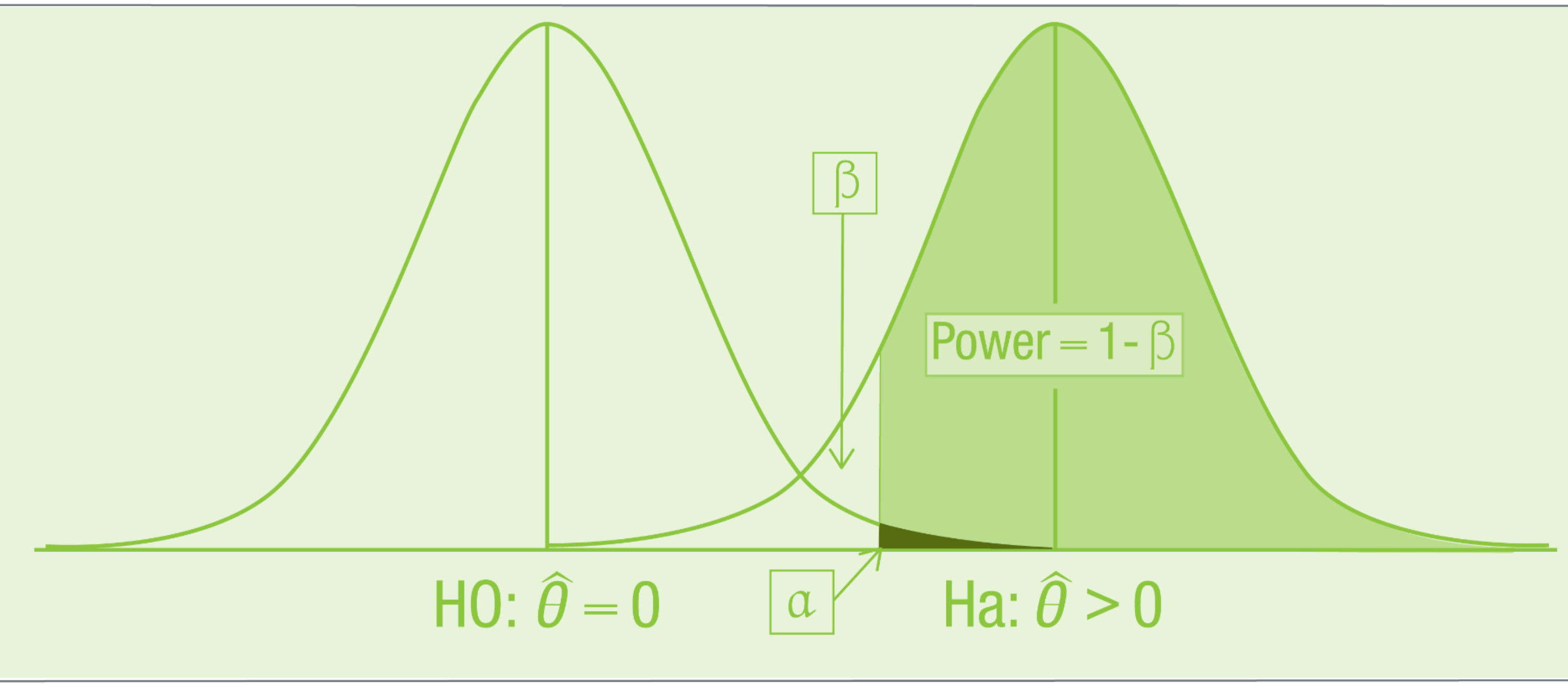
Methods and Procedures for Trend Analysis of Air Quality Data



Air

Land

Water

Biodiversity

Alberta

Methods and Procedures for Trend Analysis of Air Quality Data

Thompson Nunifu and Long Fu

This publication can be found at: https://open.alberta.ca/publications/9781460136379.

Comments, questions, or suggestions regarding the content of this document may be directed to:
Ministry of Environment and Parks, Environmental Monitoring and Science Division
10th Floor, 9888 Jasper Avenue NW, Edmonton, Alberta, T5J 5C6
Email: EMSD-Info@gov.ab.ca
Website: environmentalmonitoring.alberta.ca

For media inquiries, please visit: alberta.ca/news-spokesperson-contacts.aspx.

Recommended citation:

Nunifu, T. and L. Fu. 2019. Methods and Procedures for Trend Analysis of Air Quality Data. Government of Alberta, Ministry of Environment and Parks. ISBN 978-1-4601-3637-9. Available at: https://open.alberta.ca/publications/9781460136379.

Published August, 2019

ISBN: 978-1-4601-3637-9

Alberta’s Environmental Science Program

The Chief Scientist has a legislated responsibility for developing and implementing Alberta’s environmental science program for monitoring, evaluation and reporting on the condition of the environment in Alberta. The program seeks to meet the environmental information needs of multiple users in order to inform policy and decision-making processes. Two independent advisory panels, the Science Advisory Panel and the Indigenous Wisdom Advisory Panel, periodically review the integrity of the program and provide strategic advice on the respectful braiding of Indigenous Knowledge with conventional scientific knowledge.

Alberta’s environmental science program is grounded in the principles of:

- *Openness and Transparency.* Appropriate standards, procedures, and methodologies are employed and findings are reported in an open, honest and accountable manner.
- *Credibility.* Quality in the data and information are upheld through a comprehensive Quality Assurance and Quality Control program that invokes peer review processes when needed.
- *Scientific Integrity.* Standards, professional values, and practices of the scientific community are adopted to produce objective and reproducible investigations.
- *Accessible Monitoring Data and Science.* Scientifically-informed decision making is enabled through the public reporting of monitoring data and scientific findings in a timely, accessible, unaltered and unfettered manner.
- *Respect.* A multiple evidence-based approach is valued to generate an improved understanding of the condition of the environment, achieved through the braiding of multiple knowledge systems, including Indigenous Knowledge, together with science.

Learn more about the condition of Alberta’s environment at: environmentalmonitoring.alberta.ca.

Acknowledgements

The project was a collaborative effort between the Lower Athabasca Regional Plan (LARP) Technical Team and the Standards, Quality, and Innovation team of Alberta Environment and Parks' Environmental Monitoring and Science Division (EMSD). This report benefited tremendously from the input of the LARP Technical Team members regarding the needs and specifics of implementation of a new trend analysis approach. Dr. Prabal Roy from the Northern Region played an instrumental role in the initiation of the project by drafting the project plan, organizing meetings and presentations, managing the project in collaboration with Dr. Nunifu and maintaining constant communication with the LARP technical team to keep the members updated on the progress of the project. Dr. Roy conducted the initial test of the new trend analysis method and provided valuable feedback regarding user experience. The LARP Technical Team was instrumental in the initiation of this project by identifying the need for a tool that could be used in the analysis of short-term trends. Their numerous contributions by way of inputs and feedback to the project planning, sharing of results through presentations are much appreciated. We also acknowledge the contribution of other Standards, Quality and Innovation team members especially, David Powers and Steven Vuong, who helped in reviewing the initial draft of the report and also tested the trend analysis tool. We thank Dr. Charlie Loewen for his thorough review of the initial draft of the manuscript and the valuable feedback, comments, and suggestions. This report also received tremendous support from our Executive Director, Dr. Garry Scrimgeour. Dr. Scrimgeour encouraged the team to document the performance of models in this report to support the work of EMSD and other groups in the Department.

The authors also thank the external reviewer for their technical review and feedback, which have enhanced this work.The reviewer holds a Ph.D. in biometrics (forest modeling) and has more than 25 years of applied research experience in advanced statistics and modeling. The reviewer has authored and co-authored 120 scientific and technical publications, and have made numerous presentations at various venues including invited and keynote speeches at international scientific conferences in the United States, Japan, Portugal, Colombia and Chile.

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

ACF	Autocorrelation Function
AEMERA	Alberta Environmental Monitoring Evaluation and Reporting Agency
AEP	Alberta Environment and Parks
AQ	Air Quality
AQMF	Air Quality Management Framework
AR	Autoregressive Process
ARCH	Autoregressive Conditional Heteroscedastic
ARIMA	Autoregressive Integrated Moving Average
CASA	Clean Air Strategic Alliance
EGARCH	Exponential Generalized Autoregressive Conditional Heteroscedasticity
EMSD	Environmental Monitoring and Science Division
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GJR-GARCH	Glosten-Jagannathan-Runkle Generalized Autoregressive Conditional Heteroscedasticity
GLS	Generalized Least Squares
IWS	Iteratively Weighted Least Squares
LARP	Lower Athabasca Regional Plan

LOESS	Nonparametric Local Weighted Regression Smoothening Technique
MA	Moving Average
MK	Mann-Kendall
NAGARCH	Nonlinear Generalized Autoregressive Conditional Heteroscedasticity
OLS	Ordinary Lest Squares
PACF	Partial Autocorrelation Function
QA/QC	Quality Assurance/ Quality Control
QGARCH	Quadratic Generalized Autoregressive Conditional Heteroscedasticity
RMSE	Root Mean Square Error
RUGARCH	A statistical package in R-program used for the analysis of GARCH models.
SQI	Standards, Quality and Innovation
TFPW	Trend Free Pre-whitening
US EPA	United States Environmental Protection Agency
WBEA	Wood Buffalo Environmental Association
WLS	Weighted Least Squares

1 Background and rationale

Trends of ambient air quality in Alberta Environment and Parks (AEP) have traditionally been assessed using annual summaries (e.g., annual averages or annual quantiles) of hourly data from ground monitoring stations. The goal is to determine if there is a trend in the concentrations of the air quality parameters over time. This assessment was performed using a tool developed in 2005 by staff in the Operations Division of AEP. The tool utilized the Theil-Sen method (Theil 1950, Sen 1968) to estimate trend and the Mann-Kendall method (Mann 1945, Kendall 1975) to test for the statistical significance of the estimated trend. However, Yue et al. (2002) have demonstrated that these methods (the Theil-Sen and Mann-Kendall) are affected by autocorrelation (see Section 2.3.1 for a detailed discussion). Therefore, a procedure known as the Trend Free Pre-Whitening (TFPW) described in Yue et al. (2002) was adopted and implemented in the tool to address the effect of autocorrelation in the data. When annual summaries are used, the problem of autocorrelation is significantly reduced due to the longer averaging period (1 year), thus making the task less challenging for TFPW. The tool is a Fortran-based statistical program but utilizes Microsoft Excel as its user interface to input and display outputs in the form of annual trends and related statistics. In this document, this tool will be referred to as TFPW.

A detailed analysis of air quality data was conducted to provide scientific support to the 2013 Air Quality Management Framework (AQMF) Management Response for the Lower Athabasca Regional Plan (LARP). A technical report was later released by Liu et al. (2015) based on this analysis. The 2013 AQMF Management Response, backed by the technical report (Liu et al. 2015) identified some limitations with the TFPW tool. The primary statistical concerns were related to the sample size. Specifically, using the tool to analyze data for a short-term trend was found to be problematic. Since annual averages of hourly data were used by the tool to assess trend, the statistical sample size was restricted to the number of years over which trend is being analyzed (1 data point per year). Thus for a short period (e.g., five years or less), the sample size of five or less is too small to estimate and draw a meaningful statistical conclusion about the trend. The statistical implication of such a small sample size is two-fold. First, there is not enough statistical power (i.e., the probability of detecting a trend when there is a trend) to efficiently test the null hypothesis about the trend. Second, increasing the sample size to improve statistical power by using shorter averaging periods (e.g., monthly or daily) will likely increase the autocorrelation problem. Using shorter averaging time may also introduce seasonal variations into the data which the tool is not designed to address.

As it is an annual communication to Albertans on the management response to air and surface water trigger exceedances at monitoring stations in the Lower Athabasca Region (LAR), the 2013 Management Response report recommended that the issues discussed above be addressed. A new approach and a tool with enough statistical power and the capability to analyze both short-term and long-term trends of air quality to support the ambient environmental condition reporting was recommended. The LARP technical team contacted the Standards, Quality and Innovation (SQI) team of EMSD, previously within Alberta Environmental Monitoring, Evaluation and Reporting Agency (AEMERA) in 2015, to develop a new tool. Specifically, the LARP technical team members were interested in information on short-term changes of ambient air quality to identify emerging issues and track the effectiveness of air quality management in the area and needed a tool that can estimate both short-term and long-term trends. This report documents the details of the statistical theory and analysis of the approaches of the new tool.

The request from the LARP technical team was timely, as statistical tool development was part of the SQI team plan at the time. Following the request, a project initiation meeting was held on July 16, 2015, between the LARP technical team and the SQI team to discuss details of the project. A presentation was made by the SQI team to the LARP technical team in August 2015 on the proposed process, and the project officially started in September 2015. An initial package of the new tool which is based on R-software and a draft report was completed in September 2016. This is the final report for the project documenting the scientific information considered when developing the new tool.

1.1 Project objectives

The objectives of the present study are two-fold: First, we identified and described the statistical challenges typically encountered when conducting trend analysis of air quality data. Specifically, the challenges related to autocorrelation, heteroscedasticity, and non-normality of data and considerations related to statistical power are described and defined. We also provided guidance on how to deal with statistical challenges when conducting trend analysis. Second, we identified and tested three methods using NO_2 and SO_2 data from three air quality monitoring stations in Alberta.

The three methods tested are:

1) A nonparametric approach where the Theil-Sen and the Mann-Kendall methods are applied sequentially to estimate trend and to test that the estimated trend is statistically significant, respectively.
2) A parametric approach where linear regression is used to estimate trend and the generalized autoregressive conditional heteroscedasticity (GARCH) method is used to estimate the standard error to test the null hypothesis that the estimated trend is not statistically significant.
3) A parametric approach where linear regression is used to estimate trend and the autoregressive integrated moving average (ARIMA) method is used to estimate the

standard error to test the null hypothesis that the estimated trend is not statistically significant.

The first approach (Theil-Sen and Mann-Kendall methods) is the approach currently used in the existing tool for trend analysis. This approach was selected to test the tool's applicability to monthly summaries of hourly data and to demonstrate how the challenges (seasonality and autocorrelation) introduced by using these monthly summaries can be addressed. The other reasons for selecting this approach is the popularity of the Theil-Sen and the Mann-Kendall methods in environmental science and their attractive statistical properties; these are nonparametric methods and do not require any distributional assumptions (e.g., Wang and Swail 2001, Yue et al. 2002,). As a non-parametric approach, it also provides a good reference for comparison with the other two parametric approaches tested in this report.

The second and the third approaches are similar. They are both linear regression (parametric) methods except that they use different methods to estimate the standard error of trend for subsequent testing of the null hypothesis. GARCH and ARIMA were chosen for this study because in theory, they are statistically more robust and superior over other standard error estimation methods under linear regression analysis of time series data (e.g. Bollerslev 1986, Nelson 1991, Ding et al. 1993, Zakoïan 1994).

The criteria used for comparing the three approaches are: 1) Confidence interval width of the trend estimates, 2) Fit statistics measured by mean square prediction error and 3) statistical power measured by the minimum size of monthly trend each method can detect using two years of measurements (24 months). The results of this study should be treated as initial and preliminary findings.

1.2 The organization of this report

The critical issues and statistical challenges of trend analysis are discussed in Section 2. Section 3 presents an overview of methods suitable for trend analysis and an in-depth discussion of the methods presented in this report. Specifically, the ARIMA and GARCH regression analysis methods are discussed in detail. This discussion includes how each method addresses the statistical issues and challenges discussed in Section 2. Section 4 provides guidance on how to conduct trend analysis of air quality data using the methods discussed in Section 3. The procedures discussed in the guidance are used to develop a statistical tool in the R programming language and will be demonstrated and made available to end users. Section 5 presents the results of a case study that applied the methods discussed in Section 3 on NO_2 and SO_2 data from arbitrarily selected monitoring stations from the Wood Buffalo Environmental Association (WBEA) Airshed.

2 Issues and challenges in trend analysis

2.1 Why is trend analysis important

The trend of the ambient concentration of a specific environmental pollutant can be described in a variety of ways including cyclical, seasonal, monotonic or step-wise, and may reflect patterns due to seasonal variation, emissions inputs from anthropogenic activities, or natural events. From an environmental management perspective, a cumulative increase in the ambient concentration levels of an air quality parameter over time may indicate the presence of an input emission source of the parameter into the ambient environment. This cumulative increase may be small and gradual requiring the analysis of long-term data to detect. Also a decision has to be made as to whether the estimated trend is statistically significant (Hirsch et al. 1982, Hirsch and Slack 1984, Esterby 1993). Therefore, the choice of a suitable statistical method for trend analysis with enough power to support this decision is important.

Understanding the trends of ambient air quality parameters may assist in answering fundamental questions such as: i) are the concentrations of key air quality parameters increasing or decreasing over time? ii) Are anthropogenic activities contributing to the ambient concentrations of these air quality parameters? The answers to these questions and others may determine if corrective measures are needed under the Land Use Framework (AEP 2008). As corrective measures should be timely, proportional and appropriate, efforts must be made to avoid fixing a false positive (Type I error) or failure to detect an existing trend (Type II error). This further underscores the importance of making the right statistical inference about the trend.

2.2 Sources of uncertainties

A major challenge in trend analysis is error or uncertainties in the data. Uncertainties play a major role in reducing the statistical power of trend analysis. Major sources of uncertainties are the dynamics of the geophysical environment (i.e. variability due to the atmospheric and meteorological cycles) and data capture activities (including uncertainties due to measurement, sampling, sample treatment and analysis), and may increase the variability in the data. The increased variability may impact statistical inference about the trend by reducing the statistical power of the analysis.

Other issues such as insufficient quantity of data and the repeated nature of continuous air quality data may further exacerbate this problem and further reduce statistical power. As continuous air quality data in Alberta are reported on an hourly basis at various stations across the province, the data are time series. The individual data points in such series could be correlated resulting in a statistical phenomenon called autocorrelation. Autocorrelation is a statistical problem that affects the statistical inference about the trend by impacting the estimated standard error of the estimated trend (Yue et al. 2002). Combined with increase uncertainty from other sources, autocorrelation can further reduce the statistical power of trend analysis.

2.3 Statistical challenges

The focus of this section is on statistical issues that may arise from the nature of air monitoring data and how these issues may affect the statistical estimation of a trend and the subsequent statistical inferences.

2.3.1 Autocorrelation

In standard regression or time series analysis textbooks (e.g. Johnston 1984, Harvey 1990, Kariya and Kurata 2004), autocorrelation refers to the correlation between the error terms from different time periods of a regression model fitted to a time series data. Autocorrelation is simply a correlation coefficient such as Pearson correlation coefficient (ρ). However, instead of the correlation between two different variables, autocorrelation is between two values of the same variable collected or observed at different times, *t* and *t-k*, where *t-k* is the k^{th} lag of *t*. The measure of autocorrelation is used to assess the randomness or independence of individual values in the data series, a key assumption needed for most univariate statistical analysis of the data. Thus the presence of autocorrelation is an indication that the values in the data are not independent.

There are several causes of autocorrelation. In air monitoring data a pollution signal that may cause the concentration of a specific pollutant to increase may take time to die out causing successive measurements of the pollutant to be correlated. Autocorrelation could also be caused by a factor that is correlated with the pollutant concentration thereby causing the individual values of the pollutant to be correlated. The most common cause of autocorrelation in air pollution data is cyclical variations such as diurnal or seasonal variations. The individual values measured on the same phase of the cycle tend to be correlated. Autocorrelation may also exist for observations collected at different spatial locations. In such cases, observations collected at locations that are either near each other or have similar levels of a factor that is correlated with the pollutant, may be correlated. This is referred to as spatial autocorrelation. In this document, however, autocorrelation is used to refer to correlation of observations in time (temporal autocorrelation). In this case first-order autocorrelation refers to the correlation between an observation and the first lag (lag 1) observation in the series. Similarly, second-order autocorrelation refers to the correlation between an observation and the second lag (lag 2) observation in the series.

Autocorrelation may affect the analysis and interpretation of environmental data (e.g., Tiao et al. 1990, Weatherhead et al. 1998, 2000). If this dependence (correlation) is positive, there is a high chance of concluding that there is a positive trend in the concentration when in fact there may be no trend. Statistically, it can be shown that positive autocorrelation can result in smaller standard error of estimate of a trend than the actual standard error (e.g., Tiao et al. 1990, Weatherhead et al. 1998, 2000, and Yue et al. 2002). The result is that the test statistic for the null hypothesis may be inflated. This inflation of the test statistic has been shown to increase nonlinearly with the strength of the autocorrelation (Yue et al. 2002). The consequence of the inflated test statistic could be an increase in the false positive rate, which in this case is an increase in the rate of false detection of trend. The converse is also true; negative autocorrelation may inflate the standard error estimate leading to false acceptance of the null hypothesis.

The use of linear regression (parametric) analysis requires (in addition to independence) that environmental data be normally distributed and have uniform variance across time. These requirements have contributed indirectly to the popularity of the use of non-parametric analysis such as the Theil-Sen trend estimator (Theil 1950, Sen 1968) and the Mann-Kendall trend test (Mann 1945, Kendall 1975). The enormity of literature on the use of these non-parametric methods is a testament to this fact. However, studies have also shown that autocorrelation can affect the results of these non-parametric methods. False rejections of the null hypothesis (false positives) are more likely in the trend analysis of a dataset with autocorrelation, and the likelihood increases with increasing strength of the correlation when the Mann-Kendell trend test is used (Kulkarni and von Storch 1995, Yue et al. 2002). The variance of a trend estimated by the Theil-Sen test was shown by Yue et al. (2002) in a simulation study to increase with the magnitude of autocorrelation, which if not accounted for can result in gross underestimation of the standard error.

The issue of autocorrelation has resulted in the use of pre-whitening, an approach adopted to remove autocorrelation before implementing Theil-Sen slope estimation and Mann-Kendall test for trend analysis (e.g., von Storch 1995, von Storch and Zwiers 1999, Wang and Sail 2001, Yue et al. 2002, Zhang and Zwiers 2004). While this is considered by many to be an attractive solution to problems related to autocorrelation, it has also been noted to tend to “wash-out” the trend in the data (Yue et al. 2002). Recursive pre-whitening originally proposed by Kulkarni and von Storch (1995) has been demonstrated by Wang and Swail (2001), and Zhang and Zwiers (2004) to be an effective alternative to the original pre-whitening method. Kulkarni and von Storch (1995) and Yue et al. (2002) have demonstrated that false rejection of the null hypothesis occurs at a rate close to the nominal Type I error rate of 5% when recursive pre-whitening is applied to data with the first-order autocorrelation before the Mann-Kendall trend test is conducted.

2.3.2 Heteroscedasticity

Heteroscedasticity is a statistical term used to describe non-uniform variance of the residuals from the statistical fit of a trend. The problem of non-uniform variance is associated with parametric methods, such as linear regression analysis that relies on a statistical reference distribution to test the null hypothesis of no trend. The issue of heteroscedasticity is typically reflected in the residual variance not being constant for all values of the independent variable (i.e., time). In environmental sciences, periods of episodic conditions may result in not only elevated levels of the pollutant concentration but may also cause large temporal fluctuation in the concentrations, resulting in higher variance. An added challenge for air quality data is that the variance may not follow a definite temporal pattern, making it difficult to model with a simple variance function.

Standard regression textbooks such as Neter et al. (1983), Johnston (1984), Graybill and Iyer (1994), and Kariya and Kurata (2004) show that heteroscedasticity affects the variance estimates of a regression slope (e.g., trend). The poor variance estimate for trend may result in a poor estimate of the test statistic, which can result in an incorrect inference about the trend.

Non-parametric methods such as Mann-Kendall trend test (Mann 1945, Kendall 1975) and Theil-Sen slope (Theil 1950, Sen 1968) are not affected by the issue of heteroscedasticity. In parametric regression analysis, however, procedures such as weighted least squares (WLS) and iteratively weighted least squares (IWLS) (e.g., Cleveland 1979, Carroll and Rupert 1988, Cleveland and Durvlin 1988, Ryan 1997) are typically used to address heteroscedasticity. More sophisticated techniques such as the generalized least squares which addresses both autocorrelation and heteroscedasticity (e.g., Johnston 1984, Kariya and Kurata 2004) can be used. More recently the generalized autoregressive conditional heteroscedasticity (GARCH) approach has been implemented for trend analyses in environmental sciences to address both heteroscedasticity and autocorrelation (e.g., Wang et al. 2005, Wang et al. 2006, Chen et al. 2008, and Modarres and Ouarda 2013).

2.3.3 Non-normality

Another statistical challenge is that the distributions of air quality data are typically not normal, but instead are usually skewed distributions (Georgeopoulos and Seinfeld, 1982). This skewness is attributed to the less frequent occurrence of poor air quality episodes. In some cases, such as in nitrogen dioxide (NO_2) where concentrations show strong seasonal variation, the data may show bi-modal distributions.

The normality assumption is perhaps the least challenging issue compared with the other two discussed above for the following reasons:

1) The normality assumption in linear regression is required to use the normal distribution or the small sample version, the student's t-distribution as the reference distribution for hypothesis testing, but this requirement applies to the distribution of the regression residuals. Although the original pollution data may be non-normal, the residuals tend to be approximately normal as they are conditional on the independent variables thus making this requirement less relevant.

2) It takes a considerable departure from normality to invalidate the hypothesis test, which hardly happens with regression residual especially when the sample size is large. More details can be found in standard linear regression textbooks (e.g., Neter et al. 1983, Johnston 1984).

Regardless, the previous practice has been to avoid the use of regression analysis altogether in favor of nonparametric methods.

2.3.4 Statistical power

Statistical power refers to the ability of a statistical analytical method to detect an effect if any exists. In statistical terms, power is the probability of rejecting an incorrect null hypothesis. For trend analysis of air quality data, a higher statistical power implies an increased probability of detecting an existing effect (trend) and thus rejecting the null hypothesis of no trend. Statistical power is affected by several factors. One factor is the effect size. In trend analysis, the target effect is the trend, and large trends (increases or decreases in magnitude) are much easier to detect. In the words of Cohen (1988), "all null hypotheses, at least in their 2-tailed forms, are false, meaning, whatever we are looking for is always going to be there – it might be there in such small quantities that we are not bothered about finding it". Thus, thinking of effect size in environmental sciences often involves addressing a fundamental question: is the effect significant enough to warrant further investigation? An answer to this question may help in deciding the choice of a Type I error rate that is considered acceptable.

A second factor that affects the statistical power is the sample size. A relatively large sample size is expected to improve the estimates of the parameters and their associated standard errors. The improved standard errors estimates help refine the probability of detecting an effect.

Another factor is the probability of wrongly concluding that there is an effect (trend) when there is none. This probability is the Type I error or the Type I error rate (designated as *α*). The Type I error rate is commonly set at 5% in environmental sciences to indicate the probability of rejecting a correct null hypothesis. If the Type I error rate is set high, then there is an increased chance of concluding that there is a trend. Thus, a higher Type I error rate artificially increases the statistical power. Statistical power and the probability of accepting an incorrect null hypothesis (known as a Type II error, represented by β) are complementary, such that $\text{power} = 1 - \beta$ (see Figure 1 below). In Figure 1, increasing the Type I error rate implies a decrease in Type II errors and hence increases the statistical power.

The last factor to discuss is the error or the conditional variation from month-to-month (or any chosen time resolution). In the trend analysis of environmental data, various sources of data capture errors, as well as autocorrelation, contribute to this variation and may mask a trend, making it difficult to assess its significance. Larger variations may require larger amounts of data (sample size) to assess trends of a given magnitude.

The question of how many years of data are needed to detect a trend of a given size has been studied by Tiao et al. (1990), and Weatherhead et al. (1998, 2000). These studies developed a relationship that ties together all the factors discussed above into Equation 1 below. Figure 1 is the graphical representation of this relationship (Tiao et al. 1990). Equation 1 relates the number of years of monitoring (i.e. sample, *n*) needed to detect trend of magnitude (effect size, *θ*) at a 95% confidence interval (Type I error rate of 5%), while risking a Type II error rate of β, given that month-to-month variation (*σ*) is observed and the first order autocorrelation is assumed with autocorrelation coefficient, *ρ*. Mathematically, this relationship is given by:

$$n = \left[\frac{(2+z_\beta)\sigma}{|\theta|}\sqrt{\frac{1+\rho}{1-\rho}}\right]^{2/3} \qquad \text{(Equation 1)}$$

Where: z_β is the β-percentile of the standard normal curve; the lower bound of the alternate hypothesis is $-z_\beta$, such that $P(Z < -z_\beta) = \beta$.

Based on this relationship, the number of years needed to detect an annual trend *θ* at a 95% probability can be calculated by assuming a 90% statistical power (meaning β = 10% and $z_{0.10} \approx 1.3$), thus giving the relationship:

$$n = \left[\frac{3.3\sigma}{|\theta|}\sqrt{\frac{1+\rho}{1-\rho}}\right]^{2/3} \qquad \text{(Equation 2)}$$

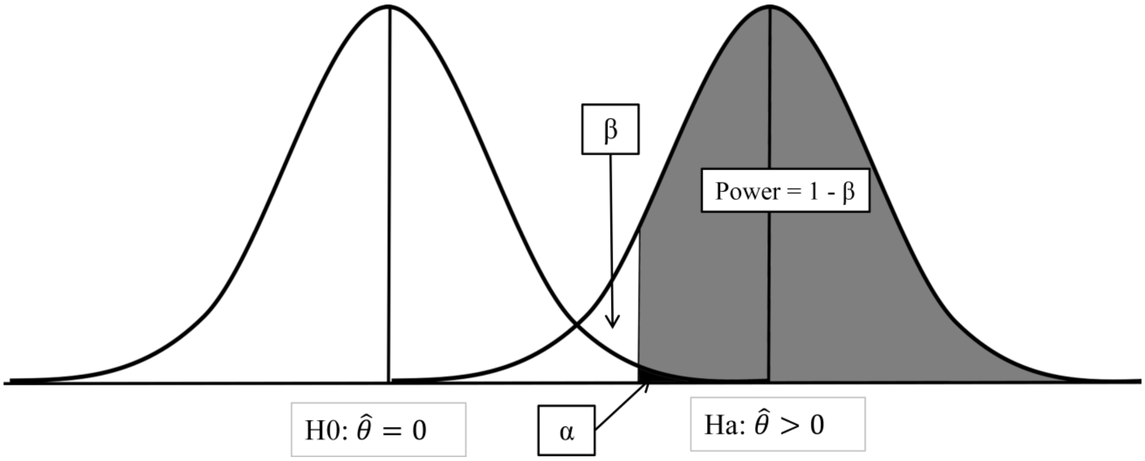


Figure 1- The probability distribution of estimated effect $\hat{\theta}$ under the null hypothesis H0: $\hat{\theta} = 0$ (left curve) and alternate hypothesis HA: $\hat{\theta} > 0$ (right curve) with an illustration of the probability for detecting trend (power = 1- β). Modified from Tiao et al. (1990).

Please note that the assumption used to generate Figure 1 above is that both the null and alternative hypothesis are normally distributed. While this assumption may not be tenable in some cases where statistical power analysis may be needed, Figure 1 serves to illustrate the relationships between the factors that affect statistical power.

3 Statistical principles of selected methods for trend analysis

The general concept of trend analysis is that the environmental data series is a function of a time-dependent trend and a random or stochastic variation (error). Mathematically, this may be written as:

$$Y_t = f(t) + \varepsilon_t \qquad \text{(Equation 3)}$$

Where:

Y_t, $t = t_1, t_2, t_3, \ldots, t_n$, is the data series. This may be the concentration of a specific pollutant in the air;

$f(t)$ is a fixed function of time;

ε_t is the stochastic random variable or the noise term.

The null hypothesis is that the data series $f(t)$ is not dependent on time but is constant. The alternative hypothesis is that $f(t)$ is not constant over time but is some monotonic function of time. The idea of trend analysis is to decide on a test statistic and verify that the term ε_t satisfies the assumptions of the test statistic. If not, adjustments are made to fulfil the test assumption, and the test statistic is constructed, and the null distribution is derived (Yue et al. 2002). Thus the various methods of trend analysis in environmental science literature all aim at testing the hypothesis that $f(t)$ in Equation 3 is a constant, or at least changing in a monotonic but negligible manner, such that $f(t)$ can be regarded as a constant, assuming the ε_t are well-behaved. The inability of the ε_t terms to meet test assumptions has resulted in different methods of trend analysis. While several methods exist in the scientific literature for analyzing trend, the focus of this report is on the linear regression and the nonparametric methods (Theil-Sen and Mann-Kendall). Specifically, the linear regression methods discussed in this report are the generalized autoregressive conditional heteroscedasticity (GARCH) and the autoregressive integrated moving average (ARIMA).

3.1 Nonparametric methods

3.1.1 The Mann-Kendall test

The use of nonparametric methods such as Mann-Kendall analysis (Mann 1945, Kendall 1975) and Theil-Sen trend estimator (Theil 1950, Sen 1968) for trend analyses in environmental sciences is very popular (e.g. Kulkarni and von Storch 1995, Wang and Swail 2001, Yue et al. 2002, Zhang and Zwiers 2004). This is because environmental data are often highly skewed, have censored or missing data due to detection limits, or have significant outliers due to the occurrence of environmental extremes. Thus, nonparametric methods, which do not assume any distribution, are often the most suitable in such circumstances for analyzing trends (e.g., Lins and Slack 1999, Douglas et al. 2000, Zhang et al. 2001, Yue et al. 2002). Amongst these nonparametric methods, the Mann-Kendall test and the Theil-Sen slope are the most frequently used for analyzing trends in environmental sciences. These methods are currently implemented in tandem in the current tool used by AEP for trend analysis.

The Mann-Kendall test also called Kendall's tau test due to Mann (1945) and Kendall (1975), is the rank-based nonparametric test for assessing the significance of a trend. The assumption here is that a sample of data Y_i, i =1, 2, 3. . . , n are independent and identically distributed (IID) i.e., the data points are a collection of independent observations with no trend. The alternative is that a monotonic or some other form of trend exists in the data; i.e., conditional on time or any sequence in which the data points were collected, a trend can be observed in the data. On that basis, Kendall's tau statistic, which measures some interdependence amongst the observations, is calculated and its statistical significance assessed. The Kendall’s tau statistic in itself is not a measure of trend, but rather an indicator of whether a trend exists.

Kendall’s tau statistic is based on rank sums and calculated by:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(Y_j - Y_i)\,; \quad j > i \qquad \text{(Equation 4)}$$

Where:

$$\operatorname{sgn}(\mathrm{x}) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{cases} \qquad \text{(Equation 5)}$$

The distribution of *S* (calculated above) under the null hypothesis (H_0) is symmetrical and is normal in the limit as *n* becomes large (Wang and Swail, 2001). Under H_0, the mean of *S* is zero and, in case of no ties (e.g., no multiple values for the same sampling time), the variance of *S* is given by:

$$V_S^2 = \frac{n(n-1)(2n+5)}{18} \qquad \text{(Equation 6)}$$

3.1.2 The Theil-Sen slope

The Theil-Sen method (Theil 1950, Sen 1968), which is based on Kendall’s rank correlation, is used to estimate trend and is commonly used in combination with the Mann-Kendall test to provide both an estimate and a test for trend (e.g., Yue et al. 2002). The Theil-Sen method considers measurements Y_1, Y_2, Y_3, . . ., Y_n of an environmental variable (say the concentration of a pollutant) taken at times t_1, t_2, t_3, . . ., t_n, where $t_1 \le t_2 \le t_3 \le . . . \le t_n$ (the time intervals do not necessarily have to be equal), as independent observations. The gradient D_k, (*k=1,2,3,. . . N*), for each *N* pairs of observations taken at times t_j and t_i such that $1 \le i \le j \le n$ and $(t_j - t_i) > 0$, can be calculated as:

$$D_k = (Y_j - Y_i)/(t_j - t_i) \qquad \text{(Equation 7)}$$

The estimate of trend ($\hat{\beta}$) in the data series Y_1, Y_2, Y_3, . . ., Y_n can then be calculated as:

(Equation 8)

$$\hat{\beta} = \begin{cases} D_{((N-1)/2)+1} & \text{if } N \text{ is odd,} \\ \left(D_{N/2} + D_{(N/2)+1}\right)/2 & \text{if } N \text{ is even.} \end{cases}$$

The above represents an empirical nonparametric calculation of the median of D_k. The $(1 - \alpha)$ confidence interval for $\hat{\beta}$ may be calculated as follows (Wang and Swail 2001):

1) Compute M_1 and M_2 using the estimate (V_S) from the Mann-Kendall test described above and the $(1 - \alpha/2)$ quantile of the standard normal distribution ($Z_{(1-\alpha/2)}$) as:

$$M_1 = \frac{\left(N - Z_{(1-\alpha/2)}V_S\right)}{2} \qquad and \qquad M_2 = \frac{\left(N + Z_{(1-\alpha/2)}V_S\right)}{2}, \qquad \text{(Equation 9)}$$

2) Determine the order statistics D_{M_1} and D_{M_2+1} as the lower and upper $(1 - \alpha)$ confidence limits respectively from the collection of the N gradients (D_k).

An alternative approach for calculating the confidence interval of a Theil-Sen slope is to use a bootstrap simulation method (e.g., Wilcox, 2005). The recent version of the TFPW uses the bootstrap approach to estimate confidence intervals of the Theil-Sen slope estimates of the trend. Details of the bootstrap simulation procedure are presented in Section 4.

3.1.3 Dealing with autocorrelation in nonparametric trend analysis

The null hypothesis in the Mann-Kendall test is that the data points are independent and randomly ordered and, as discussed above, the Theil-Sen slope is also calculated based on the same assumption. However, the effect of autocorrelation causes individual observations in the data to be dependent, thus violating this assumption.

Positive autocorrelation has been shown to increase the probability of detecting a trend in a Mann-Kendall test when in fact none may exist. Hamed and Rao (1998) investigated this problem and derived a theoretical relationship to calculate the variance of the Mann-Kendall trend test statistic for autocorrelated data. They then proposed a modified non-parametric trend test, which is suitable for autocorrelated data, based on the modified value of the variance of the Mann-Kendall trend test statistic.

Earlier work by Kulkarni and von Storch (1995) and a simulation study by Yue et al. (2002) have also demonstrated that positive autocorrelation may result in increased false positives for both the Theil-Sen and Mann-Kendall methods. Based on the work by Yue et al. (2002), the recursive or iterative pre-whitening procedure has been summarized in the following steps:

1) Estimate the slope ($\hat{\beta}$) of β of the series X_t using the Theil-Sen method (Sen 1968) and then use the estimated slope ($\hat{\beta}$) to de-trend the data series X_t with:

$$\theta_i = X_t - \hat{\beta}t \qquad \text{(Equation 10)}$$

2) The de-trended data series θ_i is used to estimate the first order correlation coefficient $\hat{\rho}$.

3) The estimate of $\hat{\rho}$ obtained from step 2 is used to pre-whiten X_t:

$$X_t' = (X_t - \hat{\rho}X_{t-1})/(1-\hat{\rho}) \qquad \text{(Equation 11)}$$

4) Re-estimate a new trend parameter $\hat{\beta}$ of the pre-whitened series using the Theil-Sen estimator.

5) Repeat steps 2 to 4 until a new estimate of $\hat{\rho}$ is less than 0.05 and there is no significant change in $\hat{\beta}$, as specified in the initial criteria. The final Theil-Sen estimate $\hat{\beta}$ is taken as the estimate of trend. The Mann-Kendall trend test is then applied to the final pre-whitened series to test the null hypothesis of no trend in the data series.

3.2 Linear regression analysis

Regression analysis is one of the conventional methods of analyzing trends in environmental data (including contaminants). The regression method estimates a variable, often the concentration of a pollutant as a well-defined function of time with all applicable standard regression assumptions duly tested and accounted for, to appropriately test the null hypothesis of no trend. The linear regression method is the most common and assumes that $f(t)$ in Equation 3 is a linear function of time, thus resulting in Equation 12 below:

$$Y_t = \alpha + \beta t + \varepsilon_t \qquad \text{(Equation 12)}$$

Where:

Y_t is the environmental observation at time t;
α is a regression parameter representing the intercept;
β is a regression parameter (slope) that estimates the trend of the environmental observations;
ε_t is the random error (white noise) term associated with each observation.

Given that all the assumptions are met and that the trend is indeed linear, Equation 12 may be fitted to the environmental observations by ordinary least squares (OLS) to test the hypothesis that the trend (β) is statistically significant. From purely statistical perspectives, the null hypothesis of no trend can only be tested using Equation 12 if the following conditions are met:

1) The relationship between the environmental data series Y_t and time is linear in the parameters α and β; i.e., Equation 12 is the true specification of the relationship between the environmental observations and time.

2) The error (ε_t) is assumed to be distributed normally with mean = 0 and variance = σ^2; i.e., ε_t ~ IIDN(0, σ^2). The reference distribution used to test the null hypothesis of no trend is Student's t-distribution, which is a small sample version of the standard normal curve, with mean 0 and a sample-based variance of s^2/n. An important characteristic here is symmetry – Student's t-distribution (like the standard normal) is symmetric about the mean, 0. Thus, skewness can result in the lack of symmetry.

3) The error (ε_t) is assumed to be an independent random variable, i.e., there is no correlation (no autocorrelation) amongst the ε_t's. The idea here is that the error term in the current period t is not dependent on either the error term in the previous period (t - 1), the error term in the following period (t + 1), or on error term in any other period (t + i).

4) The error (ε_t) is assumed to be identically distributed, i.e., ε_t has a constant variance σ^2 across time (homoscedasticity).

5) Time is a non-stochastic variable, i.e., time is not recorded in error by the measurement or monitoring instrument. If time is recorded in error, i.e., if instead of t, t + ξ is recorded, where ξ is random with an unknown distribution, then t is no longer non-stochastic, and the standard linear OLS estimation does not apply.

Data from natural systems seldom meet these assumptions, resulting in the OLS approach not being used. If these assumptions are not met, the OLS estimate of the trend (β) may still be unbiased but may no longer be efficient, as the standard error of the estimate for β is poor. Consequently, any conclusions drawn about the trend may be wrong. Other estimators of linear regression had to be explored; usually requiring that some adjustment is made to the data. As already discussed above, air quality data are typically associated with the statistical issues which tend to violate the assumptions described previously.

A common approach in linear regression estimation is to transform the data to achieve normality and uniform variance. For instance, taking the log-transformation of a lognormal dataset will result in a dataset that is normally distributed. While the interpretation of the estimated coefficients may not be straightforward, the transformation will allow for valid hypothesis testing and will either confirm that the estimated trend in the original units is either statistically significant or not. An example of a common transformation that can be used for trend analysis of air quality data is the log-linear transformation of the form:

$$\log(Y_t) = \alpha + \beta t + \varepsilon_t \qquad \text{(Equation 13)}$$

This equation can be used to transform the series Y_t to achieve normality. In such a case, the literal interpretation of trend β no longer applies; a unit increase in time no longer results in a β unit increase in Y_t, but instead, a new value $Y_{t+1} = Y_t e^{\beta}$ is created. Thus, a unit increase in time does not result in a constant increase in Y_t anymore. A consequence of a log-linear transformation of Y_t is to infer a nonlinear trend between Y_t and time, although if β is sufficiently small, the trend will be approximately linear for the most part. Thus, data transformation to meet the above statistical assumptions may sometimes result in problems with the interpretation of the results.

3.2.1 Dealing with autocorrelation and heteroscedasticity: The generalized least squares and related approaches

The effects of autocorrelation and heteroscedasticity can be solved by using an estimate $\hat{\rho}$ of the coefficient of autocorrelation and a variance function in a generalized least squares (GLS) framework. Details of the implementation of GLS and related methods can be found in standard linear regression textbooks (e.g., Neter et al. 1983, Johnston 1984, Graybill and Iyer 1994, Kariya and Kurata 2004, Baltagi 2008). The goal is to obtain an efficient estimate of the standard error of the regression coefficient β (trend) using the coefficient of autocorrelation ($\hat{\rho}$) and the residual variance function, in order to correctly test the null hypothesis of no trend. On that basis, several related methods have been developed over the years to address this issue. We describe briefly, some of the common procedures below.

Although the methods may be different in nomenclature, they address similar issues to improve the reliability of hypothesis testing about the regression coefficients: autocorrelation and heteroscedasticity. The weighted least squares and iterative weighted least squares (e.g., Cleveland 1979, Carroll and Rupert 1988, Cleveland and Durvlin 1988, Ryan 1997) are regression estimation methods designed to address the issue of non-uniform variance using a residual variance function. The generalized least squares methods (e.g., Johnston 1984, Kariya and Kurata 2004, Baltagi 2008) and mixed effects modeling (e.g., Pinheiro and Bates 2000) define more generalized error structures using estimates of the coefficient of autocorrelation and residual variance function to address both autocorrelation and non-uniform variance. In both cases, it is assumed that the residual variance can be approximated by a simple function which can be applied with the estimate of the coefficient of autocorrelation to address both autocorrelation and heteroscedasticity.

Another class of regression estimators worthy of mentioning is the robust estimators. Robust estimators have been used extensively to address issues with outliers in regression analysis (e.g., Rousseeuw and Leroy 1987, Ryan 1997, 2008). Other research works have developed robust estimators to improve statistical inference in the presence of heteroscedasticity (e.g., Tofallis 2008) and autocorrelation (Hardin and Hilbe 2012). Also, robust estimators such as robust minimax designs (Wiens 1998, 2000) have also been developed to improve statistical inference in model misspecification and heteroscedasticity in linear regression analysis. These are all intended to improve inferences on the regression parameters in Equation 12, which in this case includes time trend β.

3.2.2 Dealing with autocorrelation and heteroscedasticity: the ARCH, GARCH and ARIMA approaches

In some cases such as in stock markets, the non-uniform residual variance (heteroscedasticity) may be difficult to approximate with a simple variance function. The extremely volatile nature of stock market data series often give rise to non-uniform variance with no defined pattern, making it difficult to model using the traditional generalized least squares regression estimators. The Autoregressive Conditional Heteroscedastic (ARCH) models (Engle 1982) and the various forms of Generalized Autoregressive Conditional Heteroscedastic (GARCH) models (e.g. Bollerslev 1986, Nelson 1991, Ding et al. 1993, Zakoïan 1994) were developed to address the volatility (extreme non-uniform of variance) associated with the wide variation of trading prices in stock markets over time

GARCH models have been widely applied in econometrics for volatility modeling and was originally proposed by Engle (1982) as Autoregressive Conditional Heteroscedasticity (ARCH) for modeling the conditional variance of a financial time series. Bollerslev (1986) further developed the idea by including a lagged conditional variance term. This term helped smoothen the conditional heteroscedastic function to produce the GARCH models.

GARCH models have rarely been applied in environmental sciences until recently by Wang et al. (2005), Wang et al. (2006), Chen et al. (2008) and Modarres and Ouarda (2013). Wang et al. (2005) applied GARCH-type models for analyzing streamflow data from the northeastern Tibet Plateau of China and discussed the advantages of GARCH models for addressing conditional heteroscedasticity in daily and monthly streamflow data. Chen et al. (2008) developed an algorithm based on the GARCH type models for modeling time series of 10-day streamflow of the Wu-Shi River in Taiwan and showed that GARCH models are superior for modeling time series compared to the traditional linear time series models. In Canada, Modarres and Ouarda (2013) used the GARCH approach for modeling daily streamflow data from the Matapedia River in the province of Quebec. They compared the GARCH models to the autoregressive integrated moving average (ARIMA) models and concluded that GARCH models are superior to the ARIMA models for modeling conditional heteroscedasticity.

The relevance of this method for analyzing air quality data is that the non-uniformity in the variance (heteroscedasticity) in air quality data may not follow a specific pattern, but can vary widely within short periods (like stock market) in response to changes in ambient environmental conditions. These events could be episodic events such as wildfires and may cause significant variation in the concentration of a specific pollutant during those events. Addressing this type of heteroscedasticity for air pollution data to provide efficient estimates of standard errors for hypothesis testing require the capability of methods such as ARCH and GARCH.

The basic idea of applying GARCH to linear regression models is as follows:

1) If we assume that a linear regression model is:

$$Y_t = \alpha + \beta t + v_t \qquad \text{(Equation 14)}$$

Where Y_t, α, β, and t are as defined before (see Equation 3), and preliminary analyses show that the assumptions for regression analysis that the error term v_t is IIDN(0, σ^2) are not met; this suggests that v_t is not well-conditioned and that some adjustments need to be made.

2) Because air quality data are usually time series, the error term v_t will likely follow the autoregressive process (i.e., autocorrelation may be present). Thus, the qth order autoregressive process (AR(q)) is used to model v_t as:

$$v_t = \sum_{i=1}^{q} \rho_i v_{t-i} + \varepsilon_t \qquad \text{(Equation 15)}$$

Where ρ_i is the ith order coefficient of autocorrelation and v_{t-i} is the ith lag of v_t, i = 1, 2, 3, . . ., q.

The error term ε_t in Equation 15 now consists of independent observations but with non-uniform variance (σ_t^2).

3) The GARCH approach is to model the conditional variance (σ_t^2) of ε_t as:

$$\varepsilon_t = \sigma_t \epsilon; \qquad \epsilon \sim \mathrm{IIDN}(0,1);$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{V} \gamma_i \varepsilon_{(t-i)}^2 + \sum_{j=1}^{M} \theta_j \sigma_{t-j}^2. \qquad \text{(Equation 16)}$$

Where ω, γ_i and θ_j are parameters to be estimated. The above formulation is called the GARCH (V, M) model. Both autocorrelation and heteroscedasticity are modelled in the same process resulting in better estimates of the standard error of β. This enables a more robust hypothesis testing about β. Combined with the autoregressive process in step (2) above, the modelling framework is called the AR(q)-GARCH (V, M). This can be implemented both in standard statistical packages such as SAS (SAS Institute Inc. 2005) and in R (R Core Team 2018). Note that the difference between the ARCH formulation as initially proposed by Engle (1982) and the GARCH models is the third term on the right-hand side of Equation 16. The ARCH models do not include the term $\sum_{j=1}^{M} \theta_j \sigma_{t-j}^2$ in Equation 16. Also, note that Equation 16 is only one version of the GARCH formulation and is comparatively the simplest. A few examples of more complex versions of GARCH typically used in extremely volatile series such as stock markets are: the nonlinear asymmetric GARCH (NAGARCH) (Engle and Ng 1993), the exponential GARCH (EGARCH) (Nelson 1991), the quadratic GARCH (QGARCH) (Sentana 1995), and the Glosten-Jagannathan-Runkle GARCH (GJR-GARCH) (Glosten et al. 1993).

A class of models popular in univariate time series analysis and commonly known as the Box-Jenkins models (Box and Jenkins 1968) are good candidates for modeling residuals in a regression analysis framework. These classes of models, known as the autoregressive integrated moving average (ARIMA) models, use the combination of the autoregressive (AR) process, time differencing of the series (integration (I)) and moving average (MA) to model and make forecasts of time series. These are written as ARIMA(*p, d, q*), where *p* is the order of the autoregressive process, *d* is the time interval for differencing to make the series stationary, and *q* is the moving average window. The most common in this class of models are the autoregressive moving average models (ARMA) models since in most applications the time series is differenced to achieve stationarity before analysis.

Referring to Equation 14 above, ARIMA can be applied to model the residual component (v_t) as follows:

$$\begin{cases} v_t \; = \displaystyle\sum_{i=1}^{p} \rho_i v_{t-i} + \varepsilon_t \quad (AR(p)) \\ \varepsilon_t \; = \displaystyle\sum_{i=0}^{q} \gamma_i \varepsilon_{t-i} \qquad\quad (MA(q)) \end{cases} \qquad \text{(Equation 17)}$$

In Equation 17 above, the autoregressive process models the autocorrelation of v_t while the MA component is used to smoothen the white noise component of v_t (ε_t) to achieve some form of uniformity in the variation of the white noise. The similarity of the formulation in Equation 17 with Equation 16 should be noted; the MA(*q*) in Equation 17 and the ARCH component of Equation 16 are similar, except that the ARCH part of Equation 16 models the residual variance while the MA(*q*) component models the residuals.

3.3 Summary

Whether parametric or nonparametric, each method has its strengths and drawbacks primarily due to the nature of environmental data – the presence of uncertainties due to sampling and measurement errors, distributional issues, the occurrence of extreme values, autocorrelation and sometimes uneven data censorship due to variable minimum detection or reporting limits.

Autocorrelation is the primary statistical challenge in trend analysis that the methods presented in this report will address. The parametric approach (Tiao et al. 1990, Weatherhead et al. 1998, 2000), and the nonparametric methods of the Mann-Kendall test (Mann 1945, Kendall 1975) used with the Theil-Sen estimator of a trend (Theil 1950, Sen 1968) address autocorrelation differently. However, both methods are intended to provide statistically robust estimates and inferences about the trend. Pre-whitening the data series before conducting nonparametric analysis helps obtain robust estimates of trend test statistics for nonparametric methods, while the linear regression methods model the residual errors to address both autocorrelation and heteroscedasticity to obtain efficient estimates and tests for trend.

4 Guidance for trend analysis of air quality data

4.1 The purpose of this guidance

This section provides general guidance on how to conduct trend analysis and provides specifics on how to use the three approaches discussed above to conduct trend analysis. Also, based on the need for shorter-term analysis of trends, this section also provides guidance on how to use monthly data and what issues to expect and how to deal with these issues if monthly data are used for trend analysis.

4.2 Data preparation

The primary source of air quality data used in this report is the AEP airdata website (formerly known as the Clean Air Strategic Alliance (CASA) data warehouse). The data for each station considered for trend analysis must be checked to ensure that there are no obvious data gaps or errors. Since most station data stored in this database would have undergone the required QA/QC checks, the flags associated with the values should be checked to resolve any data issues before rejecting or accepting the value. The hourly data must be compiled to obtain monthly averages using the following suggested rules:

- 50% of the total hourly observations must be available before a particular day’s observations are included in determining the monthly average. This requirement may be a “low bar” to cross when compared to the data completeness criteria of 75%, typically used in Alberta for daily data. But as the objective is to estimate the monthly average, 50% represent a very descent sample size.
- For a particular month to be included for trend analysis, the month must have valid observations (including values below detection limits) available for at least 50% of all hours in the month. If not, the month should be assigned a missing value but included in

the dataset. Note, 50% is considered a fairly good sample for estimating monthly averages.

- Hourly missing values are treated as missing observations, while values below detection limits may be replaced with 0. This is already the case for the data from continuous air monitoring stations that are stored in the airdata warehouse. As this guidance is for continuous monitoring data, replacing non-detects with 0 is not expected to have a significant impact on trend estimates for average concentrations. As non-detects are more likely in months with low concentrations and less likely for months with high concentrations of the target pollutant, replacing them with 0 should not distort the trend for the average concentrations. Moreover, with the exception of SO_2 which is typically low at non-industrial sites where non-detects are more likely to occur at higher percentages, non-detects are typically low for other major pollutants such as NO_2 such that their impact on the data is minimal.
- It is desirable to first aggregate data from hourly to daily averages before determining monthly averages. This is to ensure that all daily observations are equally weighted.
- Using the data completeness criteria above may result in missing data for some months. In such cases, decision on whether to include the affected year should be made based on whether and how the missing could impact the trend estimate. If the number and pattern (random, continuous or systematic) of missing data in a particular year could significantly impact the trend then the whole year’s data should be excluded from the analysis. We recommend that only continuous yearly worth of data be used for trend analysis; annual (continuous 12 months) data gaps should be avoided. As a guide missing monthly data should not exceed 30% (4 months) if they are consecutive months or 50% (6 months) if the pattern is random.

It is highly preferable to use whole year (from January to December) values. If partial years are used, the calendar month needs to be identified and use it as the starting value. For example, if the time window starts from May 2005, the starting value of the time variable is 5; subsequent months are then numbered in that order. ***The position of the month in the year determines its season and is very important for the method described in this document for modeling seasonality***. If a particular month has no valid record, that month should be assigned a missing value but included in the data. However, if a specific tool such as the “*OPENAIR*” package in R is to be used, requirements for the proper use of the tool should be followed. The suggestions herein are applicable for the tool being packaged based on this report.

4.3 Preliminary data analysis

The statistical issues present in the dataset should be identified before the trend analysis is conducted. It is recommended that more than one method be used for trend analysis so that the results can be compared. If statistical inferences made from the different methods are consistent, i.e., both of them rejecting or accepting the null hypothesis, then the statistical decision about trend will have high confidence. Note, because the pre-whitening step is based on the first order autocorrelation, the nonparametric method may produce less reliable results for a higher order of autocorrelation. The pre-whitening step is to “wash-out” autocorrelation and may be successful in doing so if the first-order autocorrelation is the issue. In most cases, monthly averaging time periods may be enough to limit the autocorrelation problem to a first-order autocorrelation problem.

The first task is to examine the data for any unusual distributions and the presence of any apparent outliers and influential values. Simple histogram plots may help display the distribution of the data; a bi-modal pattern might appear due to strong seasonality, particularly for NO_2. The presence of outliers may result in statistical contamination of the data and may lower the statistical power to detect a trend. However, this will likely be a lesser issue in air quality data, especially if hourly data are rolled-up into monthly data. However, if there is a significant amount of outliers in the monthly average themselves, the dataset may need to be investigated further for issues that might have caused such extreme values.

US EPA (2006, page 115-120) presents an extensive discussion of diagnostic tests to identify potential outliers and guidance on how to select a statistical test for outliers. A step-by-step procedure for conducting an outlier test using selected methods is also presented. However, for this guidance, no detailed discussion of outlier analysis is included, as outliers are not currently considered as a serious issue for trend analysis of air quality data when monthly averages are used. The amount of hourly data averaged to obtain monthly data (744) is large enough to significantly reduce the impact of outliers if they are only a few data points. Furthermore, data stored in the EMSD airdata warehouse go through significant QA/QC checks to ensure that errors are identified and removed, corrected or flagged.

As a general guide however, monthly average data points suspected to be outliers should be included for further analysis unless there are good reasons recorded in the EMSD airdata warehouse to indicate that those values are not valid data points. Values suspected to be outliers may be influential values reflecting the occurrence of unusual environmental and not necessarily erroneous data points, and thus should not be discarded. If a monthly average has been confirmed to be an error, it should be designated as a missing value. Simple methods such as graphical time series plots of monthly averages may help identify suspected outliers for further investigation.

4.3.1 Accounting for seasonality

Ambient concentrations of air quality parameters tend to have seasonal patterns, which adds more variability to the data and lowers the statistical power for detecting a trend. While nonparametric smoothing methods such as LOESS (local regression) (Cleveland 1979) can be used to model and account for seasonality, the harmonic regression typically used in time series analysis to examine the cyclical behavior of time series is recommended. The goal is to decompose the time series to remove the seasonal (harmonic) component of the series. Given that the LOESS smoothing method is flexible and is heavily data dependent, there is an increased likelihood that part of the trend may be removed through the de-seasonalizing process, especially if the smoothening is not applied correctly. Thus the resulting de-seasonalized series may show a trend that is smaller than the actual trend in the original series.

As presented in Tiao et al. 1990 and Weatherhead et al. (1998, 2000), the combination sine and cosine functions are used to model seasonality (S_t) shown in Equation 18:

$$S_t = \sum_{j=1}^{\lambda} \left\{ \beta_{1,j} sin\left(\frac{2\pi jt}{12}\right) + \beta_{2,j} cos\left(\frac{2\pi jt}{12}\right) \right\} \qquad \text{(Equation 18)}$$

Where: λ is the number of sine and cosine terms needed to completely account for seasonality (a maximum of 4 may be enough according to Weatherhead et al. (1998, 2000)), $\beta_{1,j}$ and $\beta_{2,j}$ are the parameters associated with the j^{th} sine and cosine terms respectively.

4.3.2 Test for autocorrelation

The next and perhaps the most important task was to test for autocorrelation. Since monthly data is being used as opposed to annual data, the likelihood of autocorrelation is higher. Most standard statistical packages, such as the R-software, can do this test using an autocorrelation function (ACF) or partial autocorrelation function (PACF). Most statistical packages are capable of producing the Durbin-Watson statistic (Durbin 1969), a test statistic for autocorrelation. When there is no autocorrelation, this statistic is approximately equal to 2. As a guide, however, values between 1.5 and 2.5 usually indicate there is no autocorrelation. Values smaller than 1.5 indicate positive autocorrelation while values greater than 2.5 indicate negative autocorrelation. The statistical package may include probability of significance of the Durbin-Watson statistic to help with the statistical inference on autocorrelation.

The following procedure can be followed to test for autocorrelation manually using simple statistical tools, such as Microsoft Excel:

1) Deseasonalize and de-trend the monthly data using simple linear regression of the form:

$$Y_t = \alpha + S_t + \beta t + v_t \qquad \text{(Equation 19)}$$

Where: α is the regression intercept, S_t is seasonality term (Equation 18), β is the time coefficient (trend) of the data and v_t is the error term assumed to be autocorrelated. De-seasonalization is a critical step, as monthly data is affected by seasonal variations. The parameters α, β, $\beta_{1,j}$ and $\beta_{2,j}$ can be estimated by the preliminary regression fit of Equations 19.

