Science of the Total Environment

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

--Manuscript Draft--

Manuscript Number:	
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:	Accurate and flexible trend assessment methods are valuable tools for describing historical changes in 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 of monitoring data and trends of interest. Here, we propose a novel application of generalized additive models (GAMs) to model seasonal and multi-decadal changes in a long-term monitoring dataset of 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 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 of the 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.
Suggested Reviewers:	Jereme Gaeta Jereme.Gaeta@wildlife.ca.gov
	Sam Bashevkin sam.bashevkin@deltacouncil.ca.gov
	Will Smith william_e_smith@fws.gov
	Christina Burdi Christina.Burdi@wildlife.ca.gov
	Michael Beakes mbeakes@usbr.gov

Opposed Reviewers:	James Hagy hagy.jim@epa.gov Previously reviewed the manuscript.

Cover Letter

TAMPA BAY ESTUARY PROGRAM
PARTNERSHIP FOR A HEALTHY BAY

May 7th, 2021

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

We are pleased to submit 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.

Our manuscript presents a novel application of generalized additive models (GAMs) to describe seasonal and multi-decadal changes in monitoring data, using the USGS San Francisco Bay water quality data as a case study. Conclusions from trend analyses are often used to inform management decisions that can affect long-term quality of environmental resources. Many environmental monitoring programs collect temporally resolved but irregular time series data to quantify these trends. Mismatches between the scales of monitoring and analysis or management questions presents statistical challenge. Existing methods (e.g., non-parametric trend tests, time series decomposition) may not accurately describe these trends, particularly as it relates to the propagation of uncertainty from incomplete data and estimates from trend analysis models. Our proposed methods estimate seasonal averages in a response variable with GAMs 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. By doing so, more accurate trend assessments are obtained.

Our approach has broad appeal because the problems it addresses are not unique to our examples and we provide a supplemental software package for others to apply our developed techniques. For example, the COVID19 pandemic created challenges for many monitoring agencies due to the logistics of sampling to ensure the safety of laboratory and field personnel. As a result, data gaps during 2020 due to missed sampling events as agencies adapted to new safety protocols are now common features of many long-term datasets. The analysis methods we provide in our manuscript can accommodate data gaps in trend assessment by providing an accurate estimate of uncertainty from missing data, while still allowing a comparison of inter-annual trends between years with complete data. This will have value by ensuring that trends can still be accurately assessed when comparing data at the height of the pandemic in 2020 with previous or future years.

We are confident that readers of Science of the Total Environment will find this manuscript informative and appreciate the opportunity to publish our work in this venue.

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

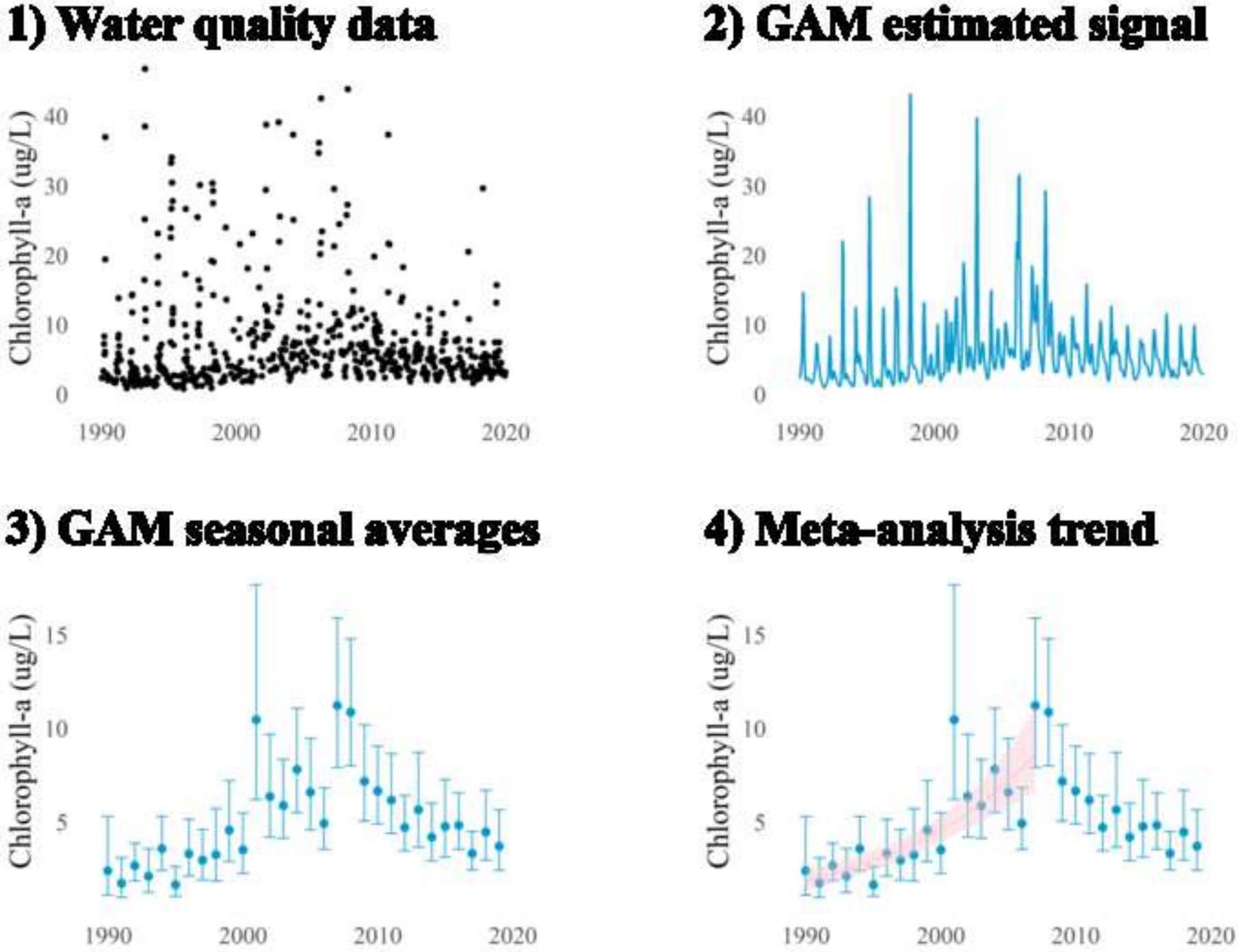
Ariella Chelsky (ariellac@sfei.org), San Francisco Estuary Institute, Richmond, Ca

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

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

Accurate and flexible trend assessment methods are valuable tools for describing historical changes in 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 of monitoring data and trends of interest. Here, we propose a novel application of generalized additive models (GAMs) to model seasonal and multi-decadal changes in a long-term monitoring dataset of 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 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 of the 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 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. Many trend analysis methods lack the flexibility to evaluate one to many 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.

