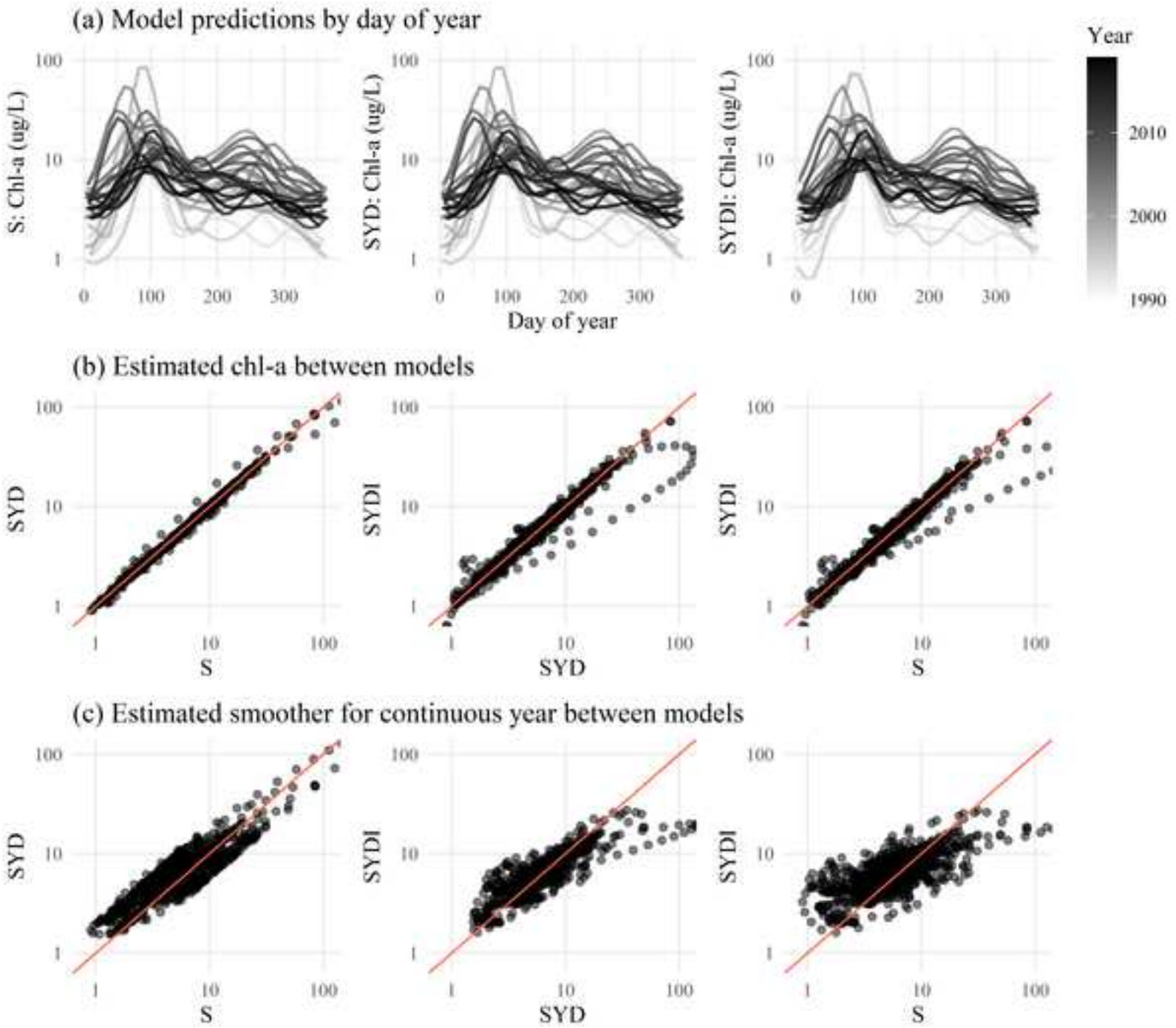


Figure 4

Click here to access/download;Figure;Fig4.tiff

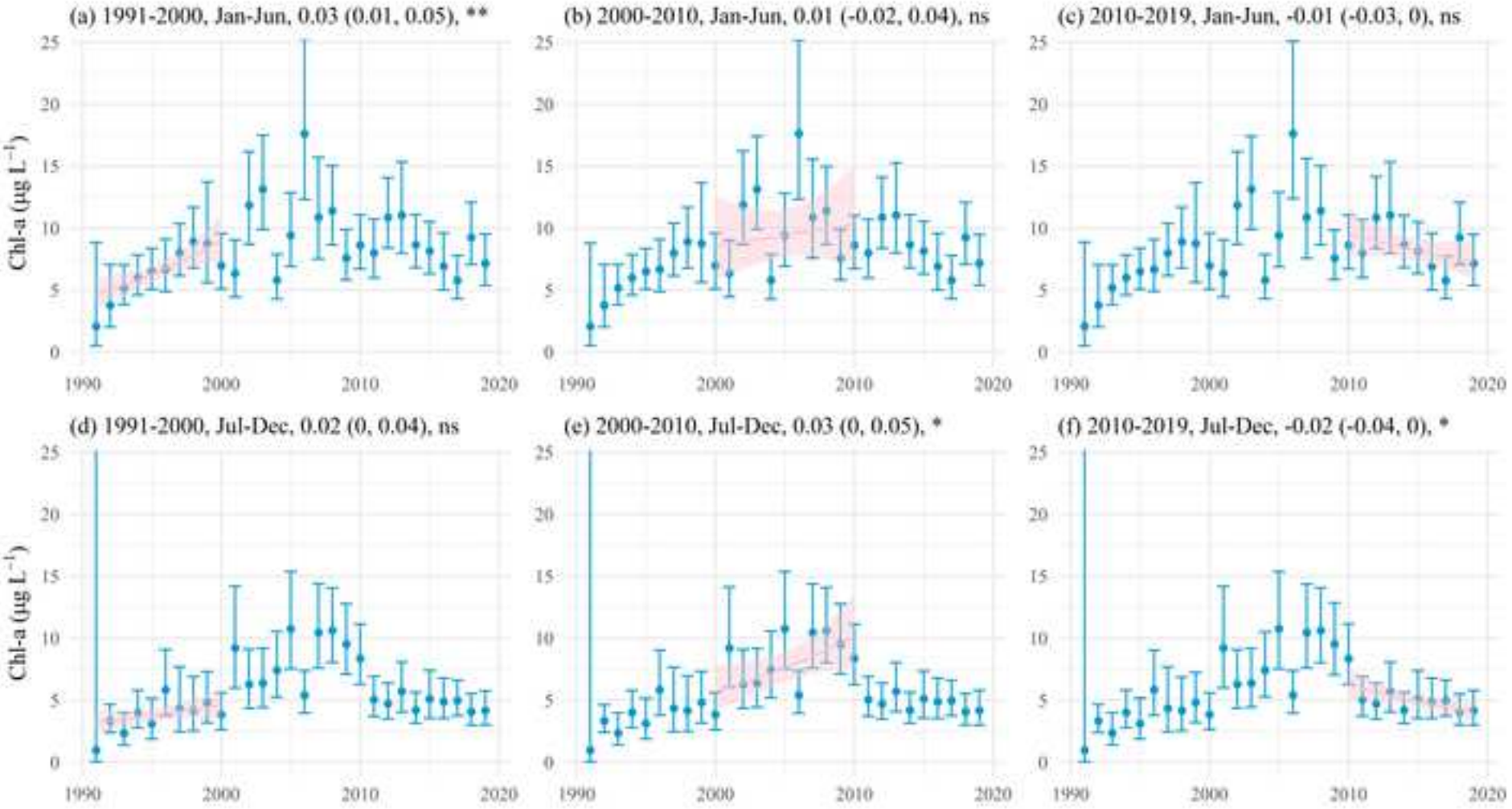


Figure 5

Click here to access/download;Figure;Fig5.tiff

