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

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

Last manuscript build 2021-02-04 09:10:23

# 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 analysis methods. 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 are a general method for modelling a response variable as the additive sum of one or more smoother splines fit to predictor variables. These models have shown promise in other applications to water quality monitoring data 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-meta regression analyses 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 for an a priori period of time within each year from the GAM results, including an accurate assessment of standard error within each year where the seasonal average is estimated. Finally, we demonstrate how trends can be quantified with mixed-meta regression analyses that evaluate changes in the seasonal averages between years with full propagation of uncertainty in the estimates. Overall, this approach leverages the ability of GAMs to identify an optimal amount of smoothing from the raw data and uses those results to accurately assess trends that minimize the likelihood of false positive or negative results, which is particularly important for time series with missing observations or varying sample effort over time.

# Introduction

Accurate quantification of trends must consider variation at different temporal scales regardless of the question of interest, such that 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 (Urquhart et al. [1998](#ref-Urquhart98), Cumming et al. [2006](#ref-Cumming06), Forbes and Xie [2018](#ref-Forbes18)). 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. In this paper, we develop methods to estimate across-year trends of within-year features, such as a 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 (Hirsch et al. [1982](#ref-Hirsch82), Helsel et al. [2020](#ref-Helsel20)). Wan et al. ([2017](#ref-Wan17)) 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, only use time as a descriptor for trends, 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](#ref-Cleveland90), Cloern and Jassby [2010](#ref-Cloern10), Stow et al. [2015](#ref-Stow15)). 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](#ref-Wan17)), and it does not usually estimate standard errors to allow hypothesis testing (but see Hafen [2010](#ref-Hafen10)). STL methods may also over-simplify trends into absolute 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](#ref-Cloern10), Navarro et al. [2012](#ref-Navarro12)).

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 (Hirsch et al. [2010](#ref-Hirsch10), Beck and Hagy [2015](#ref-Beck15), Beck et al. [2018](#ref-Beck18b)). 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 localized multi-polynomial smoothing methods, with specific application to descriptive 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 that uses Monte Carlo estimates of false positive rates from the model results (Hirsch et al. [2015](#ref-Hirsch15)). Although a useful addition to the original method (Hirsch et al. [2010](#ref-Hirsch10)), the approach requires extensive resampling as a post-hoc application to 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 our fourth and last 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 (He et al. [2006](#ref-He06), Morton and Henderson [2008](#ref-Morton08), Pearce et al. [2011](#ref-Pearce11), Haraguchi et al. [2015](#ref-Haraguchi15), Murphy et al. [2019](#ref-Murphy19)). In statistics, the evolution of non-parametric regression methods has largely converged on GAMs, rather than kernel smoothing methods used by both STL and WRTDS. Kernel smoothing and spline smoothing are closely related, and a key challenge for each is to determine the appropriate degree of smoothing. For example, WRTDS can potentially give results similar to the spline-based smoothing methods used in GAMs, although at higher computational expense and with the limitation that uncertainty estimates are not readily obtainable from the original method (Beck and Murphy [2017](#ref-Beck17)). WRTDS also requires specifying the desired degree of smoothing via the “windows” that define the local weighting at each point in the time series. The windows in WRTDS are conceptually similar to the kernel (or bandwidth) used in more conventional smoothing methods. A challenge of WRTDS is that there is no clear rule that can be applied to determine the best choice of window size. The associated tradeoff between over- and under-smoothing is a hallmark of these approaches.

GAMs have various advantages compared to kernel smoothing methods and are therefore used for non-parametric regression smoothing in many fields. The “basis functions” used to formulate GAMs can be customized based on expected patterns in the data. Examples include 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 likelihood estimates or optimization of out-of-sample prediction error, which address a key concern around methods like WRTDS that do not have analogs for choosing optimal degrees of smoothing. 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 that can be optimized more quickly than other approaches (Wood [2017](#ref-Wood17)).

GAMs have been applied previously to evaluate trends in water quality time-series from long-term monitoring programs (Haraguchi et al. [2015](#ref-Haraguchi15), Murphy et al. [2019](#ref-Murphy19)). For example, Murphy et al. ([2019](#ref-Murphy19)) used GAMs to decompose water quality time series from Chesapeake Bay into long-term and seasonal trends (Murphy et al. [2019](#ref-Murphy19)) and test trend hypotheses between two points in time. 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 [2003](#ref-Wood03), [2017](#ref-Wood17)) rather than being constrained *a priori* for different time scales, and prioritizes full incorporation of uncertainty. Other studies of environmental time-series with GAMs have addressed the use of transformed response data (Yang and Moyer [2020](#ref-Yang20)), serial correlation in high resolution data (Morton and Henderson [2008](#ref-Morton08), Yang and Moyer [2020](#ref-Yang20)), and quantifying time lags in relationships between response and predictor variables (Lefcheck et al. [2017](#ref-Lefcheck17)).

Our motivating problem has several characteristics that are 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 to time series data. This paper develops GAMs as a distinct application that can meet these requirements while addressing key limitations of seasonal Kendall tests and the more complex STL and WRTDS methods. We do so with the explicit goal of quantifying interannual trends in seasonal averages.

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? Are there significant across-year trends in the spring phytoplankton bloom or baseline chlorophyll concentrations during periods of low productivity between blooms? We provide examples illustrating how these questions can be addressed using methods with GAMs. We also provide an approach for using meta-analysis methods (Gasparrini et al. [2012](#ref-Gasparrini12), Sera et al. [2019](#ref-Sera19)) after signal extraction with GAMs. This approach is new to environmental trend-detection problems. For this step, we provide methods using meta-analysis for isolating seasonal trends with reasonable certainty from GAM results and evaluating these trends between years.

# 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 in the US state of California. Major freshwater inputs enter the system through the Sacramento-San Joaquin Delta complex upstream of Suisun Bay, where the combined inflow from both rivers is approximately 28 km per year. 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. An estimated 73.8 metric tons dy of inorganic nitrogen are discharged into the Bay, primarily from wastewater (Novick and Senn [2014](#ref-Novick14)). Agricultural runoff from the upper watershed contributes 30 metric tons dy of nitrogen to the SFE via the Delta.

Nitrogen and phosphorus levels in SFE usually exceed concentrations that cause eutrophication in other estuaries. However, SFE has demonstrated resistance to eutrophication, which has been attributed to high suspended sediments 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 (Cole and Cloern [1984](#ref-Cole84), Alpine and Cloern [1988](#ref-Alpine88), Jassby [2008](#ref-Jassby08), Kimmerer and Thompson [2014](#ref-Kimmerer14), Lehman et al. [2017](#ref-Lehman17)). Renewed interest in understanding the potential for nutrient loading to negatively affect water quality has occurred for South Bay, where harmful algal blooms (HABs), elevated summer-fall chlorophyll concentrations, and low dissolved oxygen concentrations began around 1999 (Figure 1) (Cloern et al. [2020](#ref-Cloern20)). 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](#ref-Cloern16), Schraga et al. [2020](#ref-Schraga20)). Chl-a on GFF filters was analyzed fluorometrically after extraction in 90% acetone. 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.

## GAM application

We implemented the GAMs in three stages. First, we used a GAM to estimate a smooth pattern of variation 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](#ref-Gasparrini12), Sera et al. [2019](#ref-Sera19)). Meta-analysis uses response data that includes standard errors (uncertainties) as needed to address our questions. We developed two mixed effects meta-analysis approaches that provided 1) a simple comparison of whether seasonal features differ across years, and 2) an estimation of short-term linear trends on time scales chosen by the analyst.

### First-stage analysis: GAM estimation

We used four different GAMs to estimate the long-term trend and uncertainty from the time series data, each with the potential to achieve a similar fit to the data but with a slightly different model structure (Table 2). Models are shown in the notation of the mgcv R package as formulas for the gam function (Wood [2017](#ref-Wood17)).

GAMs smooth the raw data across time to separate variability in the response variable associated with an explanatory factor, independent of noise inherent in the raw data. 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 of cont\_year), and num\_knots\_Y is the number of knot or “connections” along the spline that influence curvature. Log-transformed chl-a was used for all analyses to meet assumptions of normally-distributed residuals.

The optimal level of smoothing in mgcv is determined by a penalty on the net curvature of the spline (Wood [2004](#ref-Wood04)). Smoothing was determined using generalized cross-validation (GCV, as implemented in mgcv), which approximately minimizes out-of-sample prediction error. To allow GCV (or other alternatives) to work as intended, a potential upper limit on the number of knots determined by the analyst must be sufficiently large. Results should not be sensitive to the number of knots; if they are, the number of knots should be increased. We chose a sufficiently large number of knots for num\_knots\_Y (Model S) as 12 times the number of years in the time series that was modeled. This created the potential, as determined through GCV, to have one knot per month as an approach to both prevent under-fitting the observed data and to accurately estimate the seasonal signal within a year. If the data were too sparse, the model will have insufficient degrees of freedom to fit 12 knots per year, in which case the number of knots was reduced by one knot per year until sufficient degrees of freedom were available (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, given 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 upper limit for 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 seasonal pattern and changes the interpretation of the cont\_year spline to represent smoothed deviations from the average within-year 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 model, whereas the sum of components for each model produce comparable results between models. Our goal is to extract signals (i.e., seasonal estimates) from the fitted time series and different allocation of penalties among the splines in each model is unimportant for this task. However, interpreting differences in fit between model SYD and models S or SY can be problematic because the penalties for smoothing splines based on curvature are heuristic (Wood [2017](#ref-Wood17)). 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 fits 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 consider data from the same period in other years by imputing an average seasonal pattern with the doy spline. However, an interpretation of these imputations is challenging. For example, the spring chl-a peak is a notable feature every year in 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) + s(doy, bs = "cc") + 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 s(cont\_year), 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 formulaically from model S to a greater extent than model SYDI, but all of the splines use the same inputs to smooth the same data. The splines in cont\_year and doy will likely not capture as much variation in model SYDI compared to model SY 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, this outcome is difficult to achieve for reasons explained below.

The distinct aspect of model SYDI is the anticipation that within-year fluctuations will vary smoothly from year to year. For example, if one year has an especially early and large magnitude spring bloom, the years before and after would be expected also to have earlier and larger than average spring blooms. For the raw SFE data, and chl-a dynamics in many estuaries, this is not expected to be the case. An early and large spring bloom could easily be followed the next year by a late and relatively small spring bloom. If the interaction spline has many knots, then the degree of smoothness across years can be estimated to be very low, allowing fluctuations within years to vary strongly across years. However, if that is the case, the conceptual motivation for model SYDI may not be supported by the data. Given that observed bloom dynamics for most systems do not vary smoothly across years, the potential benefits of using model SYDI could be outweighed by using simpler models. Moreover, the spline and penalty structure differ from the other models, making it difficult to evaluate their relative performance (Table 3).

For models SYD and SYDI, the allocation of knots between the different splines can be chosen by the analyst so long as it is recognized that different choices will arbitrarily lead to related differences in allocation of variation among the splines. In models S, SY, and SYD, the number of knots in the cont\_year term need to be sufficiently large. In model SYDI, a very large number of knots in both the cont\_year spline *and* in both dimensions of the interaction spline is impossible to achieve. Allowing the cont\_year spline to include many knots would somewhat defeat the purpose of the interaction spline.

Murphy et al. ([2019](#ref-Murphy19)) used spline formulations related to those proposed here, but are insufficient for our needs of extracting and quantifying seasonal trends. Murphy et al. ([2019](#ref-Murphy19)) have a “gam0” with only s(doy) and linear cont\_year terms, a “gam1” like our SYD, and a “gam2” like our SYDI. Their gam0 is *a priori* not of interest for our data because of an assumption of constant seasonality and a linear annual trend. Compared to our application of SYD and SYDI, their implementation of “gam1” and “gam2” allocated knots differently, leading to a different interpretation of results. For “gam1” (our SYD), they set a maximum number of knots in the s(cont\_year) term of 2/3 times the number of years, whereas we adopt a different approach using a number of knots equal to 12 times the number of years, or one per month. Accordingly, they interpret this spline as separating a long-term (or low-frequency) trend from other patterns, whereas we use it to separate intra-annual changes in season from noise at a finer temporal scale. In the “gam2” model (our SYDI model), Murphy et al. ([2019](#ref-Murphy19)) do not explicitly consider the number of knots in the interaction spline. In both cases, they effectively use their choice of number of knots in different spline components as an *ad hoc* allocation of variation in the data to different components based on previous interpretations of water quality dynamics in the system, while noting that additional explanatory variables are necessary to explain the residual variability or autocorrelation in the data. Murphy et al. ([2019](#ref-Murphy19)) acknowledge that incomplete modeling of fluctuations in the data may inflate their Type I error rates for estimating temporal changes, but they account for that by incorporating additional hydrologic variables in a subsequent phase of their models in a step-wise manner. For the methods presented here, which focus first on fully identifying temporal change, we seek to avoid inflating Type I error rates by increasing the upper limit for the knots. Finally, Murphy et al. ([2019](#ref-Murphy19)) 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), making differences in AIC unimportant. In our example, large differences in AIC only reflect inadequate choice of knots in one or more splines, which should be avoided.

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, S is identical to SY and is not shown). Models SY, 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 models using only the continuous year smoother (s(cont\_year), as an example), the model fits differ substantially because of structural differences and differences in the penalties applied to the basis functions. 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 SFE. This model was chosen because of the relatively quicker processing time to fit the model, while providing nearly identical explanatory power as compared to the other models (Table 3).

### Second-stage analysis: Uncertainty propagation from estimated GAMs to seasonal features

In the second-stage analysis, we estimated a seasonal average, such as the mean spring chl-a concentrations and the uncertainty in each year. We define as the average for the period of interest (e.g., seasonal average) in year , as an estimate of , and as the estimated standard error of . The season includes days. For simplicity, the following text omits subscript .

Point estimates of response values for the fitted GAM take the form , where is the vector of parameter estimates and is a model matrix of explanatory variables, including spline basis function values. Vector includes both fixed effect parameters and spline parameters, and 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 dec\_year = , then would have a row with in the first column (for the intercept parameter), (for the linear time trend) in the second column, and an evaluation of each spline basis function at in the remaining columns. The number of spline basis functions is related to the number of knots. Note that can be any time, not necessarily the time of an observation.

To obtain a vector, , of fitted point estimates for every day in a season, 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 and , an estimate of the covariance matrix of the sampling distribution of . The scalar standard errors of are the square roots of the diagonal elements of , wheras the off-diagonal elements are the correlations among the elements of . Since parameter estimates are correlated, the covariance of is .

The estimated seasonal average was calculated from the vector of daily values for each of the days in the season of interest with , where is a row vector with all values equal to . The variance of is and standard error is . Each of these estimates are from on approximate multivariate normality of the sampling distribution of . Murphy et al. ([2019](#ref-Murphy19)) provided a similar approach for comparing multi-year averages of month-scale spline values for sets of years at the beginning vs. end of a time-series.

### 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 within-year features such as seasonal water quality, characterized by the within-year means () and their standard errors () that we estimated in stage two of the analysis. In the examples, we considered two periods of interest for chl-a, January-July and August-December, which are relevant to phytoplankton bloom phenology in SFE.

This analysis provided a direct answer to the question: Is there a significant linear trend across a group of years in a seasonal average, where the time-scale of the trend is chosen by the investigator? 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 are , , with their associated standard errors, . is the number of years of the study.

A meta-analysis mixed effects model is useful to estimate linear trends when each observation has a unique and quantified standard error, which is the case with our estimates and . 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](#ref-Sera19)):

where is the intercept, is the year, is the slope, is the random effect for year , and is the residual for year . Accordingly, the seasonal average for year is . The “residual”, , represents estimation error in , namely . The residuals are assumed to be independent and normally distributed with mean 0 and variance , where the latter is estimated from the calculations above. The random effect, , is the difference between and and is considered the “residual” in the sense of unexplained variation not due to the estimation error. The random effect is assumed a normal distribution with mean 0 and unknown variance, , to be estimated.

We estimated the model (equation (??)) using the *mixmeta* package in R (Sera et al. [2019](#ref-Sera19)). 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. Following meta-analysis theory, evaluates this model as a linear mixed effects model with some variance components fixed and others estimated (). The default estimation method for meta-analysis models is restricted maximum likelihood (REML).

This method was applied to a chosen sequence or “window” of years for estimating the linear trend. This approach evaluated whether there was a significant linear trend and provided an estimated rate of change (slope) in chl-a over a chosen series of years. Using the meta-analysis considered statistical properties derived from all the data in the years of interest (i.e., the seasonal averages and their uncertainty). Years with more data have smaller standard errors for compared to years with less data. The meta-analysis explicitly includes these differences in uncertainty magnitudes.

### 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 approximate ten year blocks from 1991 to 2000, 2000 to 2010, and 2010 to 2019. As an extension of these methods, a chosen sequence of years was applied across the time series to assess periods of time within which seasonal trends could be significant or not (e.g., spring average trend from 1990-2000, 1991-2001, etc.). Although the initial year sequence and seasonal period to evaluate is chosen by the analyst, applying the method in a moving window approach reduces some of the ambiguity around points in the time series when trends may be changing with reasonable certainty. The moving window approach applies the meta-analysis model from left to right in the time series across the seasonal averages to obtain the slope and significance of the estimated trend. We chose a centered ten-year window where the model estimates are based on results at equal half-window widths to the left and right of a given year. Although a left- or right-centered window could also be applied, we limit the analysis to a centered window to demonstrate the concept.

Finally, trend results from the mixed-meta regression method for each season and different time periods were compared to alternative trend analyses to demonstrate how different and potentially misleading conclusions about trends can arise from methods that insufficiently account for propagation of uncertainty. As stated above, mixed-meta regression allows for full consideration of uncertainty in trend assessments by explicitly incorporating standard error of averages from results obtained from the GAMs and it is expected that more generalized methods that do not account for uncertainty may lead to different conclusions. Moreover, the mixed-meta analyses depend on GAM predictions to describe an estimated long-term signal in the observed time series. Trends assessed on observed data may include substantial noise independent of any “canonical” signal derived from GAMs. Trend estimates from mixed-meta regression applied to GAM seasonal averages were compared to 1) trends from ordinary least squares (OLS) regression applied to seasonal averages from observed data and 2) trends from OLS regression applied to GAM seasonal averaged. The comparisons were demonstrated with select examples where differences were pronounced and then applied to all stations.

### Back-transformation of model results

Model results were back-transformed from log-space to aid in the interpretation of trends (Bradu and Mundlak [1970](#ref-Bradu70), Duan [1983](#ref-Duan83)). Back-transformation was accomplished using equation (2) for model predictions, estimates of mean values, and endpoints of confidence intervals from GAM results, such that:

where the back-transformed, expected value of the response variable (chl-a) is a function of the mean value in log-space and a dispersion estimate from the model. The dispersion is obtained directly from the fitted GAM object as an estimate of the variance of the residuals divided by the model degrees of freedom.

Slope estimates that were used to determine significance of a seasonal trend across years from the mixed-meta analysis were not back-transformed. While it is possible to back-transform model predictions, estimates of the mean values, and endpoints of confidence intervals, it is not possible for slope estimates in log-space that can change in arithmetic space across the values of the independent variable. As such, all results reporting slope estimates as rates of change in chl-a per year were reported at the scale of the model in log-space.

### wqtrends R package

The wqtrends R package developed by the authors includes all methods described above. The package is available for download at <https://tbep-tech.github.io/wqtrends>. A full vignette describing installation and use is also available in the link provided.

# Results

## Model performance and observed results

Model predictions for chl-a trends across all stations had an average R-squared value equal to 71% (Table 4) and ranging from 59% (station 22) to 78% (station 18). GAM predictions from north to south on the longitudinal axis showed more pronounced annual and seasonal changes in chl-a towards the more southern stations (Figure 4). 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 South Bay 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 (Figure 4).

## Trend estimates

To demonstrate the differences in results that can be obtained with the trend tests, estimates of linear trends across years for different seasons are shown for station 34 (Figure 5). All plots show trends in seasonal averages within ten-year windows, whereas rows a-c show estimates from January to June and rows d-f show estimates from July to December. The seasonal trend analyses showed that January to June chl-a increased (log chl-a slope 0.03 g L yr, 0.01-0.06 95% confidence interval) form 1991 to 2000, whereas a trend for the same period in July to December was not observed. Chl-a also increased from 2000 to 2010, but only for July to December (log 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 in fact estimated to be moving in opposite directions. Figure 6 shows results from similar analyses as those in figure 5, but applied to all stations.  
Mixed-meta regression analyses applied to seasonal averages showed that increases (based on ) 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, whereas chl-a decreased for 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.

## Trend comparisons

Results from a ten-year moving window comparison of seasonal trends provided additional context on when significant changes were occurring at each station (Figure 7). 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 more trends in the July to December period. Additionally, trends in recent years in both seasonal periods were not observed for the more northern stations. For both seasonal periods, a change from increasing to decreasing chl-a occurred at most stations around 2007.

Results showing trend estimates from mixed-meta regression on GAM seasonal estimates provided different conclusions than those from either OLS regression through seasonal averages from observed data (Figure 8 row 1) or OLS regression through GAM estimates (row 2). Figure 8a shows trend estimates for station 36 for April to June averages from 1991 to 2000. Only the mixed-meta regression results shows a trend in this example (based on ). The observed regression results (top plot) and average regression on GAM estimates (middle plot) did not identify a trend. Figure 8b shows trend estimates for station 22 for October to December averages from 2000 to 2010. Unlike the first example, the top two figures show trends, whereas the bottom plot for the mixed-meta regression analysis does not show a trend because of added uncertainty in the averages provided by the GAMs. In both cases, only the mixed-meta regression results provide accurate trend estimates because of full propagation of uncertainty across methods.

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 9). The slope estimates from the OLS models applied to the observed data were understandably more variable than the slope estimates from the GAM averages and mixed-meta methods, with much larger slopes observed especially at the more southern stations in the January to March period. Slope estimates from the OLS models applied to the averages from the GAMs as compared to the mixed-meta results were more similar, excluding some of the slope estimates for the southern stations. Differences in significance of trends between the OLS models applied to the GAM averages and the mixed-meta analyses were also observed, reflecting the ability of the former to more accurately assess significance of trends by accounting for uncertainty in the average estimates.

# Discussion

The use of the GAM results with second and third stage analyses to assess trends is a new approach that previous trend analysis methods cannot sufficiently address. Although the development of our approach was motivated by specific questions regarding assessment of seasonal changes over time, the application has advantages over more conventional methods that inadequately account for time series characteristics of water quality data from long-term monitoring programs. In particular, missing observations or irregular sampling can complicate trend assessment and comparison of trends between locations that may differ in sampling design (Junninen et al. [2004](#ref-Junninen04), Racault et al. [2014](#ref-Racault14)). 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 (Hirsch et al. [1982](#ref-Hirsch82), Helsel et al. [2020](#ref-Helsel20)). 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. 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 likelihood of incorrect conclusions from test results.

Our results demonstrated that model structure (i.e., types of smoothers) was less important than allowing the model sufficient freedom to estimate the trends over time. This is an important conclusion that can provide guidance on how GAMs should be used to model time series from long-term water quality monitoring data. 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). This approach leverages the ability of GAMs to objectively estimate smoothed trends across years by identifying an optimal level of smoothing with the mgcv R package using generalized cross-validation to extract an underlying signal in the observed data (Wood [2004](#ref-Wood04), [2017](#ref-Wood17)). Specifying an upper limit for the number of knots that can potentially be used to fit different smoothers is critical to this approach. A smaller limit can lead to under-smoothing and an insufficient characterization of trends that risks incorrect conclusions from using these models with the second and third stage of analyses.

Our examples in Figures 8 and 9 demonstrate how different conclusions can be obtained if propagation of uncertainty from raw observations across methods is unaccounted for in trend assessment. These comparisons demonstrate a common approach where raw observations may be aggregated prior to use of more conventional trend analysis methods. 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 can lead to increased Type I or II error rates depending on characteristics of the raw data. A false negative result (Type II) was shown in Figure 8a where a trend was not shown for the OLS regression results, but was observed for the mixed-meta regression results. Conversely, a false positive result (type I) was shown in Figure 8b where the OLS regression results showed a trend, whereas the mixed-meta results did not. The potential danger of inflated error rates was further demonstrated at a larger spatial scale across all stations in Figure 9.

Variability of sampling effort among years is a common characteristic of environmental monitoring data, which further increases the potential for incorrect conclusions when existing methods have been used. Averages may be skewed in a particular direction if annual estimates are based on a handful of observations from select months (e.g., summer only, Fouquet [2012](#ref-Fouquet12)). The use of GAMs to fit the long-term trend will reduce the potential of limited observations in a particular year skewing estimates of annual or seasonal averages. More importantly, limited observations in a year will be reflected in the standard error estimates derived from the GAM, which has direct implications for how uncertainty is treated in the mixed-meta regression analyses (Sera et al. [2019](#ref-Sera19)). As a result, trend assessments from mixed-meta regression are accurate for data with unequal sampling effort where other methods may fail by providing estimates that are biased or have inaccurate or incomplete estimates of uncertainty.

The underlying cross-validation methods used by GAMs in the mgcv package also reduce the decisions required by an analyst 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](#ref-Hirsch10), Wan et al. [2017](#ref-Wan17)). Prior to the implementation of generalized cross-validation in GAMs, these decisions were also required, producing fitted results that are arbitraily under- or over-smoothed to the raw observations. This is especially problematic when for policy analysis or regulatory decisions if the results change based on arbitrary decisions of the analyst. Because these decisions are no longer 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 validity of conclusions based on trends from mixed-meta regression analyses applied to the seasonal GAM averages. We acknowledge that the third stage analyses require explicit user inputs on year periods to define trends partially impose arbitrary decisions that can influence interpretation of results. Although there are undoubtedly many scenarios where years of interest can be chosen objectively (e.g., regulatory compliance periods, time since management intervention, etc.), a more general question of “when” changes occur independent of user decisions is also important to address. For example, trends could be assessed based on five or ten year moving windows, but which result should be used to inform decisions? There may be additional methods to develop using objective criteria to more accurately identify inflection points or other important periods where changes occur independent of a user choice.

The trend assessment approach does not address causality in the observed trends. In this regard, the approach is like other trend assessment methods where the focus is on understanding direction, magnitude, and confidence in changes in water quality variables over time. However, a logical follow-up question is “what” is driving 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](#ref-Wood02), Zuur et al. [2009](#ref-Zuur09)). In such cases, considerations of model structure can have direct implications on conclusions regarding potential causality. As noted above, model structure in describing the long-term trend component was irrelevant for describing long-term trends, although a distinction here is made in the objective of the analysis. Our goal herein was to describe chl-a changes relative to time where the predictors were variations on a general theme (e.g., season vs. year), whereas using GAMs with different predictors assess potential causality is a different application with alternative goals. Therefore, using our approach to assess causality will require considerations of model structure given how GAMs could be used to assess different questions.

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 South Bay (Cloern and Schraga [2016](#ref-Cloern16), Schraga et al. [2020](#ref-Schraga20)). 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 (Jassby et al. [2002](#ref-Jassby02), Cloern et al. [2007](#ref-Cloern07)). The open-source R package developed for this manuscript can be used for these analyses to provide additional insight into potential drivers of water quality change in SFE and other estuarine systems.

# Supplement

A note about the select argument in mgcv: When select = TRUE is included, the comparison between models S and SY changes. This option tells mgcv to penalize the coefficient of the linear terms in the spline. This would be appropriate if cont\_year was an explanatory variable subject to variable selection, but it is irrelevant if including both a linear and spline term for cont\_year. If select = TRUE is used, models S and SY would still be effectively equivalent, but AIC selection would suggest that one model is superior. This result would be an artifact of the choice in model SY to include a linear trend in cont\_year both as a separate term and as part of the spline, with the latter subject to penalization.

# 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.

# Figures

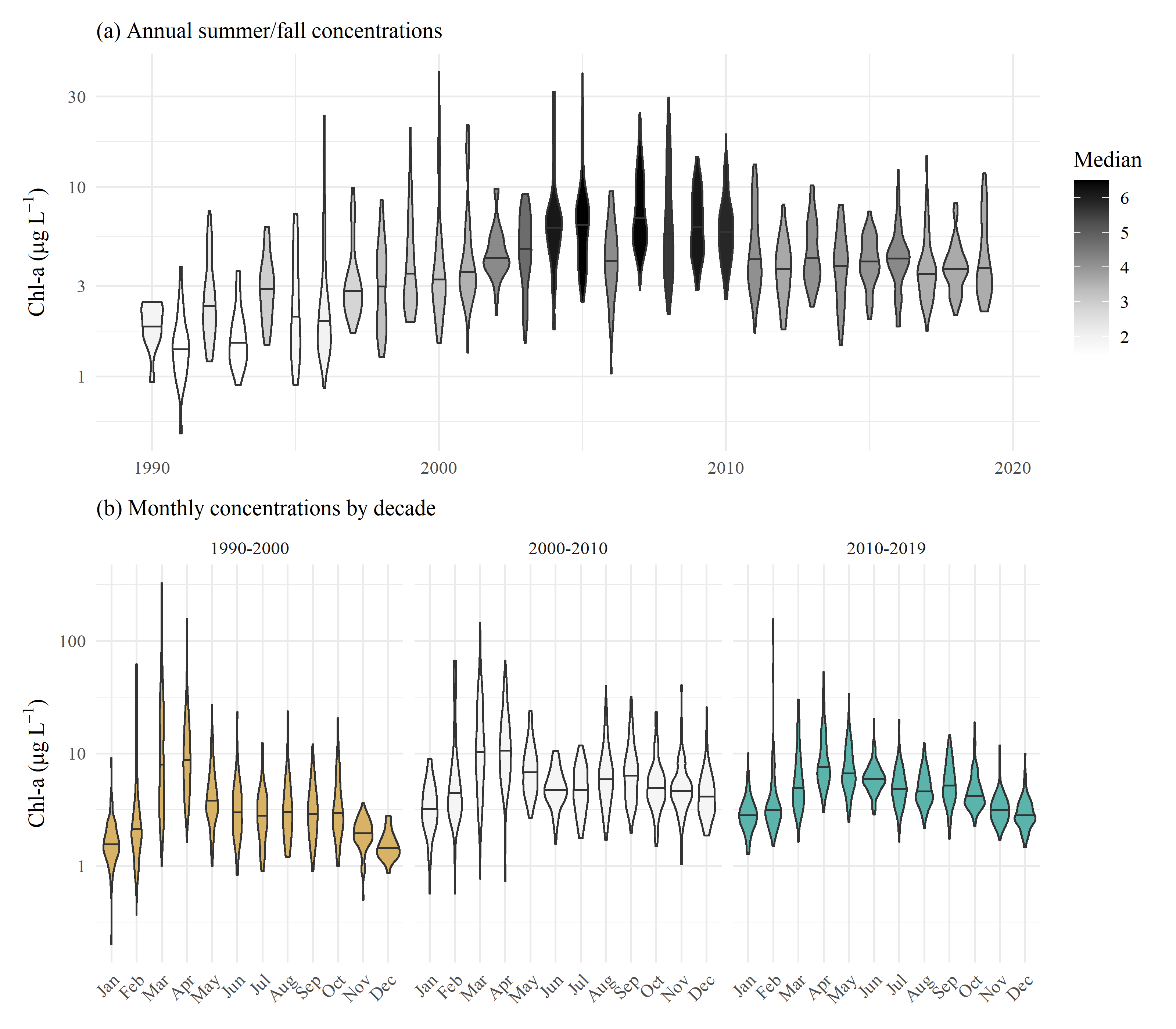


Figure 1: Observed chl-a concentrations for all stations in central and south San Francisco Estuary (18-36, Figure 2), with (a) annual summer/fall concentrations (Aug - Dec) and (b) monthly concentrations by decade.

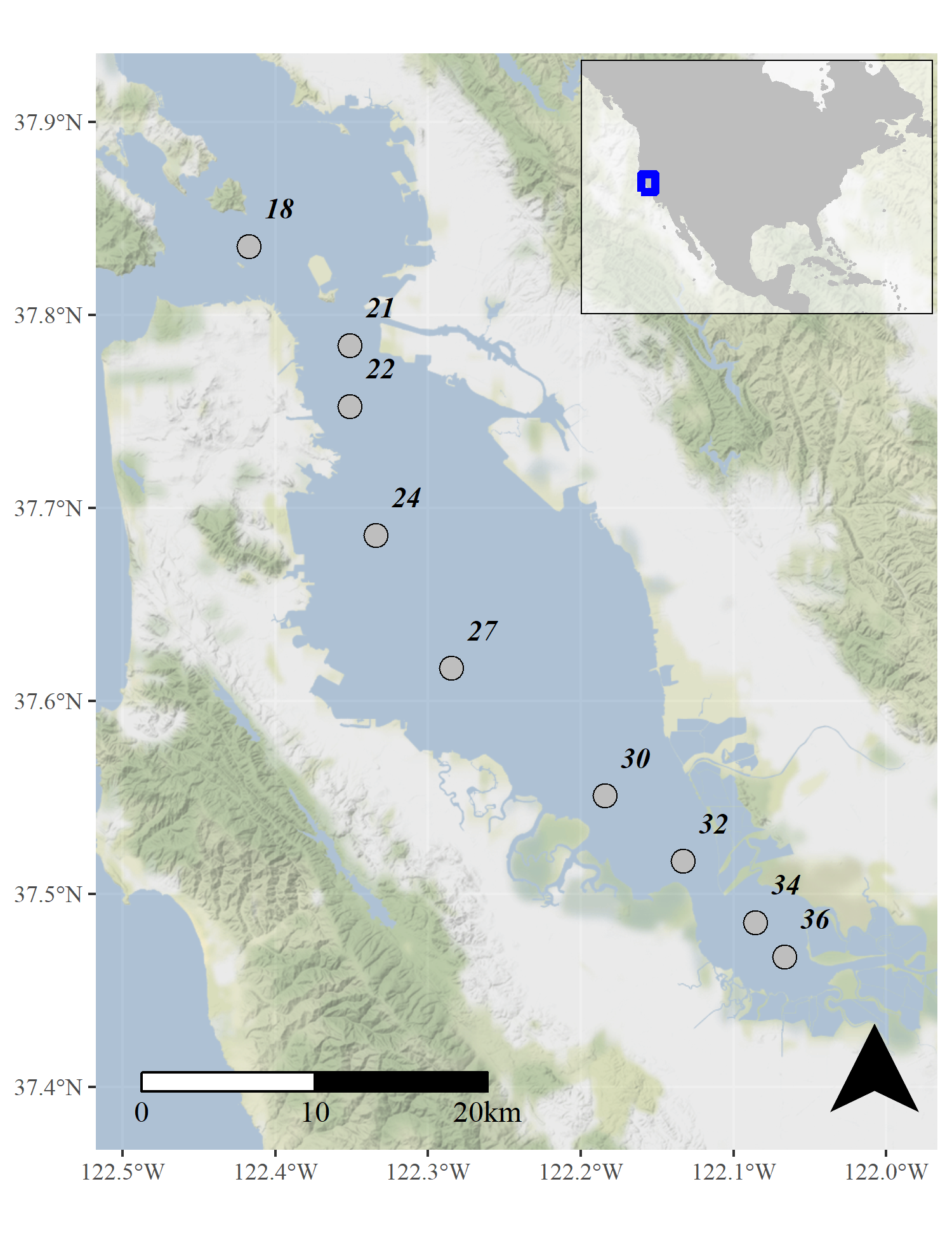


Figure 2: Station locations in the central and south San Francisco Estuary used for analysis. See Table 1 for station descriptions. Full dataset described in Schraga et al. ([2020](#ref-Schraga20)).

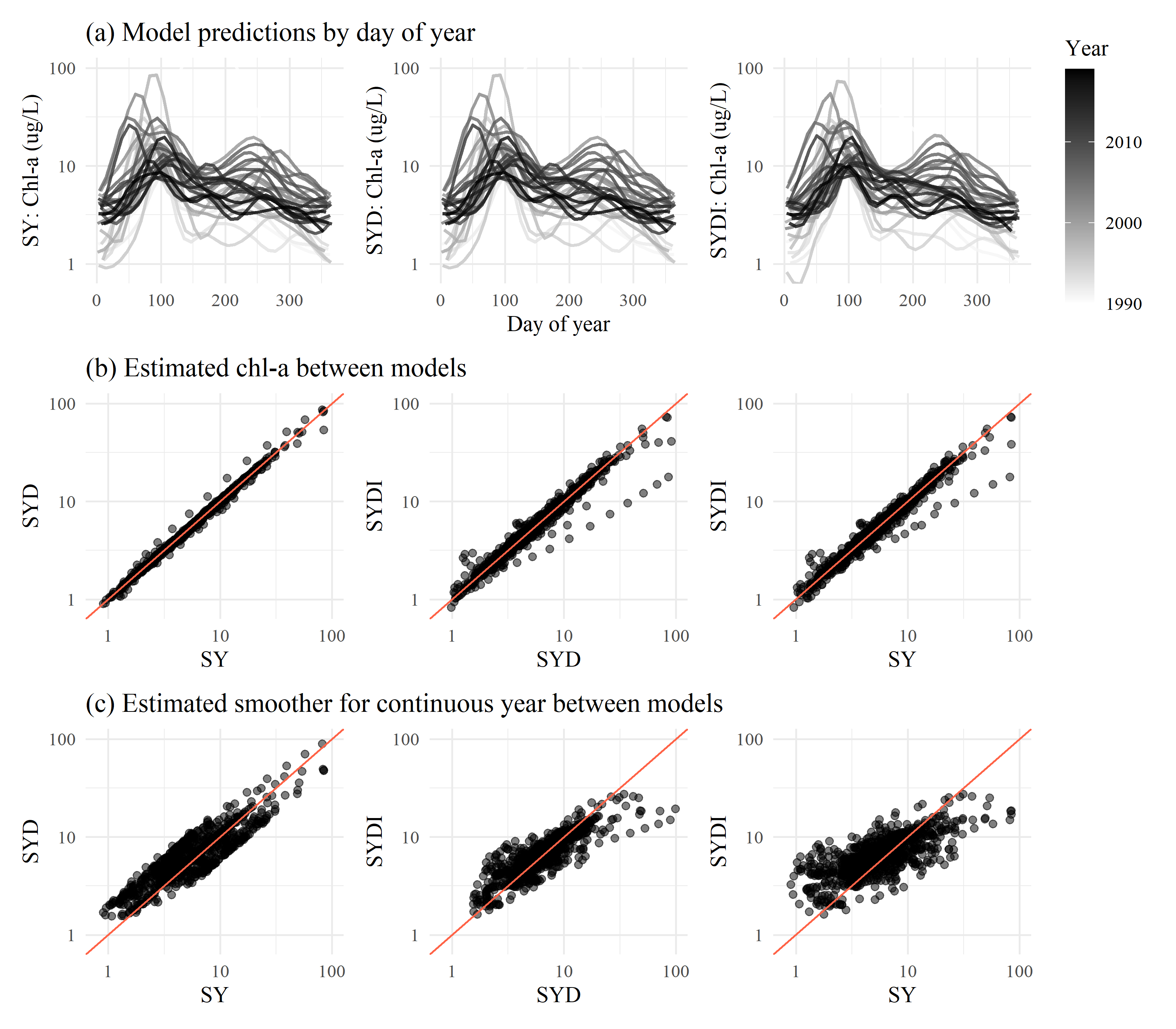


Figure 3: GAM output of estimated chl-a for models SY, SYD, and SYDI. Model S is identical to SY and is not shown. Plots in (a) show model predictions by day of year with separate lines for each year. Plots in (b) show pairwise comparisons of predicted chl-a between the models and plots in (c) show the 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 comparable except for a few observations at extreme values(a), but they vary in the penalties applied to the basis functions for any particular smoother depending on which additive components are included in each model (b). The 1:1 lines are in red to facilitate comparisons.

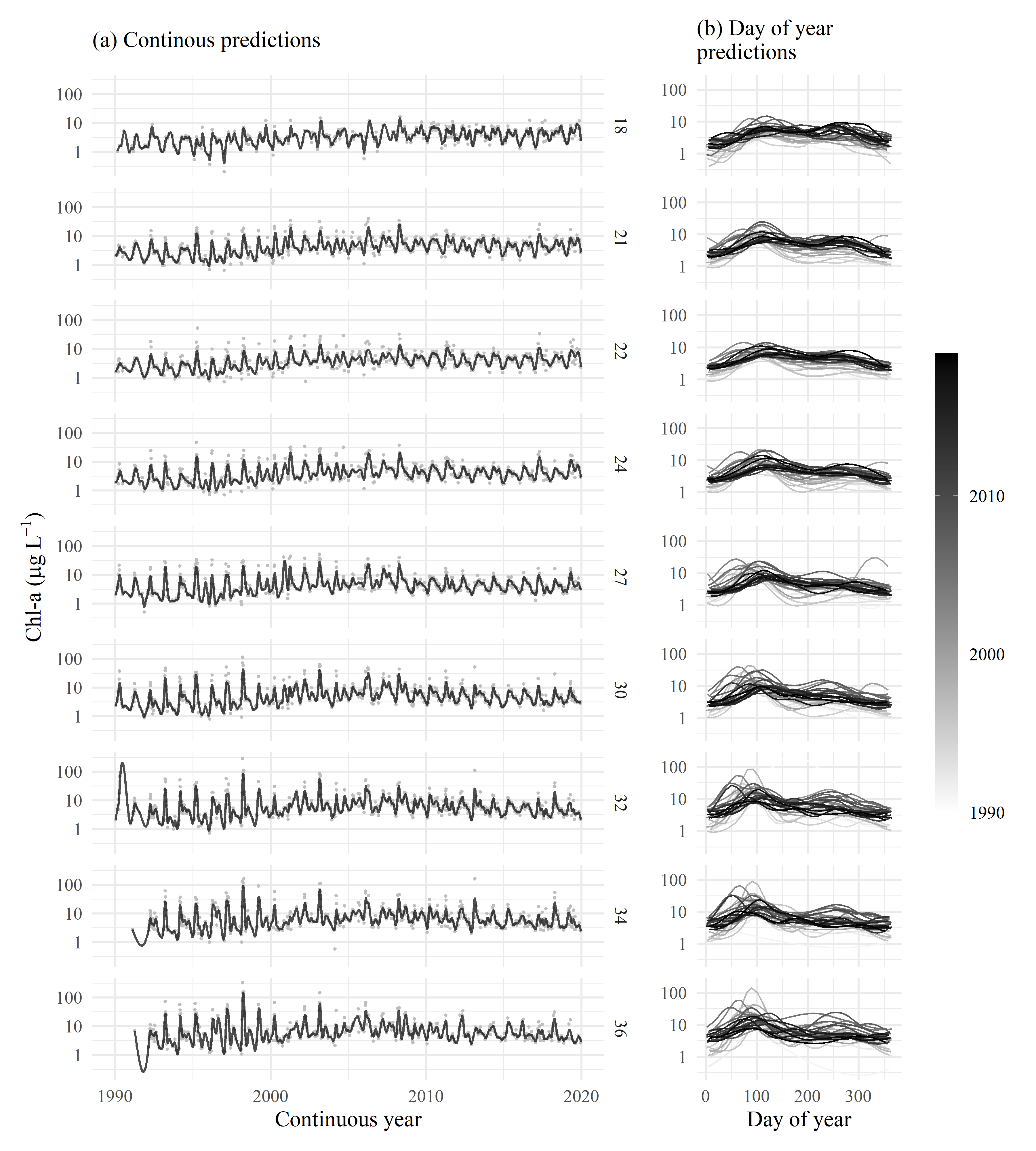


Figure 4: GAM predictions for all stations from north to south for model S. The results show (a) predictions across the time series and (b) predictions by day of year. Observed data in (a) are shown with the gray points. Station locations are in Figure 2.

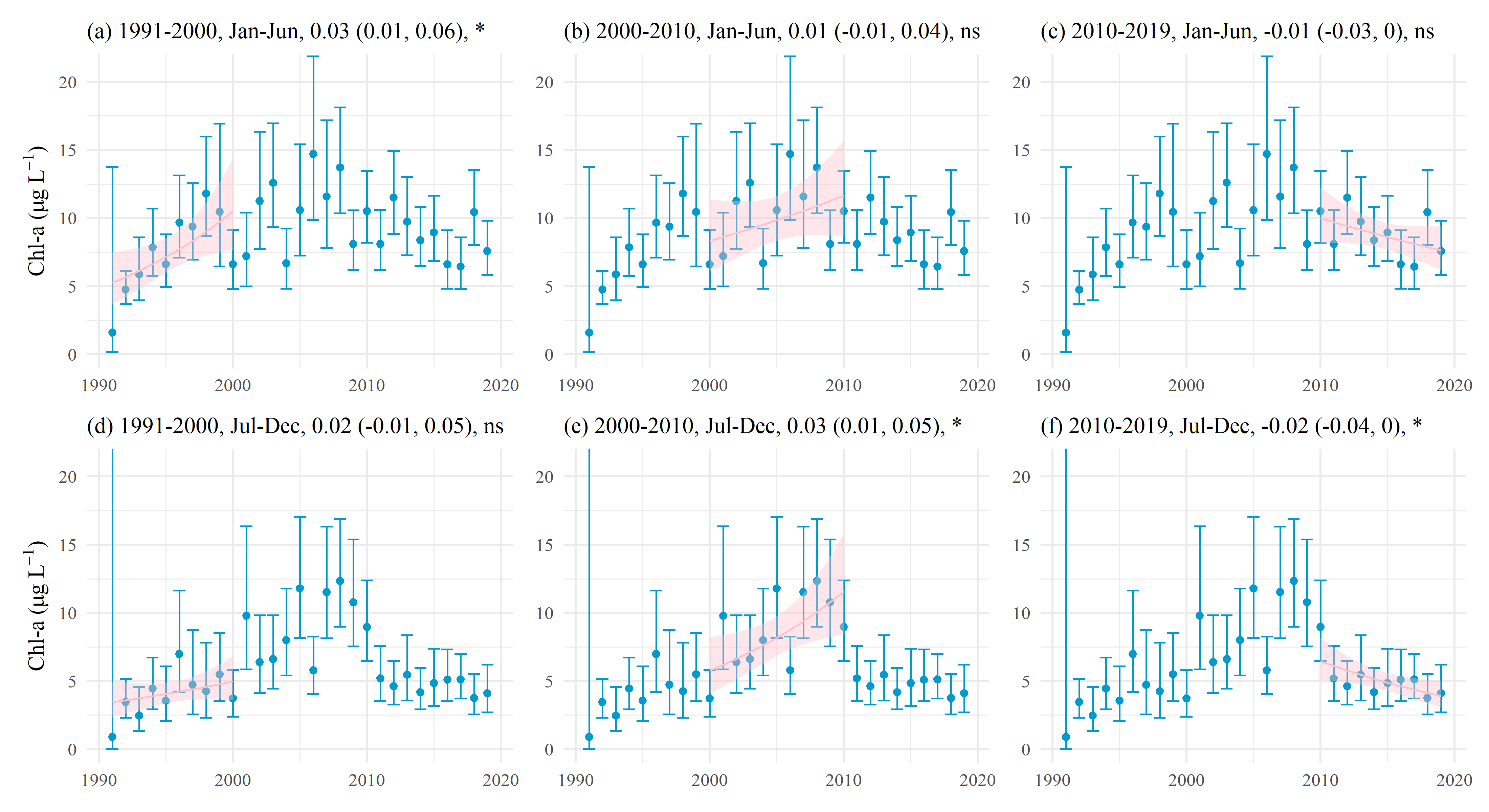


Figure 5: Examples of seasonal averages and trend estimates using results of GAM predictions for station 34. Plots (a), (b), and (c) show trend estimates for January through June and (d), (e), and (f) show trend estimates for July through December. The trend lines estimate the rate of change of chl-a per year, reported as the log-slope (+/- 95 % confidence interval) in the plot title. ns: not significant at = 0.05, \* p < 0.05, \*\* p < 0.005

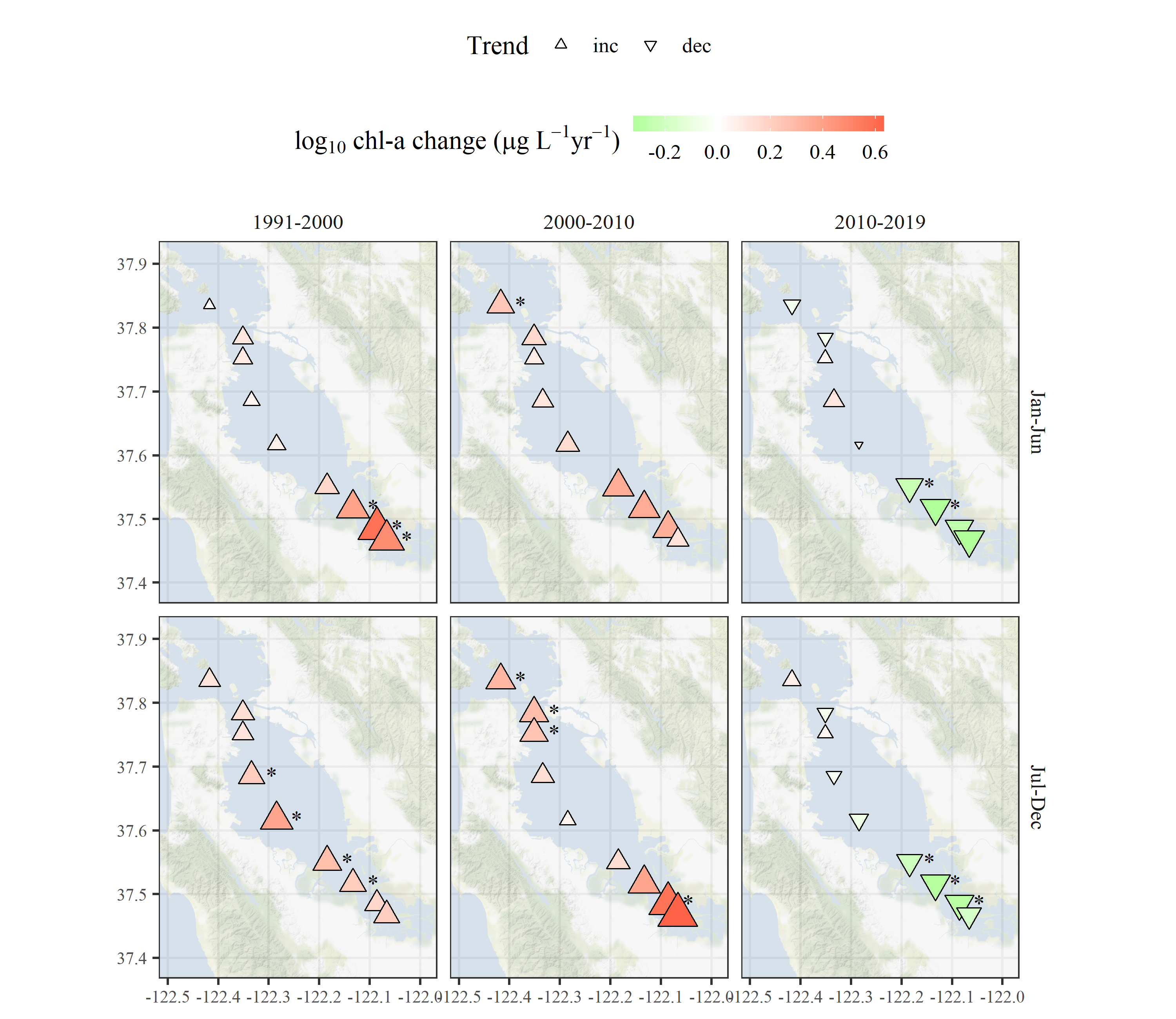


Figure 6: Trend estimates across seasons by decade for chl-a at each station. Point type, shape, and color represent the direction and magnitude of an estimated trend as the log slope for chl-a concentration per year. Trends with are marked with an asterisk. All results are from Model S.

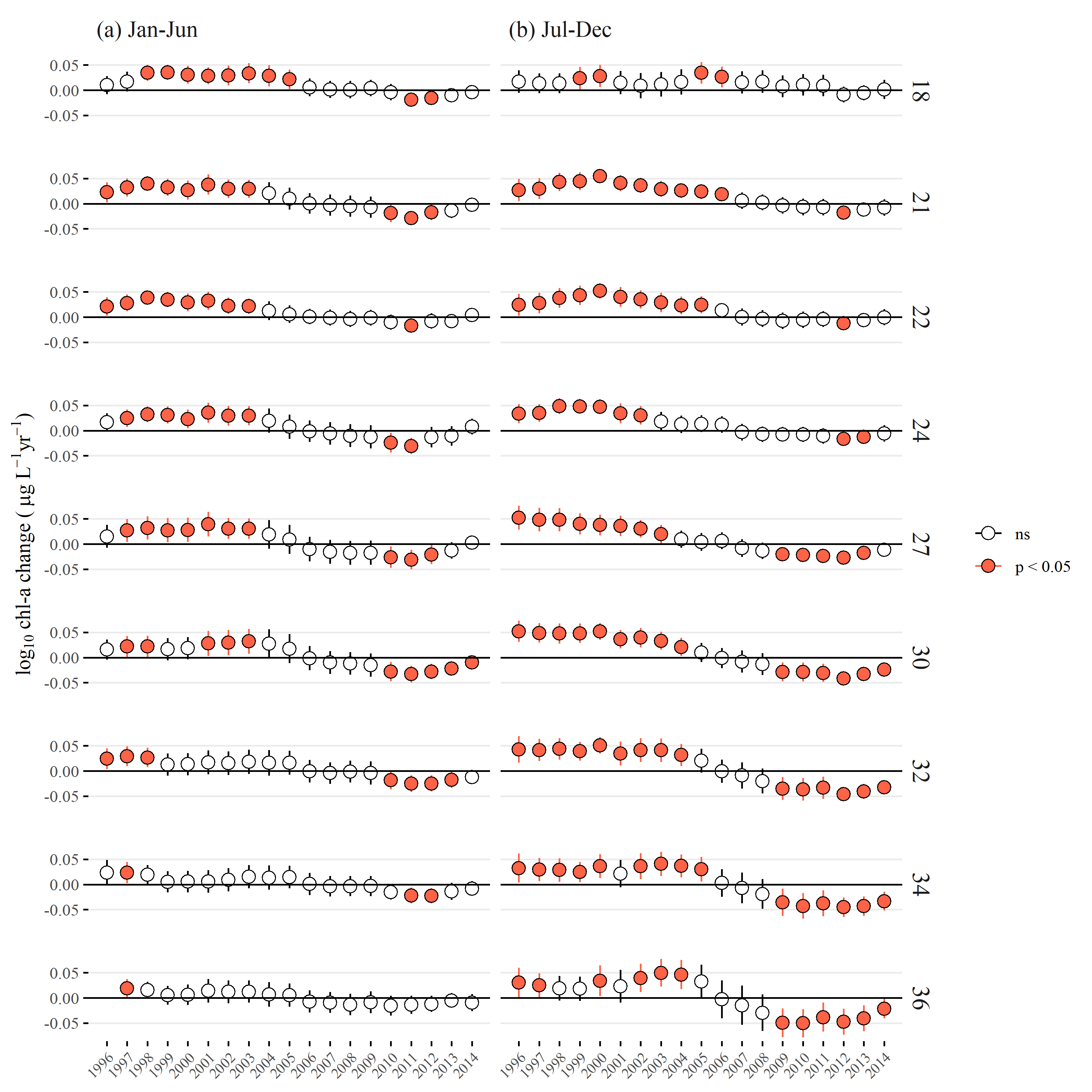


Figure 7: Estimates of log chl-a change per year from applying the mixed-meta analysis across the time series for each station. Stations are arranged top to bottom from north to south. Plots in (a) show estimates for seasonal averages from January to June and plots in (b) show estimates for seasonal 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 after each year. Estimates prior to 1996 and after 2014 are not available becuase of an incomplete ten year record for estimating the trend. Estimates in a year that are significant are shown in red.

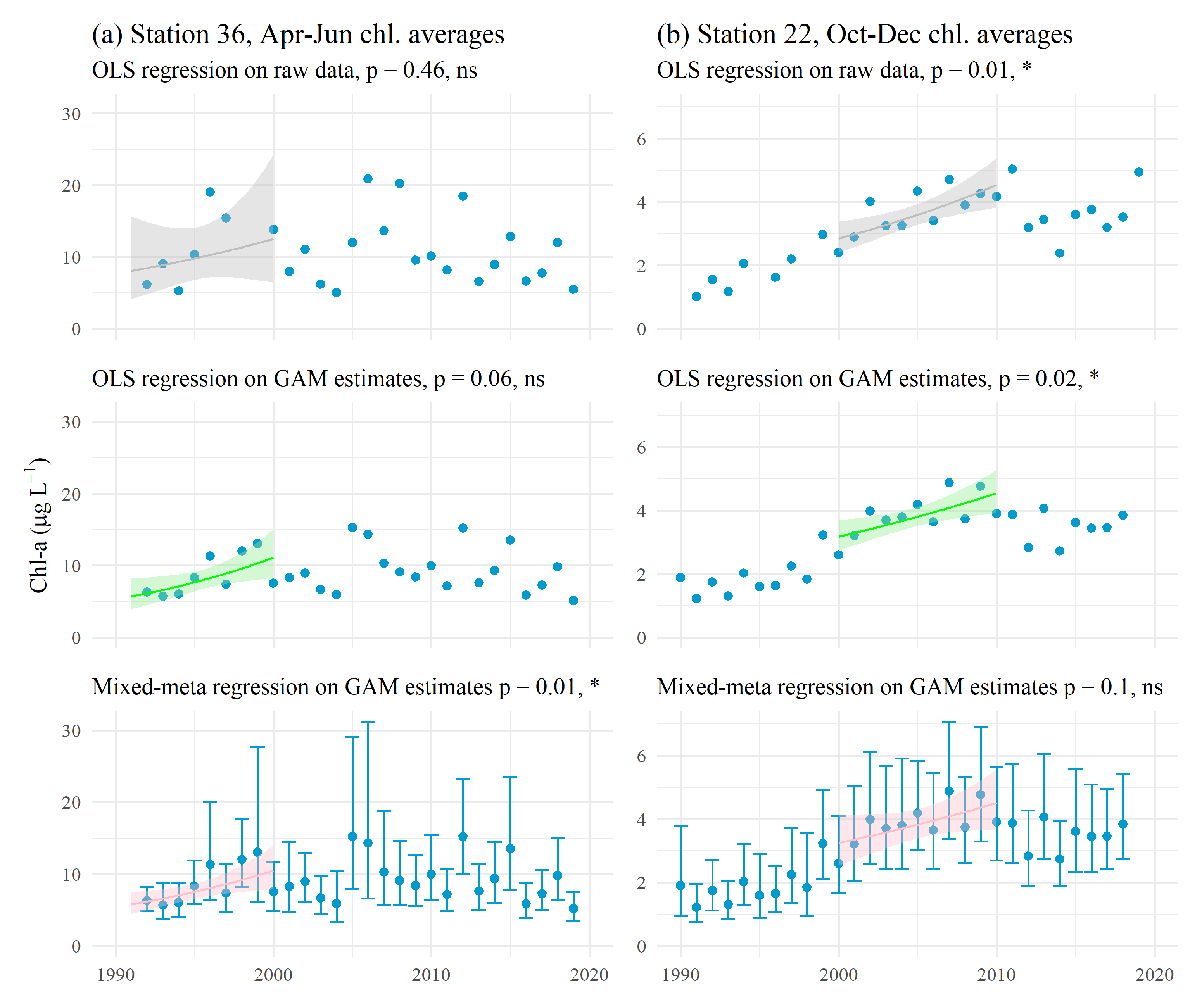


Figure 8: Trend estimate comparisons for three models applied to seasonal averages of chl-a in different annual periods at two example stations. The first row shows OLS (ordinary least squares) regression applied to annual averages of chl-a from the observed data, the second row shows OLS regression applied to annual averages of chl-a from the GAM, and the third row shows mixed-meta regression applied to the annual 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.

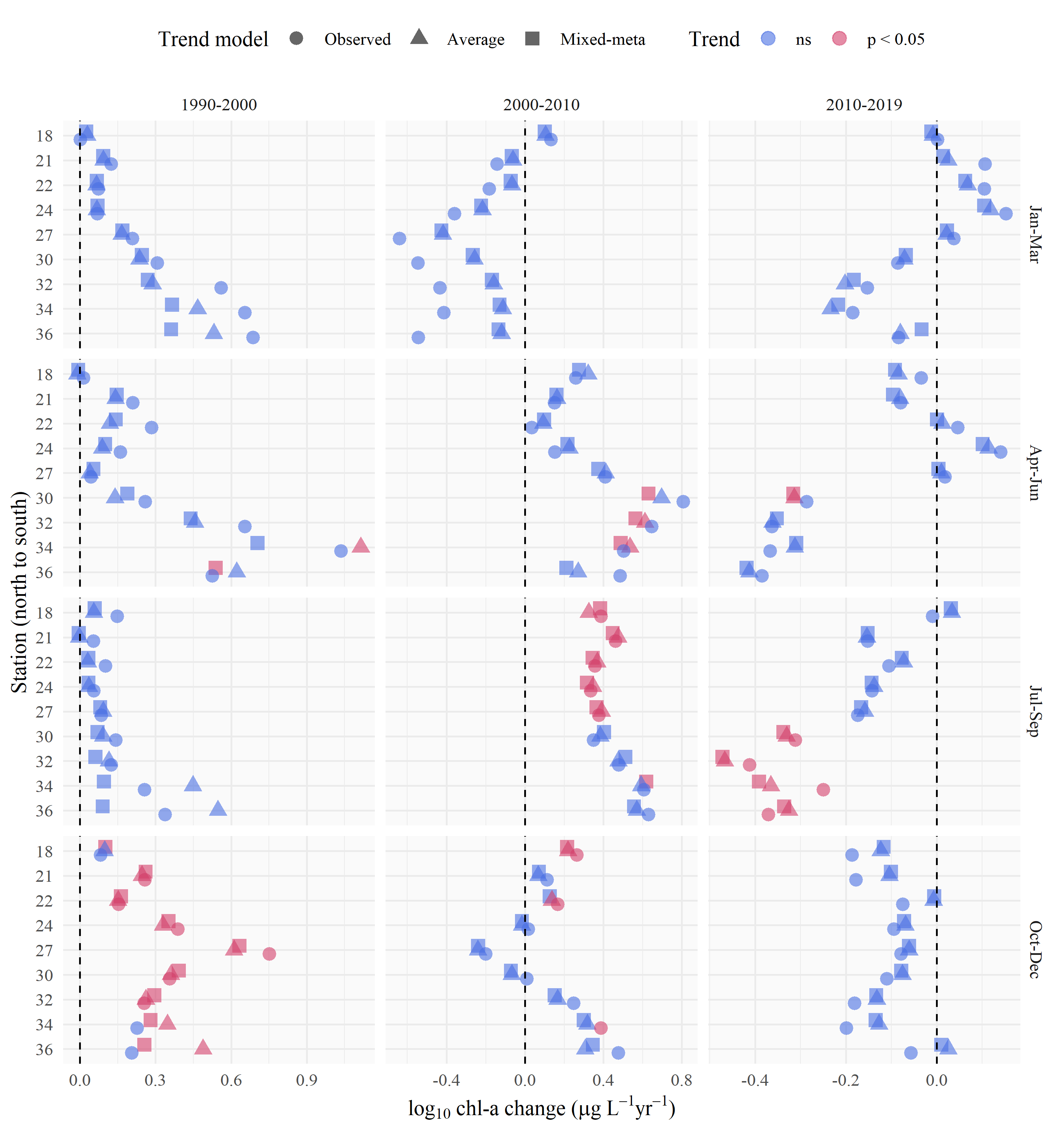


Figure 9: Trend estimate comparisons for three models applied to seasonal averages of chl-a in different annual periods at each station. The “observed” trend model is based on a linear fit to the annual averages of chl-a from the observed data, the “average” trend model is based on a linear fit to the annual averages of chl-a from the GAM model, and the “mixed-meta” trend model is based on a mixed-meta regression model fit to the annual averages of chl-a from the GAM model. Values for each model are the log-slope estimates as annual change per year within each season, with color denoting significant trends.

# Tables

Table 1: Station locations, sample sizes, and summary values (median, minimum, maximum) for chl-a (g L) . Rows 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 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, 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 comparable. 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 | GCV | R2 | smoother | edf | F | p-val |
| S | 0.06 | 0.74 | s(cont\_year) | 242.23 | 6.27 | \*\* |
| SY | 0.06 | 0.74 | s(cont\_year) | 242.23 | 6.27 | \*\* |
| SYD | 0.06 | 0.74 | s(cont\_year) | 229.33 | 3.88 | \*\* |
|  |  |  | s(doy) | 8.07 | 0.11 | \*\* |
| SYDI | 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 generalized cross-validation scores (GCV) and r-squared values.

|  |  |  |
| --- | --- | --- |
| station | GCV | R-squared |
| 18 | 0.04 | 0.78 |
| 21 | 0.04 | 0.70 |
| 22 | 0.05 | 0.59 |
| 24 | 0.04 | 0.69 |
| 27 | 0.05 | 0.72 |
| 30 | 0.05 | 0.74 |
| 32 | 0.06 | 0.74 |
| 34 | 0.07 | 0.68 |
| 36 | 0.07 | 0.73 |

# References

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

Beck, M. W., and J. D. Hagy III. 2015. Adaptation of a weighted regression approach to evaluate water quality trends in an estuary. Environmental Modelling and Assessment 20:637–655.

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

Beck, M. W., and R. R. Murphy. 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.

Bradu, D., and Y. Mundlak. 1970. Estimation in lognormal linear models. Journal of the American Statistical Association 65:198–211.

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

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

Cloern, J. E., A. D. Jassby, J. K. Thompson, and K. A. Hieb. 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., and T. S. Schraga. 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., T. S. Shcraga, E. Nejad, and C. Martin. 2020. Nutrient status of San Francisco Bay and its management implications. Estuaries & Coasts 43:1299–1317.

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

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

Duan, N. 1983. Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association 78:605–610.

Forbes, D. J., and Z. Xie. 2018. Identifying process scales in the Indian River Lagoon, Florida using wavelet transform analysis of dissolved oxygen. Ecological Complexity 36:149–167.

Fouquet, C. de. 2012. Environmental statistics revisited: Is the mean reliable? Environmental Science and Technology 46:1964–1970.

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

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

Haraguchi, L., J. Carstensen, P. C. Abreu, and C. Odebrecht. 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., S. Mazumdar, and V. C. Arena. 2006. A comparative study of the use of GAM and GLM in air pollution research. Environmetrics 17:81–93.

Helsel, D. R., R. M. Hirsch, K. R. Ryberg, S. A. Archfield, and E. J. Gilroy. 2020. Statistical methods in water resources. Page 458. 2nd editions. U.S. Geological Survey Techniques; Methods, book 4, chapter A3, version 1.1, Reston, Virginia.

Hirsch, R. M., S. A. Archfield, and L. A. De Cicco. 2015. A bootstrap method for estimating uncertainty of water quality trends. Environmental Modelling and Software 73:148–166.

Hirsch, R. M., D. L. Moyer, and S. A. Archfield. 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., J. R. Slack, and R. A. Smith. 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., J. E. Cloern, and B. E. Cole. 2002. Annual primary production: Patterns and mechanisms of change in a nutrient-rich tidal ecosystem. Limnology and Oceanography 47:698–712.

Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen. 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38:2895–2907.

Kimmerer, W. J., and J. K. Thompson. 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., D. J. Wilcox, R. R. Murphy, S. R. Marion, and R. J. Orth. 2017. Multiple stressors threaten the imperiled coastal foundation species eelgrass (*zostera marina*) in Chesapeake Bay, USA. Global Change Biology 23:3474–3483.

Lehman, P. W., T. Kurobe, S. Lesmeister, D. Baxa, A. Tung, and S. J. Teh. 2017. Impacts of the 2014 severe drought on the Microcystis bloom in San Francisco Estuary. Harmful Algae 63:94–108.

Morton, R., and B. L. Henderson. 2008. Estimation of nonlinear trends in water quality: An improved approach using generalized additive models. Water Resources Research 44:W07420.

Murphy, R. R., E. Perry, J. Harcum, and J. Keisman. 2019. A Generalized Additive Model Approach to evaluating water quality: Chesapeake Bay case study. Environmental Modelling & Software 118:1–13.

Navarro, G., I. Caballero, L. Prieto, A. Vázquez, S. Flecha, I. E. Huertas, and J. Ruiz. 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.

Novick, E., and D. Senn. 2014. External nutrient loads to San Francisco Bay. San Francisco Estuary Institute, Richmond, CA.

Pearce, J. L., J. Beringer, N. Nicholls, R. J. Hyndman, and N. J. Tapper. 2011. Quantifying the influence of local meteorology on air quality using generalized additive models. Atmospheric Environment 45:1328–1336.

Racault, M. F., S. Sathyendranath, and T. Platt. 2014. Impact of missing data on the estimation of ecological indicators from satellite ocean-colour time series. Remote Sensing of Environment 152:15–28.

Schraga, T. S., E. S. Nejad, C. A. Martin, and J. E. Cloern. 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., B. Armstrong, M. Blangiardo, and A. Gasparrini. 2019. An extended mixed-effects framework for meta-analysis. Statistics in Medicine 38:5429–5444.

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

Urquhart, N. S., S. G. Paulsen, and D. P. Larsen. 1998. Monitoring for policy-relevant regional trends over time. Ecological Applications 8:246–257.

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

Wood, S. N. 2003. Thin-plate regression splines. Journal of the Royal Statistical Society (B) 65:95–114.

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

Wood, S. N. 2017. Generalized additive models: An introduction with r. Page 476. 2nd editions. Chapman; Hall, CRC Press, London, United Kingdom.

Wood, S. N., and N. H. Augustin. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157:157–177.

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

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed effects models and extensions in ecology with r. Page 574. Springer-Verlag, New York, New York.