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 related non-parametric approaches have been used for decades in water quality trend assessments to identify monotonic changes over several years while accounting for the predictable patterns among seasons (Helsel et al., 2020; Hirsch et al., 1982). Wan et al. (2017) showed that non-parametric approaches have been the most commonly used methods in long-term water quality trend analysis despite critical limitations. For

descriptive decomposition of long-term monitoring data, they assume seasonal patterns within years do not change, require regularly spaced or balanced data, do not include additional predictors 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 and direction of trends, they apply only to narrow goals.

The seasonal trend decomposition using loess (STL) decomposes a time series into additive components of a long-term trend, a seasonal pattern, and residuals (Cleveland et al., 1990; Cloern and Jassby, 2010; Stow et al., 2015). While useful and widely applied, this method also has important limitations. STL decomposition does not incorporate explanatory variables besides time, it is defined more as an algorithm of statistical steps than as a coherent statistical model (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 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). 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 (GAMs) are central to this paper and form the basis of the fourth method to separate fluctuations on different time scales. GAMs combine one or more smoothing splines to model patterns in data and may be seen 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 cyclic splines, which can be used to model seasonal patterns, and low-dimensional interactions. 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. 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 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 herein generalizes 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 (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 describe and demonstrate the proposed methods by analyzing water quality monitoring data from 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 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 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 and evaluated between years using meta-analysis methods. This approach is new to environmental trend-detection problems and is provided in the *wqtrends* R package 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. Its watershed covers 200 thousand km^2 in the US state of California. Major freshwater inputs enter the system through the Sacramento-San Joaquin Delta complex upstream of Suisun Bay. Salinity ranges from 0 to 15 ppt in the northern subembayments and from 5 to 35 ppt in southern subembayments closer to the Pacific Ocean, depending on the tidal cycle, effluent discharge from wastewater treatment plants, and stormwater runoff (Cloern and Jassby, 2012). An

estimated 73.8 metric tons dy^{-1} of inorganic nitrogen are discharged into the Bay, primarily from wastewater (Novick and Senn, 2014). Agricultural runoff from the upper watershed contributes 30 metric tons dy^{-1} of nitrogen to the SFE via the Delta.

Nitrogen and phosphorus levels in the SFE usually exceed concentrations that cause eutrophication in other estuaries. However, the SFE has demonstrated resistance to eutrophication, which has been attributed to high concentrations of suspended sediment that reduce light penetration in the water column, low residence time caused by vigorous river flushing, and removal of primary producers by abundant suspension feeding bivalves (Alpine and Cloern, 1988; Cole and Cloern, 1984; Jassby, 2008; Kimmerer and Thompson, 2014; Lehman et al., 2017). The Regional Water Quality Control Board has showed renewed interest in 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 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. 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 uncertainty. Second, we calculated a feature of interest from the estimated GAM, along with its propagated uncertainty. For this example, the seasonal averages were extracted, whereas other features could be the timing or magnitude of a seasonal peak, but those are not developed here. 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 smooth the raw data across time. Although they can achieve similar fits, they do so by partitioning variation in the time series differently (Table 2). We discuss all four to clarify their relationships and interpretations. Models are shown in the notation of the `mgcv` R package as formulas for the `gam` function (R Core Team, 2020; Wood, 2017).

The simplest GAM for this purpose is expressed as:

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

where `y` is the time series of interest, such as chl-a, `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 a smoothing spline (in this case as a function of `cont_year`), and `num_knots_Y` is the number of knot or "connections" along the spline.

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 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`, 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`:

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

This model is mathematically equivalent to model S (Table 3). The 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 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 our approach, 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)`

where `doy` is "day-of-year" (i.e., Julian date, a count starting January 1 for each year), `bs = 'cc'` indicates that the spline will be cyclic (constrained to start and end at the same value), and `num_knots_D` is the number of knots for the `doy` spline. 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 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 and instead underestimates 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))`

where `ti()` specifies a tensor-product spline for a surface that varies smoothly as a function of both `cont_year` and `doy`. The number of knots 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 need for sufficient knots can be satisfied either by sufficiently large values for `num_knots_Y_ti` and `num_knots_D_ti` or a sufficiently large value for knots in `num_knots_Y` and `num_knots_D`, 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 likely not capture as much variation in model SYDI compared to model S given the fewer knots that are available to the former. The `ti` term represents an interaction by allowing the pattern in `cont_year` to vary by `doy` and vice-versa. The interaction term in model SYDI provides an appearance that this model is fundamentally different from those provided by the other models. However, models S, SY, and SYD all 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, which is unlikely for the SFE data and chl-a dynamics in many estuaries because bloom size typically varies between years. Thus, the conceptual motivation for model SYDI and its practical application are not supported for this analysis.

Murphy et al. (2019) used spline formulations for Chesapeake Bay water quality related to those proposed here, but for different goals and with different handling of smoothness. They evaluated a "`gam0`" with only `s(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. allowed a maximum number of knots in the `s(cont_year)` term 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)` 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 visually compare chl-a estimates from models SY, SYD, and SYDI to emphasize that similar fits can be achieved by all of the presented models (Figure 3, 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 for the arguments `num_knots_y`, `num_knots_D`, `num_knots_Y_ti`, and `num_knots_D_ti` for each model. 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 the

continuous year smoother (`s(cont_year)`), the fits differ substantially because of how each model allocates variation to the splines. 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` 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, with the 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 $\mathbf{X}$ 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 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 $\hat{\Sigma}_{\hat{\mathbf{y}}} = \mathbf{X}\hat{\Sigma}_{\widehat{\beta}}\mathbf{X}^T$.

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 $\hat{\sigma}^2_{\hat{\mu}} = A^T\hat{\Sigma}_{\hat{\mathbf{y}}}A$ and 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 question: Is there a significant linear trend across a group of years in a seasonal average? For example, is 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

329 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
330 study.

331 A mixed-effects meta-analysis model can estimate linear trends when each observation has an
332 associated standard error, which is the case with our estimates $\hat{\mu}_t$ and $\hat{\sigma}_{\hat{\mu},t}$. Differences in
333 standard errors, which may result from different monitoring effort between years, are explicitly
334 considered in the analysis. The model can be expressed using notation similar to Sera et al.
335 (2019):

336 $$\hat{\mu}_t = \beta_0 + \beta_t t + b_t + \epsilon_t \qquad (1)$$

337 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
338 the residual for year t. Accordingly, the seasonal average for year t is $\mu_t = \beta_0 + \beta_t t + b_t$. The
339 "residual," ϵ_t, represents estimation error in $\hat{\mu}_t$, namely $\hat{\mu}_t - \mu_t$. The residuals are assumed to be
340 independent and normally distributed with mean 0 and variance $\hat{\sigma}^2_{\hat{\mu},t}$, where the latter is estimated
341 from the calculations above. The random effect, b_t, is the difference between μ_t and $\beta_0 + \beta_t t$
342 and is considered the "residual" in the sense of unexplained variation not due to the estimation
343 error. The random effect follows a normal distribution with mean 0 and variance, σ_b^2, to be
344 estimated.

345 We estimated the model (equation (1)) using the *mixmeta* package in R (Sera et al., 2019).
346 Results from *mixmeta* have a similar interpretation as those from regression analysis, but
347 parameter estimates and their standard errors incorporate the known standard errors of the
348 response values. The default estimation method for *mixmeta*, restricted maximum likelihood
349 (REML), was used. The meta-analysis models were applied to a chosen sequence or "window"
350 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 of within-year seasons are relevant to phytoplankton bloom phenology in the SFE (Cloern et al., 2020). The moving-window approach applied 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 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. This analysis was then applied to all stations.

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 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 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 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 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 seasonal averages showed that increases (based on $p < 0.05$) for the January to June period were observed at stations 32, 34, and 36 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, 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 in the July to December period. Additionally, chl-a at the more northern stations has not changed in recent years for both seasonal periods. 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 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 without uncertainty propagation (Figure 7 row 2). Figure 7a shows trend estimates for station 36 for January to July

411 averages from 1991 to 2000. Only the meta-analysis regression results show a trend in this
412 example (based on $p < 0.05$). The OLS regression on observed estimates (top plot) and OLS
413 regression on GAM estimates (middle plot) did not identify trends. Figure 7b shows trend
414 estimates for the same station for July to December averages from 2000 to 2010. Unlike the first
415 example, only the top figure shows a trend, whereas the bottom two plots do not show trends. In
416 both cases, only the meta-analysis results give reliable conclusions because of full propagation of
417 uncertainty across methods. Even in cases where the p-value threshold is not of interest, the
418 confidence intervals from the alternative methods will be inaccurate.

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

427 # Discussion

428 Propagation of uncertainty from within-year features of estimated GAMs to across-year trends
429 using mixed-effects meta-analysis is a new approach that can address different questions than
430 previous methods. Our approach has several advantages over more conventional approaches for
431 analysis of water quality data from long-term monitoring programs. GAMs are capable of
432 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 between year pairs, 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. Our assessment of trends using OLS regression applied to seasonal averages from the raw observations is effectively like averaging results within a year and applying a simple Kendall test. In many cases the results may be similar, but 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 potential for incorrect conclusions at specific monitoring locations and at much larger spatial scales across all stations.

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" signal of actual trends that are minimally influenced by process or observation error present in the raw data.

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 the trend after the trend has been adequately described. This information has 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 (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.

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.

Acknowledgments

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 Nutrient Technical Workgroup and Steering Committee. We thank James D. Hagy III for reviewing an earlier draft of 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., 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

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

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). 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

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

624 Urquhart, N.S., Paulsen, S.G., Larsen, D.P., 1998. Monitoring for policy-relevant regional trends
625 over time. Ecological Applications 8, 246–257. https://doi.org/10.1890/1051-
626 0761(1998)008[0246:MFPRRO]2.0.CO;2

627 Wan, Y., Wan, L., Li, Y., Doering, P., 2017. Decadal and seasonal trends of nutrient
628 concentration and export from highly managed coastal catchments. Water Research 115, 180–
629 194.

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

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

635 Wood, S.N., 2003. Thin-plate regression splines. Journal of the Royal Statistical Society (B) 65,
636 95–114. https://doi.org/10.1111/1467-9868.00374

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

640 Yang, G., Moyer, D.L., 2020. Estimation of nonlinear water-quality trends in high-frequency
641 monitoring data. Science of The Total Environment 715, 136686.
642 10.1016/j.scitotenv.2020.136686.

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

645

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 station 36. 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.

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 (`s()`).

GAM	Additive components	Details
`S`	`s(cont_year)`	A single smoother over a continuous year variable
`SY`	`cont_year + s(cont_year)`	A linear continuous year variable and a single smoother over a continuous year variable
`SYD`	`cont_year + s(cont_year) + s(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)`	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

Figure1

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

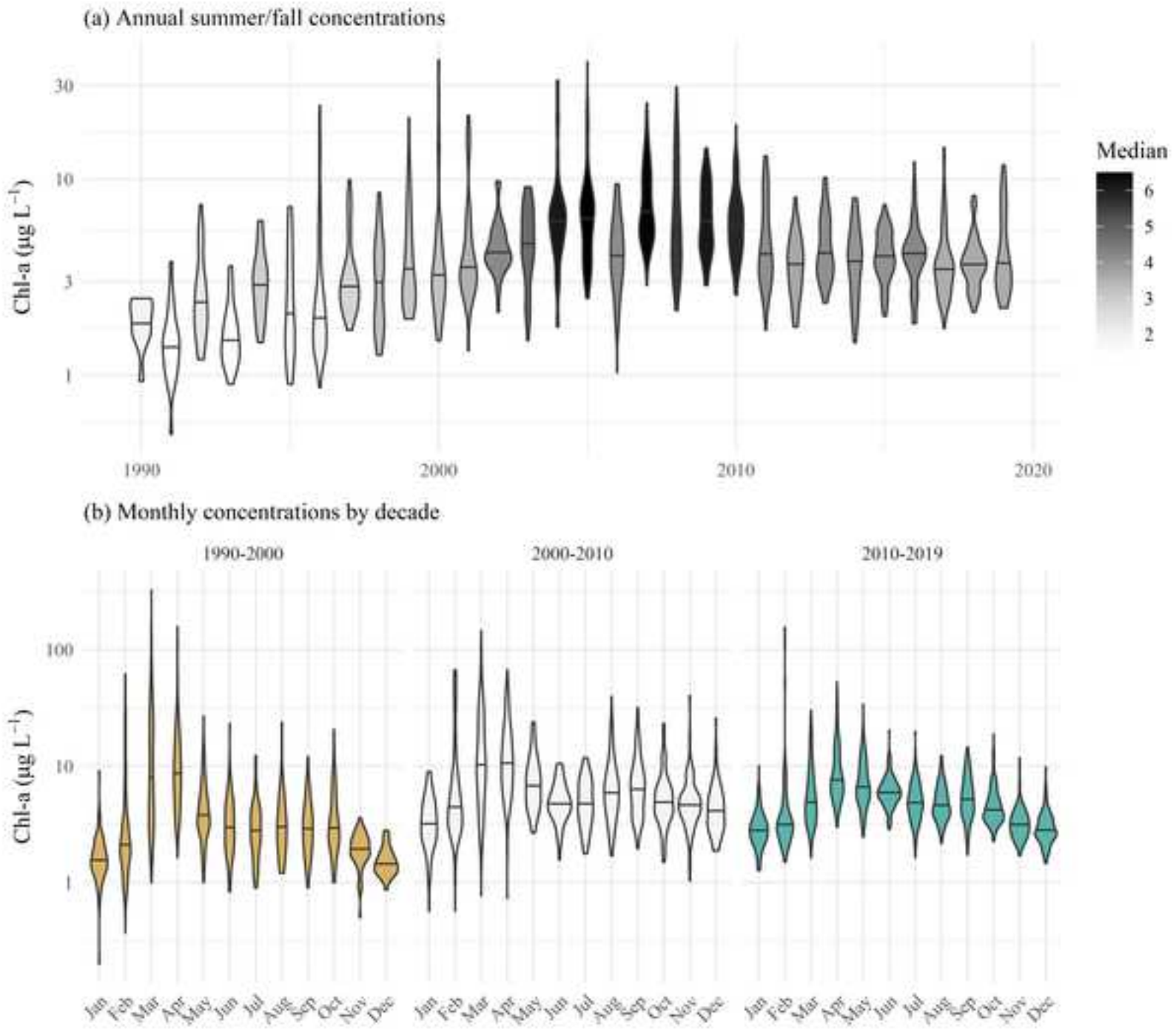


Figure2

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