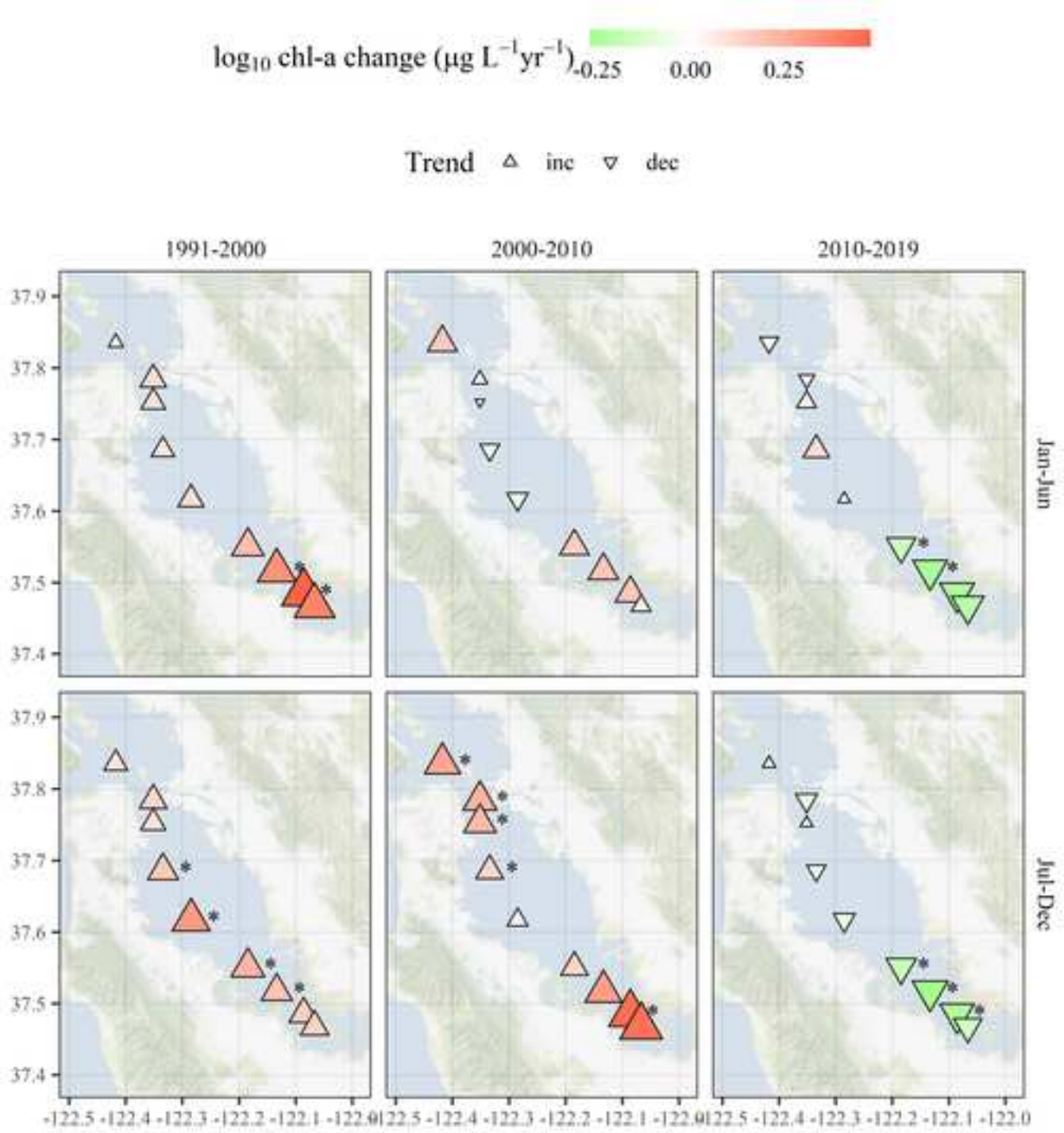


Figure 6

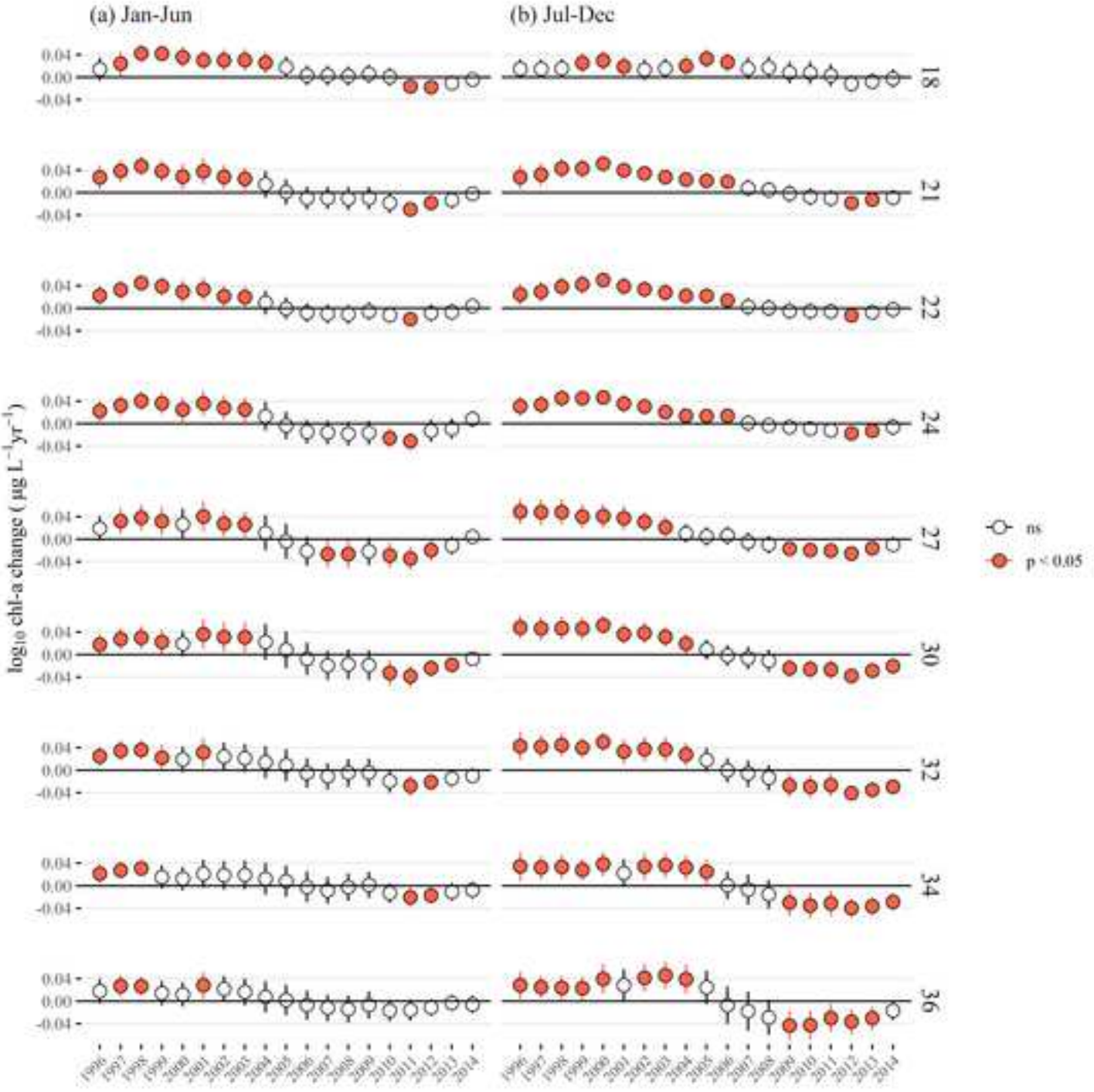


Figure 7

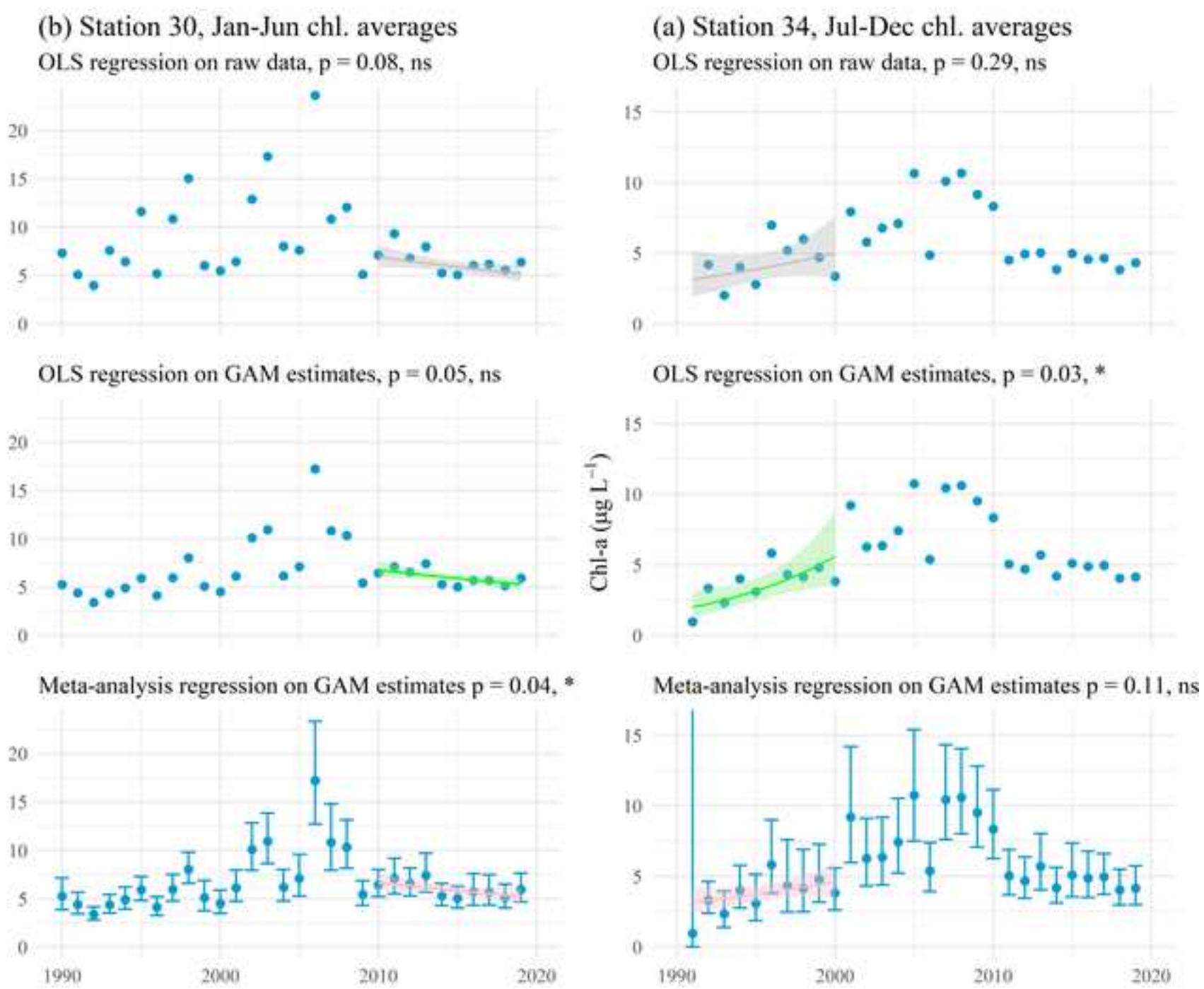
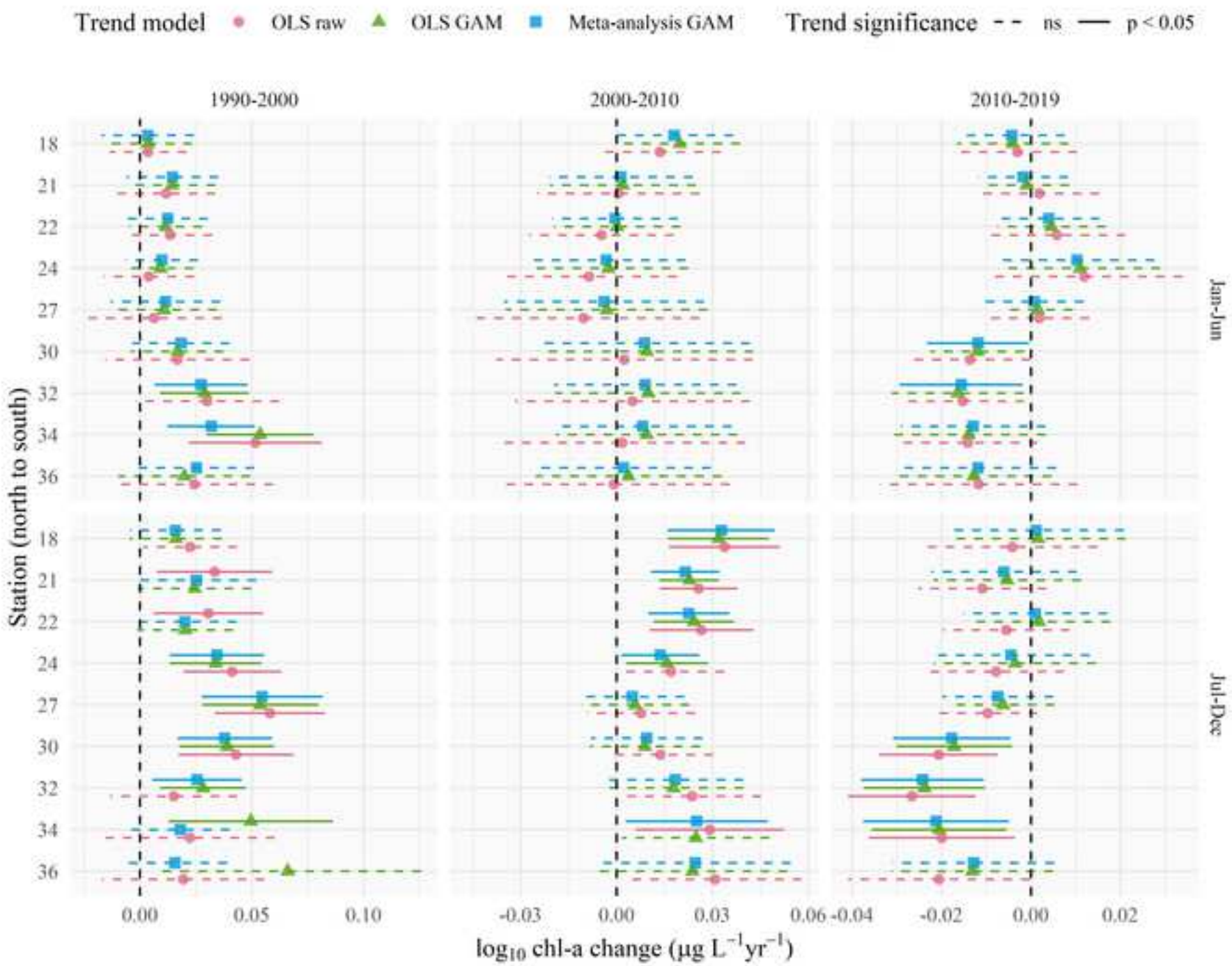


Figure 8

Click here to access/download;Figure;Fig8.tiff



Credit Author Statement

Marcus W. Beck: Conceptualization, Data Curation, Formal analysis, Methodology, Software, Writing – Original Draft, **Perry de Valpine**: Conceptualization, Methodology, Software, Writing – Original Draft, **Rebecca Murphy**: Conceptualization, Writing – Review & Editing, **Ian Wren**: Conceptualization, Writing – Review & Editing, **Ariella Chelsky**: Conceptualization, Writing – Review & Editing, **Melissa Foley**: Conceptualization, Project administration, Writing – Review & Editing, **David B. Senn**: Conceptualization, Data Curation, Project administration, Writing – Review & Editing

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: