Trend detection of atmospheric time series: incorporating appropriate uncertainty estimates and handling extreme events

Kai-Lan Chang[1,2], Martin G. Schultz[3], Xin Lan[1,4], Audra McClure-Begley[1,4], Irina Petropavlovskikh[1,4], Xiaobin Xu[5], Jerry R. Ziemke[6,7]

[1]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, CO, USA

[2]NOAA Chemical Sciences Laboratory, Boulder, CO, USA

[3]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany

[4]NOAA Global Monitoring Laboratory, Boulder, CO, USA

[5]Key Laboratory for Atmospheric Chemistry, Institute of Atmospheric Composition, Chinese Academy of Meteorological Sciences, China Meteorological Administration, Beijing, China

[6]NASA Goddard Space Flight Center, Greenbelt, MD, USA

[7]Morgan State University, Baltimore, MD, USA

[*]kai-lan.chang@noaa.gov

Abstract

This paper is aimed at atmospheric scientists without formal training in statistical theory. Its goal is to, 1) provide a critical review of the rationale for trend analysis of the time series typically encountered in the field of atmospheric chemistry; 2) describe a range of trend-detection methods; and 3) demonstrate effective means of conveying the results to a general audience. Trend detections in atmospheric chemical composition data are often challenged by a variety of sources of uncertainty, which often behave differently to other environmental phenomena such as temperature, precipitation rate, or stream flow, and may require specific methods depending on the science questions to be addressed. Some sources of uncertainty can be explicitly included in the model specification, such as autocorrelation and seasonality, but some inherent uncertainties are difficult to quantify, such as data heterogeneity and measurement uncertainty due to the combined effect of short and long term natural variability, instrumental stability, and aggregation of data from sparse sampling frequency. Failure to account for these uncertainties might result in an inappropriate inference of the trends and their estimation errors. On

the other hand, the variation in extreme events might be interesting for different scientific questions, for example, the frequency of extremely high surface ozone events and their relevance to human health. In this study we aim to, 1) review trend detection methods for addressing different levels of data complexity in different chemical species; 2) demonstrate that the incorporation of scientifically interpretable covariates can outperform pure numerical curve fitting techniques in terms of uncertainty reduction and improved predictability; 3) illustrate the study of trends based on extreme quantiles that can provide insight beyond standard mean or median based trend estimates; and 4) present an advanced method of quantifying regional trends based on the inter-site correlations of multi-site data. All demonstrations are based on time series of observed trace gases relevant to atmospheric chemistry but the methods can be applied to other environmental data sets.

1. Introduction

Chandler and Scott (2011) defined a trend as the "long-term temporal variation in the statistical properties of a process, where 'long-term' depends on the application." Since long-term is a relative concept, attempts to detect possible trends might be made before a time series is of sufficient length for accurate trend detection, because in many circumstances the necessary length of the time series is not known beforehand. Under these circumstances, the trend detection might be less reliable when dealing with the large complexities of atmospheric chemical composition measurements, e.g. the combined effect of spatial and temporal variability, instrument detection levels, and/or the influence of extreme events. Therefore the statistical models that ignore the underlying complexities produce under-represented estimation errors and biased trend estimates, providing either an over-interpretation of noisy data or inconsistent results for scientific assessment (Tong, 2019). These circumstances can be avoided if the atmospheric chemistry research community is familiar with a range of acceptable statistical approaches and their correct application.

Sound scientific assessment relies on good statistical practices. Whereas various trend detection techniques often arrive at similar answers with respect to estimated slopes or offsets, the uncertainties estimated by these different techniques vary widely. Because scientists assess the robustness of a trend through the associated uncertainty estimate, it is critical that an appropriate trend detection technique is applied.

An appropriate reported uncertainty is as important as the trend estimate, and a trend value without a properly derived uncertainty estimate provides no useful information for scientific assessment. Even though widely applied regression-based approaches always report the standard error (i.e. uncertainty) associated with each regression coefficient (e.g. trend value), the uncertainties can be unrealistically narrow if the model is applied incorrectly. The statistical model can be inappropriate if, 1) the model assumptions are violated; and/or 2) the model specifications are not adequate. If the model assumptions are not met, the result might be unreliable; if the model specifications are either mis-fitted, under-fitted (oversimplifying the reality) or over-fitted (using too many predictors to describe unimportant variation), the result is not representative.

Atmospheric scientists interested in understanding methods of time series analysis and trend detection can turn to a wide range of text books and review articles for guidance (Brockwell and Davis, 1987; Hamilton, 1994; Chatfield, 2000; Lütkepohl, 2005; Durbin and Koopman, 2012; Box et al., 2015; Shumway and Stoffer, 2017), with some sources focusing on environmental time series (Chandler and Scott, 2011), meteorology (Wilks, 2011) or climate change (Von Storch and Zwiers, 2001). However, none of these references focus on atmospheric chemistry, which may leave atmospheric chemists unaware of the most appropriate statistical methods for analyzing time series of trace gases. This paper is aimed at atmospheric chemists to show how trend analysis can be improved if appropriate techniques are applied, and to encourage the uptake of statistical thinking (i.e. not relying on a single approach).

Figure 1 shows the monthly mean time series of several trace gases measured at surface level from Mauna Loa Observatory, Hawaii (19.5°N and 155.6°W; 3397 m above sea level) (Oltmans and Komhyr, 1986; Thoning et al., 1989; Dlugokencky et al., 2020). This example demonstrates that the data characteristics and variability can vary widely among different chemical species, so a single set of trend techniques would have difficulty addressing the range of factors that contribute to uncertainty. The data characteristics of these time series can be summarized as follows: 1) *Seasonality*: methane (CH_4) and carbon dioxide (CO_2) have a regular seasonal cycle with a lower variability, carbon monoxide (CO) and ozone (O_3) have an erratic seasonal cycle with a higher variability, and nitrous oxide (N_2O) and sulfur-hexafluoride (SF_6) have no seasonal cycles because they do not interact with the biosphere and lack efficient sink mechanisms in the troposphere; 2) *Magnitude of data variability*: Strong increasing tendencies are obvious for methane, CO_2, N_2O and SF_6 even without a quantification of the trends, while trend detection for CO and ozone is challenging due to erratic seasonal variations and apparently weak changes over time; 3) *Nonlinearity*: Even though both methane and CO_2 show increasing trends, methane has a leveling-off in the early 2000s, which is not seen in CO_2; 4) *Autocorrelation*: the steady variation of CO_2 indicates that observations in the past are correlated with current observations, even if data are separated by several years (long memory, Barassi et al. (2011)), while a much shorter correlation range is found for ozone at the same location (short memory) (see later analysis) ; 5) *Extreme events*: An interesting aspect of CO and ozone is that the extreme events have changed over time; the high extremes seem to show a stronger decrease than the low extremes for CO, but the change of the extreme events for ozone is rather uncertain. Therefore, trend detection of the extreme quantiles could also be explored with appropriate techniques.

This paper is outlined as follows: Section 2 reviews the challenges in trend detection of atmospheric time series. Section 3 describes the framework of trend detection techniques. Several demonstrations of these methods are presented in Sections 4 to 6. Section 4 examines the data characteristics and autocorrelation associated with different chemical species measured at MLO (although we focus on trace gases, the methodology applies to aerosols too). Section 5 applies the quantile regression method to study the changes in extreme events, which provides additional insight to the commonly calculated mean or median trends. Section 6 demonstrates a method for deriving regional mean and extreme quantile trends from an extensive ozone monitoring network in the Southwest USA. This advanced statistical technique, when applied to a large ensemble of time series data, not only provides more concrete evidence of the trends, but it also gives more robust and consistent results regarding quantile changes. Section 7 discusses additional advanced trend detection techniques relevant to this study. The paper concludes in Section 8 with discussions on the effectiveness of various trend detection techniques.

2. Review of challenges in trend detections of atmospheric time series

Various complexities are associated with the trend detection of atmospheric time series. The fundamental statistical principles of trend detection place the emphasis on the magnitude of the trend and its associated error, sample size and autocorrelation (Tiao et al., 1990; Weatherhead et al., 1998). These principles are designed to provide sufficient (or minimum) evidence and require that the underlying assumptions are fulfilled and that model residuals are uncorrelated. Explanation of the variability is a more difficult task than trend detection, because it requires identification of all (or the most important) sources of the variability and the proper quantification of each attribution (Stott et al., 2010; Hegerl and Zwiers, 2011). To achieve the goal of appropriate attribution of data variability, we need to identify the best correlation between the observations and each covariate (i.e. a variable that is possibly predictive of the data variability) via a sequential process of variable selections and model comparisons; these processes ensure that the resulting model is adequate (neither under-fitted nor over-fitted). Additional techniques are also available for describing common phenomena such as changing magnitude of variability or varying seasonal cycle over time (Cleveland et al., 1990).

A further important aspect for atmospheric composition trends is the detection and/or quantification of trend changes. This is especially relevant in the policy arena to determine the efficacy of certain air quality measures (Box and Tiao, 1975), or to examine if the changes can be attributable to other natural or human-caused factors (Reinsel et al., 2005; Friedrich et al., 2020a).

In addition to being one of the most variable trace gases (as shown in the Introduction), ozone's extreme values are of particular interest to the research community and regulatory agencies. For example, epidemiologists might use the daily maximum 8-hour (MDA8) average to quantify human exposure to ozone pollution (Turner et al., 2016), or the number of days per summertime period in which the MDA8 exceeds 70 ppb to assess the frequency of high ozone events (the latter metric does not produce a continuous response, and an adjustment for the count data needs to be made by using generalized linear or additive models (Chang et al., 2017)). For regulatory purposes, the United States Environmental Protection Agency uses the annual 4th highest MDA8 ozone value at a monitoring site when identifying regions that comply with the National Ambient Air Quality Standards for ozone, while Europe's ozone target value is based on the 26th highest MDA8 value of a year (see also Fleming et al. (2018)).

It is crucial to recognize that extreme values are data characteristics and not outliers. The formal definition of an outlier can be considered to be a data point that shows a substantial deviation from other data points, so it is reasonable to suspect that this data point is generated by a different process or mechanism (Hawkins, 1980; Aggarwal, 2015). Although this statement is qualitative, it suggests the extreme values should not be seen as outliers if their occurrence can be justified through scientific explanation. The interest in extreme ozone values introduces an important consideration for this study: naturally occurring extreme values are not equivalent to outliers which are defined as erroneous values. Extreme values may indeed contain important information that is relevant for trend analysis. Specifically, non-parametric methods like the often-applied Sen-Theil estimator, do not distinguish between outliers (presumably due to instrumental error) and data points that simply present larger deviations from the median (presumably due to natural variability); as a result this method ignores up to 29% of the data set. However, those neglected data in the Sen-Theil estimator can be put to good use in estimating changes of extreme events. As discussed later, quantile regression is designed to efficiently provide trend estimates based on multiple quantiles (not just the median) with a single specification.

An additional trait of atmospheric chemistry observations is the measurement uncertainty, associated with instrumental and sampling conditions, and/or instrument calibration. For example, balloon-borne ozonesondes operated by NOAA's Global Monitoring Laboratory (GML) have a typical sampling frequency of once per week, and therefore aggregated monthly means and standard deviations (or errors) are based on only 4 or 5 observations. These aggregated time series are often considered to be highly uncertain due to low sampling frequency combined with inherent natural variability (Saunois et al., 2012; Chang et al., 2020).

All the uncertainties discussed above have an impact on the trend estimate and/or its estimated error, and therefore each factor must be considered carefully to avoid biased or inappropriate conclusions. Even though a large set of complications can be introduced by data measurement methods, e.g., instrumental stability/accuracy (Ambrosino and Chandler, 2013; Weatherhead et al., 2017b; Von Brömssen et al., 2018), representativeness of the measurements (Weatherhead et al., 2017a), and sampling frequency (Chang et al., 2020), these issues are beyond the scope of this study. Here we focus on various approaches to study the characteristics of the available data, for a detection of trends in single time series and multi-site data.

3. Statistical Methods

The methodology is organized as follows: discussion of relevant factors, construction of the statistical relationships, possible extensions, and an approach to report the robustness of trends.

3.1. Ingredients in a trend detection model

To represent the data variability in an atmospheric time series, we need to identify the relevant factors that can potentially affect the trend detection. For example, we can decompose a time series into several components as follows:

$$\text{obs} = \text{trend} + \text{seasonal cycle} + \text{covariates} + \text{error}, \tag{1}$$

where the first three terms link ozone to the long-term change, cyclic seasonal pattern, or external variability (these variables are referred as covariates and their magnitude or correlation with ozone can be measured by regression coefficients), and the last term represents the model residuals. The feature of this regression models include: 1) the trend and seasonal components can be further expressed as various forms, e.g. the trend component can be a line, a piecewise linear function (i.e. change point analysis), or any other nonlinear shape deemed appropriate, and the seasonal cycle can be a combination of sine, cosine or any periodic functions; 2) this model remains linear and additive (nonlinear regression is not considered here), even if the trend and seasonal components are nonlinear; and 3) the relevant covariates should be considered by the data characteristics and scientific question to be addressed. For example, the addition of a meteorological adjustment might be important for short-term trend detection and reducing the magnitude of uncertainty (Camalier et al., 2007; Wells et al., 2021), or in multi-site data we can use a spatial-referenced covariate to account for inter-site correlation (Chandler and Scott, 2011).

3.2. Investigation of statistical relationships by different methods

In contrast to the identification of important components discussed above, the methods introduced in this section focus on how the statistical relationships can be explored by several different approaches. To simplify the scenario, the demonstration is made through a basic equation for the linear trend detection of a time series, y_t, that can be expressed as $y_t = \beta_0 + \beta_1 t + N_t, \quad t = 1, \cdots, T$, which involves an intercept β_0, a slope β_1 and residual series N_t. Even though this structure looks simple, there are several methods available for estimation of these coefficients (e.g., based on median, trimmed mean, weighted mean... etc). These methods can be classified into 2 categories:

- Classical nonparametric approaches: these approaches often place the emphasis on no assumption being made for the data distribution, and therefore are usually median-based methods. The Sen-Theil estimator and Siegel's repeated medians are the most common nonparametric methods (Theil, 1950; Sen, 1968; Siegel, 1982). The Sen-Theil estimator finds the overall median change by calculating all pairwise differences between observations. The Siegel's repeated medians method finds the overall trend in 2 steps: 1) for each observation, a median change is calculated from the median of the pairwise difference against all the other observations; and 2) the overall trend is then assigned to the median among all these median values. Thus the Siegel's estimator is a more robust and computationally expensive variant of the Sen-Theil estimator. Neither the Sen-Theil nor Siegels methods involve any numerical optimization, instead they assign the trend from pairwise differences or individualized medians.

- Regression based approaches: the fundamental optimization of a simple linear equation is achieved by finding the optimal coefficients for (β_0, β_1) that minimizes the following loss functions:

$$\sum_{t=1}^{T}(y_t - \beta_0 - \beta_1 t)^2 \quad \text{for the mean estimator of the coefficients,}$$

$$\sum_{t=1}^{T}|y_t - \beta_0 - \beta_1 t| \quad \text{for the median estimator of the coefficients.}$$

  The first equation is called the ordinary least squares (OLS), and the second equation is called the least absolute deviations (LAD).

The OLS estimator is notoriously vulnerable to aberrant outliers. Therefore, several adjusted techniques are available for avoiding the influence of aberrant outliers. (It should be noted that traditional simple and multiple linear regressions are mainly based on OLS):

1) *LTS (least trimmed squares, Rousseeuw (1985))*: the LTS is designed to minimize the residual sum of squares over a subset of data, and exclude potential outliers from the fit (which is determined by the numerical optimization) .
2) *LMS (least median of squares, Rousseeuw (1984))*: this approach replaces the "sum" in least squares criterion with the median of squared residuals in the loss function:

$$\operatorname*{median}_{t}\{(y_t - \beta_0 - \beta_1 t)^2\}.$$

   By replacing the sum with the median, the influence of outliers on the optimization can be eliminated.
3) *WLS (weighted least squares)*: the WLS gives lower weights to the observations with higher uncertainties, since high uncertainty is often associated with extreme observations (although the appropriate weights are often difficult to quantify). If the weights w_t for each time t are supplied, the loss function for OLS can be modified as:

$$\sum_{t=1}^{T} w_t^2(y_t - \beta_0 - \beta_1 t)^2.$$

WLS is one of the remedies for the violation of homoscedasticity (see Appendix S2). This approach is particularly useful when repeated measurements are available, as long as the variance at each time point can be properly quantified. The time series in this paper do not have the associated variance series, therefore we use the inverse of monthly variance (derived from long-term mean series in each month) for data weighting (Schwartz, 1994).

4) *Ridge regression (Hoerl and Kennard, 1970)*: this method prevents the problem of overfitting to the outliers or noisy observations by altering the loss function as:

$$\sum_{t=1}^{T}(y_t - \beta_0 - \beta_1 t)^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \text{ where } \|\boldsymbol{\beta}\|_2^2 = \beta_0^2 + \beta_1^2 \leq c < \infty, \tag{2}$$

the second term is a parameter λ associated with L_2 (Euclidean) norm $\|\cdot\|_2$ which constrains the regression coefficients within reasonable ranges. Even though it looks like a simple adjustment, this approach essentially introduces one of the most important concepts in modern statistics, i.e., regularization (or roughness penalty). The regularization is an iterative process to filter out the noisy variations from systematic patterns in the data structure. In the current setting, we only have 2 parameters that need to be determined, but if we choose to extend the model, such as replacing the linear term with a nonlinear Loess (the locally weighted smoothing (Cleveland et al., 1990)), the result will be many undetermined (hyper-)parameters. The regularization technique can prevent overfitting to aberrant outliers and unrealistic wiggles caused by the noisy observations, and ease the multicollinearity if multiple covariates are required for explaining the data variability (Tikhonov et al., 2013). Note that equation (2) is presented as an illustration, and the regularization does not have to apply to the intercept.

5) *Lasso (least absolute shrinkage and selection operator, Tibshirani (1996))*: this method replaces L_2 norm (Euclidean distance) with L_1 norm (absolute-value distance) in equation (2), i.e., $\|\boldsymbol{\beta}\|_1 = |\beta_0 + \beta_1|$. In the multivariate setting L_1 norm outperforms L_2 norm in terms of variable selection, as L_1 norm tends to reduce the model complexity and selects fewer covariates (Leng et al., 2006).

6) *QR (quantile regression, Koenker and Zhao (1996))*: The QR differs from the techniques above, as it is an optimization-based approach to find the quantile trend (in addition to the median) by minimizing the following loss

function:

$$\sum_{t:y_t \geq \beta_0+\beta_1 t}^{T} q \left|y_t - \beta_0 - \beta_1 t\right| + \sum_{t:y_t < \beta_0+\beta_1 t}^{T} (1-q) \left|y_t - \beta_0 - \beta_1 t\right|,$$

where q is the quantile. When $q = 0.5$, the solution is equivalent to the least absolute deviations (LAD, labelled as QR-50th in this study). Numerically QR is a natural approach to quantify quantile changes other than the median.

Even though we use a simple linear equation for the above demonstrations, these can be extended to the multivariate case (Equation (1)), i.e., we can stack up all of the temporal indices and covariates into a matrix $\mathbf{X}_t$ with a corresponding coefficients vector $\boldsymbol{\beta}$, shortening the trend equation to $y_t = \mathbf{X}_t\boldsymbol{\beta} + \epsilon_t, t = 1, \ldots, T$. One can replace $y_t - \beta_0 - \beta_1 t$ in the loss functions with $y_t - \mathbf{X}_t\boldsymbol{\beta}$ in any of the regression based approaches. It should be noted that the classical nonparametric approaches do not involve any numerical optimizations and loss functions, thus this multivariate extension does not apply to those methods.

Further information on these methods is provided in three appendices in the supplemental material. Appendix S1 gives a historical context explaining why these techniques were developed and how they took advantage of increasing computing power. Appendix 2 discusses fundamental assumptions related to the OLS and how other robust techniques can be an alternative. Appendix S3 describes how the autocorrelation can be accounted for in regression based approaches.

3.3. Incorporation of various complexities in suitable methods

When discussing the assumptions and formulation of trend detection models, the distinction between various relevant factors (as described in Section 3.1) and robust techniques (as described in Section 3.2) is poorly documented. The standard textbooks for time series analysis often place the primary focus on how to account for the relevant factors, e.g., the Box-Jenkins methodology, which is based on the class of autoregressive moving average (ARMA) models and their extensions (Brockwell and Davis, 1987; Hamilton, 1994; Von Storch and Zwiers, 2001; Lütkepohl, 2005; Chandler and Scott, 2011; Durbin and Koopman, 2012; Box et al., 2015), and is (mostly) built in the class of OLS/GLS models. Whereas several different robust techniques have been proposed in parallel by other schools of thought in the statistical community, we can now combine the autocorrelation and covariates into a more advanced technique that is resistant to the impact of outliers and the non-normally distributed error term, instead of relying on the GLS models (which are less resistant to these impacts).

In the meantime, some fields of environmental research have developed different opinions regarding the calculation and hypothesis test of trends, such as the application of the slope from the Sen-Theil method and the p-value from the Mann-Kendall test in water quality research (Hirsch et al., 1982; Gilbert, 1987; Helsel and Hirsch, 2002). The classical nonparametric approaches have not been adopted within the realm of formal statistical education (e.g. the references listed above), not only because these approaches cannot incorporate the relevant factors naturally, but also because they treat the data samples in a rather wasteful way. Even though these approaches are not affected by extreme values, ignoring extreme values implies that a portion of the data will have no influence on the trend estimator. To acknowledge the value of all observations (including the sampling frequency and temporal coverage behind it), we do not recommend the Sen-Theil or Siegel's estimators, because they automatically ignore up to 29% or 50% of the data without even checking to see if those data are actually outliers. If such a large portion of data is presumed to be problematic, data quality control should be performed before making any attempt at trend analysis.

Based on the above arguments, the regression-based methods provide an unparalleled advantage over classical nonparametric approaches, because their capabilities are designed to continually evolve as analysts tackle more complex and larger datasets than ever before, facilitated by inexpensive modern computer resources. In addition to the Box-Jenkins methodology used to deal with autocorrelation, and harmonic functions used to deal with repeated seasonal patterns, several useful extensions are available:

- The identification of a change point of the trends is an important topic, especially if there are known factors or interventions which could induce a change of trends in the time series. Typically, a meaningful trend detection of an atmospheric time series requires at least a few decades of data (Weatherhead et al., 1998), so in general we do not expect the actual trends to be highly nonlinear. When a turnaround of trends or sub-seasonal patterns are required, we can extend a linear trend and a regular seasonal cycle, i.e. from $y_t = \beta_0 + \beta_1 t + \gamma \sin(2\pi \frac{\text{Month}}{12}) + \eta \cos(2\pi \frac{\text{Month}}{12}) + N_t$, to a combination of piecewise trends and higher frequencies of harmonic functions as follows:

$$y_t = [\beta_0 + \beta_1 t + \beta_2 \max(t - t_c, 0)] + \left[\sum_{q=1}^{Q} \gamma_q \sin(2\pi \frac{q\text{Month}}{12}) + \eta_q \cos(2\pi \frac{q\text{Month}}{12})\right] + N_t, \tag{3}$$

  where β_2 is an adjustment of trends after a change point occurred at a time

t_c, and Q controls the frequency of harmonic functions. Examples for analyzing such problems are provided by previous studies (Reinsel et al., 2002; 2005). In addition to a piecewise linear function, we could directly specify regression spline functions (analogous to a seasonal-trend decomposition by the Loess smoother) to represent the non-linear trends, without assuming any form of nonlinearity in advance (e.g., polynomials) (Wood, 2006).

- Until now, the focus has only been placed on the trend detection of a single (aggregated) time series. However, analysis of an ensemble of multiple correlated time series at the same time is also desirable in some cases, e.g., data in close proximity are commonly more similar, and a measure to borrow this similarity can often offer a better quantification of ensemble trends and their associated uncertainty (Park et al., 2013; Chang et al., 2020). Whereas the relationship between different time series can be highly nonlinear or very complex (e.g. spatial variability), the class of generalized additive models (GAM, Hastie and Tibshirani, 1990; Wood, 2006) allows incorporation of spatial variability and complex interactions as covariates in the trend model (Augustin et al., 2009; Chang et al., 2017; Wood et al., 2017). In a situation where we have a collection of time series from multiple sites in meaningful spatial proximity, such as the ozone monitoring network across the southwestern USA, we can also modify the trend model as:

$$\text{obs} = \text{trend} + \text{seasonal cycle} + \text{spatial inhomogeneities} + \text{error},$$

  in order to account for potential spatial inhomogeneities (see Section 6). Therefore, this type of modeling approach can be very flexible. Since a large amount of parametrization is usually required to capture the potential spatial variability or any other nonlinear relationships (which can be estimated by a linear combination of various basis functions), the regularization to avoid overfitting is a built-in routine for the model fitting of GAM (Wood, 2006).
- The trend estimation can be made for either mean or specific quantiles (Koenker and Hallock, 2001; Fasiolo et al., 2020), including a single time series or multiple time series from a monitoring network.

These extensions make the regression-based methods more efficient and satisfactory than the traditional nonparametric approaches.

Before selecting a trend detection technique based solely on its basic description, we emphasize the importance of examining which uncertainties have been taken into account by the different techniques. Regression based methods are often considered to be passive learning tools, that can only handle the specific task specified by the model formulation, and nothing more. For example, a regression model can handle seasonality and autocorrelation only if we explicitly specify these issues in the model formulation. Therefore, if additional forcings, such as atmospheric circulations or meteorological conditions, are considered to be critical to identify the trend and its uncertainty, they should be specified in the models (techniques cannot be a surrogate for these covariates). It is always good practice to inspect residuals for any "suspicious patterns" and use this information to adapt the statistical model if necessary (e.g. Guillas et al. (2006)).

The above discussion has two key messages for the analyst: 1) specification of the trend model, e.g. identification of relevant covariates, should be motivated by the scientific question to be addressed; the techniques only help us with improving quantification of trends and their uncertainty; and 2) we should not judge a method and its result only by its name or basic descriptions; application of advanced techniques, e.g. implementation of overfitting prevention through the GAM, does not mean that autocorrelation or any other important factors relevant to trend detection have been taken into account. Instead, evaluation should be made by carefully inspecting any factors that are deemed important for the trend analysis.

3.4. Using signal-to-noise ratio to assess the robustness of trends

To assess the uncertainty of the trend estimate, in the past, a common rule for trend detection has been to label a trend as "statistically significant" if the magnitude of the estimated trend is greater than two standard errors from zero, which corresponds to a p-value less than a threshold of 0.05. If a trend did not pass this test then it was labeled as "statistically insignificant". Recent recommendations have called for abandonment of the phrase "statistical significance", e.g. Amrhein et al. (2019); Tarran (2019), supported by the special issue "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$" in the peer-reviewed journal, The American Statistician (`https://www.tandfonline.com/toc/utas20/73/sup1#`). This recommendation is based on the fundamental concept that "statistical significance was never meant to imply scientific importance" (Wasserstein et al., 2019) and "scientific conclusions [...] should not be based only on whether a p-value passes a specific threshold" (Wasserstein and Lazar, 2016).

The advice from Wasserstein et al. (2019) is to abandon the use of the phrase, "statistically significant" and simply report the p-value for all trend calculations; any conclusion that a trend is scientifically meaningful should be accompanied by a thoughtful evaluation and discussion of the data. Wasserstein et al. (2019) also recommend that researchers consider using alternate statistical methods to replace or supplement p-values. Following this advice we consider the signal-to-noise ratio (SNR, i.e. the ratio between the magnitude of the trend and its sigma uncertainty (standard error)) in addition to slope, confidence interval and p-value when evaluating a trend (see Section 6). This method allows us to distinguish a strong trend with a low uncertainty (i.e. a higher ratio) from a strong trend with a high uncertainty (i.e. a lower ratio). A higher SNR indicates stronger confidence in the resulting trend detection. Likewise, one could imagine a situation in which greater confidence is placed in a trend with low magnitude but very low uncertainty (high ratio), compared to a trend with high magnitude and high uncertainty (low ratio).

4. Quantifying autocorrelation and uncertainty in different chemical species

4.1. Quantifying autocorrelation

We continue our exploration of the data characteristics presented in Figure 1 by first examining the autocorrelations in different trace gases measured at Mauna Loa Observatory and reported as monthly means; we then compare various fits to those time series. Figure 2 shows the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the different trace gases based on monthly means (after deseasonalization). ACF finds the correlation of any time series with its lagged values (i.e. the correlation is 1 at lag 0 by definition, and decreases afterwards). PACF finds the correlation after excluding the variations that can be explained by the previous lag(s), and therefore PACF plots typically have a spike at lag-1, which indicates a large portion of the higher-order autocorrelations can be explained or represented by the lag-1 correlation. Except for CO and ozone, the other gases have a slow decay ACF, but have a single spike at lag-1 in the PACF. The presence of such a spike suggests that autocorrelation persists for a period of time (over 24 lags or months in the figure), but this behavior can be well represented by an AR(1) process. The ACF for CO and ozone reveals a substantial drop, in contrast to other trace gases. PACF shows a different pattern for CO and ozone, with CO having an oscillation between positive and negative numbers in the first 6 lags. This oscillation indicates that considerable seasonal variations remain after deseasonalization, requiring a more complex component of covariate(s) or error structure to account for the sub-seasonality. Ozone shows weak lag-2 correlation in PACF: the following will examine the impact of various lags of the autoregressive model on the trends and their uncertainty.

The first part of Table 1 reports the fitted trend value and 2-sigma uncertainty for CO and ozone by various lags of the autocorrelation process (with a regular seasonal cycle and a single linear trend included in the model, i.e. the M1 model setting discussed in the next paragraph). From a statistical point of view, the OLS estimate of the trend value (which does not account for autocorrelation) remains unbiased in the presence of autocorrelation (i.e. the estimate is not systematically different from the truth). However, the autocorrelation does result in underestimated uncertainty for the OLS, with the OLS estimators having the lowest uncertainty for both ozone and CO. Due to a stronger autocorrelation in the CO time series, the magnitude of increased uncertainty is also larger than that for ozone. In terms of the model comparisons by R^2 and mean-square error, no substantial improvement was found by increasing the lag of autoregressive process for either CO or ozone (not shown), but the maximal signal-to-noise ratio is achieved by AR(2) process for CO and AR(1) process for ozone.

4.2. Exploring different complexities of data characteristics

To illustrate the different levels of complexity between the trace gas time series, we further compare several model specifications listed as follows (we do not show the results for N_2O and SF_6):

M1: **fixed seasonality** + linear trend,

M2: **fixed seasonality** + **nonlinear trend**,

M3: **fixed seasonality** + **nonlinear trend** + varying seasonality,

M4: **fixed seasonality** + **nonlinear trend** + **varying seasonality**,

where the bold fonts indicate that regularization has been applied to this component. Whereas the seasonal cycle is essential in time series modeling, different approaches to estimate this term do not have a noticeable impact on the results of the estimates of the other terms in the model (Weatherhead et al., 1998), including the regularization. The varying seasonality component essentially represents the short-term variability (with respect to the long-term trend). Trend detections based on similar decompositions of a time series can be commonly found in the literature (e.g. Boleti et al. (2018; 2020)). These equations are built hierarchically by changing or adding a single component only. Since the variations for CO_2 and methane are relatively steady, a simple approach is expected to capture the most variability. The fitted results from models M1 and M2 for CO_2 and methane are shown in Figure 3. Even though the CO_2 record shows a slight departure from the straight line, it highlights the potential acceleration of increase in recent years. The distinction between linear and nonlinear fits (specified by the penalized cubic regression splines) is more obvious for the methane record due to a pause of trends in the early 2000s. Both trends increase monotonically and show no sign of turnaround, and therefore we might conclude that the linear approximation of the methane trend (54.8 [$\pm$5.9] ppb per decade over the period 1983-2019) provides an adequate description of the trend, with a proviso that a leveling-off period occurred in the early 2000s, and thus the rates of increase in the other periods are higher than the average.

Note that the variabilities of CO_2 and methane are considered to be less variable, not only due to a lack of complex interannual variations, but also because the magnitudes of the trends are much stronger than their seasonality. When applying models M1 and M2 to the CO and ozone records (upper panel of Figures 4 and 5), we see even with the nonlinear trend, the fitted results cannot adequately capture the seasonal peaks and troughs which show large departures from the regular seasonal cycle. Therefore, the next step is to investigate if further curve-fitting techniques, such as varying seasonality over time (Ambrosino and Chandler, 2013), can improve the quality of the fit and, more importantly, the trend detection. However, it is meaningless to pursue a perfect fit without proper scientific interpretations of the model specifications. To avoid overfitting, we illustrate the fits without and with regularization for the varying seasonality (from models M3 and M4, respectively).

The effect of regularization is displayed in the lower panels of Figures 4 and 5: the fit from M3 indeed captures many peaks and troughs, but only minor differences can be seen from the fits between M2 and M4, especially for the trend component. Therefore, certain information metrics are needed for quantifying those model fits. We use three metrics to assess the quality of the fit: 1) R^2: coefficient of determination; 2) MSE (mean-square error): the overall mean squared residual between model fitted and observed values; 3) GCV (generalized cross validation): the mean squared error in a leave-one-out test. Lower MSE and GCV indicate a better fit. However, a low MSE accompanied with a high GCV often indicates severe overfitting, because it implies when we randomly remove one data point and refit the model under the same setting, the new model will have a very poor performance when predicting this training point (also known as poor generalizability).

The second part of Table 1 reports these three metrics for CO and ozone (the models M5-M8 will be discussed later in this section): 1) the fit from M2 is better than M1 for all metrics, because the general nonlinearity is taken into account, even though the level of nonlinearity looks minor for both CO and ozone; 2) the fit from M3 shows a substantial improvement on R^2 and MSE over other models, but also has the worst GCV score, which implies severe overfitting as discussed above; 3) the only difference between M3 and M4 is the application of regularization on the varying seasonality term, with M4 achieving a good balance between fidelity (low MSE) and complexity (low GCV). Overall, M4 is the best model in terms of curve fitting and representing the underlying process, it does not necessarily have a strong impact on the trend estimate. In this case a linear approximation of the trends (e.g. from M1) seems to be adequate, even though it might leave plenty of room for improvement.

4.3. Incorporating meteorological covariate(s)

At this stage, we only consider the very basic components that are relevant to trend detection (i.e. autocorrelation and seasonality), and the novel technique to capture the irregular seasonality. However, there is also a different approach to improve the model predictability: incorporation of relevant covariates (e.g., meteorological variables). There is a clear physical basis for taking this approach as previous work has shown correlation between ozone and temperature (Rasmussen et al., 2012; Pusede et al., 2015), and for the specific case of Mauna Loa, ozone trends have been shown to differ between dry and moist air masses (Gaudel et al., 2018). Instead of using a varying seasonality component (i.e. M4) to account for the irregular part of the time series, we further specify different models that extend from M2 via:

M5: **fixed seasonality** + **nonlinear trend** + dewpoint,

M6: **fixed seasonality** + **nonlinear trend** + relative humidity,

M7: **fixed seasonality** + **nonlinear trend** + temperature,

M8: **fixed seasonality** + **nonlinear trend** + dewpoint + relative humidity + temperature.

In this example the regularization aims to avoid overfitting by functional components (e.g., nonlinear trends and seasonality: these terms are approximated by the spline functions), thus we do not apply the regularization to the linear term (i.e., the correlation between ozone and a meteorological variable is only measured by a single regression coefficient). Determination of the best (sub)set of covariates is also known as the variable selection, the conventional approach relies on the statistical significance and p-value of a given regression coefficient, or relies on the Lasso technique to directly rule out the unimportant covariate(s) (however, for an illustrative purpose, we do not adopt this approach here). Since we do not use the p-value as the sole piece of evidence for evaluating a trend (Wasserstein and Lazar, 2016), we use the same metrics listed above to assess the model fits. From the statistics in Table 1, the dewpoint (M5) is the most important variable to explain ozone variability at MLO (Gaudel et al., 2018), followed by relative humidity (M6) and temperature (M7). Once the dewpoint is accounted for, the inclusion of one or two additional covariates (e.g. M8) does not substantially improve the model fit.

The fitted result of M5 (dewpoint) is shown in the upper panel of Figure 6, revealing substantial improvement with respect to M2 in Figure 5. More importantly, with the meteorological adjustment, the nonlinear component from M2 is almost degenerated to a line by the regularization, which indicates that a consideration of nonlinearity is not required in this case. From the summary ozone statistics in the second part of Table 1, we can see that inclusion of dewpoint as a covariate reveals an almost linear trend, and produces a lower GCV score than either the nonlinear fit from M2 or the more complicated numerical optimization from M4.

The lower panel of Figure 6 also compares the residual series from M2, M4 and M5; except for an overlapping single spike in the late 1970s, we can see a similar error pattern between M4 and M5, thus the complex approach from M4 might have detected the signal of meteorological phenomena. Therefore, inclusion of essential covariates is the key to improving model predictability, rather than searching for a numerical method that may not be meaningful from a physical or scientific perspective. Nevertheless, if the essential covariates are unknown, the novel technique might be useful to identify potential signals out of the residuals.

To quantitatively summarize the trends, we replace the nonlinear components in M5-M8 with linear trends and report the results in the third part of Table 1 (using AR(2) process for CO and AR(1) process for ozone). For CO, M5 outperforms M1 in terms of R^2, MSE and GCV, but the trend estimate and uncertainty are almost identical; whereas under the same circumstance for ozone, the trend uncertainty is substantially reduced from M5 (i.e., incorporation of dewpoint variation). Therefore the trend detection and quantification are a rather complex problem (the method works for ozone, but it doesn't provide any advantage for CO).

This example shows that different levels of complexity influence trend detection of atmospheric time series, including: 1) the magnitude of autocorrelation could have a strong impact on the trend uncertainty; 2) trend detection is a different task from curve fitting, so pursuing a high R^2 value or a perfect fit through the numerical method is not the primary goal for trend detection. Also, a model selected by a single information metric (e.g., maximal R^2 or minimal MSE value) does not imply that the model is appropriate; and 3) the novel technique is only useful when we specify the appropriate model, requiring us to consider the model's implications for bad inference (fitting non-meaningful changing seasonality) and good inference (finding that nonlinearity of trends can be attributable to meteorological variability). We made this demonstration by showing that a linear fit (analogous to a GLS routine) with an appropriate model formulation can outperform the nonlinear fit with a complex numerical optimization (via a GAM framework).

In terms of trend detection, even though the linear trends in this section show a departure from zero at the 95% confidence level regardless of autocorrelation or covariates, this outcome is simply due to the signal being much stronger than the noise. Rather than limiting this analysis to just one or a handful of time series, which may result in an incomplete or biased view of the impact of autocorrelation, Appendix S4 in the supplementary material provides a demonstration of the impact of autocorrelation on short-term trend detection. The demonstration relies on 1,728 globally distributed time series based on monthly tropospheric column ozone values detected by the OMI/MLS satellite instruments (from October 2004 through December 2019) (Ziemke et al., 2019), and it clearly shows that substantial discrepancies arise when ignoring autocorrelation.

5. Using quantile regression to explain the changes in extreme events

The previous section showed that a perfect fit to a time series using a numerical method is not a solution for trend detection, rather relevant covariates might be the key for improving model predictive power. However, the complex variability of an atmospheric time series, such as ozone, can not always be attributable to specific factors, and can also be subject to measurement uncertainty. Whereas several trend detection techniques are able to describe the central tendency of a time series, usually represented by mean or median based slope estimates, consideration of changes in the extreme values (e.g. 5th or 95th percentiles) should also be a part of trend analysis, as the central and extreme tendencies are complementary components of an atmospheric time series (Simon et al., 2015; Gaudel et al., 2020). An effective method for quantifying trends across the range of observations (e.g. low, median and high values) is quantile regression.

As a demonstration of quantile regression we focus on long-term surface ozone time series from three remotely located monitoring sites (Cooper et al., 2020a;b): the coastal site of Mace Head, Ireland, the high elevation site of Mt. Waliguan in central China, and Schwarzwald-Sued in a low elevation forested region of southwestern Germany. These time series are at least 20 years in length (i.e., extend back in time to at least 1995), and are deseasonalized in order to focus on the irregular part of the time series (Cooper et al., 2020a). These three sites were selected because their central tendency is relatively linear (as illustrated by the Loess smoother), which facilitates the comparison of the change in extreme quantiles with respect to the central tendency. Note that the low and high percentile ozone trends at MLO are relatively consistent with the mean trends (with respect to the selected sites above), so the results are not shown here.

Figure 7 shows the monthly anomaly series from the three sites. To demonstrate the unique capability of quantile regression, we also fit several trend estimates from different techniques (and the Loess smoother for an indication of variability on shorter time scales). As described earlier in this paper, autocorrelation results in underestimated trend uncertainties but does not result in biased trend estimates (thus the lines from OLS and GLS are almost identical). Even though some trend estimates could be more sensitive to outliers or extreme values, with sufficiently long time series (and no aberrant outliers), most techniques yield similar trends, particularly those techniques that are designed to avoid the influence of outliers by using median slope estimates (e.g. Sen-Theil, Siegel, QR-50th and LMS), by removing the most extreme data (e.g. LTS), or by implementing regularization (e.g. Lasso and ridge regression). It should be noted that only the LMS estimator shows a visible difference from the other estimators at Mace Head and Mt. Waliguan, presumably because the LMS estimator can be unstable in response to small changes in the data (Hettmansperger and Sheather, 1992). Nevertheless, since all of these techniques aim to derive trends that are representative of the central tendency of the time series, none are suitable for the investigation of extreme events.

The quantile regression provides a natural extension to estimate the trend at any specific quantiles (in addition to the QR-50th for the median change in Figure 7). For example, we show the quantile trends and their uncertainty (accounting for autocorrelation) from the 5th to the 95th percentile for all three sites in Figure 8. The primary indication of these plots is that the changes in different percentiles can be inconsistent with the mean or median trends, especially for the extreme percentiles, thus it is desirable to include these estimators of extreme percentiles to convey our extended knowledge beyond the central tendency. The distribution of the quantile trends at Mace Head shows that the mean trend estimator is stronger than the median estimator, and consistently stronger than the estimators for all percentiles greater than the 40th percentile. Because the estimators for the 5th and 10th percentiles are stronger than the mean estimator, we can conclude that the positive mean trend is largely driven by the strong increases of the lower percentiles. Similarly, the increasing mean trend at Mt. Waliguan could be driven by the strong enhancements of the high percentiles (Lefohn et al., 2017), and the decreasing trend at Schwarzwald-Sued could be driven by the strong decline of high percentiles, although the uncertainty of the quantile trends mostly overlaps with the uncertainty of the mean trends.

In addition to the quantile linear trend analysis demonstrated above, we further show that the change point analysis can also be carried out by quantile regression (Equation (3), but only applied to deseasonalized anomalies). Figure 9 shows the ozone anomaly series measured at Zugspitze, Germany (47.4°N, 11.0°E, 2800 m). The primary feature of this time series is that it has a clear (overall) upward trend and a relatively steady trend before and after the late 1990s (Cooper et al., 2020a). We fit quantile piecewise trend models to the time series, and we can see how the changes vary at different quantiles. The largest turnaround can be found in the change of trend at the 5th percentile, and the upward trend at the 95th percentile since the late 1970s has paused. Nevertheless, the overall mean trend does not show a substantial decrease after 1997. This is another example that demonstrates how the statistical relationships can be explored through quantile regression.

6. Deriving common mean and quantile trends in multi-site data

Trend analysis of a collection of multiple time series has become a necessary task for scientific assessments nowadays, due to the availability of a variety of monitoring data from local to regional-scale networks. Such analysis has two main purposes: 1) compare trends from different locations, and 2) derive common trends within a network, to enable the comparison of trends between different networks.

A direct approach to achieve the first purpose is to fit a model to each time series independently, but in reality the lengths of the time series are often different and the spatial coverage of a network can change over time. In order to truncate the data to a (minimum) common period, a portion of data is often wasted. Also, this approach might not explore the full potential of available information. For example, none of the sites show a strong trend, but a high agreement of the trends is observed across all sites. Under this circumstance the small signal among all sites is expected to be representative. Therefore, a joint statistical inference of multiple sites is a better option to deliver a more reliable conclusion.

The irregular distribution of monitoring stations in space is an obvious reason that a common trend can not be derived properly and representatively by calculating a simple average. Given that urban surface ozone or other pollutants can be sensitive to localized emissions (e.g. traffic), the data variability and trends from neighboring locations might be different, which introduces additional spatial inconsistencies. Due to these inherent inhomogeneities, as well as the fact that a network can consist of hundreds or thousands of monitoring sites, approaches that do not account for spatial inhomogeneities will yield unreliable results.

The final goal of this paper is to describe methods for quantifying regional scale trends based on observations from large and widespread monitoring networks. For this demonstration, a collection of daily surface ozone time series from 168 monitoring stations across the southwestern United States (California, Nevada and Arizona) was downloaded from the Tropospheric Ozone Assessment Report (TOAR) database (Schultz et al., 2017; TOAR database, 2017). To reduce the complexity of the problem, we use all maximum daily 8-hour averages (MDA8) limited to the warm season (April to September) to derive the regional trends (i.e., around 183 data points per year for each station) over 2000-2014 using all 168 stations.

A preliminary data visualization is shown in Figure 10 by comparing the mean and quantile trends and their SNR values derived from each individual site. We can see that the pattern of the 95th percentile trends tends to be negative with strong SNR, and the magnitude of negative trends is reduced for the mean and the median MDA8 values, whereas both trends and SNR values for the 5th percentile are centered around zero. This figure illustrates why the multi-site trend analysis is complicated, due to the highly variable local trends. The first two rows of Figure 11 further display the regional 5th, 50th and 95th MDA8 distributions during 2000-2002 and 2012-2014 (several techniques are available for this type of analysis, see the study by Heaton et al. (2019); details are beyond the scope of this paper. Here we use Gaussian process approximation through the quantile GAM (Fasiolo et al., 2020)). Figures 10 and 11 show that a general reduction can be expected for the 50th and 95th percentiles over the study period, and the next step is to investigate the sub-regional variations and explicitly quantify the regional trends.

6.1. *Investigating sub-regional variations*

To compare the trends from different sub-regions, we further approximate the 5th, 50th and 95th regional MDA8 distributions on a $0.1^\circ \times 0.1^\circ$ grid covering the monitoring network for each year, and derive the trend estimate based on the GLS-AR1 model in each grid cell (we can also directly apply quantile regression to all MDA8 values; the result will be similar, but it requires much more computational power due to a far greater sample size). The results are shown in the third row of Figure 11: at the 95th percentile negative trends dominate across most of the region, at the 50th percentile the negative trends are of a lower magnitude and there are a few additional spots with positive trends, while the results are mixed across the region at the 5th percentile.

The map view of SNR for MDA8 ozone trends and uncertainties is shown in the fourth row of Figure 11 . When the ratio exceeds a value of ± 2, the signal of the trend is twice as large as the estimation uncertainty, which corresponds to a rejection of the null hypothesis at the 95% confidence level. A continuous scale of SNR allows us to gauge our confidence in a trend based on our tolerance for noise in the time series. For example, the largest magnitude of positive trends at the 5th percentile was found over the city of Bakersfield, but the highest SNR ratio over California was found in the Los Angeles region.

The above findings and discussion demonstrate that reporting SNR is a useful endeavor for providing additional information on the trend uncertainty (especially in a map view). It efficiently characterizes the quality of the trend estimation in an objective way, without further computation. Thus reporting SNR is an effective and intuitive alternative to providing a dichotomized statement of statistical significance based on a p-value threshold, since the uncertainty cannot be dichotomized.

6.2. Deriving overall regional trends

Deriving common trends from multi-site data requires the consideration of two additional challenges (Chandler and Scott, 2011): 1) data from neighbouring sites are likely to be correlated (but not necessarily with similar trends), and 2) each site might show a unique feature due to its geographical characteristics (e.g., degree of urbanization), thus the general statistical model for multi-site data can be written as:

$$\text{obs}(s,t) = \text{trend}(t) + \text{fixed spatial field}(s) + \text{varying spatial field}(s,t) + \text{error},$$

where the first component is the regional trend, the second component represents the purely spatial field (i.e., not varying with time), the third component represents the temporally varying spatial patterns (i.e., an interaction term), and the error term follows an AR(1) process. The second and third terms address the challenges pointed out above, respectively. Therefore, even though a single trend component is used to represent the common signal regionally, the interaction term allows some deviations to the regional trends from each individual station (adjustments are made to the individual trend against the regional trend for each station). The fixed spatial field is specified through the same GAM setting described in the last section, and the varying spatial field is represented by the station-specific variations using a factor smoothing technique (without actually implementing the full spatial interpolation for each year (Chang et al., 2017; Pedersen et al., 2019)).

The upper panel of Figure 12 shows the regional trends corresponding to the mean, 5th, 50th and 95th percentiles (their values are reported in Table 3). If we simply assume all sites are independent, and calculate the regionally pooled trend estimate and standard error (by calculating an independent trend and uncertainty for reach site, and then simply taking the mean and pooled standard error, i.e., if $\sigma_{SE}(i)$ is the standard error of the fitted trend at site i, then the pooled standard error is $\sqrt{\sum_{i=1}^{n} \sigma_{SE}^2(i)/n}$), the regional mean trend and 2-sigma range will be -0.72 [$\pm$1.11] ppb per year. However, once we take into account inter-site correlations, the slope is less negative and uncertainty estimate is reduced substantially (-0.32 [$\pm$0.15]). Except for the spatial irregularity, this is also likely due to a well recognized phenomenon called preferential sampling (Diggle et al., 2010), e.g., an area with denser monitoring locations can be simply due to the fact that this area is more polluted and spatially dense measurements are desired to evaluate human exposure. Therefore a simple average of all individual trends results in biased regional trends (in this case, an overestimation of negative trends). We also observe that the magnitude of the decreasing rate in the 95th percentile is more than twice as great as the 50th percentile (and with a higher SNR). The regional trend for the 5th percentile is flat, as we expected from the result in the last section. A further demonstration is made by displaying the trend estimate for every 5th percentile (with the 1st and 99th percentiles also included) in the lower panel of Figure 12. With this amount of information, we see that the variations are transitioning smoothly from one percentile to the next (in contrast to the result from a single time series, see Figure 8) with no spike in variability, as expected.

6.3. Sensitivity of the regional trend to the sites with a stronger signal

The final experiment is devoted to a sensitivity and stability test regarding the impact of those sites with the strongest signal on the estimation of the regional trend. For the annual 95th, 50th and 5th percentile trends we sequentially removed the sites with p-values less than 0.01, 0.05 and 0.10, and refitted the statistical model to investigate the influence of the remaining sites on the regional trends. The result is shown in Figure 13: in each panel we first show the regional trend estimated using all available sites (dark red), then the resulting trend after removing the sites with p-values less than 0.01 (orange), 0.05 (light blue) and 0.10 (dark blue). The features of this plot can be summarized as follows: 1) At the 50th and 95th percentiles, since the removed sites had relatively strong negative trends, the magnitudes of the slopes of the regional trends are reduced with each iteration; 2) Even though the slopes have changed, the interannual variations remain very similar in each iteration, indicating that this statistical approach is very robust; 3) The degree to which the slopes decrease depends on the initial strength of the signal. For example, at the 95th percentile the slope is strong when all sites are used, thus the drop is also the strongest when sites are removed sequentially, but at the 5th percentile the slope is very weak from the outset, thus the result is insensitive to the removal of sites with the strongest signal (also because fewer sites are removed, see following comparison).

To assess the uncertainty of the sensitivity analysis, we provide the summary statistics for the further removal of sites according to the p-value in Table 3. For the 95th and 50th percentiles, the magnitude of trends decreases and the p-value increases with each iteration. The implication is that if the signal is strong enough (e.g. 95th percentile), we can still derive a clear regional trend even if 50% of the most representative sites are removed. For example when the individual sites with p-values less than 0.10 were removed from the analysis the remaining sites were only 38% of the original network but the regional trend clearly persisted (see the 95th percentile results in Table 3). This result is consistent with the discussion of p-values by Wasserstein et al. (2019) and demonstrates that a trend can still contain valuable information when the p-value exceeds a threshold of 0.05; this result is also consistent with the vector plot of trends and uncertainty demonstrated in the TOAR special issue (Gaudel et al., 2018; Fleming et al., 2018). In this example we have shown that an advanced modeling approach making full use of all available information enables us to properly quantify the mean and extreme quantile changes, and make robust statements about the regional variation, which is not possible when the analysis is limited to just one or a handful of sites.

7. Discussion of further advanced techniques

In the previous sections we demonstrated the trend detection of single time series by various trends techniques and of multi-site data based on the GAMs. These techniques are chosen not only because their systematic and flexible formulations allow for extensions (e.g. from linear to non-linear trends or from single time series to multi-site data), but also because their programming languages have similar syntax (see supplementary code). However, advancements in trend detection techniques are continuously evolving, and several additional developments are available and can be applied to achieve differing but appropriate perspectives.

As discussed and demonstrated previously, trends in extreme events of atmospheric compositions are of great interest. Quantile regression is a straightforward approach for practitioners since it shares similar theoretical background and implementation as traditional regression models. Other perspectives are through 1) bootstrap-based approaches (Gilleland, 2020), and 2) approaches based on the generalized extreme value (GEV) or threshold exceedance (e.g. generalized Pareto) models (Berrocal et al., 2014; Stein, 2017; Opitz et al., 2018). Bootstrap is a resampling procedure that can be used for estimating the sampling distribution about the trends and/or their uncertainty. This technique is also known for its ability to mitigate the violation of normality assumption, and for being robust to autocorrelation and heteroskedasticity in the errors (Politis and White, 2004; Gardiner et al., 2008; Noguchi et al., 2011; Friedrich et al., 2020a;b). Bootstrap-based approaches are commonly adopted by practitioners due to their simplicity. In contrast, the GEV or generalized Pareto models currently receive less attention because they involve greater mathematical complexity and require some advanced knowledge in probability theory.

In this paper we adopt the Loess or smoothing spline to capture the nonlinearity of the trends, but several other approaches are also possible. Except for simple situations, such as a turnaround or a leveling-off of the trend, it is generally difficult to interpret highly nonlinear behavior through an explicit parametric representation or a deterministic model (Chandler and Scott, 2011). Many adaptive nonlinear trend fitting techniques are available, such as state-space modeling (and its variant, dynamical linear modeling) (Petris et al., 2009; Durbin and Koopman, 2012; Laine et al., 2014), vector autoregressive modeling (Holt and Teräsvirta, 2020), empirical mode decomposition (Wu et al., 2007), signal filter technique (Thoning et al., 1989), the Gasser-Müller kernel smoothing (Gasser and Müller, 1984), the Kalman filter (Harvey, 1990; Ramos-Ibarra and Silva, 2020), and the Kolmogorov-Zurbenko filter (Rao et al., 1997; Yang and Zurbenko, 2010). It should be emphasized that even though the above techniques are able to capture the nonlinearity in the time series, not all the curve features can be considered to be a change point of the trends (see Figure 9).

Detection of change point(s) is an important topic that is only partially covered in this paper (see the review by Reeves et al. (2007)). Broadly speaking, change point analysis involves two considerations: 1) do we expect one or multiple change points? and 2) is the location of change point(s) known or unknown? These questions determine the complexity of the analysis. If the timing of a change point is expected (e.g. intervention takes effect), piecewise trends can be applied (see the example in Figure 9); if the number of change points and their locations are both unknown, some learning techniques can be applied for such identifications (Li and Lund, 2012; Fryzlewicz and Rao, 2014; Zuo et al., 2019). However, care should be taken when detection of trends and multiple change points is carried out simultaneously, since it is inappropriate to conclude a change of long-term trends based on a shorter time frame (e.g. near the beginning or end of the study record). Therefore, one should not use simple curve fitting techniques, such as polynomials, to perform change point analysis. Instead, a formal test of appropriateness and meaningfulness of change point is preferred (Friedrich et al., 2020a).

Finally, our demonstration on the analysis of multi-site data relies on a combination of trend detection and spatial modeling techniques, which account for irregularity of the spatial distribution of stations and potential spatio-temporal interactions. Under this framework, other spatial modeling approaches can serve as an alternative (Heaton et al., 2019). Additional approaches for deriving common trends from an ensemble of time series include: 1) co-integration analysis that investigates whether the average differences between two or more time series remain relatively invariant over time (Engle and Granger, 1987; Johansen, 1988; Pfaff, 2008); (2) principal component analysis that extracts as much of the data variability as possible (Estrada and Perron, 2017); and (3) rolling window regression that mitigates the biases resulting from time series with different lengths or mild instances of missing observations (Lang et al., 2019).

8. Conclusions

This paper gives an overview of current statistical knowledge for atmospheric composition trend detection and analysis. We make a distinction between the numerical optimizations (behind the statistical methods) applied to trend estimation and the scientifically relevant factors that should be considered when stating a level of confidence for trend detection. Techniques alone are not the spirit of trend detection, but are supporting tools that help us to tackle the numerical issues, such as the influence of outliers, non-normally distributed residuals, or the risk of overfitting. Beyond the basic and indispensable components for the trend detection (e.g., autocorrelation and seasonality), we also show that an appropriate model formulation with simple GLS routines can outperform a model fitted by complex numerical optimization via a GAM framework. Therefore, the technique itself cannot be used as a replacement for the essential covariates in the trend model (or used as justification for taking them into account).

Note that the above statement is limited to trend detection of a time series. If the analysis problem involves any sort of prediction (e.g. predicting ozone at unobserved locations or forecasting ozone levels), the application of novel techniques, such as machine learning techniques, remains a promising approach (Kleinert et al., 2021; Leufen et al., 2021).

Decades ago robust statistics based on median values were developed for minimizing the impact of aberrant outliers in the data (i.e., assuming the worst case scenario), the cause of which are beyond the experience or knowledge of the data analyst. However, today those aberrant outliers can now be tracked and ruled out by quality control and database management methods (Schultz et al., 2017), and therefore the problem of aberrant outliers is hardly an issue any more (but the identification of possible anomalies is still one of the most challenging problems for the research community (Foorthuis, 2021)). Under the circumstance that the aberrant outliers are removed and the data record is sufficiently long, most techniques can describe the central tendency properly and give similar trend estimators (either mean or median based estimator), but this also implies these estimations cannot be used to represent the change of the extreme events. When data are distributed remotely from other points, but believed to be valid observations (e.g. part of natural variability), conventional regression models may have difficulty addressing this extreme data variability. Alternatively, we can seek to investigate the changes of the extreme events, with quantile regression being a natural solution to provide this estimation. In this paper we illustrate how the analysis of extreme quantile changes can provide additional insight to the mean or median based estimators, and can reveal the impact of the extreme events on the central tendency of the trend.

Based on our comparison of trend detection methods, the classical nonparametric methods (i.e. Sen-Theil and Siegel's repeated medians) are not recommended for routine use, because even though the aberrant outliers (and erroneous data) are ruled out, these estimators still treat the remaining extreme values as outliers which are omitted from the trend estimation. Instead the following methods are preferable as they account for as much information and data variability as possible in an objective way. To accommodate the possibility of autocorrelation and covariates, the class of GLS models remains a good foundation and flexibility for incorporating different sources of uncertainty and different advanced modeling approaches, such as the basis function representation of complex functional form in GAM. In addition to the central tendency of time series represented by the GLS estimator, quantile regression also provides insight regarding the extreme quantiles, which can have very different trends compared to the median or mean trend, and maintains the flexibility for incorporating autocorrelation and covariates into the models. These recommendations are made because this set of techniques can be learnt under the similar statistical framework and can therefore be extended to address additional complexities with less effort. However, other approaches discussed in Section 7 might also be appropriate, as long as the relevant factors are properly accounted for.

We used a collection of multiple surface ozone time series in the southwestern United States to illustrate a regional-scale assessment of trends, based on both the regional mean and quantile trends. Analyzing a large data set with hundreds or thousands of monitoring sites simultaneously is a common challenge in the atmospheric sciences. The information in each station can be thought of as a piece of a puzzle, some are informative, and some are ambiguous, but if we can put the pieces together into a bigger picture, the volume of information will be maximized, and the result will be compelling.

References

Aggarwal CC. 2015. *Outlier analysis (2nd ed.)*. Springer.

Ambrosino C, Chandler RE. 2013. A nonparametric approach to the removal of documented inhomogeneities in climate time series. *J Appl Meteorol Clim* **52**(5): 1139–1146. doi: 10.1175/JAMC-D-12-0166.1.

Amrhein V, Greenland S, McShane B. 2019. Scientists rise up against statistical significance. *Nature* **567**: 305–307. doi:10.1038/d41586-019-00857-9.

Augustin NH, Musio M, von Wilpert K, Kublin E, Wood SN, et al. 2009. Modeling spatiotemporal forest health monitoring data. *J Am Stat Assoc* **104**(487): 899–911. doi:10.1198/jasa.2009.ap07058.

Barassi MR, Cole MA, Elliott RJR. 2011. The stochastic convergence of CO_2 emissions: a long memory approach. *Environ Resour Econ* **49**(3): 367–385. doi:10.1007/s10640-010-9437-7.

Berrocal VJ, Gelfand AE, Holland DM. 2014. Assessing exceedance of ozone standards: a space-time downscaler for fourth highest ozone concentrations. *Environmetrics* **25**(4): 279–291. doi:10.1002/env.2273.

Boleti E, Hueglin C, Grange SK, Prévôt AS, Takahama S. 2020. Temporal and spatial analysis of ozone concentrations in Europe based on timescale decomposition and a multi-clustering approach. *Atmos Chem Phys* **20**(14): 9051–9066. doi:10.5194/acp-20-9051-2020.

Boleti E, Hueglin C, Takahama S. 2018. Ozone time scale decomposition and trend assessment from surface observations in Switzerland. *Atmos Environ* **191**: 440–451. doi:10.1016/j.atmosenv.2018.07.039.

Box GEP, Jenkins GM, Reinsel GC, Ljung GM. 2015. *Time series analysis: forecasting and control (5th ed.)*. John Wiley & Sons.

Box GEP, Tiao GC. 1975. Intervention analysis with applications to economic and environmental problems. *J Am Stat Assoc* **70**(349): 70–79.

Brockwell PJ, Davis RA. 1987. *Time series: theory and methods*. Springer.

Camalier L, Cox W, Dolwick P. 2007. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmos Environ* **41**(33): 7127–7137. doi:10.1016/j.atmosenv.2007.04.061.

Chandler R, Scott M. 2011. *Statistical methods for trend detection and analysis in the environmental sciences*. John Wiley & Sons.

Chang KL, Cooper OR, Gaudel A, Petropavlovskikh I, Thouret V. 2020. Statistical regularization for trend detection: An integrated approach for detecting long-term trends from sparse tropospheric ozone profiles. *Atmos Chem Phys* **20**: 9915–9938. doi:10.5194/acp-20-9915-2020.

Chang KL, Petropavlovskikh I, Cooper OR, Schultz MG, Wang T. 2017. Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia. *Elem Sci Anth* **5**: p.50. doi:10.1525/elementa.243.

Chatfield C. 2000. *Time-series forecasting*. CRC press.

Cleveland RB, Cleveland WS, McRae JE, Terpenning I. 1990. STL: A seasonal-trend decomposition. *J Off Stat* **6**(1): 3–73.

Cooper OR, Schultz MG, Schröder S, Chang KL, Gaudel A, et al. 2020a. Multi-decadal surface ozone trends at globally distributed remote locations. *Elem Sci Anth* **8**: p.23. doi:10.1525/elementa.420.

Cooper OR, Schultz MG, Schröder S, Chang KL, Gaudel A, et al. 2020b. TOAR data collection of “Multi-decadal surface ozone trends at globally distributed remote locations”. doi:10.34730/e792cad833174ebcafd9f052711e5660.

Diggle PJ, Menezes R, Su Tl. 2010. Geostatistical inference under preferential sampling. *J Roy Stat Soc C* **59**(2): 191–232. doi:10.1111/j.1467-9876.2009.00701.x.

Dlugokencky EJ, Crotwell AM, Thoning KW, Mund JW. 2020. Atmospheric methane from quasi-continuous measurements at Barrow, Alaska and Mauna Loa, Hawaii, 1986-2019, Version: 2020-08. doi:10.15138/ve0c-be70.

Durbin J, Koopman SJ. 2012. *Time series analysis by state space methods*. Oxford University Press.

Engle RF, Granger CWJ. 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica* pp. 251–276.

Estrada F, Perron P. 2017. Extracting and analyzing the warming trend in global and hemispheric temperatures. *J Time Ser Anal* **38**(5): 711–732. doi:10.1111/jtsa.12246.

Fasiolo M, Wood SN, Zaffran M, Nedellec R, Goude Y. 2020. Fast calibrated additive quantile regression. *J Am Stat Assoc* pp. 1–11. doi:10.1080/01621459.2020.1725521.

Fleming ZL, Doherty RM, von Schneidemesser E, Malley CS, Cooper OR, et al. 2018. Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health. *Elem Sci Anth* **6**(12). doi:10.1525/elementa.291.

Foorthuis R. 2021. On the nature and types of anomalies: a review of deviations in data. *Int J Data Sci Anal* doi:10.1007/s41060-021-00265-1.

Friedrich M, Beutner E, Reuvers H, Smeekes S, Urbain JP, et al. 2020a. A statistical analysis of time trends in atmospheric ethane. *Clim Change* **162**(1): 105–125. doi:10.1007/s10584-020-02806-2.

Friedrich M, Smeekes S, Urbain JP. 2020b. Autoregressive wild bootstrap inference for nonparametric trends. *J Econom* **214**(1): 81–109. doi:10.1016/j.jeconom.2019.05.006.

Fryzlewicz P, Rao SS. 2014. Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. *J Roy Stat Soc Ser B* pp. 903–924. doi:10.1111/rssb.12054.

Gardiner T, Forbes A, Mazière Md, Vigouroux C, Mahieu E, et al. 2008. Trend analysis of greenhouse gases over Europe measured by a network of ground-based remote FTIR instruments. *Atmos Chem Phys* **8**(22): 6719–6727. doi:10.5194/acp-8-6719-2008.

Gasser T, Müller HG. 1984. Estimating regression functions and their derivatives by the kernel method. *Scand J Stat* pp. 171–185.

Gaudel A, Cooper OR, Ancellet G, Barret B, Boynard A, et al. 2018. Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation. *Elem Sci Anth* **6**(39). doi: 10.1525/elementa.273.

Gaudel A, Cooper OR, Chang KL, Bourgeois I, Ziemke JR, et al. 2020. Aircraft observations since the 1990s reveal increases of tropospheric ozone at multiple locations across the Northern Hemisphere. *Sci Adv* pp. aba8272. doi::10.1126/sciadv.aba8272.

Gilbert RO. 1987. *Statistical methods for environmental pollution monitoring*. John Wiley & Sons.

Gilleland E. 2020. Bootstrap methods for statistical inference. Part II: Extreme-value analysis. *J Atmos Ocean Technol* **37**(11): 2135–2144. doi:10.1175/JTECH-D-20-0070.1.

Guillas S, Tiao GC, Wuebbles DJ, Zubrow A. 2006. Statistical diagnostic and correction of a chemistry-transport model for the prediction of total column ozone. *Atmos Chem Phys* **6**(2): 525–537. doi:10.5194/acp-6-525-2006.

Hamilton JD. 1994. *Time series analysis*. Vol. 2. Princeton.

Harvey AC. 1990. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Hastie TJ, Tibshirani RJ. 1990. *Generalized additive models*. Vol. 43. CRC press.

Hawkins DM. 1980. *Identification of outliers*. Vol. 11. Springer.

Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, et al. 2019. A case study competition among methods for analyzing large spatial data. *J Agric Biol Environ Stat* **24**(3): 398–425. doi:10.1007/s13253-018-00348-w.

Hegerl G, Zwiers F. 2011. Use of models in detection and attribution of climate change. *Wiley Interdiscip Rev Clim Change* **2**(4): 570–591. doi:10.1002/wcc.121.

Helsel DR, Hirsch RM. 2002. *Statistical methods in water resources*. Vol. 323. US Geological Survey Reston, VA.

Hettmansperger TP, Sheather SJ. 1992. A cautionary note on the method of least median squares. *Am Stat* **46**(2): 79–83.

Hirsch RM, Slack JR, Smith RA. 1982. Techniques of trend analysis for monthly water quality data. *Water Resour Res* **18**(1): 107–121.

Hoerl AE, Kennard RW. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1): 55–67.

Holt MT, Teräsvirta T. 2020. Global hemispheric temperatures and co-shifting: A vector shifting-mean autoregressive analysis. *J Econom* **214**(1): 198–215. doi:10.1016/j.jeconom.2019.05.011.

Johansen S. 1988. Statistical analysis of cointegration vectors. *J Econ Dyn Control* **12**(2-3): 231–254.

Kleinert F, Leufen LH, Schultz MG. 2021. IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany. *Geosci Model Dev* **14**(1): 1–25. doi:10.5194/

gmd-14-1-2021.

Koenker R, Hallock KF. 2001. Quantile regression. *J Econ Perspect* **15**(4): 143–156. doi:10.1257/jep.15.4.143.

Koenker R, Zhao Q. 1996. Conditional quantile estimation and inference for ARCH models. *Econom Theory* **12**(5): 793–813.

Laine M, Latva-Pukkila N, Kyrölä E. 2014. Analysing time-varying trends in stratospheric ozone time series using the state space approach. *Atmos Chem Phys* **14**(18): 9707–9725. doi:10.5194/acp-14-9707-2014.

Lang PE, Carslaw DC, Moller SJ. 2019. A trend analysis approach for air quality network data. *Atmos Environ* **2**: 100030. doi:10.1016/j.aeaoa.2019.100030.

Lefohn AS, Malley CS, Simon H, Wells B, Xu X, et al. 2017. Responses of human health and vegetation exposure metrics to changes in ozone concentration distributions in the European Union, United States, and China. *Atmos Environ* **152**: 123–145. doi:10.1016/j.atmosenv.2016.12.025.

Leng C, Lin Y, Wahba G. 2006. A note on the lasso and related procedures in model selection. *Stat Sin* pp. 1273–1284.

Leufen LH, Kleinert F, Schultz MG. 2021. MLAir (v1.0) – a tool to enable fast and flexible machine learning on air data time series. *Geosci Model Dev* **14**(3): 1553–1574. doi:10.5194/gmd-14-1553-2021.

Li S, Lund R. 2012. Multiple changepoint detection via genetic algorithms. *J Clim* **25**(2): 674–686. doi:10.1175/2011JCLI4055.1.

Lütkepohl H. 2005. *New introduction to multiple time series analysis*. Springer.

Noguchi K, Gel YR, Duguay CR. 2011. Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data. *J Hydrol* **410**(3-4): 150–161. doi:10.1016/j.jhydrol.2011.09.008.

Oltmans SJ, Komhyr WD. 1986. Surface ozone distributions and variations from 1973–1984: Measurements at the NOAA Geophysical Monitoring for Climatic Change Baseline Observatories. *J Geophys Res Atmos* **91**(D4): 5229–5236.

Opitz T, Huser R, Bakka H, Rue H. 2018. INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21**(3): 441–462. doi:10.1007/s10687-018-0324-x.

Park A, Guillas S, Petropavlovskikh I. 2013. Trends in stratospheric ozone profiles using functional mixed models. *Atmos Chem Phys* **13**(22): 11473–11501. doi:10.5194/acp-13-11473-2013.

Pedersen EJ, Miller DL, Simpson GL, Ross N. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* **7**: e6876. doi:10.7717/peerj.6876.

Petris G, Petrone S, Campagnoli P. 2009. Dynamic linear models, in *Dynamic Linear Models with R*. Springer: pp. 31–84.

Pfaff B. 2008. *Analysis of integrated and cointegrated time series with R*. Springer.

Politis DN, White H. 2004. Automatic block-length selection for the dependent bootstrap. *Econom Rev* **23**(1): 53–70. doi:10.1081/ETC-120028836.

Pusede SE, Steiner AL, Cohen RC. 2015. Temperature and recent trends in the chemistry of continental surface ozone. *Chem Rev* **115**(10): 3898–3918. doi:10.1021/cr5006815.

R Core Team. 2020. R: A language and environment for statistical computing.

Ramos-Ibarra E, Silva E. 2020. Trend estimation and forecasting of atmospheric pollutants in the Mexico City Metropolitan Area through a non-parametric perspective. *Atmósfera* **33**(4):

401–420. doi:10.20937/atm.52757.

Rao ST, Zurbenko IG, Neagu R, Porter PS, Ku JY, et al. 1997. Space and time scales in ambient ozone data. *Bull Am Meteorol Soc* **78**(10): 2153–2166.

Rasmussen DJ, Fiore AM, Naik V, Horowitz LW, McGinnis SJ, et al. 2012. Surface ozone-temperature relationships in the eastern US: A monthly climatology for evaluating chemistry-climate models. *Atmos Environ* **47**: 142–153. doi:10.1016/j.atmosenv.2011.11.021.

Reeves J, Chen J, Wang XL, Lund R, Lu QQ. 2007. A review and comparison of changepoint detection techniques for climate data. *J Appl Meteorol Climatol* **46**(6): 900–915. doi:10.1175/JAM2493.1.

Reinsel GC, Miller AJ, Weatherhead EC, Flynn LE, Nagatani RM, et al. 2005. Trend analysis of total ozone data for turnaround and dynamical contributions. *J Geophys Res Atmos* **110**(D16). doi:10.1029/2004JD004662.

Reinsel GC, Weatherhead E, Tiao GC, Miller AJ, Nagatani RM, et al. 2002. On detection of turnaround and recovery in trend for ozone. *J Geophys Res Atmos* **107**(D10): ACH–1. doi:10.1029/2001JD000500.

Rousseeuw PJ. 1984. Least median of squares regression. *J Am Stat Assoc* **79**(388): 871–880.

Rousseeuw PJ. 1985. Multivariate estimation with high breakdown point. *Math Stat Appl* **8**(283-297): 37.

Saunois M, Emmons L, Lamarque JF, Tilmes S, Wespes C, et al. 2012. Impact of sampling frequency in the analysis of tropospheric ozone observations. *Atmos Chem Phys* **12**(15): 6757–6773. doi:10.5194/acp-12-6757-2012.

Schultz MG, Schröder S, Lyapina O, Cooper OR, Galbally I, et al. 2017. Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations. *Elem Sci Anth* **5**: p.51. doi:10.1525/elementa.244.

Schwartz J. 1994. Air pollution and daily mortality: a review and meta analysis. *Environ Res* **64**(1): 36–52.

Sen PK. 1968. Estimates of the regression coefficient based on Kendall's tau. *J Am Stat Assoc* **63**(324): 1379–1389.

Shumway RH, Stoffer DS. 2017. *Time series analysis and its applications: with R examples*. Springer.

Siegel AF. 1982. Robust regression using repeated medians. *Biometrika* **69**(1): 242–244.

Simon H, Reff A, Wells B, Xing J, Frank N. 2015. Ozone trends across the United States over a period of decreasing NOx and VOC emissions. *Environ Sci Tech* **49**(1): 186–195. doi:10.1021/es504514z.

Stein M. 2017. Should annual maximum temperatures follow a generalized extreme value distribution? *Biometrika* **104**(1): 1–16. doi:10.1093/biomet/asw070.

Stott PA, Gillett NP, Hegerl GC, Karoly DJ, Stone DA, et al. 2010. Detection and attribution of climate change: a regional perspective. *Wiley Interdiscip Rev Clim Change* **1**(2): 192–211. doi:10.1002/wcc.34.

Tarran B. 2019. Is this the end of "statistical significance"? *Significance* **16**(2): 4. doi:10.1111/j.1740-9713.2019.01244.x.

Theil H. 1950. A rank-invariant method of linear and polynomial regression analysis, I, II and III. *Nederl Akad Wetensch Proc* **53**: 386–392 & 521–525 & 1397–1412.

Thoning KW, Tans PP, Komhyr WD. 1989. Atmospheric carbon dioxide at Mauna Loa Observatory: 2. Analysis of the NOAA GMCC data, 1974–1985. *J Geophys Res Atmos* **94**(D6): 8549–8565.

Tiao GC, Reinsel GC, Xu D, Pedrick JH, Zhu X, et al. 1990. Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation. *J Geophys Res Atmos* **95**(D12): 20507–20517.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B* **58**(1): 267–288.

Tikhonov AN, Goncharsky AV, Stepanov VV, Yagola AG. 2013. *Numerical methods for the solution of ill-posed problems*. Vol. 328. Springer.

TOAR database. 2017. Tropospheric Ozone Assessment Report, links to Global surface ozone datasets. doi:10.1594/PANGAEA.876108.

Tong C. 2019. Statistical inference enables bad science; statistical thinking enables good science. *Am Stat* **73**(sup1): 246–261. doi:10.1080/00031305.2018.1518264.

Turner MC, Jerrett M, Pope III CA, Krewski D, Gapstur SM, et al. 2016. Long-term ozone exposure and mortality in a large prospective study. *Am J Respir Crit Care Med* **193**(10): 1134–1142. doi:10.1164/rccm.201508-1633OC.

Von Brömssen C, Fölster J, Futter M, McEwan K. 2018. Statistical models for evaluating suspected artefacts in long-term environmental monitoring data. *Environ Monit Assess* **190**(9): 558. doi:10.1007/s10661-018-6900-3.

Von Storch H, Zwiers FW. 2001. *Statistical analysis in climate research*. Cambridge University Press.

Wasserstein RL, Lazar NA. 2016. The ASA's statement on p-values: context, process, and purpose. *Am Stat* **70**(2): 129–133. doi:10.1080/00031305.2016.1154108.

Wasserstein RL, Schirm AL, Lazar NA. 2019. Moving to a world beyond "$p < 0.05$". *Am Stat* **73**(1): 1–19. doi:10.1080/00031305.2019.1583913.

Weatherhead EC, Bodeker GE, Fassò A, Chang KL, Lazo JK, et al. 2017a. Spatial coverage of monitoring networks: A climate observing system simulation experiment. *J Appl Meteorol Climatol* **56**(12): 3211–3228. doi:10.1175/JAMC-D-17-0040.1.

Weatherhead EC, Harder J, Araujo-Pradere EA, Bodeker G, English JM, et al. 2017b. How long do satellites need to overlap? Evaluation of climate data stability from overlapping satellite records. *Atmos Chem Phys* **17**(24): 15069–15093. doi:10.5194/acp-17-15069-2017.

Weatherhead EC, Reinsel GC, Tiao GC, Meng XL, Choi D, et al. 1998. Factors affecting the detection of trends: Statistical considerations and applications to environmental data. *J Geophys Res Atmos* **103**(D14): 17149–17161.

Wells B, Dolwick P, Eder B, Evangelista M, Foley K, et al. 2021. Improved estimation of trends in US ozone concentrations adjusted for interannual variability in meteorological conditions. *Atmos Environ* **248**: 118234. doi:10.1016/j.atmosenv.2021.118234.

Wilks DS. 2011. *Statistical methods in the atmospheric sciences (3rd ed.)*. Vol. 100. Academic press.

Wood SN. 2006. *Generalized additive models: an introduction with R*. CRC press.

Wood SN, Li Z, Shaddick G, Augustin NH. 2017. Generalized additive models for gigadata: modeling the UK black smoke network daily data. *J Am Stat Assoc* **112**(519): 1199–1210. doi: 10.1080/01621459.2016.1195744.

Wu Z, Huang NE, Long SR, Peng CK. 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proc Natl Acad Sci USA* **104**(38): 14889–14894. doi:10.1073/

pnas.0701020104.
Yang W, Zurbenko I. 2010. Kolmogorov–Zurbenko filters. *Wiley Interdiscip Rev Comput Stat* **2**(3): 340–351. doi:10.1002/wics.71.
Ziemke JR, Oman LD, Strode SA, Douglass AR, Olsen MA, et al. 2019. Trends in global tropospheric ozone inferred from a composite record of TOMS/OMI/MLS/OMPS satellite measurements and the MERRA-2 GMI simulation. *Atmos Chem Phys* **19**(5): 3257–3269. doi: 10.5194/acp-19-3257-2019.
Zuo B, Li J, Sun C, Zhou X. 2019. A new statistical method for detecting trend turning. *Theor Appl Climatol* **138**(1): 201–213. doi:10.1007/s00704-019-02817-9.

Contributions

Contributed to conception and design: KLC; Contributed to acquisition of data: MGS, XL, IP, XX and JRZ; Contributed to analysis and interpretation of data: all authors; Drafted and/or revised the article: KLC drafted the article while MGS, XL, IP, XX and JRZ helped with the revision; Approved the submitted and revised versions for publication: all authors.

Funding information

This work was supported in part by the NOAA Cooperative Agreement with CIRES, NA17OAR4320101. MGS acknowledges funding from ERC-2017-ADG 787576 (IntelliAQ) and the program Engineering digital futures of the Helmholtz Association's research field Information.

Acknowledgements

We would like to thank Betsy Weatherhead for inspiring this work, and Owen Cooper (University of Colorado Boulder and NOAA CSL) for helpful and thought-provoking suggestions which improved the content of this paper.

Competing interests

The authors declare no competing financial interests.

Data accessibility statement

The source of Mauna Loa Observatory data used in Section 4 is available at https://www.esrl.noaa.gov/gmd/dv/site/?stacode=MLO; The time series data used in Section 5 can be downloaded at http://doi.org/10.34730/e792cad833174ebcafd9f052711e5660; The TOAR data can be accessed at https://doi.org/10.1594/PANGAEA.876108

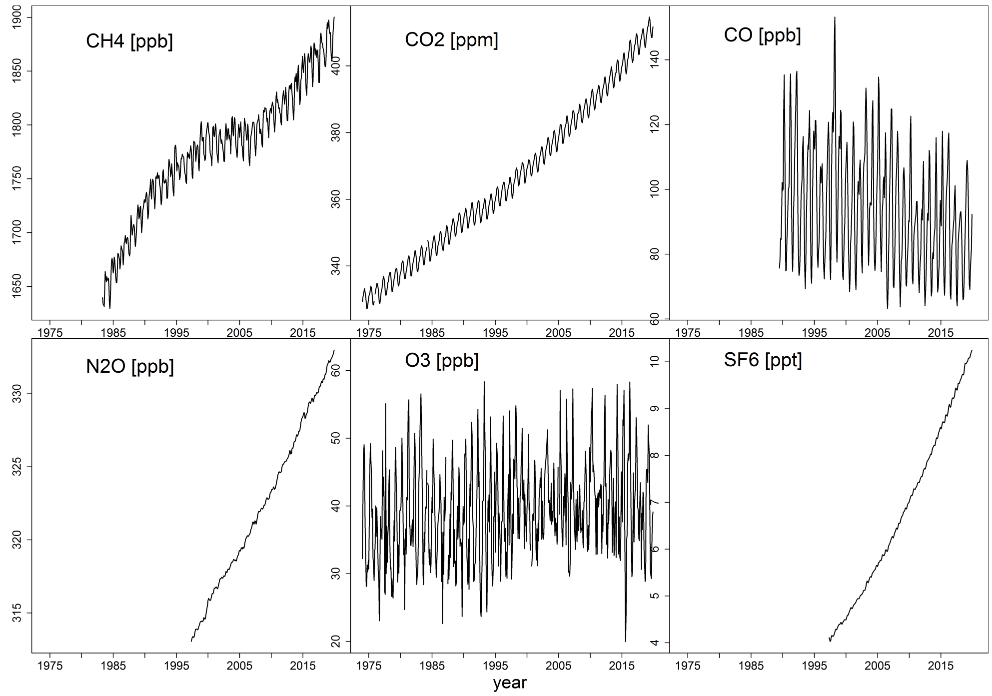

Figure 1: Monthly mean time series for different chemical species.

Trace gases are measured at Mauna Loa Observatory (MLO), Hawaii.

(Schultz et al., 2017; TOAR database, 2017); OMI/MLS satellite data are available for download at https://acd-ext.gsfc.nasa.gov/Data_services/cloud_slice/. All computations are implemented in R (R Core Team, 2020).

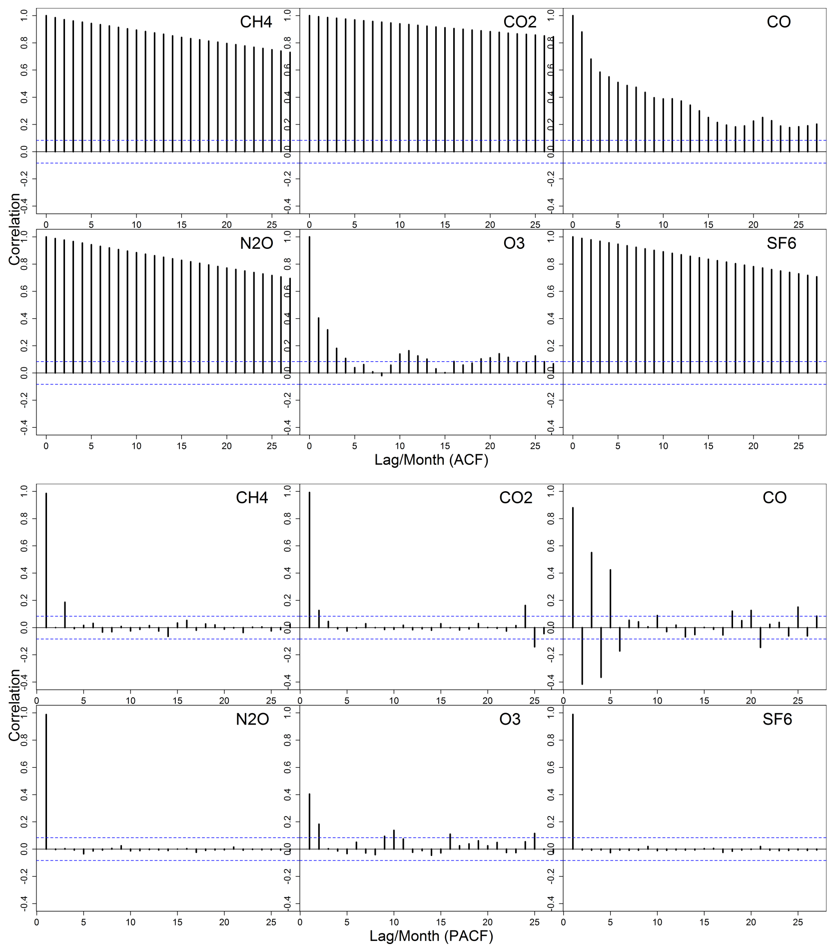

Figure 2: Autocorrelation function and partial autocorrelation function for different chemical species.

Trace gases are measured at Mauna Loa Observatory (MLO), Hawaii (after deseasonalization).

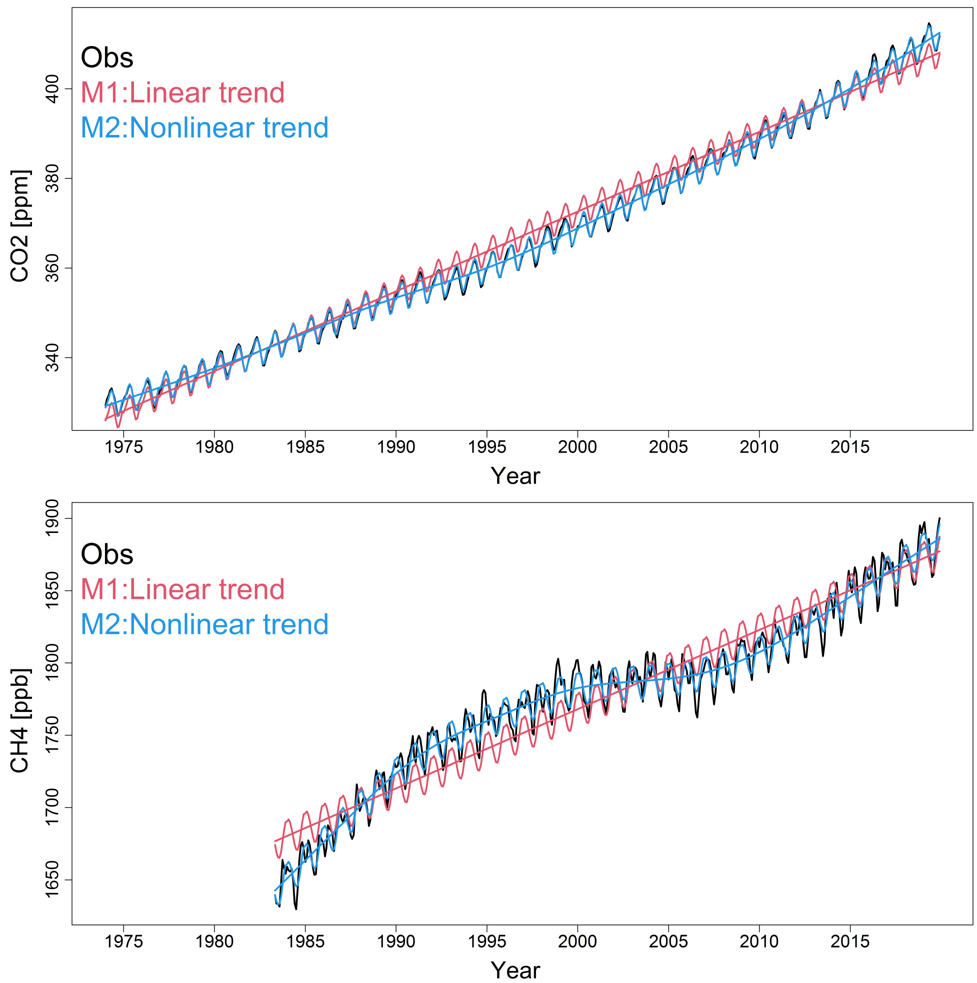

Figure 3: Linear and nonlinear fits to the CO_2 and methane time series at MLO.

The smooth curve or straight line is the trend component extracted from the full model fit.

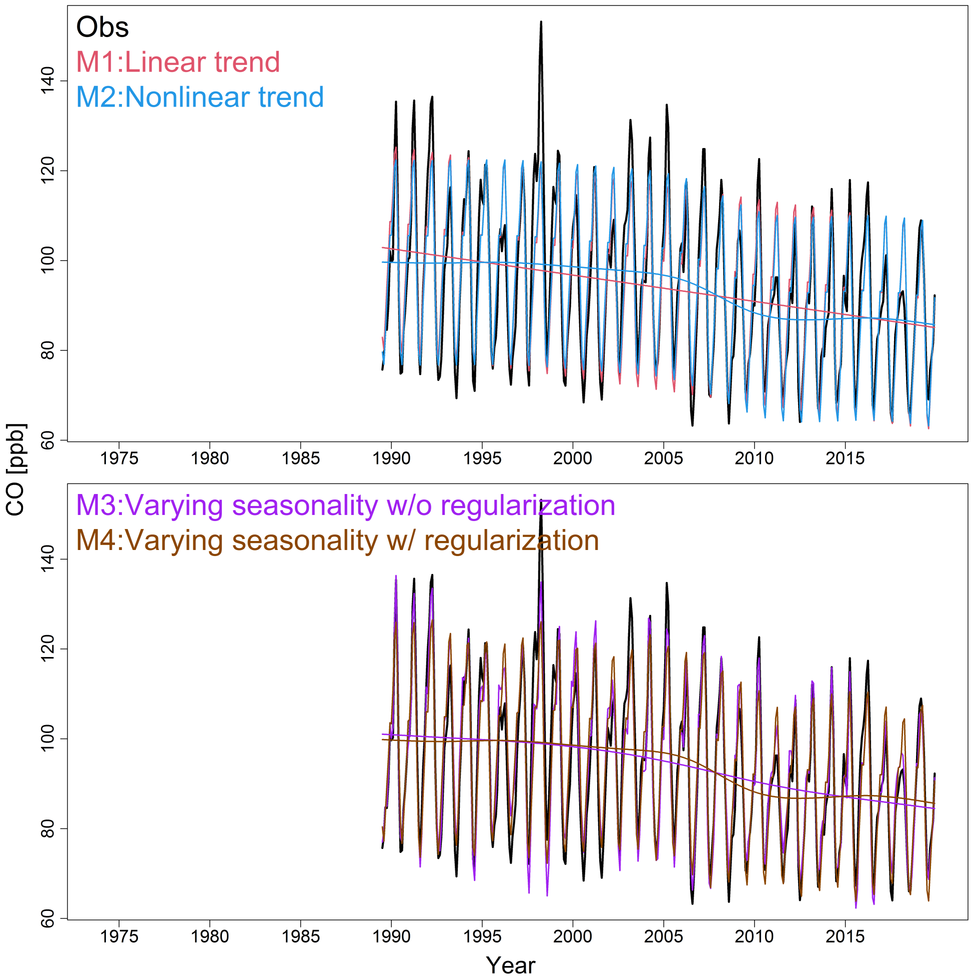

Figure 4: Various fits to the CO time series at MLO.

Various models include (top) linear and nonlinear trends, and (bottom) an additional varying seasonality component with and without regularization. The smooth curve or straight line is the trend component extracted from the full model fit.

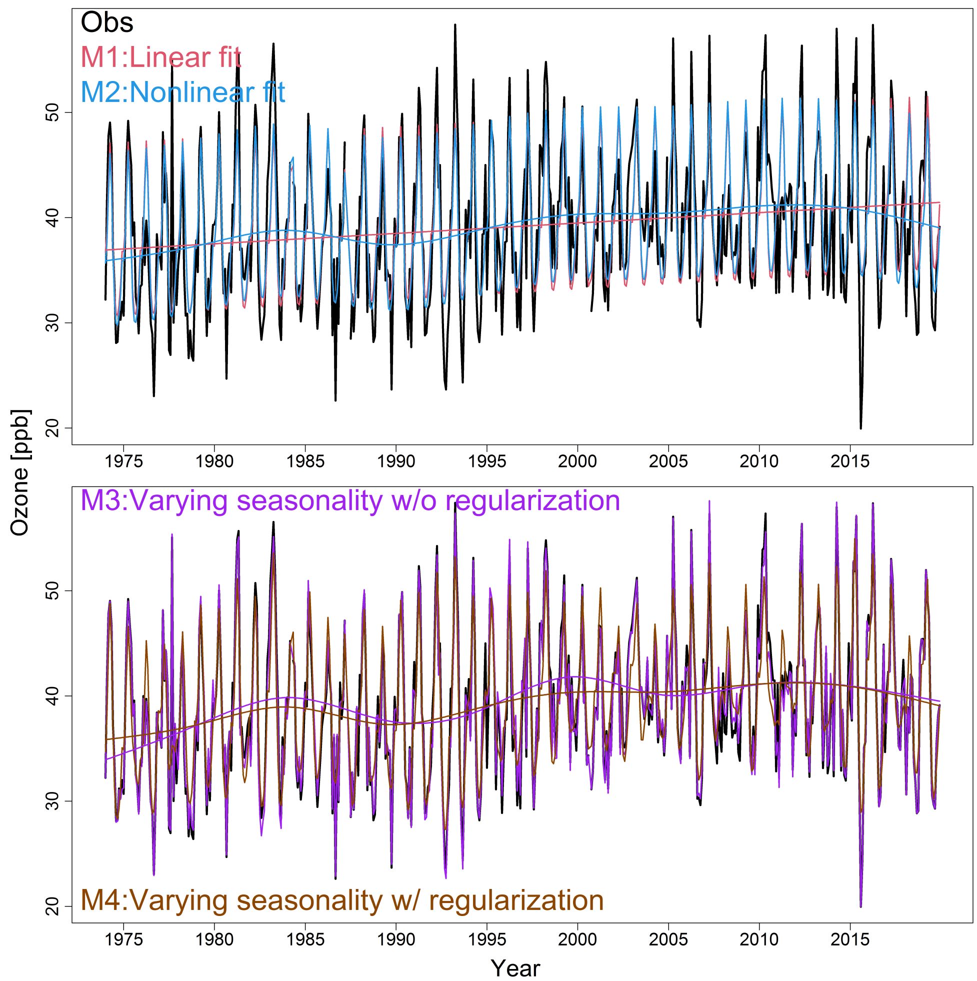

Figure 5: Various fits to the ozone time series at MLO.

Various models include (top) linear and nonlinear trends, and (bottom) an additional varying seasonality component with and without regularization. The smooth curve or straight line is the trend component extracted from the full model fit.

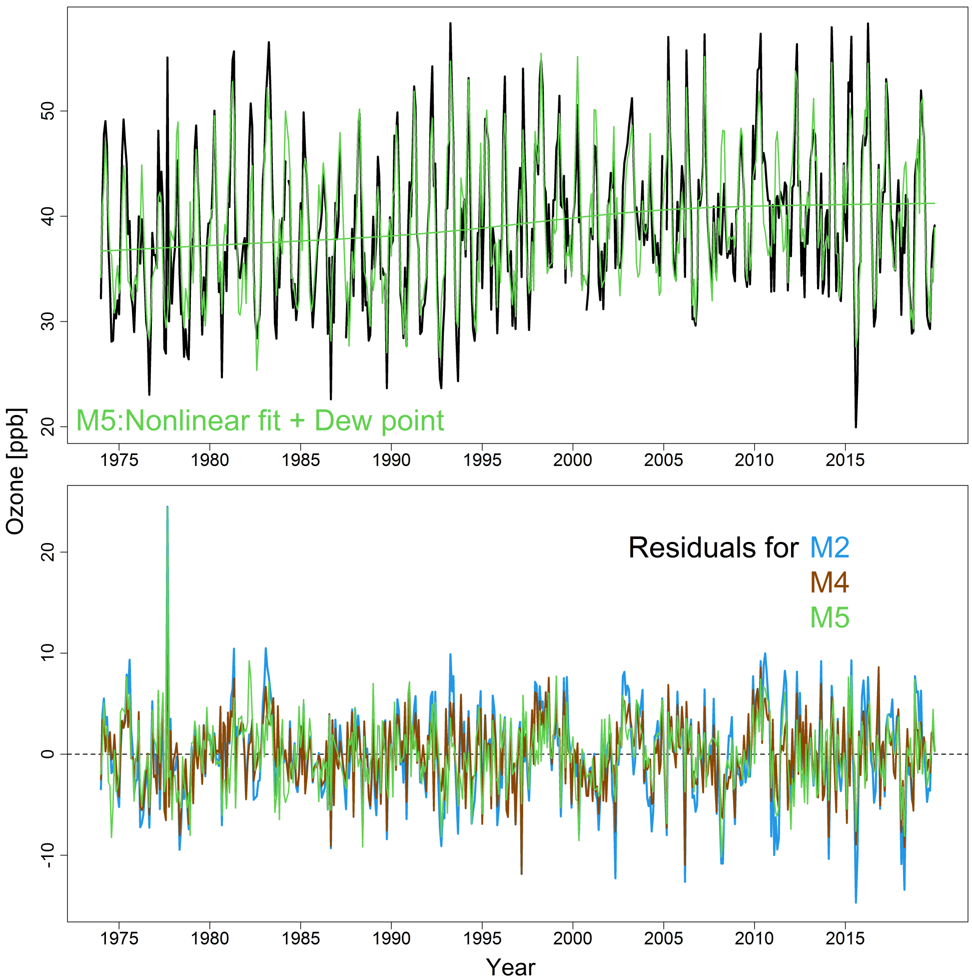

Figure 6: Model fits to the ozone time series with meteorological adjustment at MLO

The upper panel shows the observed and modeled values, and the lower panel shows a comparison of residual series from different model fits.

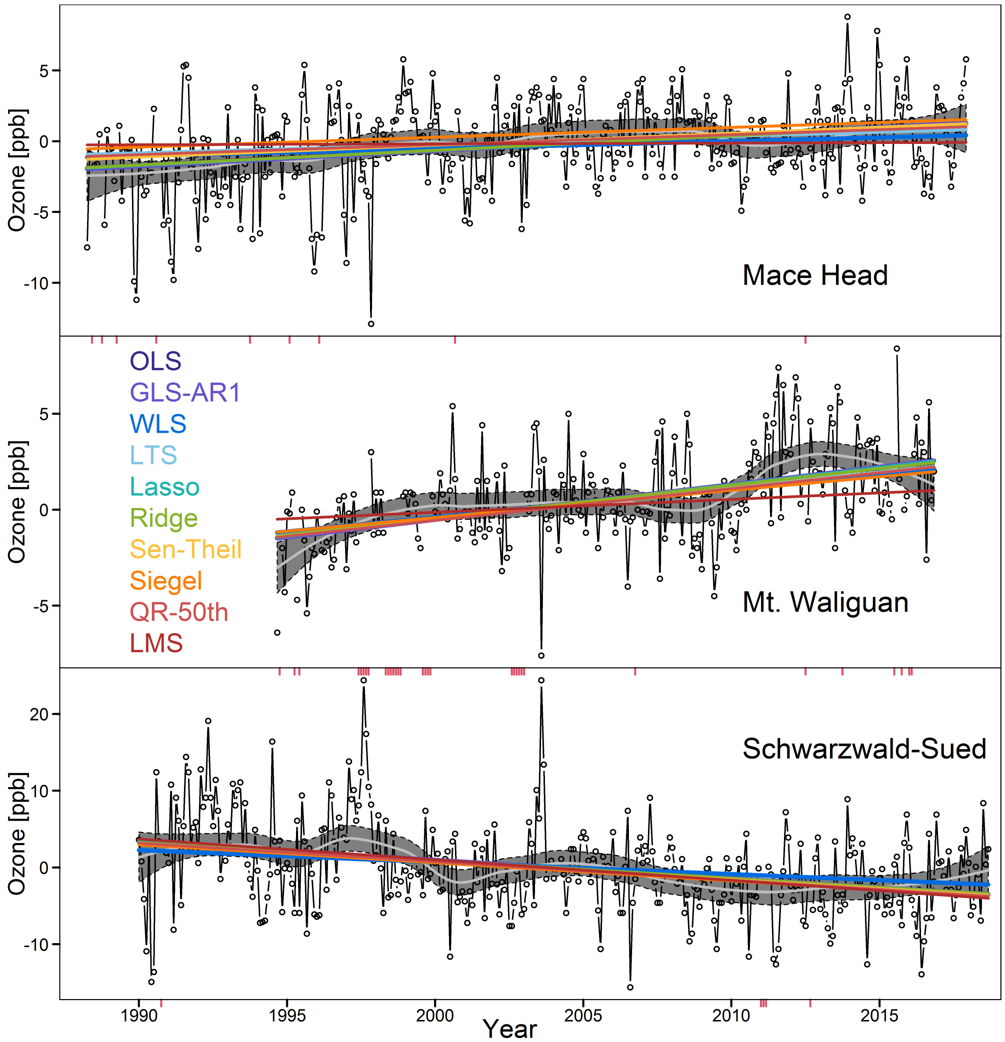

Figure 7: Monthly ozone time series and mean trends at Mace Head, Mt. Waliguan and Schwarzwald-Sued.

Regression lines from several trend detection techniques are fitted. The nonparametric Loess smoother and its 95% confidence interval is highlighted with a gray envelope to illustrate the potential tendency of the trend. Each red tick on the x-axis indicates that a monthly value is missing.

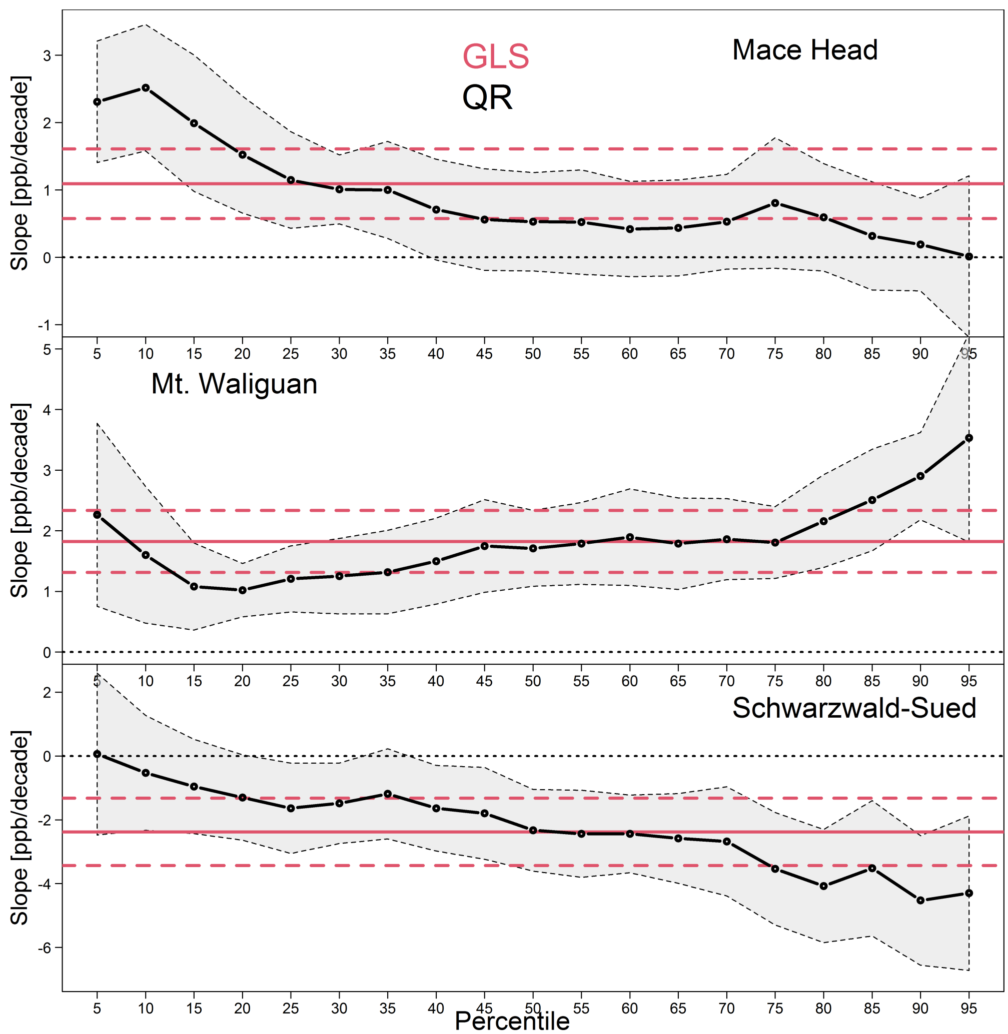

Figure 8: Various illustrations for the distributions of the 5th-95th quantile trends and the 95% confidence intervals.

Demonstrations are made for the ozone time series measured at Mace Head, Mt. Waliguan and Schwarzwald-Sued, with the trend mean value derived by GLS-AR1 model provided for reference.

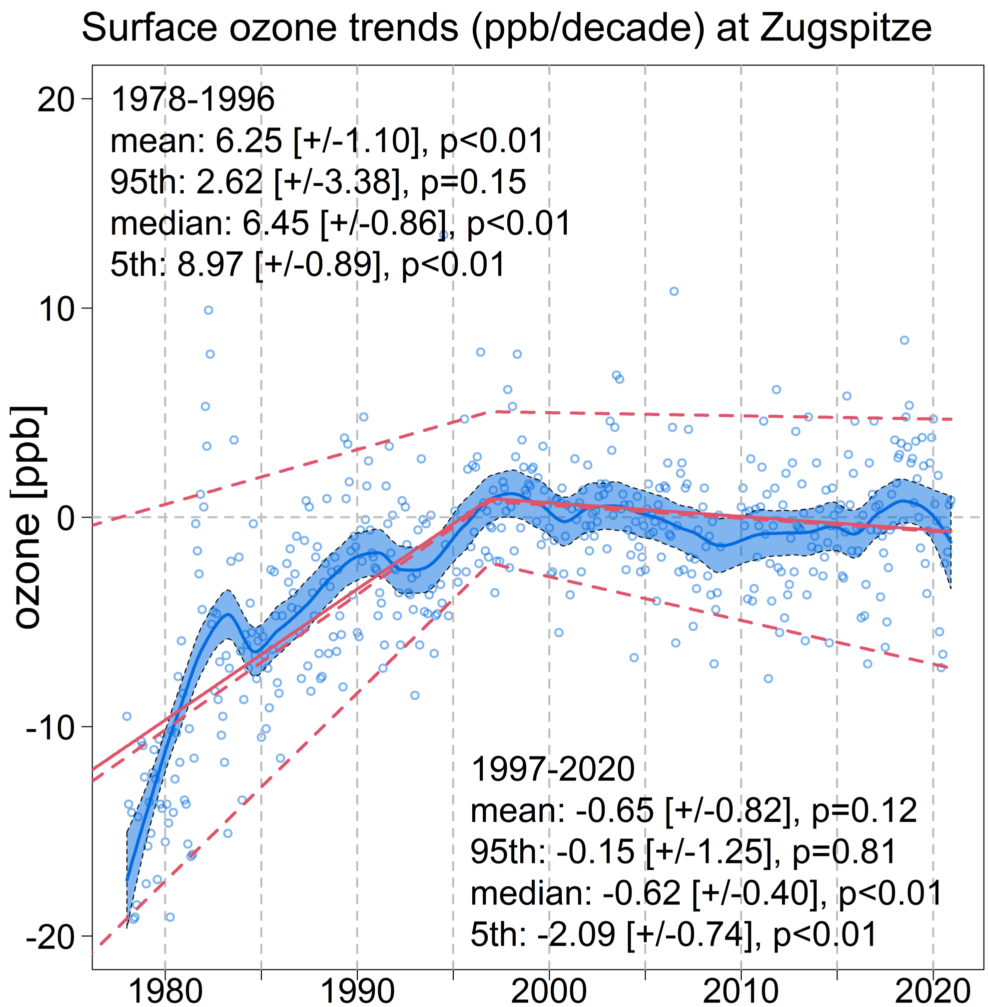

Figure 9: A demonstration of change point analysis based on quantile regression.

The ozone anomaly series is measured at Zugspitze, Germany.

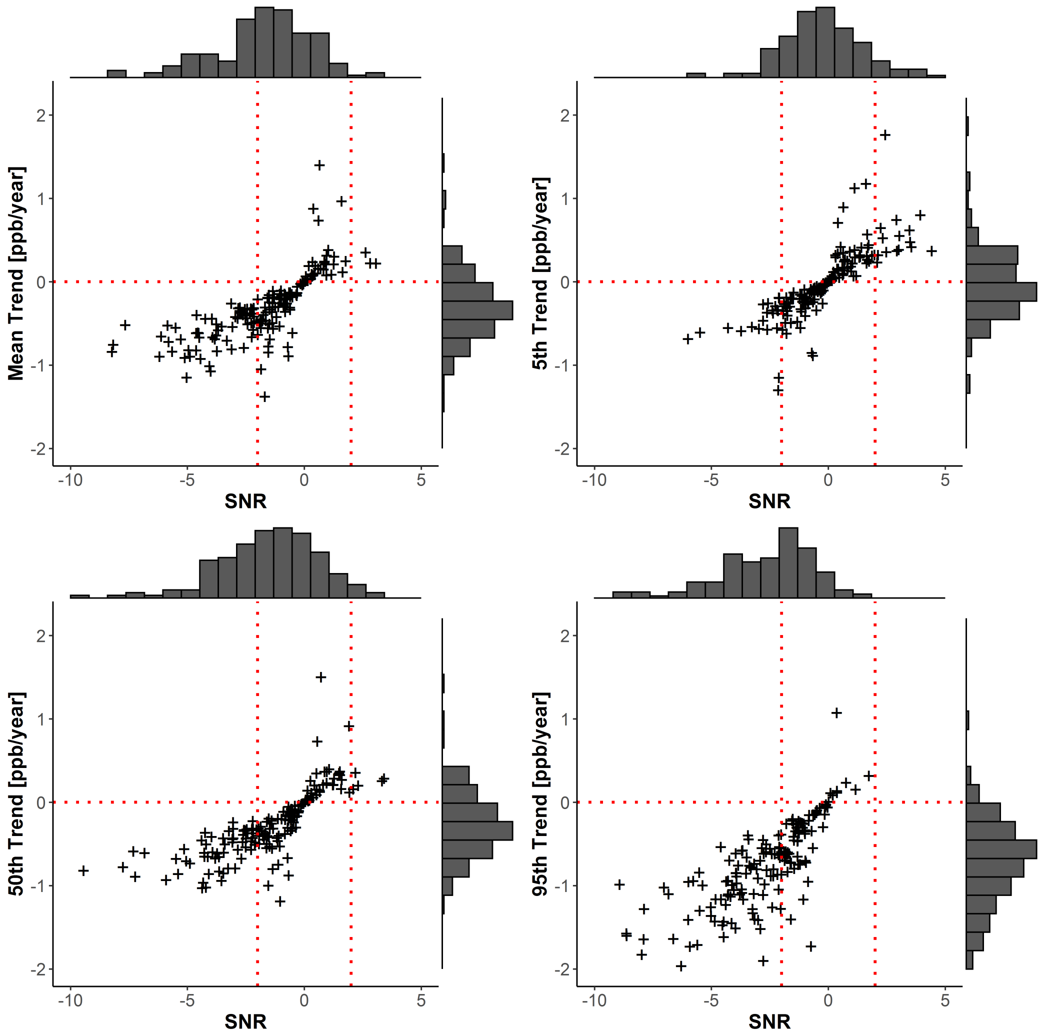

Figure 10: Scatter plots of MDA8 quantile trends and SNR values in the south-western US.

Demonstrations are made for individual time series trend analysis of the mean, 5th, 50th and 95th percentiles over 2000-2014.

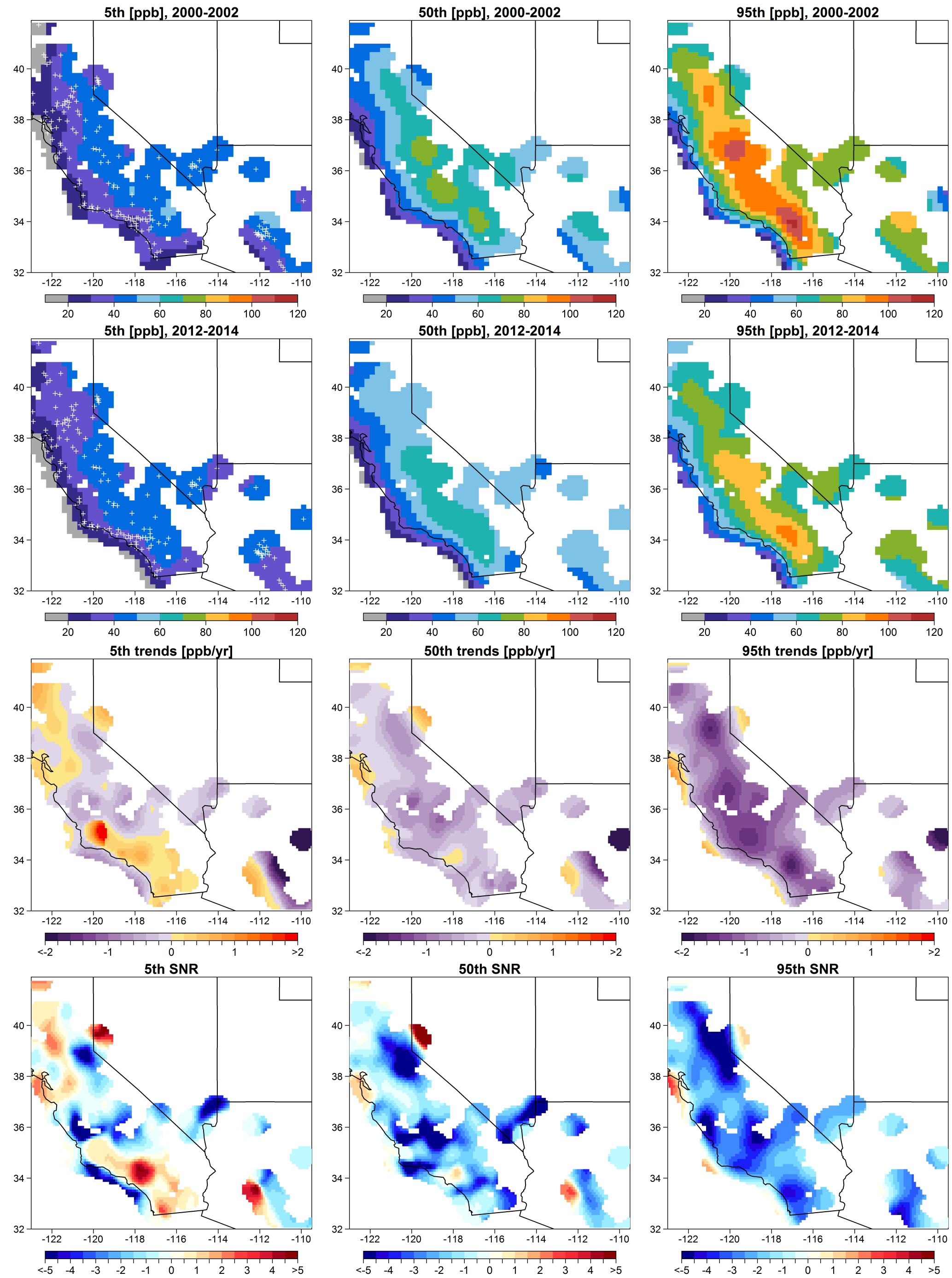

Figure 11: Quantile spatial fields over two periods, distributions of trends and SNR ratios in the southwestern US.

Demonstrations are made for MDA8 spatial distributions (in units of ppb) of the 5th, 50th and 95th percentiles over 2000-2002 (1st row) and 2012-2014 (2nd row) in the southwestern US, with corresponding spatial distributions of trends (in units of ppb per year, 3rd row) and signal-to-noise ratios (SNR, i.e. trend value divided by standard error, 4th row) over 2000-2014. White crosses represent the locations of the monitoring stations.

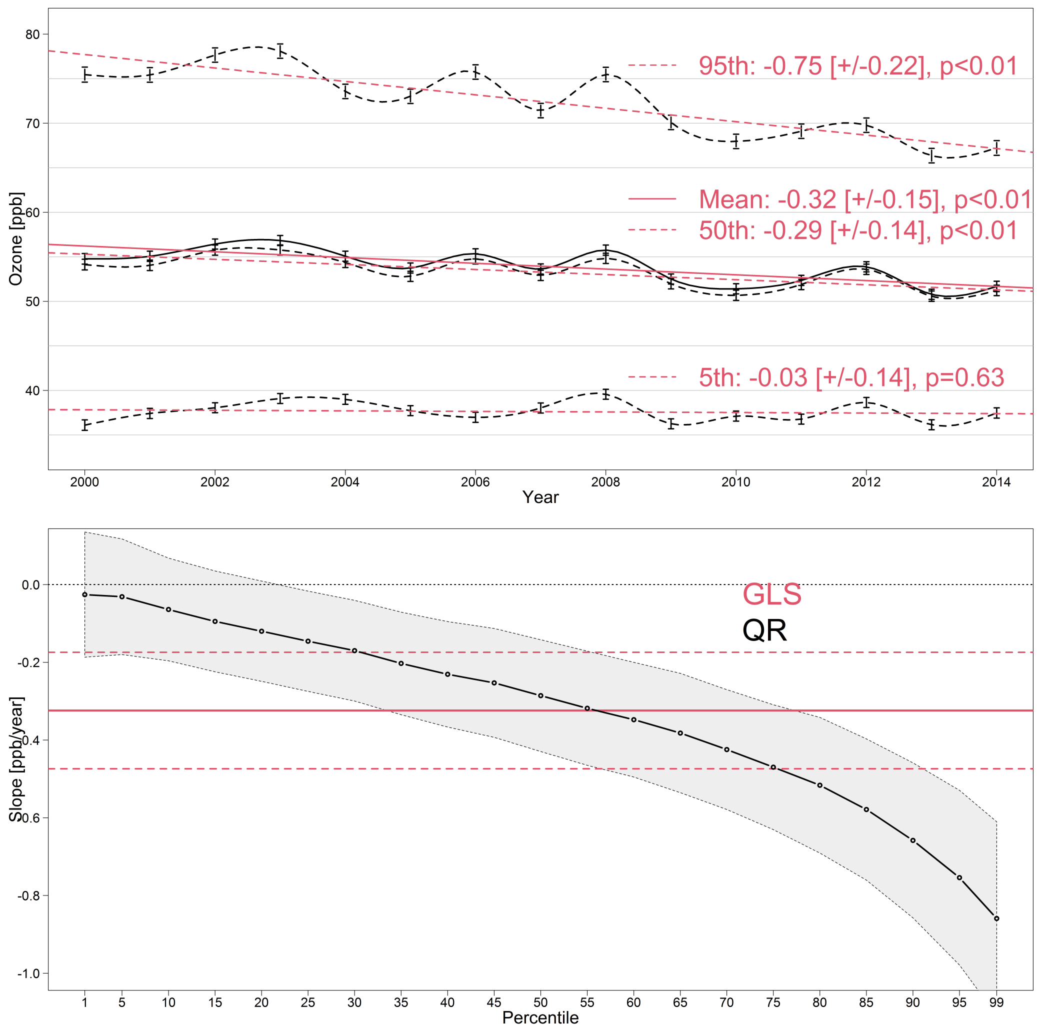

Figure 12: The estimated regional trends and the quantile distribution of regional trends in southwestern US.

The regional ozone time series and trends are estimated with respect to the mean, 5th, 50th and 95th percentiles (upper panel). Quantile distributions of regional trends are based on the 1st, 5th, 10th, ..., 90th, 95th, 99th percentiles, with the trend mean value derived by GLS-AR1 model provided for reference.

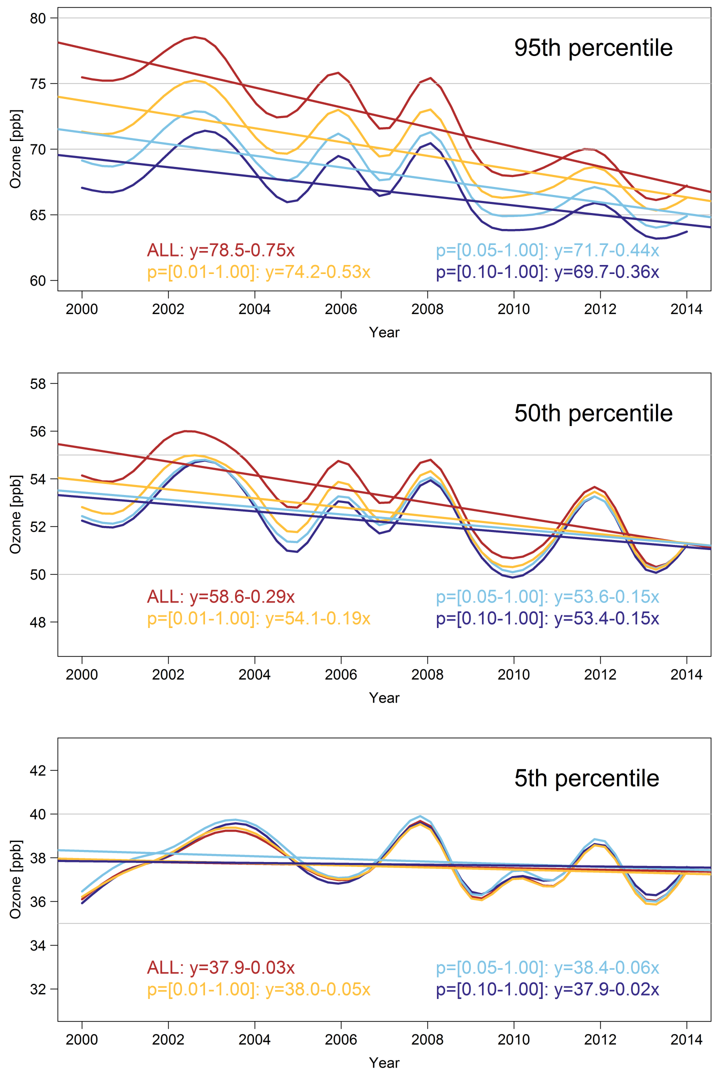

Figure 13: Impact of the representativeness of sites on trends.

Estimated long-term changes for MDA8 using all 168 sites (red), and only the sites with p-value of slope of the trend within the range of [0.01, 1.00] (orange), [0.05, 1.00] (light blue) and [0.10, 1.00] (dark blue).

Table 1: Comparison of (a) fitted trends, 2-sigma uncertainty [ppb decade^{-1}] and signal-to-noise ratio (SNR) from various lags of autocorrelations; (b) fitted quality from various fits using R^2, MSE and GCV score; and (c) linear trend estimates when incorporating different covariate(s) for CO and ozone at MLO.

(a) Statistics from various lags of autocorrelations									
		OLS	AR1	AR2	AR3	AR4	AR5	AR6	
CO	trend	-5.85	-5.68	-5.73	-5.50	-5.76	-5.49	-5.65	
	2-sigma	0.89	2.68	1.72	2.84	1.91	2.79	2.33	
	SNR	-13.12	-4.24	-6.65	-3.88	-6.04	-3.93	-4.84	
Ozone	trend	0.99	0.99	0.99	0.99				
	2-sigma	0.30	0.43	0.48	0.48				
	SNR	6.64	4.65	4.04	4.12				
(b) Fitted quality from various fits									
		M1	M2	M3	M4	M5	M6	M7	M8
CO	R^2	81.8	82.6	90.9	86.9	84.8	84.8	84.2	84.9
	MSE	55.1	50.7	27.4	39.6	46.0	45.9	47.7	45.6
	GCV	57.9	54.9	1095.5	51.4	50.3	50.3	52.1	50.4
Ozone	R^2	58.5	60.0	97.2	75.4	77.0	75.3	64.7	77.0
	MSE	20.9	20.2	1.4	12.4	11.6	12.5	17.8	11.6
	GCV	21.7	21.3	169.8	18.0	12.2	13.0	18.9	12.3
(c) Linear trend estimate with covariate(s) included									
		M1	M2	M3	M4	M5	M6	M7	M8
CO	trend	-5.73				-5.68	-5.73	-5.76	-5.64
(AR2)	2-sigma	1.72				1.71	1.71	1.71	1.73
	SNR	-6.65				-6.64	-6.71	-6.72	-6.53
Ozone	trend	0.99				1.17	0.93	0.53	1.42
(AR1)	2-sigma	0.43				0.30	0.31	0.40	0.36
	SNR	4.65				7.68	6.04	2.66	7.97

Table 2: Regional trend estimates based on the 95th, 50th and 5th percentiles of all available MDA8 values and only the sites with p-value of slope of the trend within a certain range in southwestern US.

Percentile		Intercept (ppb)	Slope (ppb yr^{-1})	2-sigma (ppb yr^{-1})	p-value	SNR	# site
95th	All sites	78.46	-0.75	0.22	<0.01	-6.82	168 (100%)
	$p = [0.01 - 1.00]$	74.22	-0.53	0.24	<0.01	-4.42	104 (62%)
	$p = [0.05 - 1.00]$	71.72	-0.44	0.24	<0.01	-3.67	83 (49%)
	$p = [0.10 - 1.00]$	69.71	-0.36	0.27	0.02	-2.67	63 (38%)
	$p = [0.15 - 1.00]$	68.27	-0.30	0.31	0.08	-1.94	54 (32%)
	$p = [0.20 - 1.00]$	67.92	-0.26	0.36	0.18	-1.44	46 (27%)
	$p = [0.30 - 1.00]$	67.72	-0.20	0.48	0.41	-0.83	33 (20%)
	$p = [0.40 - 1.00]$	65.14	-0.16	0.53	0.55	-0.60	26 (15%)
50th	All sites	55.58	-0.29	0.14	<0.01	-4.14	168 (100%)
	$p = [0.01 - 1.00]$	54.12	-0.19	0.15	0.02	-2.53	128 (76%)
	$p = [0.05 - 1.00]$	53.58	-0.15	0.15	0.06	-2.00	105 (63%)
	$p = [0.10 - 1.00]$	53.39	-0.15	0.16	0.08	-1.88	94 (56%)
	$p = [0.15 - 1.00]$	53.19	-0.12	0.15	0.15	-1.60	83 (49%)
	$p = [0.20 - 1.00]$	53.58	-0.12	0.16	0.16	-1.50	73 (43%)
	$p = [0.30 - 1.00]$	52.99	-0.10	0.15	0.21	-1.33	63 (38%)
	$p = [0.40 - 1.00]$	53.07	-0.08	0.16	0.30	-1.00	56 (33%)
5th	All sites	37.90	-0.03	0.14	0.63	-0.43	168 (100%)
	$p = [0.01 - 1.00]$	37.98	-0.05	0.15	0.54	-0.67	154 (92%)
	$p = [0.05 - 1.00]$	38.37	-0.06	0.14	0.40	-0.86	136 (81%)
	$p = [0.10 - 1.00]$	37.86	-0.02	0.14	0.78	-0.29	118 (70%)
	$p = [0.15 - 1.00]$	37.99	-0.04	0.14	0.62	-0.57	103 (61%)
	$p = [0.20 - 1.00]$	37.94	-0.04	0.15	0.59	-0.53	98 (58%)
	$p = [0.30 - 1.00]$	37.48	-0.03	0.14	0.65	-0.43	84 (50%)
	$p = [0.40 - 1.00]$	37.22	-0.02	0.16	0.80	-0.25	72 (43%)