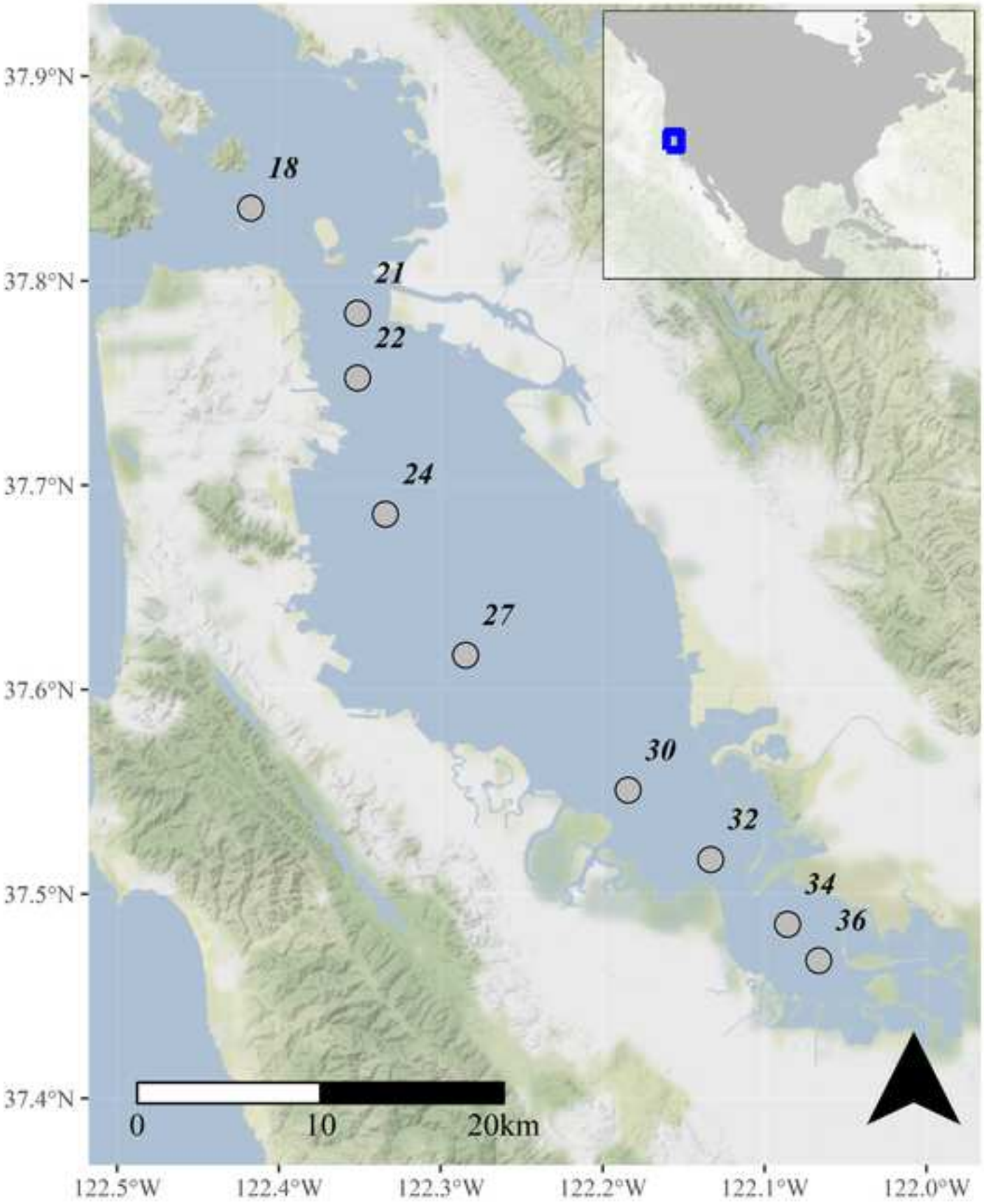


Figure3

Click here to access/download;Figure;Fig3.tiff

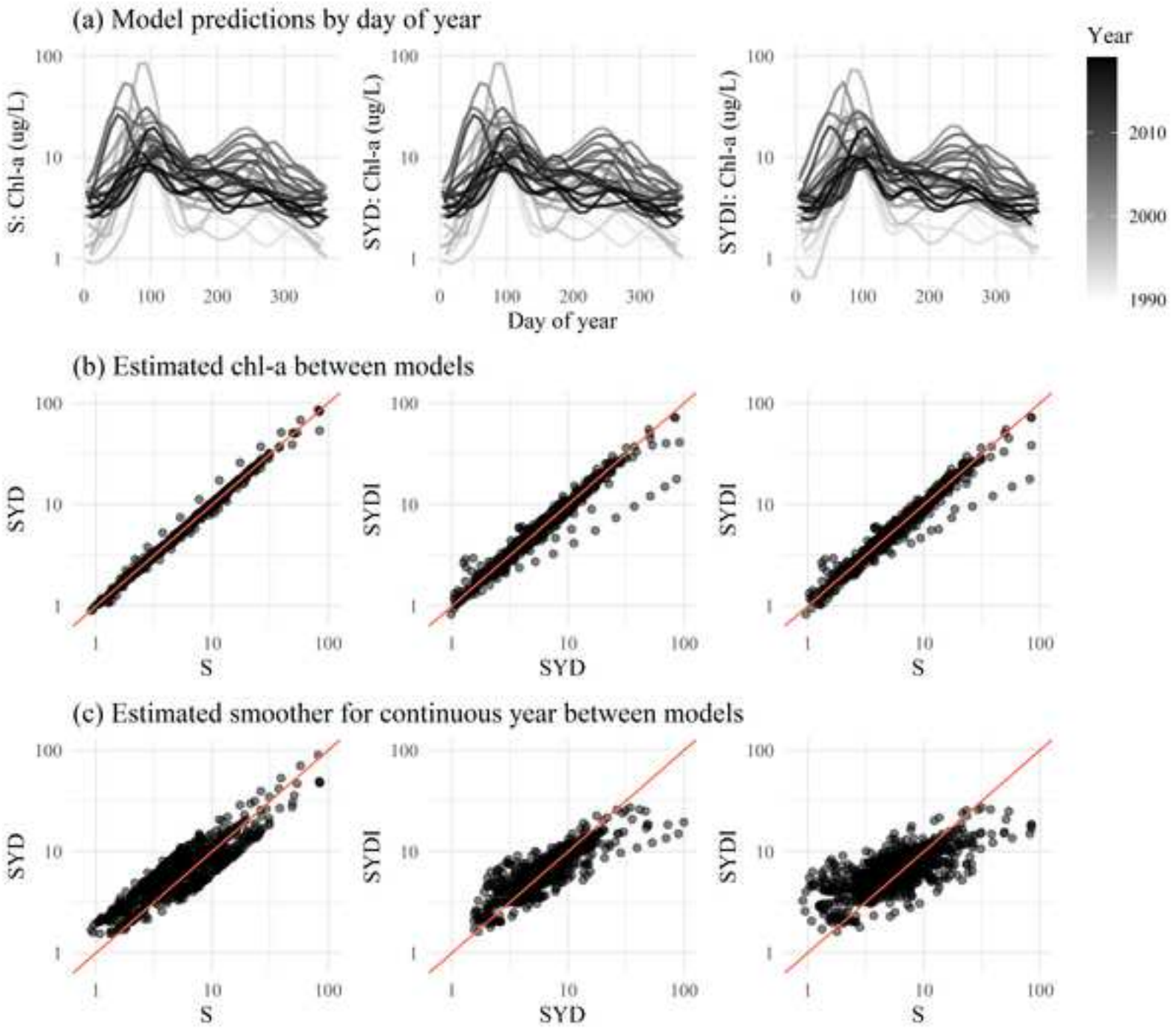


Figure4