2) Calculate the error terms $v_t = Y_t - \left(\hat{\alpha} + \hat{S}_t + \hat{\beta}t\right)$. Note that the “hat” designations are meant to show that the above variables are estimates.

3) Calculate the coefficients of autocorrelation ρ_i by estimating the regression equation given in Equation 15 above.

For monthly data, a maximum of *q* = 4 may be enough, and for the case study reported in the later section, only the first order autocorrelation coefficient was statistically significant in most of the datasets. Therefore for most analyses, the first order autocorrelation will be sufficient.

4) Regardless of the presence and the nature of autocorrelation, it is recommended that more than one trend analysis method be used so that the results can be compared. The diagnostics should only confirm the implementation details to use for each method, e.g., in the case of GARCH and ARIMA, the autocorrelation diagnostics should provide guidance on which values of q to use in the autoregressive part as shown in Equation 15 and Equation 17. If both first and second order autocorrelations are significant, the q should be set to 2. Note, the current setup of Theil-Sen/Mann-Kendall method requires the diagnoisis of autocorrelation for pre-whitening to be implemented, the order of autocorrelation is not needed.

4.3.3 Test for heteroscedasticity

This test is useful for only the parametric option. Although often subjective, graphical plots of the error terms (v_t) against independent variables (in this case time) or the approximate predicted terms $\hat{Y}_t = \hat{\alpha} + \hat{S}_t + \hat{\beta}t$ are the easiest way to assess non-uniform variances. If the pattern of the scatter plot of v_t against, for example, $\hat{Y}_t$ or *t* does not show noticeable non-uniformity (e.g., see Figure 2(a)), then there is no heteroscedasticity. Figure 2(b) shows a clear case of non-uniform variance, as the points appear to be more widely scattered with increasing time.

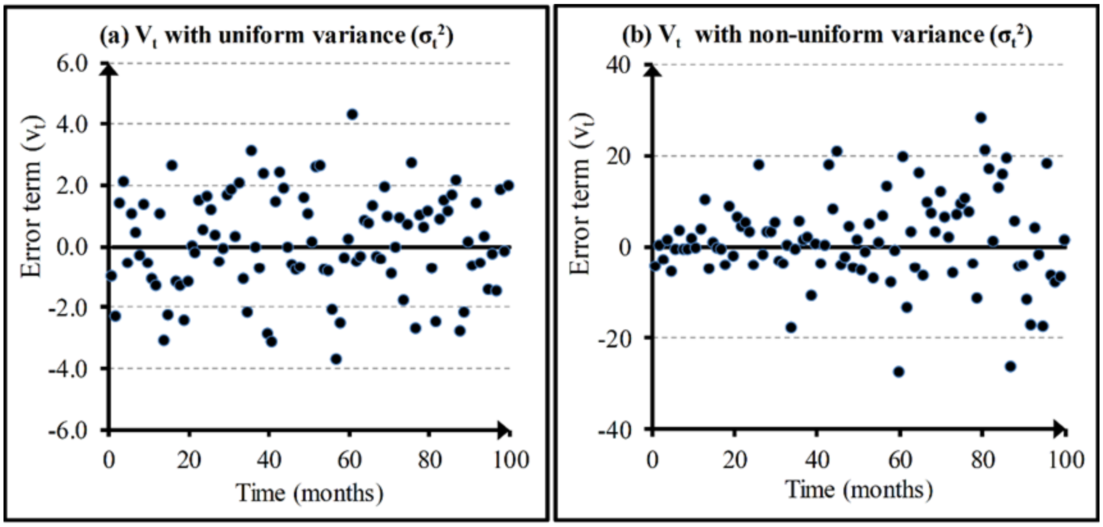


Figure 2- Variation of error variance with the independent variable time. Figure 2(a) shows the case of uniform variance with the points scattered almost evenly across time, while Figure 2(b) shows the case of non-uniform variance, where the scatter of the points appear to be increasing in width with time. Note that the data presented in this Figure were created for illustration; they are not real air quality data.

While plots like Figure 2 may be enough to test for non-uniform variance, more sophisticated statistical tests do exist in some standard statistical packages. For example, the modified Breusch-Pagan test (Breusch and Pagan 1979) and the White test (White 1980) are both available in SAS. Note that the plots like those in Figure 2 may not show a clear picture of non-uniform variance if the sample size is small, in which case, the test can be done by including an option in the command syntax in the statistical package being used.

Briefly, the Breusch-Pagan test can be conducted as follows using the null hypothesis of homoscedasticity (no heteroscedasticity):

i) Fit Equation 19 to the data using OLS and estimate the regression residuals ($\hat{v}_t$);

ii) Regress ($\hat{v}_t{}^2$) on the k independent variables in the original regression equation (i.e., Equation 19) and obtain the goodness of fit statistic R^2. Note: for trend analysis, time and seasonality variables are the independent variables;

iii) The Breusch-Pagan test can be done using either F-statistic or the Lagrange Multiplier (LM) statistic. Calculate the F-statistic as: $F = \frac{R^2(n-(k+1))}{k(1-R^2)}$ if the F-statistic is chosen for the test or the Lagrange Multiplier (LM) statistic as: $LM = nR^2$, if the LM statistic is chosen for the test; where n is the sample size;

iv) Calculate the p-value for the F-statistic using F-distribution with (k, n-(k+1)) degrees of freedom, if the F-statistic is used for the test or the p-value for the LM statistic using χ^2 distribution with k degrees of freedom, if the LM statistic is used for the test;

v) If the p-value is smaller than a critical probability value set prior to the analysis (e.g., 0.05) then the null hypothesis of homoscedasticity is rejected, i.e., there is heteroscedasticity.

The steps for conducting the White test are similar to the Breusch-Pagan test outlined above except that, in step ii) above, $\hat{v}_t{}^2$are regressed on the independent variables as well as their squares and cross-products to obtain the R^2. This is equivalent to regressing $\hat{v}_t{}^2$ on the predicted dependent variable $\hat{Y}$ and its square $\hat{Y}^2$ to determine the R^2. The R^2 is then used to complete the White test following steps iii) to v) above. Note that the Breusch-Pagan and the White test both use R^2 to calculate the test statistic. The only difference is how the R^2 is obtained.

4.4 Nonparametric trend analysis

As described in Section 3, the nonparametric option described in this document is a combination of the Mann–Kendall test (Mann 1945, Kendall 1975) and the Theil-Sen trend estimator to provide both the test for the presence of trends as well as the magnitude and confidence limits of the trend. Although details of the principles and theories of these two methods have been discussed in Section 3, additional practical details of the two methods are presented below.

4.4.1 The Mann-Kendall trend test

Recall the discussion in Section 3.1 that for a given sample of *n* random variables (X_1, X_2, . . . , X_n), the Mann-Kendall statistic *S* is calculated by Equations 4 and 5. The distribution of *S* is symmetrical and normal if a minimum sample size is met (Wang and Swail 2001). It must be emphasized here that adequate sample size is required for this condition to hold. As indicated in Yue and Wang (2004), the minimum sample size required is 8. From Section 3.1 above, the mean of *S* is zero, and the variance is given by Equation 6. Based on these properties, the test statistic can be computed (Wang and Swail, 2001, Yue and Wang 2004) as:

$$Z = \begin{cases} (S-1)/V_S & \text{if } \ x > 0 \\ 0 & \text{if } \ x = 0 \\ (S+1)/V_S & \text{if } \ x < 0. \end{cases} \qquad \text{(Equation 20)}$$

All variables in Equation 20 are defined as in Equations 4 and 5. Given that the *S* statistic is approximately normal, the test statistic Z is also approximately normal, and $Z_{(1-\alpha/2)}$ is set as the critical value; where $\Phi(Z_{(1-\alpha/2)}) = \alpha/2$ and $\Phi(\dots)$ is the standard normal cumulative distribution function and α is the significance level of the test (Hirsch et al. 1982). A conclusion of no trend is made if $|Z| < Z_{(1-\alpha/2)}$. A positive Z value indicates an increasing trend and, a decreasing trend is indicated by a negative Z value (Wang and Swail 2001). The Mann-Kendall test can be implemented manually using the above procedure and formulas in Equations 4, 5, 6 and 20.

4.4.2 The Theil-Sen slope estimator

The Theil-Sen method (Sen 1968), which is based on Kendall's rank correlation, is used to estimate trend ($\hat{\beta}$) as presented in Equations 7 and 8. Although the confidence interval of $\hat{\beta}$ can be calculated as described in Wang and Swail (2001) and presented in Section 3.1.2 above, the process of determining the confidence interval of the median presented in Gilbert (1987) may also be used. The use of bootstrap simulations (e.g., Wilcox, 2005) to calculate the confidence interval of $\hat{\beta}$ are also common in recent years. A brief outline of the bootstrap method is presented below. The recent version of the TFPW uses the bootstrap approach to estimate confidence intervals of the Theil-Sen slope estimates of trend. In this report, however the Mann-Kendall test present in Section 4.4.1 is used for statistical inference.

4.4.3 Implementing nonparametric methods

Since the Mann-Kendall method relies on data independence to produce a valid test for trend, if the preliminary tests above establish the presence of autocorrelation, then the observations are not independent, and modification has to be made. Earlier research work by Kulkarni and von Storch (1995) and a simulation study by Yue et al. (2002) have demonstrated that positive autocorrelation may result in an increased occurrence of false positives. Following the work of Yue et al. (2002), a recommended approach for implementing the iterative pre-whitening for monthly data (assuming autocorrelation has been determined to be an issue) is as follows:

1) First, because monthly data, which is affected by seasonal variation, is used for trend analysis, the data series Y_t has to be deseasonalized. This is done by subtracting the seasonality component from Equation 18 above as:

$$X_t = Y_t - \hat{S}_t = \alpha + \beta t + v_t; \\ v_t = \rho v_{t-1} + \varepsilon_t. \qquad \text{(Equation 21)}$$

Where $\hat{S}_t$ is as given in Equation 18.

2) The next step is to use the deseasonalized series X_t in Equation 21 above to estimate the slope ($\hat{\beta}$) of β using the Theil-Sen method (Sen 1968) described above. The slope estimate is then used to de-trend the data series $X_t : \theta_t = X_t - \hat{\beta} t$

3) Next, the de-seasonalized and de-trended data series θ_i is used to estimate the first order correlation coefficient $\hat{\rho}$.

4) Next, the estimate of $\hat{\rho}$ obtained from step 3) above is used to pre-whiten X_t: $X_t' = (X_t - \hat{\rho}X_{t-1})/(1-\hat{\rho})$ and a new trend parameter $\hat{\beta}$ is re-estimated using the Theil-Sen estimator.

5) Steps 2 to 4 are repeated until a new estimate of $\hat{\rho}$ is found to be less than 0.05 and there is no significant change in $\hat{\beta}$ as specified in the initial criteria. The final Theil-Sen estimate $\hat{\beta}$ is taken as the estimate of the trend. The Mann-Kendall trend test is then applied to the final pre-whitened series to test the null hypothesis of no trend in the data series.

A suggested criterion for deciding on whether the change in $\hat{\beta}$ is significant is to set a small change δ in $\hat{\beta}$, where δ is the maximum change that is considered to be negligible such that $\hat{\beta} - \delta \approx \hat{\beta}$. Once the change in $\hat{\beta}$ is less than δ, the iteration is stopped. A guide to selecting the value of δ is to examine the order of magnitude of the data being analyzed. A preliminary linear regression fit to the data may give a rough idea of the magnitude of the trend to expect. A small percentage of the estimated trend in this preliminary analysis (say 0.01%) may be a good estimate of δ. The steps for implementing iterative pre-whitening are summarized in the flowchart presented in Figure 3. The confidence interval of the slope estimate $\hat{\beta}$ may be calculated as described in Section 3.2 above or by the bootstrap simulation.

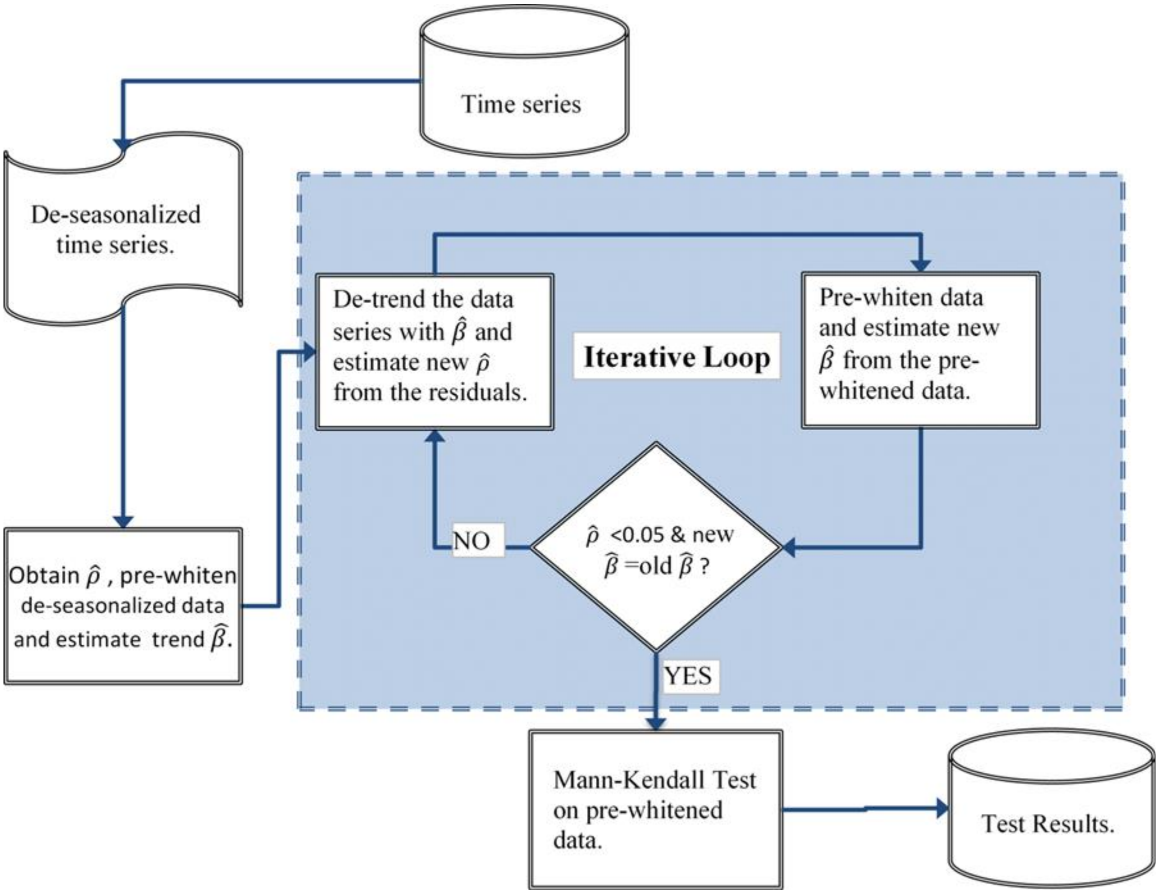

Time series
De-seasonalized time series.
Obtain $\hat{\rho}$, pre-whiten de-seasonalized data and estimate trend $\hat{\beta}$.
De-trend the data series with $\hat{\beta}$ and estimate new $\hat{\rho}$ from the residuals.
Iterative Loop
Pre-whiten data and estimate new $\hat{\beta}$ from the pre-whitened data.
NO
$\hat{\rho}$ <0.05 & new $\hat{\beta}$ =old $\hat{\beta}$?
YES
Mann-Kendall Test on pre-whitened data.
Test Results.

Figure 3- A flowchart showing the implementation process of the iterative Trend-Free Pre-Whitening method for nonparametric estimation of air quality trends used in the present study.

If a bootstrapping simulation is to be used to construct a confidence interval for the estimated Theil-Sen slope, the simulation should be done on the final pre-whitened data series. The following is a suggested procedure for implementing bootstrap simulation on the pre-whitened data series:

1) Decide on the subsample size using the minimum sample size needed to implement the Theil-Sen slope estimation successfully as a guide. Although technically a sample size as small as five can be used to implement Theil-Sen slope estimation, a smaller subsample size will likely result in a much wider confidence interval, particularly if the original sample (bootstrap population) size is much larger. The idea of bootstrapping is to calculate the Theil-Sen slopes for a sufficiently large number of subsamples and then determine the α/2 and (1 -

α/2) percentiles as the lower and upper confidence limits, respectively. While there are various suggestions in literature to help guide subsample size selection, a detailed, objective but somewhat tedious approach is:

a) Decide on a subsample size and complete the simulation;

b) Change the random seed and rerun the simulation and compare the results in a) to the results obtained using the new seed;

c) If the results are very different, then the selected subsample size is too small. Increase the subsample size to a larger number, then repeat steps a) and b) and compare the results. Stop when the results are similar. This is a judgement call, but as a guide, an overlay histogram of the dataset in a) with the dataset in b) may help judge the similarity between the two datasets.

2) Decide on the number of subsamples (replicates) to select (*N*). A recommended minimum of 100 subsamples may be used. Note: a larger number of replicates may take a long time to complete. However, for more accurate results, relatively larger numbers are recommended. The validity of results from bootstrapping is dependent on the convergence of the simulation exercise. Therefore, an adequate number of replicates are needed for the simulation exercise to converge. A key feature of bootstrapping, however, is its ability to converge rather quickly, making the number of replicates required less onerous.

3) Calculate the slope for each subsample and then the empirical α/2 and (1 - α/2) percentiles for the collection of the N slopes as the lower and upper confidence limits. If desired, the standard error of the collection of *N* slopes can also be calculated.

4.4.4 Some limitations of the nonparametric methods

While these methods provide good tests and estimates of the trend, the Mann-Kendall and the Theil-Sen methods have limitations. One noticeable limitation is the complexity introduced when both first-order and second-order autocorrelations are statistically significant. Pre-whitening would have to remove both orders of autocorrelation, and some analytical challenges may arise as a result. As there are no well-established methods of pre-whitening when both first and second order autocorrelations exist, the hope is that the second-order autocorrelation may indirectly be removed through the pre-whitening process described above.

Generally, under normal conditions where the parametric analysis assumptions are met, the parametric methods have more statistical power than the nonparametric rank-based methods. Also, the assumption of normality is needed to test the hypothesis of Mann-Kendall test. Although meeting the assumption of normality for the Mann-Kendall test might be easier because of the general stability of the Mann-Kendall statistic to skewness, sufficient sample size is needed for this assumption to hold.

4.5 Parametric methods

4.5.1 Linear regression with GARCH and ARIMA

As previously discussed, the variance in air quality data can be volatile due to complex meteorological conditions and emission sources. This variation may not follow a particular pattern thus requiring the use of a much more sophisticated approach like GARCH to handle the inherent variance in air quality data. Recall from the previous sections that the basic idea of applying GARCH to linear regression models for the trend in air quality is as follows:

1) If the preliminary analyses show the presence of autocorrelation and non-uniform variance, this suggests that the error term v_t in Equation 19 is not well-conditioned and that some adjustments need to be made.
2) Use the q^{th} order autoregressive process (AR(q)) to model v_t as given by Equation 15 and obtain the error terms ε_t which are independent observations but may have non-uniform conditional variance (σ_t^2), i.e., they may not be identically distributed. The value of q should have been decided on in the preliminary diagnostics for autocorrelation.
3) Model the conditional variance (σ_t^2) of ε_t using the GARCH (V, M) model (Equation 16).

The total number of regression parameters to be estimated will depend on the values of *M* and *V*. The steps 1) to 3) above can be implemented in either SAS (SAS Institute Inc. 2005) using the ***AUTOREG*** procedure in SAS/ETS or in the R using the ***RUGACH*** package (Ghalanos 2013) as described in the next sections.

4.5.2 Test for the order of autoregression and GARCH (*V*, *M*) effects

The process of applying AR(*q*)-GARCH(*V*, *M*) begins with a decision on the values of *q*, *V*, and *M* to specify. The parameter *q* is the order of autocorrelation to specify. For monthly air quality data, a maximum order of 2 (but 1) is typically sufficient for most analytical work, and this value would have been determined in the preliminary diagnostics section described above. If not, statistical packages, such as SAS and R, provide customized procedures for deciding on which autocorrelation coefficients to include (the value of *q*). In SAS/ETS, the stepwise autoregression procedure can be used (SAS Institute Inc. 2005).

The decision on the optimum values for *V* and *M* are not straightforward and may require some optimization routines (e.g., Lee and King 1993, Wong and Li 1995). The recommended approach in this report is to use trial-and-error starting with *V* = 1 and *M* = 1 and then examining the fitted coefficients in each case for statistical significance. However for monthly data, *V* = 1 and *M* = 1 may be enough. This simplifies Equation 16 to:

$$\sigma_t^2 = \omega + \gamma\varepsilon_{(t-1)}^2 + \theta\sigma_{t-1}^2 \qquad \text{(Equation 22)}$$

Where *ω*, *γ* and *θ* are the GARCH parameters.

4.5.3 Implementing AR(*q*)-GARCH (*V*, *M*)

The AR(*q*)-GARCH (*V*, *M*) can be implemented in most standard statistical packages such as SAS and R. In R, the AR(*q*)-GARCH (*V*, *M*) can be implemented in the ***RUGARCH*** package (Ghalanos 2013) and because R is a free software that is widely used, the discussions that follow will focus on R only. The installation of the R-package comes with detailed documentation regarding the features of the package, which can be found in the R help section. However, the general process is summarized as follows:

1) Creating variables: All variables to be used for modeling must be defined or created. Seasonality variables may be generated using Equation 18 or if some other method is chosen to deseasonalize the data, such as the LOESS process, seasonality variables may not be relevant moving forward and can be ignored. The response variable (e.g., monthly average concentration) of the pollutant (e.g., SO_2 or NO_2) is also created as discussed earlier.
2) The seasonality variables and the time variable (months) are combined into an exogenous dataset as shown in Textbox 1 below. If the data had already been deseasonalized, then the only exogenous (independent) variable is time (month).

Textbox 1: Code for creating a set of independent variables

```
##Dependent variable - e.g., Monthly average NO2 concentration:
yc <- data$no2  ##Note: data is read into R from an external source, e.g. excel

## Independent or exogenous set of variables
exog <- cbind(s1, s2, s3, s4, c1, c2, c3, c4, time)
```

3) The model is specified as shown in the sample R-code given in Textbox 2 below as follows (the format of the model specification is known as the ugarch.spec):
   a. The variance model GARCH (*V*, *M*) is defined in the **variance.model** section of the **ugarch.spec**. In this sample and for the analytical work done in this report, a much simpler variance model is specified (i.e., sGARCH).
   b. The mean model is defined to include the main regression model and also the autoregressive process (AR(*q*)) by including the external regressors (the exogenous dataset specified in step 2) above), under the **mean.model** section. The armaOrder = (*q*, *p*) specifies the autoregressive parameter *q* and the moving average parameter *p*. In the sample code below (Textbox 2), the moving average parameter is excluded by replacing *p* with 0.

c. The model distribution is specified under the **distribution.model** section; the distribution could be normal, skewed normal, etc. (the complete list is given in the R documentation of the ***RUGARCH*** package). For monthly data, normal (“norm”) may be good enough to approximate the distribution of the residuals. The skewed normal (“snorm”) option can be used if some skewness is observed in the fitted residuals.

Textbox 2: Creating fit specification for the GARCH model

```
## Fit specification: AR(q)-GARCH(V, M) with skewed normal distribution (snorm)
fit.spec   <- ugarchspec(
                    variance.model = list(model = "sGARCH", garchOrder = c(V, M)),  #V=1,
M=1
                    mean.model  = list(armaOrder = c(q, 0), include.mean = TRUE,
                              external.regressors = exog),
                    distribution.model   =   "snorm")
```

4) The model is then fitted to each set of pollutant data using the **ugarchfit** and specifying the response variable and the fit specifications developed in steps 3)a, 3)b and 3)c above (Textbox 2).

Textbox 3: Fitting the GARCH model

```
##Fitting GARCH(V, M) with skewed normal distribution
Garch_fit        <- ugarchfit(data = yc, spec = fit.spec)
Garch_fit
```

4.5.4 Some limitations of the GARCH approach

While the GARCH method is sufficient to address both autocorrelation and heteroscedasticity, the non-normality problems may not be addressed sufficiently. Although options exist in both SAS/ETS and the RUGARCH package in R (Ghalanos 2013) for specifying alternative distributions for the residuals (besides normal) to improve the standard error estimates, the conditional distribution of the residuals must be understood in order to specify the right option. The skewed-normal (snom) option is robust enough for most linear regression residual as they are typically unimodal distributions that show little to moderate skewness. Another limitation is that the GARCH requires a minimum of 100 data points to implement. Realistically, however, a sample size of less than 100 may not show the extreme volatility that GARCH was intended for; i.e., the addition of the term $\theta\sigma_{t-1}^2$ in Equation 22 in the GARCH specification was intended to model this extreme volatility. This term becomes redundant for smaller sample sizes and may lead to some illogical results in some instances (Ghalanos 2013). Thus constraints have been added in both the SAS/ETS software and the ***RUGARCH*** package in R to prevent GARCH from running if the sample size is less than 100. This may be solved by reducing GARCH (*V*, *M*) to ARCH (*V*). This is an easy adjustment to make in SAS, but not so easy in R. If using R; the recommendation is to fit the ARIMA model to the data as discussed in Section 4.5.5 below, drawing on the similarity between ARCH (Equation 16) and the ARIMA (Equation 17) models.

4.5.5 Implementing the ARIMA (*p*, *d*, *q*)

The implementation of the ARIMA method is very similar to the to the GARCH approach, but much more straightforward. This can be done with SAS or R. As with the GARCH approach discussed above, the implementation using R is described by the following steps:

1) The variables are created as done in Section 4.5.3, including a variable to model seasonality. However if some other method is used to deseasonalize the data, seasonality variables should be excluded. The response variable is also created.

2) The seasonality variables and the time variable (months) are combined into an exogenous dataset as done in Section 4.5.3. If the data had already been deseasonalized, then the only exogenous (independent) variable is time (month).

3) The model is then fitted to each set of pollutant data using the set of independent variables and the response variable. The sample R code is presented in the Textbox 4, below. The value of *p* is the order of autocorrelation, *d* is the differencing parameter (set to 0 for this study as differencing is not needed) and *q* is the moving average parameter (1 or 2 will be suitable for monthly data that has been deseasonalized, e.g. using methods described in Section 4.3.1).

Textbox 4: Code for fitting ARIMA

```
##Dependent variable - e.g. Monthly average NO2 concentration:
yc <- data$no2  ##Note: data is read into R from an external source, e.g. excel

## Independent or exogenous set of variables
exog <- cbind(s1, s2, s3, s4, c1, c2, c3, c4, time)
##Fitting the ARIMA model
arima_fit     <-  arima(yc, order = c(1,0,1), xreg = exog) #p=1 d=0 and q=1
arima_fit
```

4.6 Deriving annual trends from monthly trends

The above proposal suggests the use of monthly averages for analysis, but most conclusions about trends are made on an annual basis, so the monthly trends need to be converted to annual trends. An easier way of looking at this issue is to look at the trend as a time-dependent gradient or monthly increase (or decrease) rate. This way, the annual trend can be derived by multiplying monthly trend (β) by 12 months, i.e., year trend is 12β.

Alternatively, the independent variable (month) could be annualized by dividing by 12 (the number of months in a year) and trend analysis implemented. As trend measures the change in pollutant concentration per unit change in time, the estimated trend will represent annual changes.

4.7 Additional remarks

The ideas proposed in this report will be capable of addressing most trend analyses needs. A package (tool) has been created in R and is being finalized for user convenience.

While it is desirable to implement both parametric and nonparametric methods such that their results can be compared, one may be better in some cases than the other. For instance, when second-order autoregressive processes or higher (AR(2+)) are encountered, the nonparametric methods presented above may not be applicable, as there are no well-established methods for conducting iterative pre-whitening when AR(2+) is the problem.

As a guide, if results from two or more selected methods are inconsistent, then there are two possible problems; 1) incorrect implementation of one or both methods and 2) the conditions necessary for one method to work well is not satisfied. For the first problem, the procedure used to implement both methods must be re-checked to ensure they were properly implemented. Next, the issues investigated in Section 4.3 above should be re-checked to make sure that they were properly identified and addressed.

If both problems are addressed, and the results are still inconsistent, the parametric regression methods described here adequately address autocorrelation and heteroscedasticity and theoretically should have more statistical power than the nonparametric method. If the examination of the regression residual shows severe departures from a normal distribution (which will rarely be the case), the preference should be given to the result of the nonparametric method. The parametric regression analysis provides outputs that include test statistics and graphics such as histograms or normal q-q plots of the regression residuals to support the test for normality. These outputs can be examined for any cases of non-normality. Although no separate or additional tests for normality is necessary, the regression residuals can be saved and used to conduct normality test using simple statistical tool such as R or Microsoft excel.

5 Case studies of trend analysis

5.1 Introduction

This section presents the results of a case study that compared three trend analysis approaches:

1) A nonparametric approach that uses the Theil-Sen method to estimate a trend and the Mann-Kendall method to test the null hypothesis that the trend is not statistically significant.
2) The parametric linear regression method for estimating trend and using the GARCH method to estimate the standard error for the trend to test the null hypothesis that trend is not statistically significant.
3) The parametric linear regression method for estimating trend and using the ARIMA method to estimate the standard error for trend so that the null hypothesis can be tested.

As previously stated, the methods of trend analysis in the existing TFPW tool utilizes annual summaries of hourly air quality data to assess the trend. This implies that as many as 8,760 data points could potentially be rolled up into a single data point per year for annual trend assessment. For a short-term analysis of 2 to 5 years, these annual summaries of 2 to 5 data points do not provide an adequate sample size to assess a trend accurately. Also, using annual averages precludes the examination of an intra-annual trend. Using monthly data helps increase the sample size for trend analysis while enabling intra-annual trends to be assessed.

5.2 Data sources

Sample data for this study came from three air monitoring stations; these are Anzac, Fort McKay - Bertha Ganter, and Fort McMurray - Athabasca Valley. These stations were selected arbitrarily, and the results are for illustration only. The hourly concentrations of SO_2 and NO_2 dating back to 1996 for these stations were downloaded from the Clean Air Strategic Alliance (CASA) data warehouse (now known as the airdata warehouse on the EMSD website). The concentrations were converted from parts per million (ppm) to parts per billion (ppb) before carrying out further analysis to ensure that all of the data was in the same unit.

Examination of the data for completeness revealed data gaps in the years earlier than 2000. Therefore, the NO_2 and SO_2 data for Anzac station were chosen from February 1, 2006 to August 31 2015, while similar data for the other two stations; the Fort McKay - Bertha Ganter, and the Fort McMurray - Athabasca Valley monitoring stations, were chosen from January 1, 2000 to August 31, 2015, to ensure there are no significant gaps in the data. Thus the study used 116 months of data from Anzac and 189 months of data from each of Bertha Ganter and Athabasca Valley. The criterion for including data for specific months in the analysis was based on the availability of data; at least 50% of the total number of hours for that month at the specific station. Based on this criterion, the month of January 2006 was dropped from the Anzac dataset since NO2 data was available for only 190 hours out of the total 744 hours and for SO2, only 121 hours out of the 744 hours were available during that month, each representing less than 50% in both cases. A summary of the percentage count of available data per month and the number of months included in the analysis for each station is presented in Table 1. The percentages in Table 1 show that all monthly averages were obtained from over 50% of valid data for each month.

Table 1- Summary of percentages of hours of available valid data per month for each station and the number of months included in the analysis for each station and each parameter (NO_2 and SO_2).

Monitoring station	Number of months	Percentage of available hourly data per month		
		Minimum	Average	Maximum
NO_2 data				
Anzac	115	85.22	93.41	96.24
Bertha Ganter	188	51.34	92.45	95.16
Athabasca Valley	188	70.56	93.74	95.24
SO_2 data				
Anzac	115	85.42	94.33	95.56
Bertha Ganter	188	77.96	94.23	95.43
Athabasca Valley	188	70.70	94.01	95.43

5.3 Data evaluation and summaries

5.3.1 Data summaries

The hourly data for NO_2 and SO_2 were summarized to obtain the average concentrations for each month at all three stations. Given the vast amount of data in these summaries, time series plots were created for average concentrations, and the number of observations (hourly data) used to derive the average concentrations for NO_2 and SO_2 for each month. Table 2 presents overall summaries (minimum, mean, median and maximum) of hourly concentrations for NO_2 and SO_2 from the three stations (Anzac, Fort McKay – Bertha Ganter and Fort McMurray – Athabasca Valley).

Table 2- Overall summaries (minimum, mean, median and maximum) of hourly concentrations for NO_2 and SO_2 from the three stations.

Station Name	Summaries of hourly concentrations (ppb)				Number of Samples
	Maximum	Minimum	Mean	Median	
NO_2					
Anzac	84.00	0.00	2.83	1.40	78,440
Bertha Ganter	53.00	0.00	6.39	3.00	126,974
Athabasca Valley	183.00	0.00	9.91	7.00	128,734
SO_2					
Anzac	106.00	0.00	0.54	0.00	79,221
Bertha Ganter	184.00	0.00	1.27	0.00	129,402
Athabasca Valley	84.00	0.00	0.82	0.00	129,099

Although values below detection limits (or no-detects) are not flagged in the database, the 0 values in the tables are assumed to be non-detects. The detection limits are 1 ppb for SO_2 monitors (model 43i) and 0.4 ppb for NO_2 monitors (model 42i) (personal communication with Gary Cross, Ambient Air lead, WBEA). Note, the median values of SO_2 are 0 indicating that for

SO_2, at least 50% of the hourly measurements used for this analysis are below detection limits. The NO_2 data did not have this issue as the percentage non-detects at all 3 stations were below 10%. As the interest for this study was to estimate trend, all the non-detects was assumed to be 0 for both NO_2 and SO_2.

The time series plots for NO_2 and SO_2 are presented in Figures 4, 5, and 6 for Anzac, Bertha Ganter and Athabasca Valley, respectively. Note, the intervals on the horizontal axis (date) are in days although the axis formats are “month-year” formats. Therefore, the uneven number of days in months and leap years are the reasons for the uneven labels shown on the horizontal axis for the individual years. Nevertheless, the goal was to show a general trend not specific values. The time series plots of NO_2 and SO_2 for all stations reveal similar seasonal patterns for each parameter; NO_2 showed strong seasonal variation, in line with general expectations that winter months are expected to show relatively higher levels of NO_2. The seasonal patterns for SO_2 concentrations are not as strong as they are for NO_2.

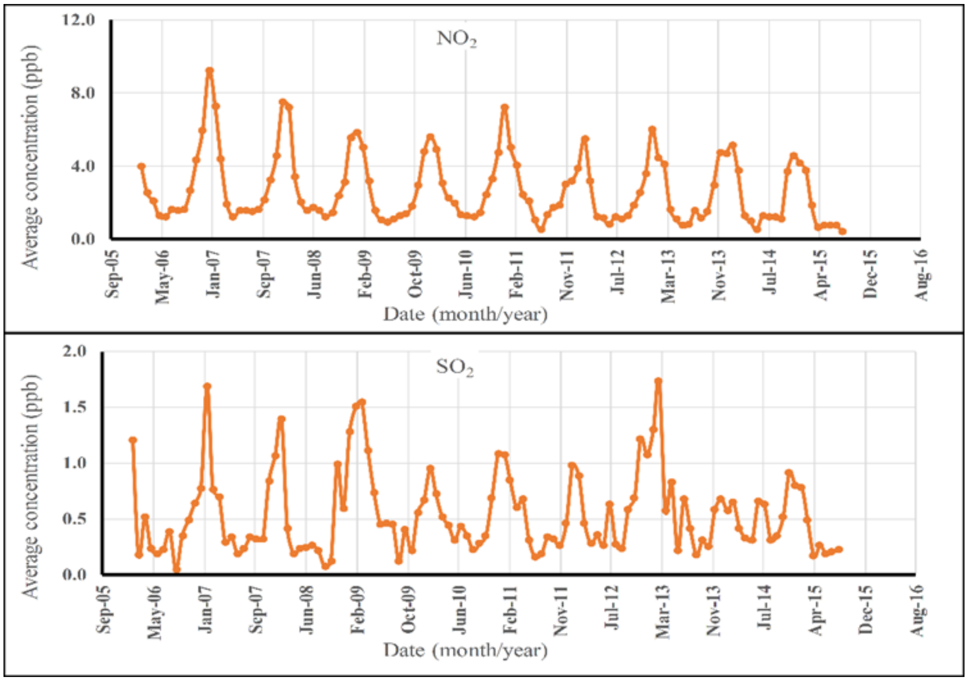


Figure 4- Time series graphs for NO_2 and SO_2 for Anzac showing monthly average concentrations (ppb) (from February 2006 to August 2015).

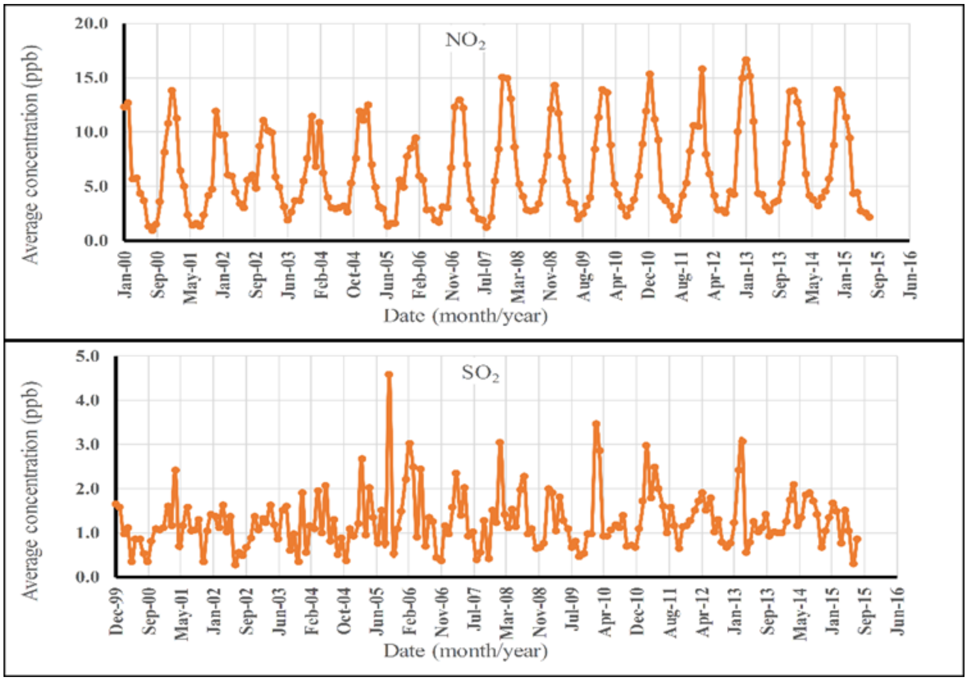


Figure 5- Time series graphs for NO_2 and SO_2 for Fort McKay - Bertha Ganter showing monthly average concentrations (ppb) (from January 2000 to August 2015).

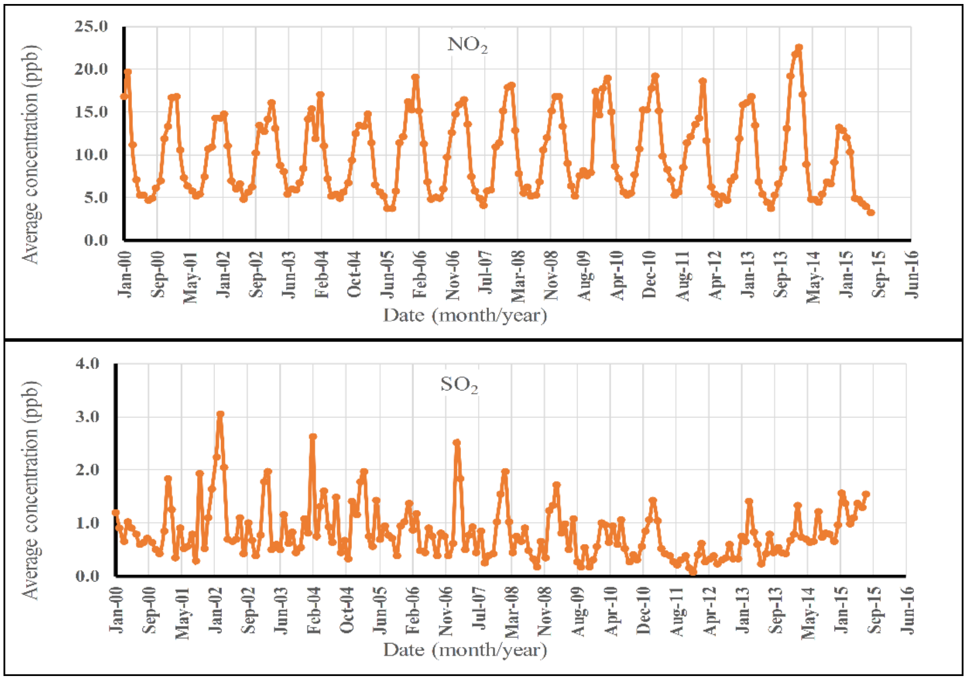


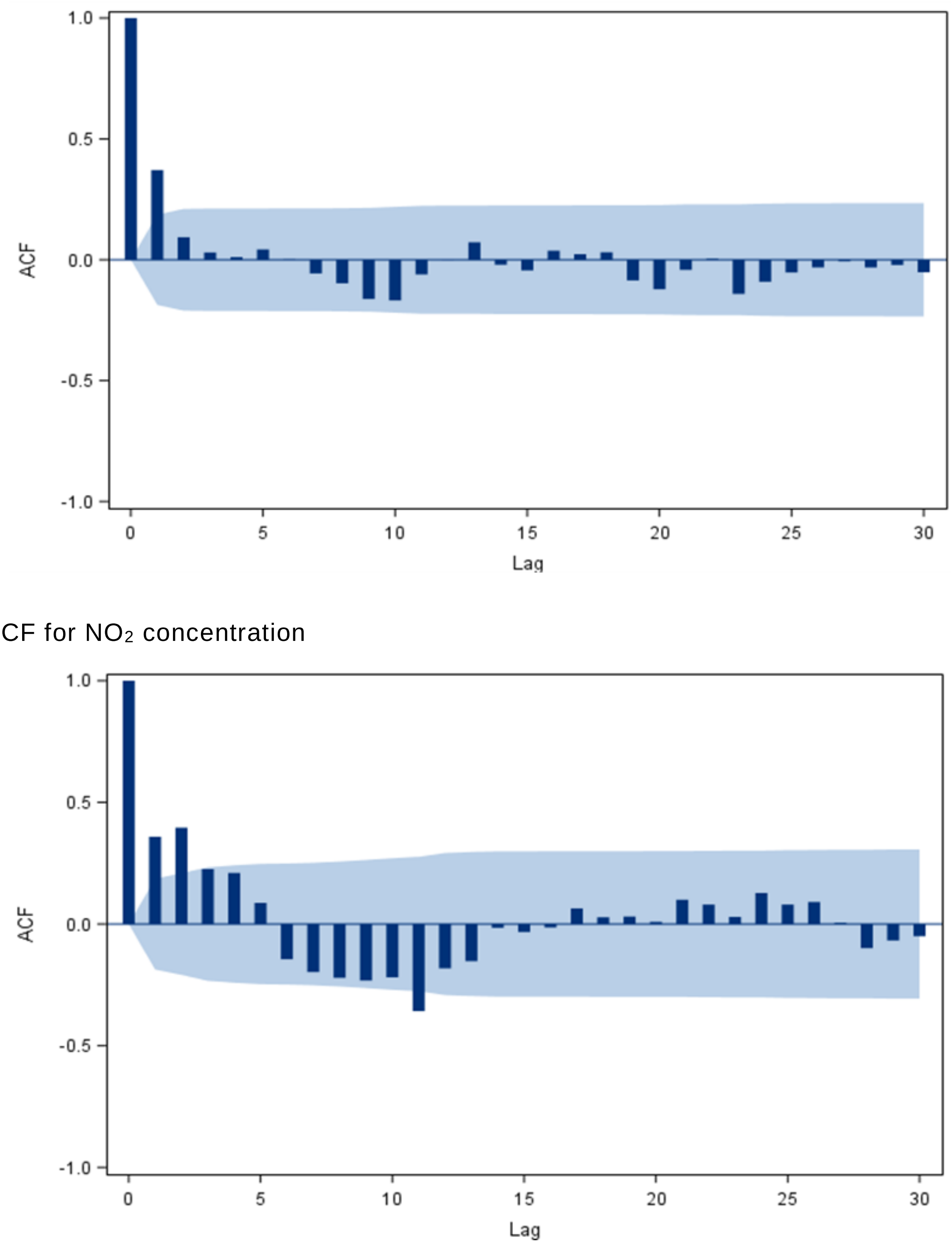
Figure 6- Time series graphs for NO$_2$ and SO$_2$ for Fort McMurray - Athabasca Valley showing monthly average concentrations (ppb) (from January 2000 to August 2015).

A visual inspection of the graphs in Figures 4, 5 and 6 does not reveal data points that deviate strongly from the general patterns of the data series, except in Figure 5(b) where a spike in SO$_2$ concentration is observed at Fort McKay – Bertha Ganter for the month of October 2005. However, this value (4.5 ppb) is not considered an unusual value when seasonal variation are taken into account. Beside there are no records in the EMSD database to suggest that the SO$_2$ data for October 2005 are erroneous. We therefore concluded that outliers are not a significant issue and no further consideration was given to outlier analysis. However, there is some evidence, particularly for SO$_2$ data series, that the temporal variations are not uniform across the period chosen for the study.

5.3.2 Test for autocorrelation

The data for the three stations were investigated for autocorrelation using the ACF function in SAS. A similar analysis can be accomplished in R. This approach is more qualitative than quantitative and presents a visually appealing illustration of the autocorrelation problem. An alternative quantitative approach is to fit Equation 15 to the time series data and then use the parameter statistics to infer the presence of autocorrelation for each lag. Although the outcome of this approach is similar the ACF analysis, the approach involves the manual calculation of each lag variable before fitting Equation 15. This approach was therefore not used here.

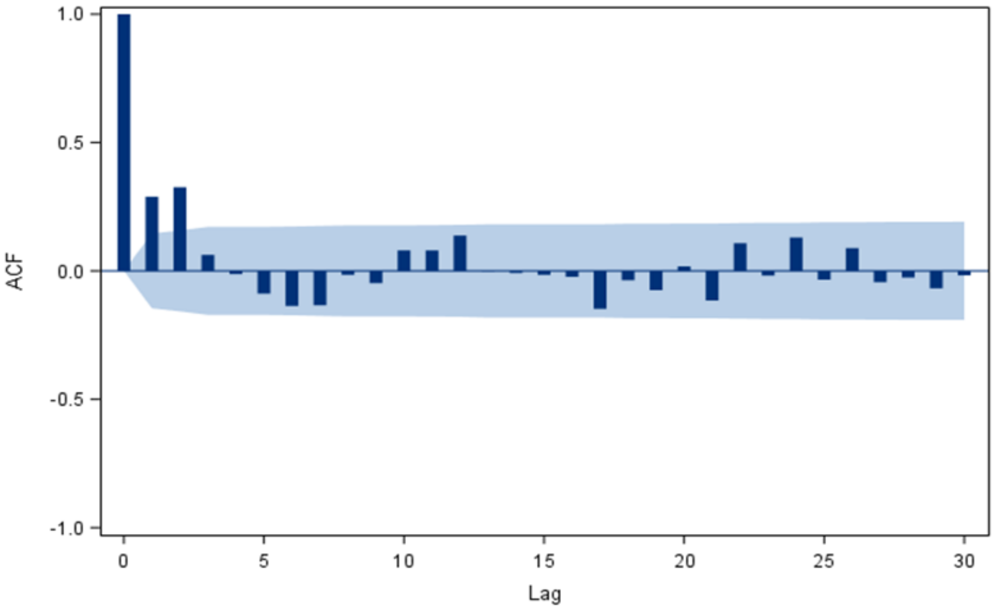
The plots of autocorrelations against the lag of the monthly concentrations are given in Figures 7, 8 and 9 for Anzac, Fort McKay – Bertha Ganter and Fort McMurray Athabasca Valley respectively. Note that although the data used for this test were deseasonalized, the same test can be performed on the raw data as described in Section 4. In each Figure, the vertical bars indicate the estimate of the autocorrelation between the current value and the lag value. Thus, the correlation value at lag 0 is the correlation between the current value and itself, and this value is 1 as shown in each graph. The second bar in each graph is the first order (lag 1) autocorrelation, while the third bar represents second-order (lag 2) autocorrelation and so on. The shaded band in each graph is the approximate 95% confidence interval for the estimated autocorrelations (each vertical bar). This shade region is approximated by $1.96/\sqrt{n}$, where n is the sample size and 1.96 if the $(100 - \alpha/2)$ percentile of the standard normal distribution. Note in SAS 1.96 is rounded to 2. The presence of autocorrelation is based on the extension of the autocorrelation bars above the shaded region and the rate of decline of the bars with increasing lag (SAS Institute Inc. 2005). In graphs where some bars extend beyond the shaded region and also decline relatively slowly indicates the presence of autocorrelation but may not necessarily show which order of autocorrelation is significant. Note that the presence of first-order autocorrelation alone may reflect higher bars that decline slowly. On that basis, all of the ACF graphs (except Figure 8 b) show evidence.



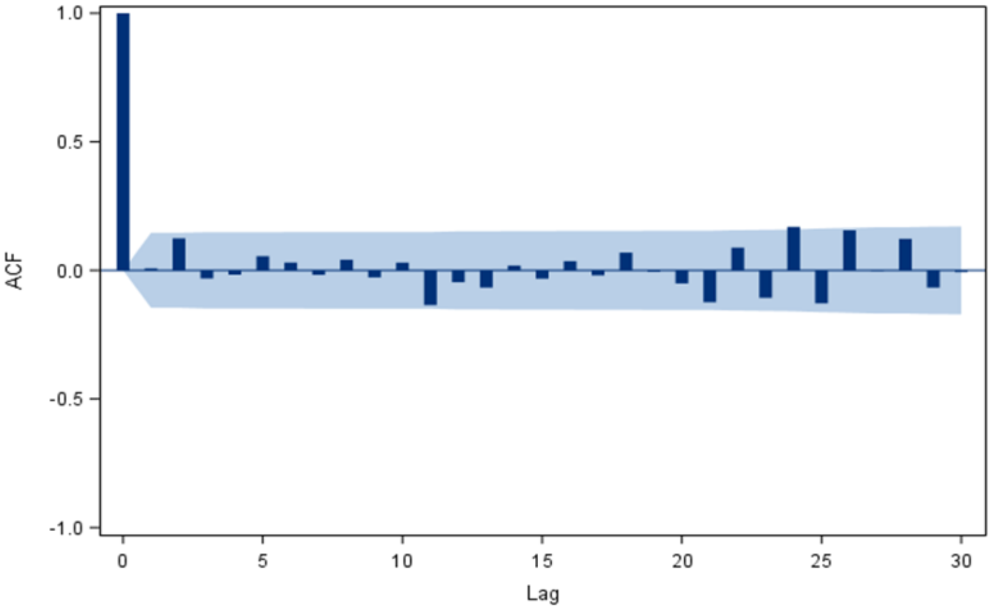
a) ACF for NO_2 concentration

b) ACF for SO_2 concentration

Figure 7- Autocorrelation functions (ACF) for NO_2 concentrations (a) and SO_2 concentrations (b) at Anzac monitoring station.

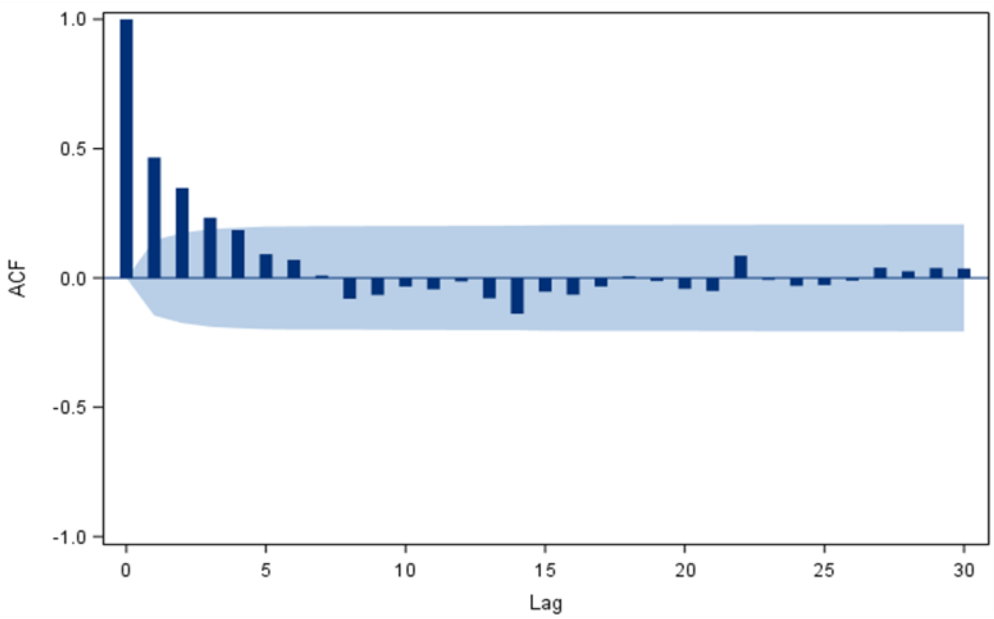


a) ACF for NO_2 concentration

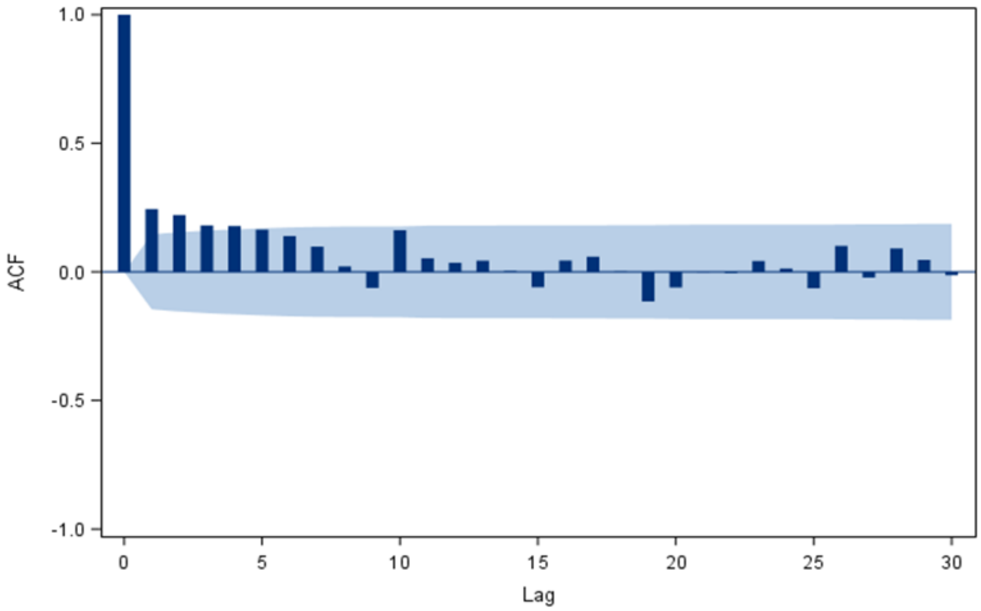


b) ACF for SO_2 concentration

Figure 8- Autocorrelation functions (ACF) for NO_2 concentrations (a) and SO_2 concentrations (b) at Fort McKay - Bertha Ganter monitoring station.



a) ACF for monthly NO_2 concentration



b) ACF for monthly SO_2 concentration

Figure 9- Autocorrelation functions (ACF) for NO_2 concentrations (a) and SO_2 concentrations (b) at Fort McMurray – Athabasca Valley monitoring station.

5.4 Data analysis

5.4.1 Modeling seasonality and de-seasonalizing the data

Due to the strong evidence of seasonality in Figure 4, 5 and 6, we characterized seasonal variation in each dataset using Fourier decomposition as in Tiao et al. (1990) and Weatherhead et al. (1998), discussed in the previous sections (see Equation 18).The complete linear regression model both for ARIMA and GARCH modeling framework is given in Equation 23.

$$Y_t = \alpha + S_t + \beta t + v_t; \qquad S_t = \sum_{j=1}^{\lambda} \left\{ \beta_{1,j} \sin\left(\frac{2\pi j t}{12}\right) + \beta_{2,j} \cos\left(\frac{2\pi j t}{12}\right) \right\} \qquad \text{(Equation 23)}$$

$$v_t = \rho v_{t-1} + \varepsilon_t.$$

All the terms in Equations 23 are defined as shown in the previous sections. Preliminary analysis was conducted on Equation 23 to determine how many sine and cosine functions are necessary to deseasonalize the data. Initially, four sine and four cosine functions (eight terms in total) were considered (i.e., λ = 4). The statistical significance of the coefficients for the eight seasonal terms were assessed at 5% probability based on the results of the preliminary regression analysis using ordinary least squares.The first four terms (two sine and two cosine terms, i.e. λ = 2) were found to be statistically significant at 5% probability and were chosen for further analysis. In expanded form, the chosen seasonality function can be written as:

$$S_t = \beta_{1,1} \sin\left(\frac{2\pi t}{12}\right) + \beta_{1,2} \sin\left(\frac{2\pi 2t}{12}\right) + \beta_{2,1} \cos\left(\frac{2\pi t}{12}\right) + \beta_{2,2} \cos\left(\frac{2\pi 2t}{12}\right) \qquad \text{(Equation 24)}$$

It should be noted that because ordinary least squares regression analysis was used in the preliminary analysis, the test statistics for the coefficients of the seasonality terms are likely inflated due to the potential effect of autocorrelation as evident in Figures 7, 8 and 9. But the potential impact on the model is to include more seasonal terms than needed, which should not impact the results of further analysis.

5.4.2 Linear regression analysis with GARCH and ARIMA

Further analysis using the ARIMA and the GARCH modeling framework used Equation 23 and the seasonality term S_t given by Equation 24. The term v_t of Equation 23 is the main focus of the ARIMA and GARCH analysis in this section. The initial step in applying AR(*q*)-GARCH(*V, M*) to model v_t is to determine the order of autocorrelation *q,* and the values of *V* and *M*. Figure 7 suggests that at least the first order autocorrelation (*q* = 1) is statistically significant. Therefore, using the initial values of *V* = 2 and *M* = 2, preliminary fitting of AR(*q*)-GARCH(*V, M*) conducted using the ***RUGARCH***-package in R show that *V* = 1 and *M* = 1 was enough to model the residuals based on 5% probability of significance. Therefore, the final model fitted for each station and each pollutant was using the GARCH framework was:

$$\begin{cases} Y_t = \ \alpha + S_t + \beta t + v_t; \\ S_t = \displaystyle\sum_{j=1}^{2}\left\{\beta_{1,j} sin\left(\frac{2\pi jt}{12}\right) + \beta_{2,j} cos\left(\frac{2\pi jt}{12}\right)\right\}; \\ v_t = \ \rho v_{t-1} + \varepsilon_t; \\ \varepsilon_t \ = \sigma_t \epsilon; \quad \epsilon \backsim \ IIDN(0,1); \\ \sigma_t^2 \ = \ \omega \ + \ \gamma\varepsilon_{t-1}^2 \ + \ \theta\sigma_{t-1}^2. \end{cases} \qquad \text{(Equation 25)}$$

A sample R-code for fitting the above model (Equation 25) is presented in the Appendix.

As mentioned earlier, ARIMA is the generalized form of the autoregressive moving average (ARMA) models, which models residual variance structure using a combination of AR and MA (see Appendix 2). Based on the results of the GARCH modeling above, the ARIMA (1, 0, 1) was used for all stations and both parameters. Equation 23 was used, but the error terms v_t were modeled differently as:

$$\begin{aligned} & Y_t = \ \alpha + S_t + \beta t + v_t; \\ & S_t = \ \sum_{j=1}^{2}\left\{\beta_{1,j} sin\left(\frac{2\pi jt}{12}\right) + \beta_{2,j} cos\left(\frac{2\pi jt}{12}\right)\right\}; \\ & v_t = \ \rho_1 v_{t-1} + \varepsilon_t; \\ & \varepsilon_t \ = \varepsilon_t + \ \gamma_1 \varepsilon_{t-1.} \end{aligned} \qquad \text{(Equation 26)}$$

Where: ρ_1, γ_1 are the autoregressive and moving average parameters, respectively. A sample R-code for fitting the above model (Equation 26) is presented in the Appendix.

5.4.3 Nonparametric methods

The seasonality parameters estimated using the linear regression with GARCH (see Appendix) were used to de-seasonalize each data series before further analysis with the nonparametric methods (Mann-Kendall test and Theil-Sen slope). In this case study, the Theil-Sen slope was calculated for each dataset as a measure to trend while the Mann-Kendall test was used to test the statistical significance of each trend, as suggested by Yue et al. (2002). However, as discussed in Section 4.4.2, confidence intervals of the Theil-Sen slopes can be estimated and their statistical significance tested using a bootstrap simulation procedure. As the Mann-Kendall method was used to provide this information, the bootstrap option was not used in this study.

Iteratively pre-whitening of each de-seasonalized data series was conducted using the steps outlined in Section 4.4.3 and presented in the flowchart in Figure 3. Each pre-whitened data series was then used to estimate Theil-Sen slope (trend) and test the statistical significance of the trend using Mann-Kendall test.

5.4.4 Comparing the nonparametric, GARCH and ARIMA methods

The approaches were compared using three different criteria: The confidence interval width, the statistical power measured by the minimum size of monthly trend each method can detect with two years of data and the root mean square prediction error as a measure of the fit of each method to the data.

The confidence interval for each method is calculated as:

$$WCI = 2 \times Z_{\alpha/2} SE$$ (Equation 27)

Where: WCI is the width of the confidence interval, $Z_{\alpha/2}$ is the $100 \cdot (1 - \alpha/2)$ percentile of the standard normal distribution and SE is the estimated standard error of trend. 5% was used as α for this comparison. From the statistical theory presented in Section 3.1, the combined Theil-Sen and Mann-Kendall methods of trend analysis does not provide comparable estimates of standard error with the other methods. Therefore, standard errors of estimates comparable to the other two methods were recovered (reverse calculated) from the test probabilities using the relationship:

$$SE = \frac{|\hat{\beta}|}{Z_p}$$ (Equation 28)

Where: SE is the derived standard error of estimate of the trend, $\hat{\beta}$ is the estimate of trend and Z_p is the $100 \cdot (1 - p)$ percentile of the standard normal curve and *p* is the test probability for a given method.

To estimate the minimum size of a monthly trend that can be detected by each method, Equation 2 was rearranged to produce Equation 29 and used to estimate the minimum monthly trend (effect size) that can be detected with two years of data.

$$\hat{\theta} = \left[\left(\frac{3.3\sigma}{n^{3/2}} \right) \sqrt{\left(\frac{1+\rho}{1-\rho} \right)} \right]$$ (Equation 29)

Where: $\hat{\theta}$ is the effect size (monthly trend), *n* is the sample size (*n* is set at 24 months), *σ* is the standard deviation of the residuals and *ρ* is the estimated coefficient of autocorrelation. The values of *σ* and *ρ* were obtained from each model outputs of the analyses in sections 5.4.2 and 5.4.3. Alternatively, *σ* may also be calculated from each model's residuals. Note that the autocorrelation estimates (*ρ*) for all three models are similar. Any differences in statistical power are driven by *σ.* Also note that results obtained using Equation 29 are extrapolations derived based on the residual statistics of the three methods, they are not from real data as the GARCH method cannot be implemented when sample size is less than 100.

The root mean square prediction error measures the goodness of fit of each method. It is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i}{n}} \qquad \text{(Equation 30)}$$

Where: RMSE is the mean square prediction error, $\hat{e}$ is the prediction error of the monthly concentration and *n* is the sample size. The data used for this analysis (error estimates) were direct outputs from the trend analysis. As this was not a validation exercise, independent data sets were not used in this exercise.

The rationale for choosing RMSE for this comparison is the fact that while two different methods give comparable estimates of trend, one method may be slightly biased and this bias can be magnified if the analysis period is much longer than what the case study did. Because RMSE combines bias estimates and variance of the residuals, it is a much better indicator of which model fits the data better. Besides Equation 30, RMSE can also be expressed in terms of bias (*B*) and variance (σ^2) as:

$$RMSE = \sqrt{\sigma^2 + B^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{e}_i - \bar{\hat{e}}\right)^2}{n} + \bar{\hat{e}}^2}$$

$$\bar{\hat{e}} = \frac{\sum_{i=1}^{n} \hat{e}_i}{n} \qquad \text{(Equation 31)}$$

Where: $\bar{\hat{e}}$ is the mean of the residuals (error) used as a measure of bias and all other variables are as defined.

5.5 Preliminary Results and Discussion

5.5.1 Trend estimates and null probabilities for the Theil-Sen/Mann-Kendall, GARCH and ARIMA

The estimates of trend and the corresponding test probabilities for the three methods are presented in Table 3 for the 3 stations: Fort McMurray - Athabasca Valley, Fort McKay - Bertha Ganter, and Anzac. As this study was just for illustration, only the parameter estimates and the test probabilities are presented. The intent of the analysis is to show how each method works and not present trend reports for these stations. Seasonality parameters are not critical to this study and are thus presented in Appendix 3.

The test probabilities presented in Table 3 are the probabilities of making a Type I error by concluding that the estimated trend by each method is statistically significant. Thus the smaller the probability, the lower the Type I error rate. The allowable Type I error rate set for this case study is 0.05. The Type I error rates for all 3 models for NO_2 at Fort McMurray Athabasca Valley are high than 0.05 (0.5101, 0.2630, and 0.8360 respectively in Table 3). Other cases of relatively high Type I error rates are: ARIMA for SO_2 at Fort McKay-Bertha Ganter (0.1526), Theil-Sen/Mann Kendall and ARIMA for SO_2 at Anzac (0.7728 and 0.3603, respectively). In all other cases, Type I error rates are less than 5%, indicating that those trends are statistically significant.

Table 3- Trend estimates and test probabilities for GARCH, Theil-Sen/Mann-Kendall and the ARIMA methods for the three monitoring stations for NO_2 and SO_2. Note, the Mann-Kendall statistics show statistical significance while the Theil-Sen slope provides trend estimates.

Monitoring Station	GARCH		Theil-Sen/MK		ARIMA	
	Monthly trend (ppb)	Probability	Monthly trend (ppb)	Probability	Monthly trend (ppb)	Probability
NO_2						
Anzac	-0.0108	0.0000	-0.0077	0.0000	-0.0064	0.0410
Bertha Ganter	0.0120	0.0000	0.0105	0.0000	0.0115	0.0001
Athabasca Valley	0.0026	0.5101	0.0047	0.2630	0.0008	0.8360

SO_2						
Anzac	0.0002	0.0000	0.0002	0.7728	-0.0008	0.3603
Bertha Ganter	0.0013	0.0234	0.0016	0.0056	0.0014	0.1526
Athabasca Valley	-0.0015	0.0005	-0.0019	0.0012	-0.0020	0.0010

The results presented in Table 3 suggest that all three methods are quite similar in their statistical conclusion about trend when applied to NO_2 data. All three tests identified significant trend for NO_2 at Anzac and Bertha Ganter and the absence of a significant trend for NO_2 concentration at Athabasca Valley base on the 5% probability of Type I error. In contrast, the significance of temporal trend varied appreciably among the three methods when applied to the concentrations of SO_2 (Table 3). The GARCH method identified a significant trend ($p << 0.001$) in SO_2 at Anzac while neither the Theil-Sen/Mann-Kendall nor the ARIMA identified significant trend ($p = 0.77$ and $p = 0.36$, respectively). At Bertha Ganter station, the GARCH and Theil-Sen/Mann-Kendall methods identified statistically significant trends ($p = 0.02$ and $p = 0.006$ respectively). The ARIMA method identified a statistically non-significant trend ($p = 0.15$). Lastly, all 3 tests identified, statistically significant trends in SO_2 concentration at Athabasca Valley (Table 3).

Based on the result presented above, the Theil-Sen/Mann-Kendall and the GARCH methods are consistent in their outcomes except for SO_2 in Anzac, where the GARCH method showed statistical significance for a trend, but the Mann-Kendall test did not. The ARIMA fit appears to be the most conservative accepting the null hypothesis of no trend in 3 out of the 6 study cases (NO_2 at Fort McMurray Athabasca Valley, SO_2 at Fort McKay Bertha Ganter and SO_2 at Anzac). But the ARIMA fit also agrees with each of the other two methods in 4 out of 6 studies cases (NO_2 at all the 3 stations and SO_2 at Athabasca valley).

Differences in the results of any two methods selected for comparison may be reflected in any of the following outcomes:

1) The trend estimates are very different but are the same sign (both are positive or negative), and the statistical conclusions are different;
2) The trend estimates are directly opposite (one is positive, and the other is negative), and the statistical conclusion of both methods is that the estimated trends are significant;
3) The trend estimates are very similar, but the statistical conclusions are different; one saying the trend is statistically significant and the other saying it is not (e.g. SO_2 at Anzac where the trend estimates are the same for the GARCH and Theil-Sen/Mann-Kendall but the GARCH method concluded trend is significant while Theil-Sen/Mann-Kendall concluded that trend is not significant, both at 5% probability); And,

4) The trend estimates are directly opposite (one is positive, and the other is negative), and the statistical conclusions are that the estimated trends are not significant (e.g., SO_2 at Anzac where the Theil-Sen/Mann-Kendall trend estimate and the ARIMA trend estimates are opposite but both concluded that there no trend at 5% probability).

If the outcome is either 1) or 2), this may indicate that one of the methods is biased. A simple overlay graph of predictions from both models and the actual values may reveal which method is biased. A quantitative bias assessment such as presented in Section 5.5.2 below can also help determine which model is biased. Also, the implementation of both methods including data preparation should be thoroughly checked. If the outcome is 3), then one method has better statistical power than the other. Check the implementation to make sure that all methods are appropriate for the dataset in question. In the example given above for SO_2 at Anzac, steps were taken to ensure that all the methods are implement properly. Therefore we can conclude that based on the GARCH test results, SO_2 trend at Anzac is statistically significant. Outcome 4) is probably nonconsequential but still needs to be investigated to ensure that proper procedures were followed in implementing both methods. If both methods are properly implemented then outcome 4 is an indication that both methods did not have enough statistical power to detect any trend.

It should be noted that the magnitudes of trend estimates presented in Table 3 are the rates at which pollutant concentrations change per month. Annual rates of change may be obtained by multiplying each estimate by 12, the number of months in a year. Alternatively, the *t* in Equation 23 above may be normalized with a quotient of 12 and used for the regression analysis. This will result in annual trend estimates.

5.5.2 Comparison of the Theil-Sen/Mann-Kendall, GARCH and ARIMA methods

Table 4 presents the widths of the 95% confidence intervals for the estimates of a trend for NO_2 and SO_2 with data from the 3 test stations. The interval widths are low which reflect the low magnitude of the monthly trend estimates presented in Table 3. A comparison of the three methods for NO_2 shows that the performance of the methods are similar. The performance of Theil-Sen/Mann-Kendall method is slightly better than the GARCH and the ARIMA methods (0.0058 vs. 0.0086 and 0.0122, respectively) at Anzac. At the Bertha Ganter station, the GARCH method performed better with confidence interval of 0.0049 followed by the Theil-Sen/Mann-Kendall method (0.0079) and the ARIMA method (0.0096). At the Athabasca Valley station, the ARIMA method is better with confidence interval width of 0.0023 compared to the GARCH method (0.0157) and the Theil-Sen/Mann-Kendall method (0.0291). In general, no method is consistently better than the other methods for all stations regarding the confidence interval of the trend estimates for NO_2.

The comparison of the three methods for SO_2 reveal a slightly different trend; with the exception of the Anzac station where the Theil-Sen/Mann-Kendall method is slightly better (0.0001 vs. 0.0002 and 0.0034 respectively for the GARCH and ARIMA methods), the GARCH method has smaller confidence interval (0.0023 for Bertha Ganter and 0.0017 for Athabasca Valley). In comparison, the confidence intervals for the Theil-Sen/Mann-Kendall method is 0.0025 for each of the two stations (Bertha Ganter and Athabasca Valley) while the ARIMA confidence intervals are slightly worse 0.0038 and 0.0146 respectively for the two stations (Table 4). Athough the GARCH is slightly better than the Theil-Sen/Mann-Kendall method for SO_2 at Bertha Ganter and Athabasca Valley, the confidence intervals for both methods are much closer for both stations, implying that the performance of GARCH over Theil-Sen/Mann-Kendall for SO_2 at these stations is marginal. But the GARCH and the Theil-Sen/Mann-Kendall methods are both better than the ARIMA method for SO_2.

Table 4- The width of the 95% confidence intervals for monthly trend estimates for NO_2 and SO_2 using the GARCH, Theil-Sen/Mann-Kendall and the ARIMA methods respectively.

Monitoring Station	GARCH	Theil-Sen/MK	ARIMA
NO_2			
Anzac	0.0086	0.0058	0.0122
Bertha Ganter	0.0049	0.0079	0.0096
Athabasca Valley	0.0157	0.0291	0.0023
SO_2			
Anzac	0.0002	0.0001	0.0034
Bertha Ganter	0.0023	0.0025	0.0038
Athabasca Valley	0.0017	0.0025	0.0146

Table 5 presents the fit statistics (RMSE) for the three methods applied to the three monitoring stations. The GARCH method appears to fit the data better with relatively smaller RMSE than the other two methods for all pollutants and for all stations (see Table 5). Although the Theil-

Sen/Mann-Kendall and the ARIMA methods are comparable in their fit statistics (RMSE) for both NO_2 and SO_2 at Bertha Ganter and Athabasca Valley (Table 5), the Theil-Sen/Mann-Kendall is much better at Anzac than the ARIMA method.

As the RMSE also include bias, residual variances were calculated and are presented in parentheses and highlighted in bold italics. Thus, the difference between the figures within and outside the parenthesis for each method is the estimate of bias. Cases in which the two values are very different (e.g., ARIMA model for Anzac NO_2) are indicative of the presence of bias. On that basis, the GARCH model has the least amount of bias for both NO_2 and SO_2 at all stations (Table 5). This finding is reflected in the trend test results for SO_2 at Anzac, where the GARCH method found a significant trend while the other two methods did not. Except for SO_2 at Anzac, the Theil-Sen/Mann-Kendall method is also less biased compared to the ARIMA method. The bias for the ARIMA model is the largest at Anzac for both SO_2 and NO_2 (Table 5). In general however, the GARCH method has the least bias, followed by the Theil-Sen/Mann-Kendall while the ARIMA method has the most bias. But all the three methods can be considered to be reasonably unbiased with the exception of the Anzac station, for practical purposes.

Table 5- The fit statistic (RMSE and residual variance) for NO_2 and SO_2 using the GARCH, Theil-Sen/Mann-Kendall and the ARIMA methods respectively. Note the residual variances are in parentheses.

Monitoring Station	GARCH	Theil-Sen/MK	ARIMA
NO_2			
Anzac	0.6346 (***0.6327***)	1.1101 (***1.1000***)	2.1913 (***1.6398***)
Bertha Ganter	1.2925 (***1.2925***)	1.3552 (***1.3550***)	1.3559 (***1.3555***)
Athabasca Valley	1.4408 (***1.4408***)	1.6360 (***1.6229***)	1.6247 (***1.6210***)
SO_2			
Anzac	0.2519 (***0.2510***)	1.2454 (***0.5747***)	1.8855 (***1.5914***)
Bertha Ganter	0.5654 (***0.5654***)	0.5682 (***0.5679***)	0.5679 (***0.5678***)
Athabasca Valley	0.3861 ***(0.3861***)	0.3956 (***0.3956***)	0.3956 (***0.3956***)

Table 6 presents the comparison of the three methods regarding the magnitude of the monthly trend (ppb) that can be detected with 24 months (2 years) of observations (assuming monthly averages are used). The values in Table 6 represent the estimated minimum size of monthly trend each method can detect. This criterion is an indicator of which method has the most statistical power for detecting a trend, where the model that can detect the smallest trend has the most statistical power. Please note that these are extrapolated based on the model statistics obtained in this study to put all models on the same level for comparison. The numbers do not indicate how a particular model will perform for 2-year trend analysis.

Based on the values in Table 6, the GARCH model has slightly better statistical power, capable of detecting smaller trends for NO_2. Except for Anzac, the ARIMA method has a slightly better power than the Theil-Sen/Mann-Kendall for NO_2. However, the values for all three models are likely within a 95% percent confidence interval of each other judging from the confidence interval widths presented in Table 4 above. For SO_2, the GARCH method has better power than the Theil-Sen/Mann-Kendall and the ARIMA methods at Anzac. At the other two stations (Bertha Ganter and Athabasca Valley) all three methods are similar, differing only at the fourth decimal place in terms of magnitude of trend each method can detect. This further confirms edge of the GARCH model in terms of statistical power over the other methods.

Table 6- The comparison of the GARCH, Theil-Sen/Mann-Kendall and the ARIMA methods respectively based on the trend each method can detect using 24 months (2 years) of measurements for NO_2 and SO_2. Note trend estimates are presented in ppb.

Monitoring Station	GARCH	Theil-Sen/MK	ARIMA
NO_2			
Anzac	0.0273	0.0534	0.0701
Bertha Ganter	0.0456	0.0528	0.0502
Athabasca Valley	0.0683	0.0748	0.0702
SO_2			
Anzac	0.0091	0.0238	0.0626
Bertha Ganter	0.0167	0.0161	0.0160
Athabasca Valley	0.0146	0.0142	0.0142

5.6 Conclusion

The results presented in this case study about the performance of each of the methods should be interpreted as preliminary findings. Additional analyses based on designed simulation studies of the three methods under various scenarios of the statistical challenges discussed in this document will better evaluate these methods and understand their merits. Although the GARCH method is better than the other two methods, such a detailed analysis are necessary to better define the domains within which each method works best. Such analyses may also examine existing alternatives and compare and evaluate their merits against the methods tested in this study.

On a preliminary basis we recommend that at least two of the methods presented in this report be used for any trend analysis exercise, preferably one of the parametric methods (GARCH and ARIMA) and the nonparametric (Theil-Sen/Mann-Kendall) so as to cross check results.

6 References

AEP. 2008. Land-use Framework. The Alberta Environment and Parks, Edmonton, Alberta, No. I/321: https://www.landuse.alberta.ca/Documents/LUF_Land-use_Framework_Report-2008-12.pdf (accessed April 25, 2018).

Baltagi, B. H. 2008. *Econometrics,* 4th Edition. Springer-Verlag Berlin Heidelberg, Germany.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. J. of Econometrics, 31: 307-327. doi: 10.1016/0304-4076(86)90063-1.

Box, G. E. P. and G. M. Jenkins.1968. Some recent advances in forecasting and control. Applied Statistics, 17: 91-109. doi: 10.2307/2985674.

Breusch, T. S. and A. R. Pagan. 1979. A Simple Test for Heteroskedasticity and Random Coefficient Variation. Econometrica, 47: 1287–1294.

Carroll, R. J. and D. Ruppert. 1988. *Transformation and weighting in regression*. Chapman and Hall, New York, USA.

Chen, C. H., C. H. Liu, and H. C. Su. 2008. A nonlinear time series analysis using two-stage genetic algorithms for streamflow forecasting. Hydrol. Process., 22: 3697–3711. doi: 10.1002/hyp.6973.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc., 74: 829-836. doi: 10.1080/01621459.1979.10481038.

Cleveland, W. S. and S. J. Devlin. 1988. Locally weighted regression: An Approach to Regression Analysis by Local Fitting. J. Amer. Statist. Assoc., 83: 596-610. doi: 10.1080/01621459.1988.10478639.

Cohen, J. 1988. *Statistical Power Analysis for Behavioral Sciences*, 2nd Edition. Lawrence Erlbaum Associates, New York, USA.

Ding, Z., C. W. J. Granger, and R. F. Engle. 1993. A long memory property of stock market returns and a new model. J. of Empirical Finance, 1: 83–106. doi: 10.1016/0927-5398(93)90006-D.

Douglas, E. M., R. M. Vogel, and C. N. Knoll. 2000. Trends in flood and low flows in the United States: impact of spatial correlation. J. Hydrology, 240: 90–105.

Durbin, J. 1969. Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals. Biometrika, 56:1–15.

Engle, R. F. and V. K. Ng. 1993. Measuring and testing the impact of news on volatility. J. of Finance. 48: 1749–1778. doi:10.1111/j.1540-6261.1993.tb05127.x.

Engle, R.F. 1982. Autoregressive Conditional Heteroscedasticity with estimates of variance of United Kingdom inflation. Econometrica, 50: 987–1008.

Esterby, S. J. 1993. Trend analysis methods for environmental data. Envirometrics, 4: 459-481.

Georgeopoulos, P. G. and J. H. Seinfeld. 1982. Statistical distributions of air pollutant concentrations. Environmental Science Technology, 16: 401-416.

Ghalanos, A. 2013. RUGARCH: Univariate GARCH models. R package version 1.0-16. http://www2.uaem.mx/r-mirror/web/packages/rugarch/index.html

Gilbert, R. O.1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Norstrand Reinhold Company Inc., New York, USA.

Glosten, L., R. Jagannathan, and D. Runkle. 1993. Relationship between the expected value and volatility of the nominal excess returns on stocks. J. of Finance, 48: 1779-1802.

Graybill, F. A. and H. K. Iyer. 1994. *Regression Analysis: Concepts and Applications*. Duxbury Press, Belmont, California, USA.

Hamed, K. H. and A. R. Rao. 1998. A modified Mann-Kendall trend test for autocorrelated data. J. of Hydrology, 204: 182-196. doi: 10.1016/S0022-1694(97)00125-X.

Hardin, J. W. and J. M. Hilbe. 2012. *Generalized estimating equations*, 2nd Edition. Chapman & Hall/CRC Press, Boca Raton, FL, USA.

Harvey, A. C. 1990. *The Econometric Analysis of Time Series*, 2nd Edition. MIT Press, Cambridge, MA, USA.

Hirsch, R. and J. Slack. 1984. A nonparametric test for seasonal data with serial dependence. Water Resour. Res., 20: 727-732.

Hirsch, R. M., J. R. Slack, and R. A. Smith. 1982. Techniques of trend analysis for monthly water quality data. Water Resour. Res., 18: 107 –121. doi: 10.1029/WR018i001p00107.

Johnston, J. 1984*. Econometric Methods,* 3rd. Edition. McGraw-Hill, New York, USA.

Kariya, T. and H. Kurata. 2004. *Generalized Least Squares*. John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Kendall, M. G.1975. *Multivariate Analysis*. Charles Griffin Co. Ltd., London, England.

Kulkarni A. and H. von Storch. 1995. Monte Carlo experiments on the effect of serial correlation on the Mann–Kendall test of trend. Meteorologische Zeitschrift 4: 82–85.

Lee, J. H. and M. L. King. 1993. A Locally Most Mean Powerful Based Score Test for ARCH and GARCH Regression Disturbances. J. of Business and Economic Statistics, 11: 17–27. doi: 10.2307/1391304.

Lins, H.F and J. R. Slack. 1999. Streamflow trends in the United States. Geophy. Res. Letters, 26:227 – 230.

Liu, Y., M. Mazur and C. Adams. 2015. Technical Supporting Document for the 2012 Air Quality Management Framework (AQMF) Management Response. Alberta Environment and Parks, Edmonton, Alberta. https://open.alberta.ca/publications/9781460128640 (accessed on April 25, 2018).

Mann, H. B. 1945. Nonparametric tests against trend. Econometrica, 13: 245-259.

Modarres, R. and T. B. M. J. Ouarda. 2013. Modelling heteroscedasticity of streamflow time series. Hydrological Sciences Journal, 58: 1–11. doi: 10.1080/02626667.2012.743662

Nelson, D.B. 1991. Conditional heteroskedasticity in asset returns: a new approach. Econometrica, 59: 347–370.

Neter, J., W. Wasserman, and M. Kutner. 1983. *Applied Linear Regression Models*. Richard D. Irwin Inc., Homewood, IL, USA.

Pinheiro, J. C and D. M. Bates. 2000. *Mixed-effects Models in S and S-PLUS*. Springer, New York, USA.

R Core Team. 2018. Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rousseeuw, P. J. and A. M. Leroy. 1987. *Robust regression and outlier detection*. Wiley, New York, USA.

Ryan, T, P. 1997. *Modern Regression Methods.* Wiley, New York, USA.

Ryan, T. P. 2008. *Modern Regression Methods,* 2nd Edition. Wiley, New York, USA.

SAS Institute Inc. 2005. SAS/STAT® 9.1 User's Guide. SAS Campus Drive, Cary, North Carolina USA.

Sen, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. J. Amer. Statist. Assoc., 63: 1379–1389. Doi: 10.2307/2285891

Sentana E. 1995. Quadratic ARCH Models. The Review of Economic Studies, 62: 639-661. doi:102307/2298081.

Theil, H. 1950. A rank-invariant method of linear and polynomial regression analysis, Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen A.53: 1397-1412. doi: 10.1007/978-94-011-2546-8_20.

Tiao, G. C., G. C. Reinsel, D. Xu, J. H. Pedrick, X. Zhu, A. J. Miller, J. J. DeLuisi, C. L. Mateer, and D.J. Wuebbles. 1990. Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation. J. Geophys. Res., 95 (D12): 20507 - 20517. doi: 10.1029/JD095iD12p20507.

Tofallis, C. 2008. Least Squares Percentage Regression. Journal of Modern Applied Statistical Methods, 7: 526–534. doi:10.2139/ssrn.1406472.

US EPA. 2006. Data Quality Assessment: Statistical Methods for Practitioners. The United States Environmental Protection Agency, Office of Environmental Information Washington, DC 20460, EPA/240/B-06/003:https://www.epa.gov/sites/production/files/2015-08/documents/g9s-final.pdf. (accessed on December 13, 2017).

von Storch H. 1995. *Misuses of statistical analysis in climate research*. In von Storch H and A. Navarra (eds)., Analysis of Climate Variability: Applications of Statistical Techniques, Springer-Verlag: Berlin; 11–26.

von Storch, H. and F.W. Zwiers. 1999. *Statistical Analysis in Climate Research*. Cambridge University Press, London, England.

Wang, W., J. K. Vrijling, P. H. A. J. M. Van Gelder, and J. Ma. 2006. Testing for nonlinearity of streamflow processes at different timescales. J. of Hydrology, 322: 247–268. doi: 10.1016/j.jhydrol.2005.02.045.

Wang, W., P. H. A. J. M. Van Gelder, J. K. Vrijling, and J. Ma. 2005. Testing and modeling autoregressive conditional heteroskedasticity of streamflow processes. Nonlinear Process in Geophysics, 12: 55–66. doi: 10.5194/npg-12-55-2005.

Wang, X. L. and V. R. Swail. 2001. Changes of Extreme Wave Heights in Northern Hemisphere Oceans and Related Atmospheric Circulation Regimes. J. of Climate, 14: 2204–2221. doi: 10.1175/1520-0442(2001)014<2204:COEWHI>2.0.CO;2.

Weatherhead, E. C., G. C. Reinsel, G. C. Tiao, C. H. Jackman, L. Bishop, S. M. Hollandsworth Frith, J. DeLuisi, T. Keller, S. J. Oltmans, E. L. Fleming, D. J. Wuebbles, J. B. Kerr, A. J. Miller, J. Herman, R. McPeters, R. M. Nagatani, and J. E. Frederick. 2000. Detecting the recovery of total column ozone. J. Geophys. Res., 105(D17): 22,201-22,210. doi: 10.1029/2000JD900063.

Weatherhead, E. C., G. C. Reinsel, G. C. Tiao, X. L. Meng, D. Choi, W. K. Cheang, T. Keller, J. DeLuisi, D. J. Wuebbles, J. B. Kerr, A. J. Miller, S.J. Oltmans, and J. E. Frederick. 1998. Factors affecting the detection of trends: Statistical considerations and applications to environmental data. J. Geophys. Res., 103: 17149–17161, doi:10.1029/98JD00995.2000.

White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48: 817–838.

Wiens, D. P. 1998. Minimax robust designs and weights for approximately specified regression models with heteroscedastic errors. J. Amer. Statist. Assoc., 93: 1440-1450. doi: 10.1080/01621459.1998.10473804.

Wiens, D. P. 2000. Robust weights and designs for biased regression models: Least squares and generalized M-estimation. J. of Statistical Planning and Inference, 83: 395 – 412. doi: 10.1016/S0378-3758(99)00102-0.

Wilcox, R. R. 2005. *Introduction to robust estimation and hypothesis testing*. Elsevier Academic Press, San Diego CA, USA.

Wong, H. and W. K. Li. 1995. Portmanteau Test for conditional heteroscedasticity, using ranks of squared residuals. J. of Applied Statistics, 22: 121–134. doi: 10.1080/757584402.

Yue, S. and C. Wang. 2004. The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. Water Resources Management, 18: 201 – 218.

Yue, S., P. Pilon, B. Phinney, and G. Cavadias. 2002. The influence of autocorrelation on the ability to detect trend in hydrological series. Hydrol. Process., 16: 1807 – 1829. doi: 10.1002/hyp.1095.

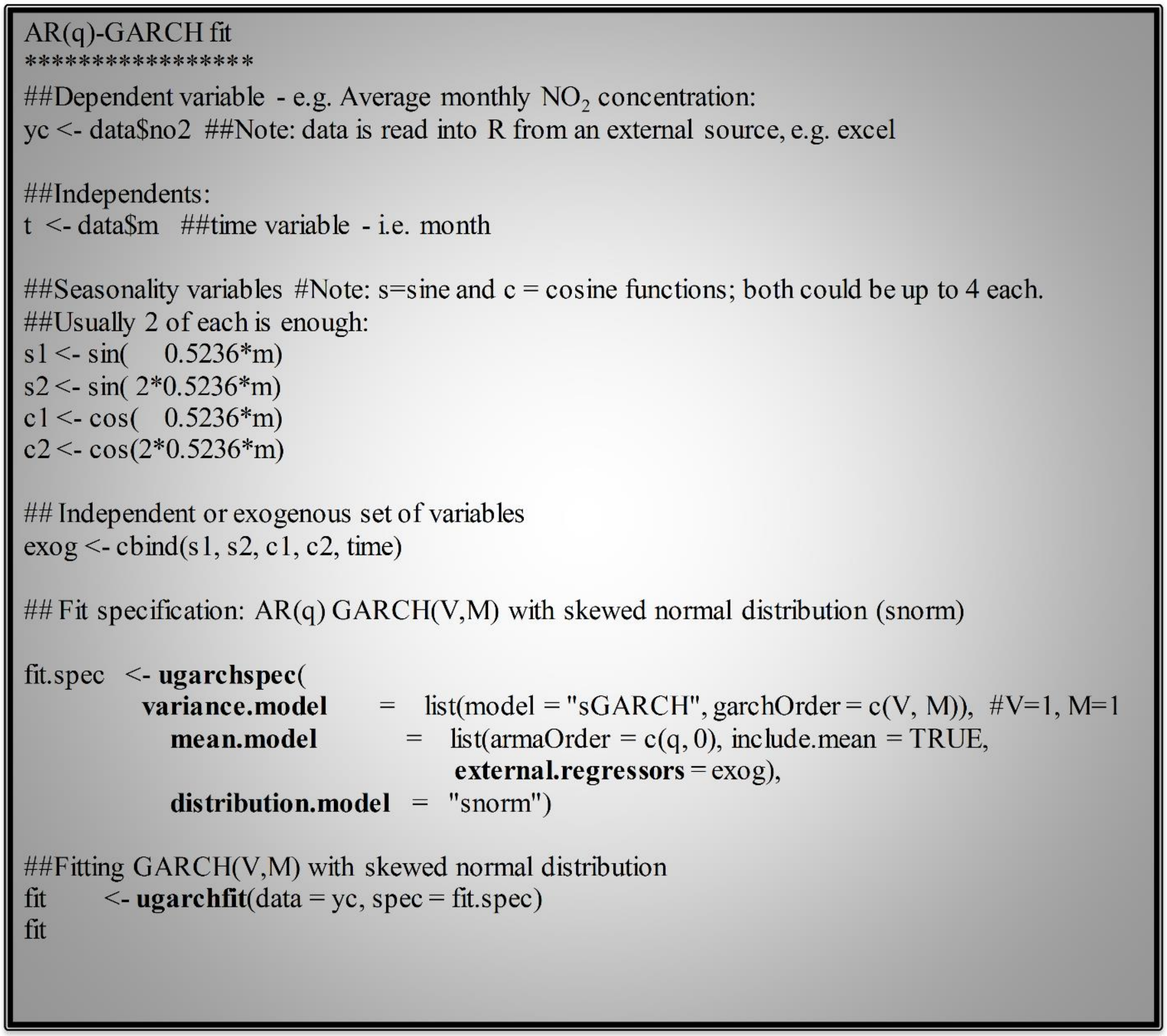
Zakoïan, J. M. 1994. Threshold heteroskedastic models. J. of Economic Dynamics Control, 18: 931–944. doi: 10.1016/0165-1889(94)90039-6.

Zhang, L., W. R. Dawes, and G. R. Walker. 2001. Response of mean annual evapotranspiration to vegetation changes at catchment scale. Water Resour. Res., 37: 701–708.

Zhang, X. and F. W. Zwiers. 2004. Comments on “Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test” by Sheng Yue and Chun Yuan Wang. Water Resour. Res., 40: W03805, doi:10.1029/2003WR002073.

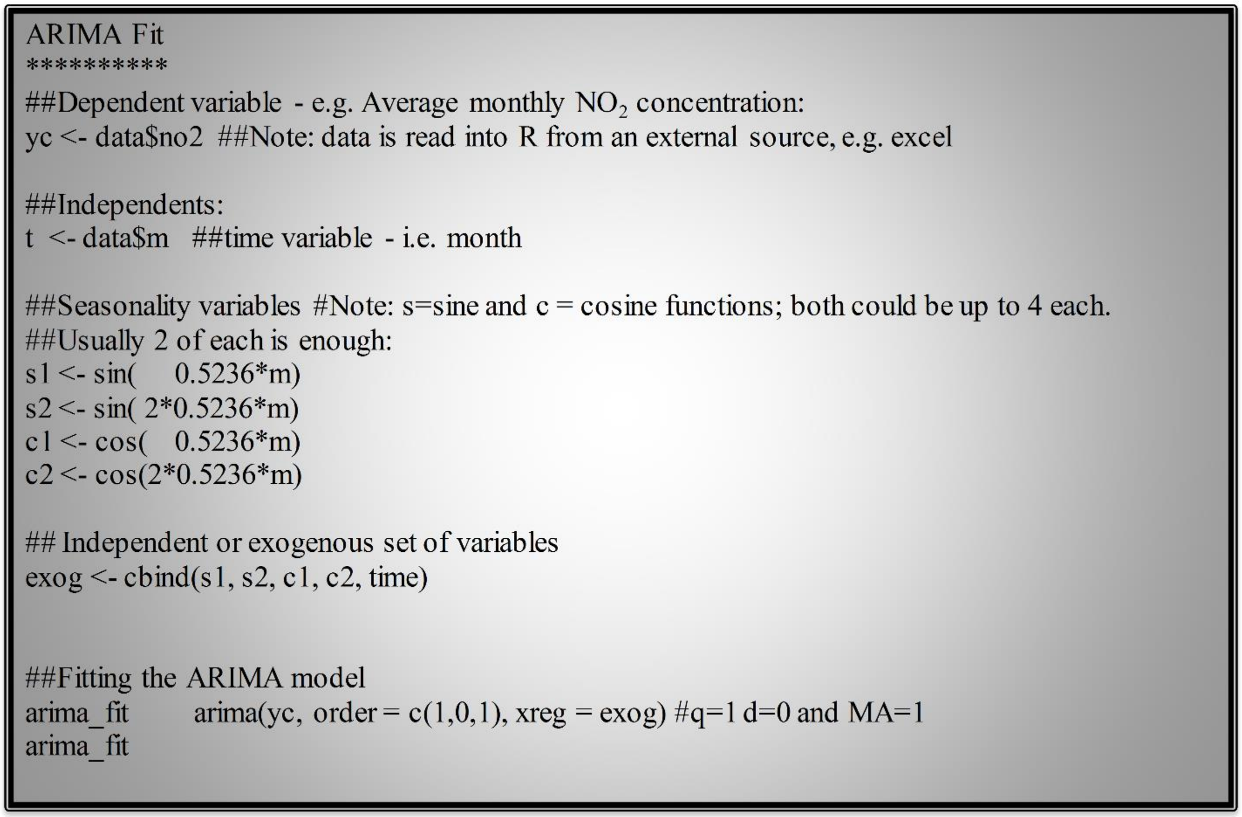
7 Appendices

R-code used for implementing GARCH modeling



```
AR(q)-GARCH fit
*****************
##Dependent variable - e.g. Average monthly NO2 concentration:
yc <- data$no2  ##Note: data is read into R from an external source, e.g. excel

##Independents:
t  <- data$m   ##time variable - i.e. month

##Seasonality variables #Note: s=sine and c = cosine functions; both could be up to 4 each.
##Usually 2 of each is enough:
s1 <- sin(     0.5236*m)
s2 <- sin( 2*0.5236*m)
c1 <- cos(    0.5236*m)
c2 <- cos(2*0.5236*m)

## Independent or exogenous set of variables
exog <- cbind(s1, s2, c1, c2, time)

## Fit specification: AR(q) GARCH(V,M) with skewed normal distribution (snorm)

fit.spec   <- ugarchspec(
              variance.model      =    list(model = "sGARCH", garchOrder = c(V, M)),  #V=1, M=1
                 mean.model           =    list(armaOrder = c(q, 0), include.mean = TRUE,
                                                external.regressors = exog),
                 distribution.model   =    "snorm")

##Fitting GARCH(V,M) with skewed normal distribution
fit        <- ugarchfit(data = yc, spec = fit.spec)
fit
```

R-code used for implementing ARIMA modeling



```
ARIMA Fit
**********
##Dependent variable - e.g. Average monthly $NO_2$ concentration:
yc <- data$no2  ##Note: data is read into R from an external source, e.g. excel

##Independents:
t  <- data$m   ##time variable - i.e. month

##Seasonality variables #Note: s=sine and c = cosine functions; both could be up to 4 each.
##Usually 2 of each is enough:
s1 <- sin(    0.5236*m)
s2 <- sin( 2*0.5236*m)
c1 <- cos(    0.5236*m)
c2 <- cos(2*0.5236*m)

## Independent or exogenous set of variables
exog <- cbind(s1, s2, c1, c2, time)


##Fitting the ARIMA model
arima_fit        arima(yc, order = c(1,0,1), xreg = exog) #q=1 d=0 and MA=1
arima_fit
```

Estimates and standard errors of the seasonal model fitted in GARCH and used for de-seasonalizing air quality time series for Theil-Sen/Mann-Kendall analysis.

Parameter	Fort McMurray Athabasca Valley		Fort McKay Bertha Ganter		Anzac	
NO_2						
$\beta_{1,1}$	0.00288	0.00025	0.00263	0.00019	1.08775	0.11526
$\beta_{1,2}$	0.00112	0.00020	0.00115	0.00016	0.78088	0.09068
$\beta_{2,1}$	0.00523	0.00025	0.00438	0.00019	1.89693	0.11527
$\beta_{2,2}$	0.00004	0.00020	0.00077	0.00016	0.29482	0.09081
SO_2						
$\beta_{1,1}$	0.00033	0.00009	0.00036	0.00009	0.27864	0.04134
$\beta_{1,2}$	0.00016	0.00008	0.00009	0.00009	0.15140	0.03443
$\beta_{2,1}$	0.00002	0.00009	-0.00012	0.00009	0.20038	0.04134
$\beta_{2,2}$	-0.00015	0.00008	-0.00019	0.00009	-0.00890	0.03453

The estimated 95% lower confidence limits (LCL) and upper confidence limits (UCL) for the monthly trend estimates of NO_2 and SO_2 using the GARCH, Theil-Sen/Mann-Kendall and the ARIMA methods respectively.

Station	GARCH		Theil-Sen/MK		ARIMA Fit	
	LCL	UCL	LCL	UCL	LCL	UCL
NO_2						
Athabasca Valley	-0.00524	0.01044	-0.00982	0.01922	-0.00039	0.00193
Bertha Ganter	0.00953	0.01447	0.00654	0.01446	0.00669	0.01631
Anzac	-0.01512	-0.00654	-0.01064	-0.00482	-0.01245	-0.00026
SO_2						
Athabasca Valley	-0.00232	-0.00068	-0.00312	-0.00068	-0.00924	0.00534
Ganter	0.00018	0.00250	0.00037	0.00291	-0.00051	0.00330
Anzac	0.00016	0.00032	0.00011	0.00025	-0.00251	0.00091