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

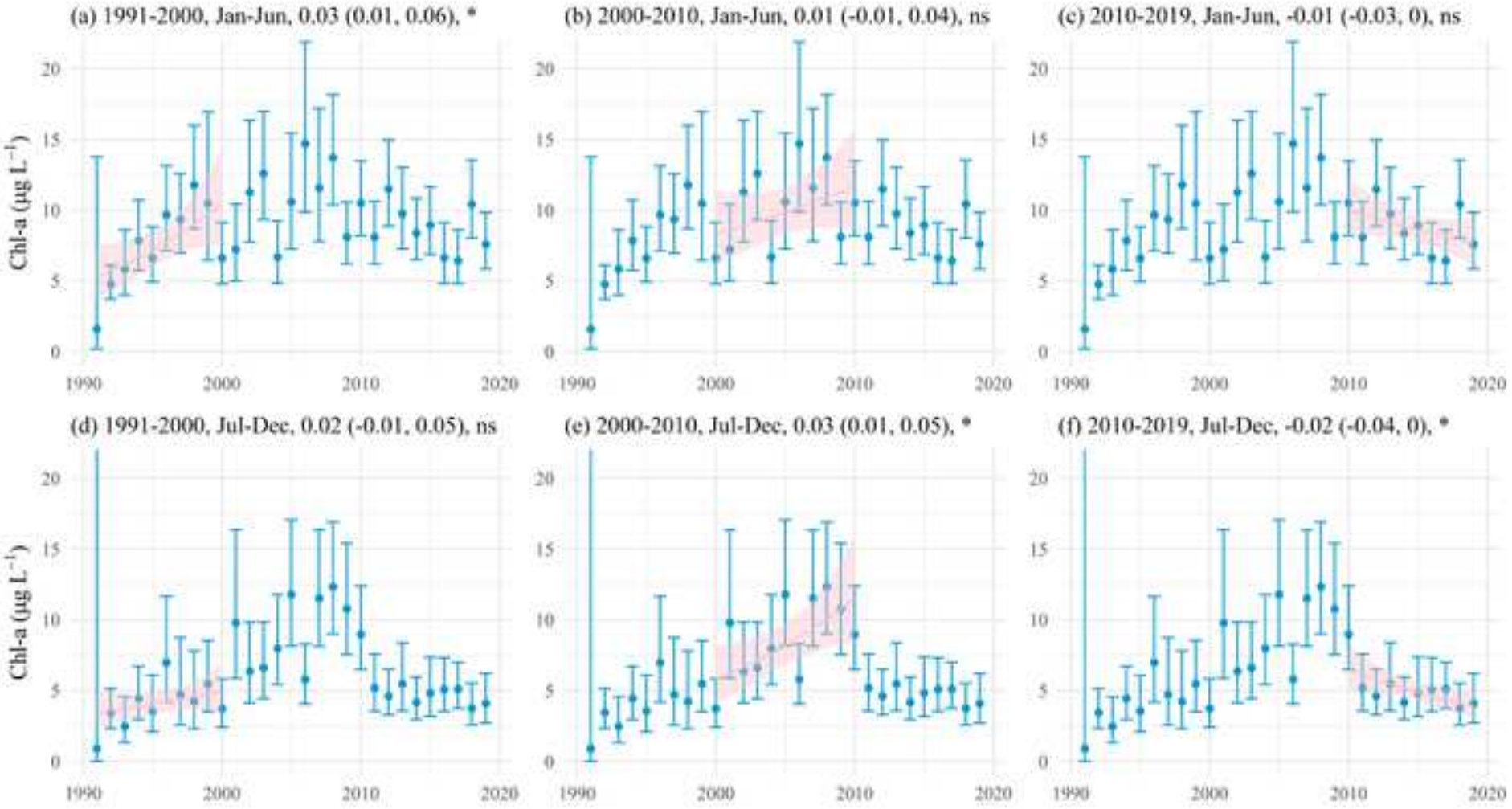


Figure5

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

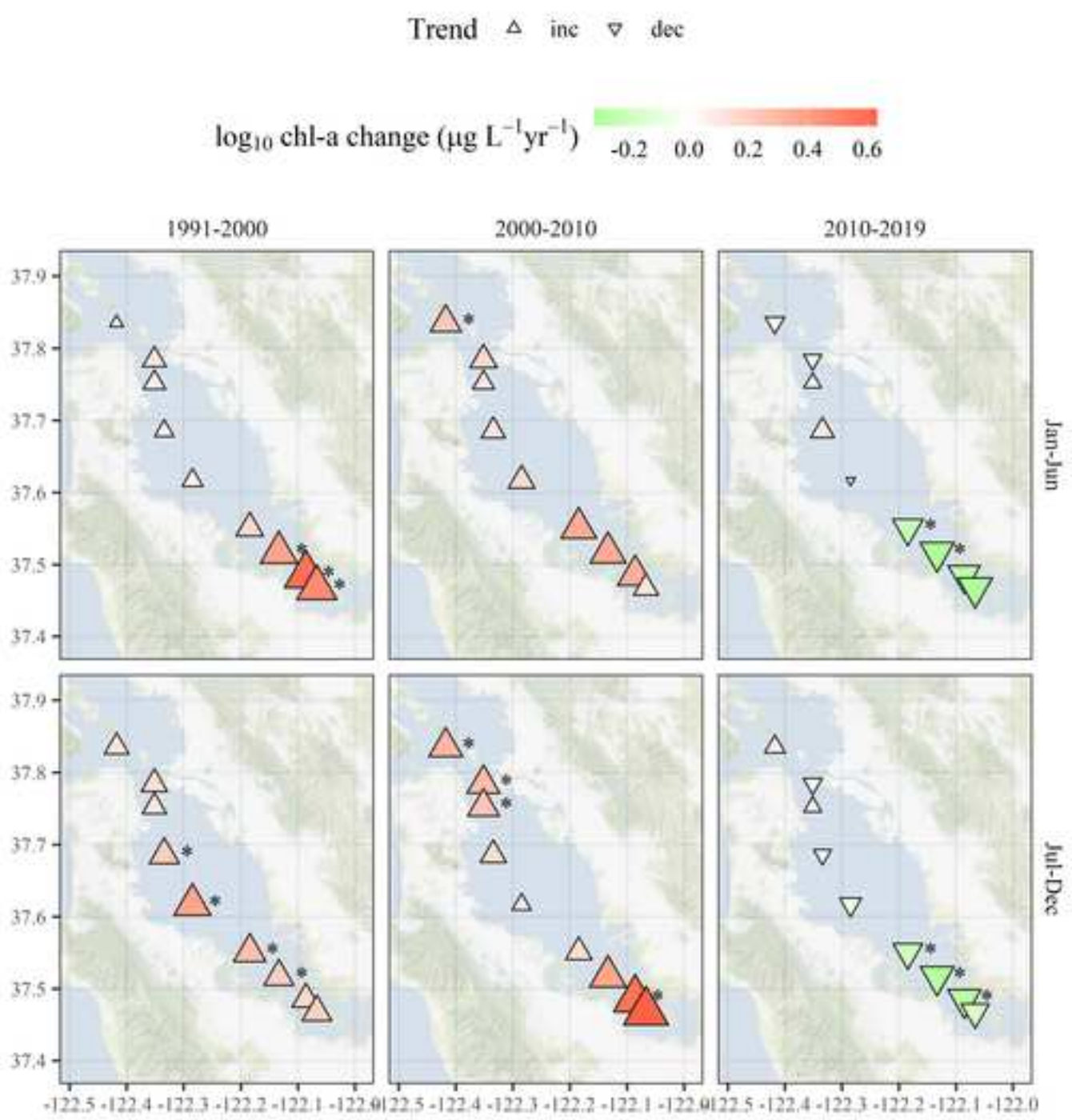


Figure6

Click here to access/download;Figure;Fig6.tiff

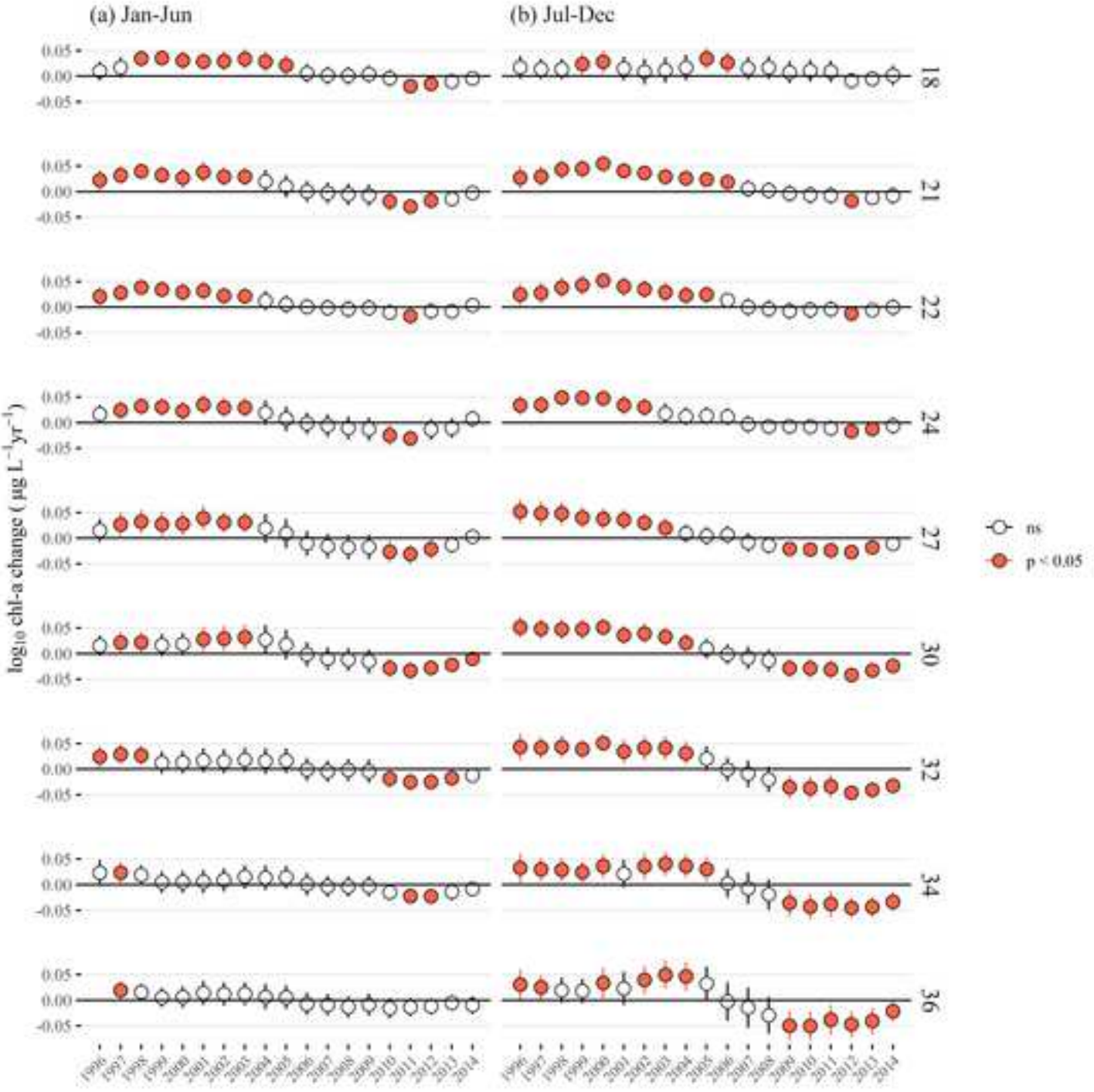


Figure7

Click here to access/download;Figure;Fig7.tiff

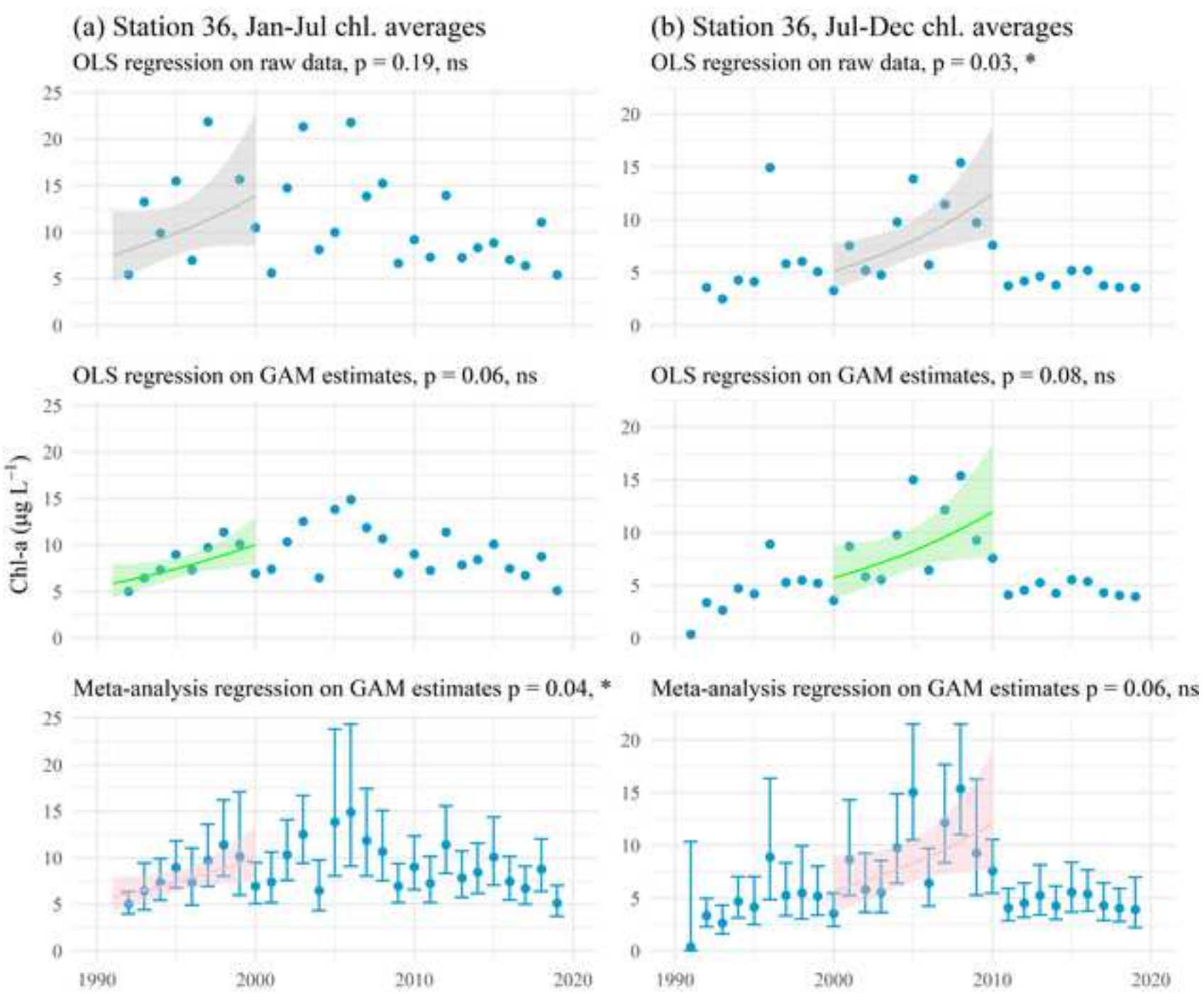
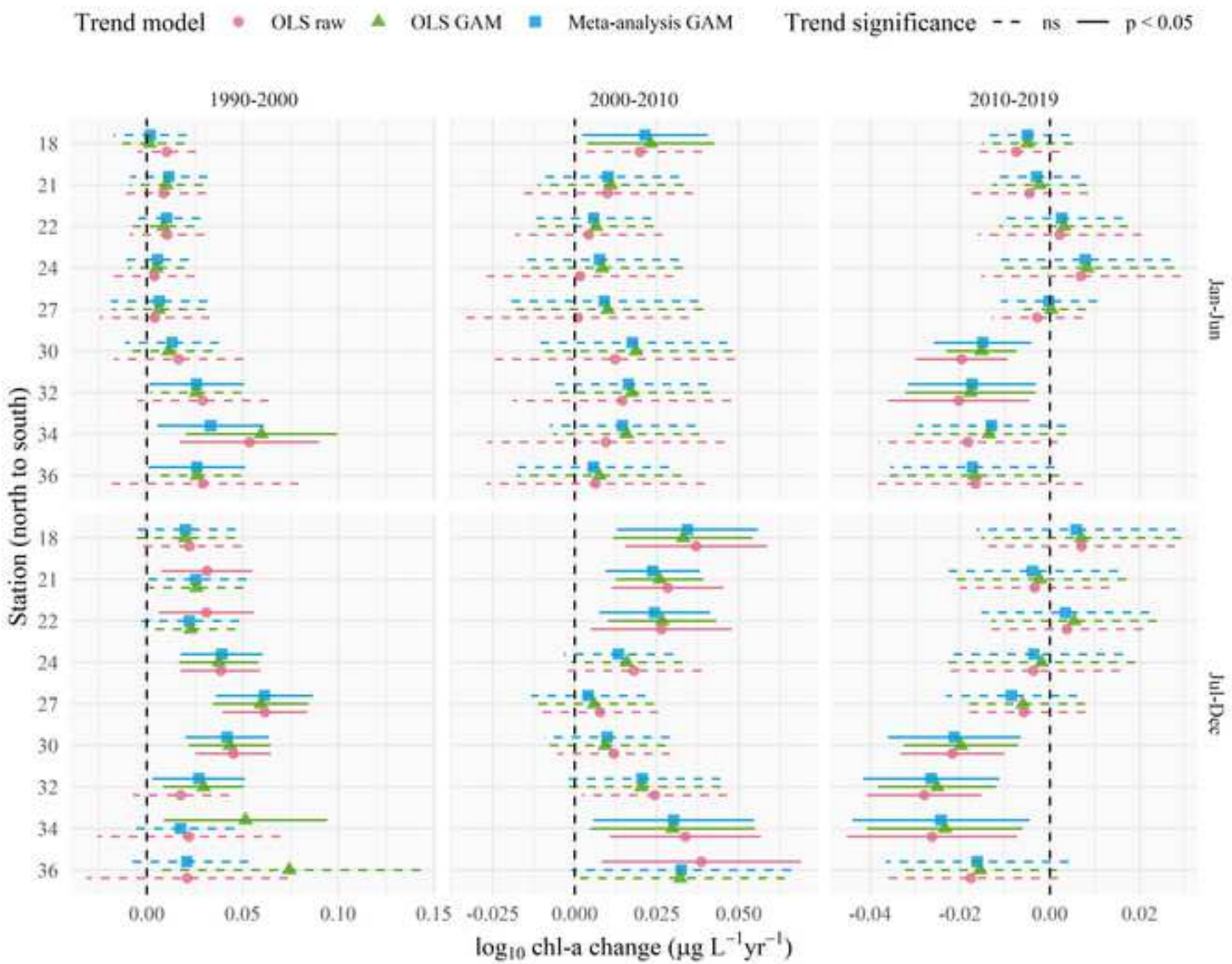


Figure8

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: