

Time series analysis – climatological applications



Xiaomeng Gu

July 28, 2021

*Thesis submitted for the degree of
Master of Philosophy
in
Statistics*

at The University of Adelaide

*Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences*



THE UNIVERSITY

of ADELAIDE

Contents

List of Tables	vii
List of Figures	ix
List of Notations	xi
List of Abbreviations	xiii
Signed Statement	xv
Acknowledgements	xvii
Dedication	xix
Abstract	xxi
1 Introduction	1
1.1 Introduction	1
1.2 Overview of Time Series Analysis	3
2 Methods	9
2.1 Stationarity	10
2.2 Ergodicity and Stability	11
2.3 White Noise	11
2.4 Non-stationarity	12
2.5 Random Walk	14
2.6 Decomposition of Series	15
2.7 Correlation	17
2.7.1 Auto-correlation	18
2.7.2 Correlogram	19
2.7.3 Partial auto-correlation	20
2.7.4 Partial correlogram	21

2.7.5	Cross correlation	22
2.7.6	Spectrum	22
2.8	Autoregressive Moving Average (ARMA) Model	23
2.9	Autoregressive Integrated Moving Average (ARIMA) Model	24
2.10	Autoregressive Fractionally Integrated Moving Average (FARIMA) Model	25
2.11	Cointegration	27
2.12	Unit Root Tests	28
2.12.1	Augmented Dickey-Fuller (ADF) test	29
2.12.2	Phillips-Perron (PP) test	29
2.12.3	Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test	30
2.13	Pre-whitening	31
2.14	Granger Causality	34
2.15	Vector Autoregressive (VAR) Model	35
2.15.1	Stationarity of the VAR models	36
2.16	Threshold Vector Autoregressive (TVAR) Model	37
3	Influence of climate indicators on rainfall in South Australia	39
3.1	Introduction	39
3.2	Data	41
3.2.1	Rainfall	41
3.2.2	Climate indicators	43
3.3	Analysis	48
3.3.1	Relationships between climate indicators	48
3.3.2	Relationship between rainfall and climate indicators	61
3.4	Conclusion	84
4	Multi-scale analysis of CO₂	87
5	Conclusion	123
A	R code	125
A.1	Section 2.3 White Noise	125
A.2	Section 2.6 Decomposition of Series	125
A.3	Section 2.13 Pre-whitening	126
A.4	Chapter 3	127
A.4.1	Linear interpolation	127
A.4.2	Deseasonalise and detrend the series	127
A.4.3	Fitting VAR models	128
A.4.4	Fitting TVAR models	129
A.5	Chapter 4	129
A.5.1	Filtering deterministic models	129

A.5.2 Fitting univariate time series models 130

Bibliography **133**

Bibliography 133

List of Tables

3.1	Statistical significances of the lagged terms, trends and the seasonal patterns with adjusted R^2 .	50
3.2	The estimated coefficients and their standard errors of the fitted regression model for the pre-whitened IOD on the lagged pre-whitened SOI.	54
3.3	The estimated coefficients and their standard errors of the fitted regression model for the pre-whitened PDO on the lagged pre-whitened SOI.	54
3.4	The estimated coefficients and their standard errors of the fitted regression models for the pre-whitened IOD on the pre-whitened PDO and the pre-whitened SOI.	55
3.5	The estimated coefficients and their standard errors of the fitted regression models for the pre-whitened PDO on the lagged pre-whitened IOD and the SOI, and conditional parameters are chosen.	56
3.6	The estimated coefficients and their standard errors of the fitted VAR(2) model for the four climate indicators, along with standard deviations (sd) of residuals and adjusted R^2 .	58
3.7	The estimated coefficients and their standard errors of the fitted VAR(2) model for the four deseasonalised and detrended climate indicators, along with standard deviations (sd) of residuals and adjusted R^2 .	59
3.8	The estimated coefficients and their standard errors of the fitted VAR(5) model for the deseasonalised and detrended series, along with standard deviations (sd) of residuals and adjusted R^2 .	60
3.9	The statistically significant coefficients and their standard errors, at the 0.05 level, of the fitted regression model for IOD on SOI.	60
3.10	The statistically significant coefficients and their standard errors, at the 0.05 level, of the fitted regression model for PDO on SOI.	61
3.11	Statistical significances of the lagged terms, trends and the seasonal patterns with adjusted R^2 .	62
3.12	Estimated coefficients and their associated standard errors of the four fitted VAR models, and conditional parameters are chosen for model 3 and model 4.	66

3.13	The standard deviations of residuals and adjusted R^2 of fitted four models for four weather stations.	74
3.14	Estimated coefficients and their associated standard errors of the two VAR(1) models for four weather stations, at the 0.05 significant level, with conditional parameters are chosen.	76
3.15	Estimated coefficients and their associated standard errors of the two VAR(2) models for four weather stations, at the 0.05 significant level, with conditional parameters are chosen.	77
3.16	Threshold values, percentage of observations below thresholds, standard deviations of residuals and adjusted R^2 s for 1 threshold TVAR(1) model with and without interaction terms, for four weather stations.	82
3.17	Estimated coefficients and their standard errors for each regime for the fitted 1 threshold TVAR(1) model, along with threshold values, standard deviations of residuals and adjusted R^2	83

List of Figures

2.1	Time series plot of a realisation of DWN process with mean 0 and variance 1.	12
2.2	Time series plot of the Airpassenger data.	13
2.3	Time series plot of the log transformed Airpassenger data.	13
2.4	Time series plot of the simulated pure random walk.	15
2.5	Time series plots showing trends and seasonal variations.	16
2.6	Decomposition of the series.	17
2.7	Time series plots of the AirPassenger data.	19
2.8	Correlogram of the detrended and deseasonalised Airpassenger data.	20
2.9	Partial correlogram of the detrended and deseasonalised Airpassenger data.	21
2.10	Spectrum of the detrended and deseasonalised Airpassenger data.	23
2.11	Correlogram of the FARIMA(1, 0.4, 0).	26
2.12	Cross correlogram between the variable $\{x_t\}$ and the $\{y_t\}$.	32
2.13	Cross correlogram of $\{u_t\}$ and $\{y_{\text{filtered}}\}$.	33
3.1	Locations of the four weather stations.	42
3.2	Time series plots of the monthly rainfall at the four weather stations.	43
3.3	Time series plot of the SOI.	44
3.4	The Indian Ocean Dipole (Bureau of Meteorology 2021 <i>b</i>).	45
3.5	Time series plot of the IOD.	45
3.6	Time series plot of the PDO.	46
3.7	Locations of 12 stations used to calculate the Southern Annular Mode (Marshall 2021).	47
3.8	Time series plot of the SAM.	48
3.9	Cross correlograms of the four climate indicators.	49
3.10	Cross correlograms of the deseasonalised and detrended climate indicators.	51
3.11	Correlograms of the deseasonalised and detrended series.	52
3.12	Cross correlograms of the pre-whitened series.	53
3.13	Time series plot of the average monthly rainfall.	62
3.14	ACF and PACF of the deseasonalised average monthly rainfall.	63
3.15	Cross correlograms between the climate indicators and average monthly rainfall.	64

3.16	Plots of the residuals obtained from model 2.	67
3.17	Correlograms of the monthly rainfall at four weather stations.	68
3.18	Correlograms and partial correlograms of the deseasonalised monthly rainfall at four weather stations.	70
3.19	Cross correlograms between climate indicators and pre-whitened rainfall at four weather stations.	72
3.20	Cross correlograms between climate indicators and pre-whitened rainfall at four weather stations.	73
3.21	Correlograms and histograms of the residuals of monthly rainfall obtained from fitted VAR(1) with linear terms only, for four weather stations.	79
3.22	Cross correlograms between the residuals of the monthly rainfall obtained from model 1.	80

List of Notations

Short form	Long form
$\{x_t\}$ or $\{y_t\}$	Time series variable
$\{w_t\}$ or $\{u_t\}$	White noise process
k	lag
γ_k	auto-covariance at lag k
ρ_k	auto-correlation at lag k
c_k	sample auto-covariance at lag k
r_k	sample auto-correlation at lag k
ρ'_k	sample partial auto-correlation at lag k
p	order of autoregressive term
q	order of moving average term
B	backward shift operator
d	differencing operator
k	lag
Δ	first difference operator

List of Abbreviations

Short form	Long form
AR	Autoregressive
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
FARIMA	Autoregressive Fractionally Integrated Moving Average
VAR	Vector Autoregressive
VARMA	Vector Autoregressive Moving Average
TAR	Threshold Autoregressive
TVAR	Threshold Vector Autoregressive
ADF	Augmented Dickey-Fuller test
PP	Phillips-Perron test
KPSS	Kwiatkowski-Phillips-Schmidt-Shin test
DWN	Discrete White Noise
acvf	auto-covariance function
acf	auto-correlation function
pacf	partial auto-correlation function
LM	Lagrange Multiplier
LBI	Local Best Invariant
SOI	Southern Oscillation Index
IOD	Indian Ocean Dipole
PDO	Pacific Decadal Oscillation
SAM	Southern Annular Model
MSLP	Mean Sea Level Pressure
SST	Sea Surface Temperature
ds	deseasonalised
dt	detrended
dsdt	deseasonalised and detrended

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed:

Date: ... 28/07 / 2021

Acknowledgements

- I would like to thank my principal supervisor, Associate Professor Andrew Metcalfe, for providing support and guidance during my candidature.
- I would like to thank Associate Professor Gary Glonek for providing insightful ideas and comments for my research.
- I thank Australian Research Council for a scholarship to pursue a Master of Philosophy.

Dedication

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Abstract

In this thesis, we review time series models and present two case studies. This first case study is an investigation into the effects of four climate indicators: the Southern Oscillation Index; the Indian Ocean Dipole; the Pacific Decadal Oscillation; and the Southern Annular Mode, on monthly rainfall, at four weather stations in South Australia. There is clear evidence that the Indian Ocean Dipole is Granger causal for rainfall. The second case study is a multi-scale analysis of atmospheric carbon dioxide, with sampling intervals of 0.5 million-year, one thousand-year, one year, and one week. The main finding is that all fitted models up to the first Industrial Revolution are borderline stable, whereas post-Industrial Revolution series shows a clear increasing trend.

Chapter 1

Introduction

1.1 Introduction

A time series is a sequence of discrete observations made over a period of time. Time series are generally considered to contain temporal information that cannot be replicated. They are important in many scientific fields as well as daily life. For instance, time series analysis includes problems in areas as diverse as finance, environmental research, medicine, criminology, political science, and social trends in TV shows or newspapers (Senin 2009). Generally, a time series is assumed to be equally spaced. However, this is not always the case especially when we have a time series that extends back for a long period. For example, the Vostok ice core data (discussed in [Chapter 4](#)) yield an unequally spaced CO₂ time series. Interpolation methods, such as linear interpolation and cubic spline interpolation, can be implemented to obtain an equally spaced time series, and also resampling method.

Many practical applications involve multiple time series. For instance, we may be interested in the temperature, humidity, amount of rainfall, and so on, over a certain period of time. The field of high-dimensional data is still under rapid development, especially for dependent data. Multivariate time series analysis consists of two major fields of study, explorative and descriptive analysis, and forecasting. The explorative and descriptive analyses are mainly focused on modelling the time series to best describe its dynamic system, and understanding the relationships between the time series variables. On the other hand, forecasting is focused on predicting future values, often based on the information obtained in the descriptive analysis. The better forecasting is helpful in future planning and decision making. Forecasting can also be classified as long-term and short-term fore-

casting. Long-term forecasting usually extend more than one year, 5-, 10-, 15-, 20-year projections are also common. Short-term forecasts are sometimes preferred rather than the long-term in the climatology disciplines, such as rainfall (e.g., Johnson & Bras 1980, Burlando et al. 1993, Toth et al. 2000). Whereas long-term predictions are more relevant to economic development and political decisions (e.g., Armstrong et al. 2010). Furthermore, forecasting analysis often involves questions of Granger causality, that is how do the past values influence the future values. Such questions are often treated within their domain rather than as part of the general discipline of time series analysis. As a result, a variety of disciplines have contributed novel ways of thinking about time series datasets (Nielsen 2021). In this research, we focused on the discipline of climatology.

Both of the descriptive and predictive analyses are based on the identification of patterns in the time series. In practice, time series often show some repeating pattern, or patterns, generally referred to as seasonality and a low frequency trend. The first stage in the time series analysis is to identify, estimate, and remove any seasonal patterns or trend. Then the deseasonalised and detrended time series is typically considered as a realisation of a stationary linear stochastic process. In particular, the variable is a linear combination, possibly infinite, of a sequence of independent random variables with a mean of zero and constant variance. Modifications to this model can then be considered if there are physical reasons or empirical evidence to support the additional complexity. Modifications include models for a changing variance, such as generalised autoregressive conditional heteroskedasticity models, fractionally differenced models for persistence, and non-linear time series.

This thesis is concerned with time series modelling, including an investigation of Granger causality between multiple time series variables, in the climatology discipline. Moreover, as we know that many real life time series exhibit non-stationary, long memory, or non-linear properties, and simple linear time series models may leave some dynamic features uncovered. Thus, we also looked into the models that can capture these properties, threshold vector autoregressive and autoregressive fractionally integrated moving average models, with two case studies.

In [Chapter 2](#) we discuss the basic concepts of time series analysis, and the widely used autoregressive integrated moving average models and their extensions.

[Chapter 3](#) is an investigation of the inter-relations between four climate indicators (Southern Oscillation Index, Indian Ocean Dipole, Pacific Decadal Oscillation, and Southern Annular Mode), and their relationships with monthly rainfall, at four weather stations, in South Australia.

In [Chapter 4](#), we present an analysis of five time series of estimated atmospheric CO₂ with

sampling intervals ranging from 0.5 million years up to one week. The main objective of this study is to look for consistent features of time series models at different scales.

In [Chapter 5](#), we give a brief overall conclusion. The conclusions of the two case studies are given at the end of their chapters.

All the analyses in the thesis are performed using the statistical software R (R Core Team 2020).

1.2 Overview of Time Series Analysis

Data are the cornerstones of time series analysis, and the innovations in time series analysis are tightly linked to the collection, record, and visualisation of data. Around 350 BC, scientists used instruments to collect weather-related data for short period of time, for instance, hourly and daily. Astronomers are the group of people that relied heavily on plotting objects, trajectories, and measurements over time, and they are masters of time series. In particular, the collection of sunspot time series was as early as 800 BC. The first time series data in medical discipline is the electrocardiograms, which were invented in 1901, to record the electrical signals in patients' heart (Nielsen 2019).

More formal mathematical analysis of time series analysis originally focused on one of two approaches, time domain analysis and frequency domain analysis. Both approaches assume that the variable is sampled at, or aggregated over, fixed time steps. The time domain approach aimed to model the current observations as the sum of a linear combination of past values and a random innovation, whereas the frequency domain focused on the frequency composition of the signal as determined by spectral analysis.

A formal mathematical approach to time series analysis was developed by Yule (1927), to study the time series behaviour of sunspots, as an application of autoregressive (AR) models. Wold (1938) developed Wold decomposition and the main result was that a stationary process can be expressed as a sum of a deterministic term and a stochastic term, where the latter term can be expressed as an infinite moving average, that is autoregressive moving average (ARMA) model. However, there was no likelihood function that enables researchers to derive maximum likelihood estimation for the parameters. Until 1970, the ARMA models became popular as Box & Jenkins (1970) proposed the Box-Jenkins method for choosing the order of the models and estimating the coefficients. The model selection criteria have been developed (e.g. Akaike 1974, Hannan 1980), as well as pattern identification methods (e.g. Gray et al. 1978) and an extended sample auto-correlation function for order determination for both stationary and non-stationary series (Tsay &

Tiao 1984).

Spectral analysis was originally used in engineering applications, and it is used to discover the periodicities in the observed data. Back to 1807, a mathematician Fourier argued that any series could be approximated as a sum of sine and cosine terms (e.g., Makridakis 1976). This idea was implemented by Schuster (1898) who used Fourier's expansions to estimate the length of periodicities that are hidden behind irregular fluctuations, and proposed the term periodogram. The periodogram is a method for estimating the power spectrum of a random signal from the observed data. Moreover, it is a graphical tool for spectral analysis and can be used to identify the dominant periods in the observed series. Schuster (1906) investigated the periodicities in the monthly sunspot data from 1749 to 1901. Nevertheless, Schuster (1906) found that if we split the data into two halves, the two periodograms behave differently, and there was no such logical explanation. Beveridge (1922) implemented periodogram analysis to discover the periodicity in wheat prices and rainfall in Western Europe. He also showed dissimilarities between first and second halves of the series, this was consistent with Schuster's findings in 1906. Additionally, his results showed periodogram is sensitive to small changes in the length of the periods, which are themselves determined by the record length. The consequence was the chance of missing significant peak, we might need to apply an extremely large number of decimal periods to avoid this. Kendall (1945) argued that periodogram can be misleading, and Bartlett (1950) proved this theoretically and showed sampling effects can distort the periodogram of a time series (e.g. Makridakis 1976). Spectrum (e.g. Bartlett 1948, 1950, Blackman & Tukey 1958), which is the smoothed periodogram, avoids the drawbacks of the classical periodogram analysis.

Quenouille (1957) had revealed the importance of analysing several related time series jointly. At that time, economists used linear regression models to explore the relationship between multiple time series. One of the features of regression is that it has a natural direction, and the direction of the regression should respect the direction of causation (Hoover 2006). However, there is nothing in the data themselves can reveal the correct direction. Granger (1969) proposed an approach to causality which is the most influential approach as his approach was data-based without direct reference to background economic theory, and for dynamic time series models. Furthermore, Granger et al. (1974) argued that using linear regression to investigate the inter-relation between multiple variables was limited because it could result in spurious claim of correlation, as detrending would not necessarily return a time series that is stationary (Granger 1981). Engle & Granger (1987) formalized the cointegrating vector approach, and coined the term. Their concept established that multiple non-stationary time series are integrated in a way that they move closely together in the long run. That is, two or more non-stationary time series share the common stochastic trend, and the linear combination of these series is a stationary process. Moreover, as the spurious correlations may not be eliminated by detrending, it

is recommended to check for cointegration of non-stationary time series.

Methods for the analysis of multivariate time series have been developed in structural specification of a vector system (Tsay 2000). However, multivariate analysis was confined to vector autoregressive (VAR) models years ago, as VAR models are easy to implement and understand, where the VAR model was introduced by Sims (1980) to model the joint dynamics and causal relationships of a set of macroeconomic variables. It then becomes popular recently in areas as diverse as biostatistics (e.g. Kirch et al. 2015), finance (e.g. Tao et al. 2011, Tsay 2013), marketing (e.g. Gelper et al. 2016), and climatology (e.g. Keng et al. 2017). Two main use of VAR models are exploring and analysing inter-relation of the time series variables, and forecasting. The performance of forecasting can be improved by using information on interactions among multiple time series (Wilms et al. 2017), this based on the notion of causality that introduced by Granger (1969). That is, if the past values of a time series variable $\{x_t\}$ helps to predict the future values of another time series variable $\{y_t\}$, then $\{x_t\}$ is Granger causal for $\{y_t\}$. VAR models were originally established for stationary time series variables, that is without trend. Researchers can implement VAR models after the process of detrending, which can be done by fitting linear regression on time. However, the importance of stochastic trends in economic variables has been discovered in 1980s, and a notion of cointegration has been introduced by Granger (1981) and Engle & Granger (1987) formalized the cointegrating vector approach, and coined the term.

The general form of the VAR model, vector autoregressive moving average (VARMA), provides better fit in practice and also results in more accurate forecasting than VAR model (Athanasopoulos & Vahid 2008) as it has more flexible covariance structure (Chan 2020). Nevertheless, Wilms et al. (2017) argued that “the VARMA model is not in general identifiable, and an identification restriction is prerequisite to estimation of it”. Furthermore, the estimation of a VARMA model is often unstable, for example in the presence of a higher-order polynomial in the corresponding objective function (Wang et al. 2020). Therefore, VAR models are commonly used as a close approximation to VARMA models.

Although the general formulation of the VAR model and its extensions are relatively recent, control engineers have been working with systems of simultaneous linear difference equations, known as a state space representations which are directly equivalent to VAR(1) models, since the early 1960s. Furthermore, a VAR(p) model can be expressed in state space form with a suitable definition of states. The state-space representation, also known as the time domain approach, is a mathematical model to model and analyse a physical system as a set of input, output and state variables related by first-order linear differential equations. A state space model is mainly consists with an observed data and the unobserved states. Both of the observed data and unobserved states are modelled to be stochastic processes. Autoregressions are a special case in which the state vector, the

previous p observations, are observed. All linear time series models have a state space representation. However, the correspondence between linear time series models and state space models are not unique (e.g., Aoki 1987). The most intensively studied state space model is the Kalman filter in 1961 (Kalman & Bucy 1961). Kalman filter was originally introduced to time series analysis for evaluating the exact Gaussian likelihood function of a model and dealing with missing values (Jones 1980). Kalman filter is an algorithm to provide optimal estimates for variables that cannot be measured directly given measurable variables over time. The Kalman filter uses a sequential algorithm for calculating least squares estimates, and Plackett published such an algorithm, in a different context, in 1950 (Plackett 1950). In time series analysis, it is used to forecasting and smoothing. It led to developments of signal extraction (e.g., Kohn & Ansley 1989, Wecker & Ansley 1983), smoothing and seasonal adjustment (e.g., Kitagawa & Gersch 1996), among others (Tsay 2000). However, this algorithm has some assumptions which can also be seen as its drawbacks, the assumption of linearity, the assumption of normality of process noise and measurement noise, and the assumption of independence between these two noises. The extended Kalman filter relaxed the linearity but we do not use it in this thesis.

It has been shown that some dynamic features of financial and economic time series cannot be explained only by linear time series models, such as asymmetry between increasing and decreasing patterns in a time series. Daniell (1946) firstly mentioned the concept of directionality for time series (e.g. Lawrance 1991), and the first formal definition of reversibility is introduced by Brillinger (1966). In the presence of directionality, non-linear time series models have better forecasting performance, and also lead to more realistic scenarios (Lawrance 1991). Therefore, non-linear time series models are attracting researchers' attention recently, as linear time series models have various restrictions and limitations in many empirical applications. An early study of directionality in non-linear time series models is threshold autoregressive (TAR) models that discussed by Tong (1978).

The threshold vector autoregressive (TVAR) models are the multivariate extension of the univariate TAR models, which extended by Tong & Lim (1980) and Tong (1983). A TVAR model is a non-linear time series model but allows for piecewise linear fitting to each regime. The general idea is that a process may have different dynamic features when the values of a variable exceed a certain threshold. For example, Aleem & Lahiani (2014) suggest that domestic prices in Mexico do not response to an exchange rate shock with monthly rate of inflation below 0.79%, but react strongly when the inflation is above that threshold value. Zheng (2013) conducted a study that use TVAR model to model the asymmetry effects of monetary policy on the US economy under the low and high financial stress regimes.

Research into time series analysis has trended to focus on econometric and financial data, and there are relatively fewer studies on other disciplines. Therefore, this thesis discussed

few popular models in the economics and applied to the climatology data. The main objective is to investigate the Granger causality using VAR and TVAR models.

Hydrologists noticed persistent auto-correlations in annual flows of the River Nile, and these can be modelled with higher order AR processes (Hurst 1951). Mandelbrot was motivated by this study, and worked on the theory of stochastic processes exhibiting heavy tails and long memory during 1960s, and he was the first to distinguish between these effects (e.g., Graves et al. 2017). The concept of fractional integration was initially suggested by Granger & Joyeux (1980) and Hosking (1984). A process follows a fractionally differenced model can be classified as a long memory process. In stationary time series, the term long memory, implies that there is non-negligible dependence between the present and all points in the past (Graves et al. 2017). Diebold et al. (1991) implemented autoregressive fractionally integrated moving average models to real exchange rates under the gold standard, and provided evidence of mean reversion (e.g., Gil-Alana & Toro 2002). The study of long memory processes is important because they exhibit non-intuitive properties where many statistical models fail to hold as a short-run proposition. We compare the fitness between short memory models (autoregressive models, autoregressive moving average models, and autoregressive integrated moving average models) and the long memory model (autoregressive fractionally integrated moving average model), with the application to the CO₂ time series at different sampling intervals. Furthermore, we look at the consistent features of the time series models at different sampling intervals.

Chapter 2

Methods

A time series is a series of observations recorded over a period of time, either discrete time or continuous time. In this thesis, we focus on discrete time series. A discrete time series of length n can be represented by $\{x_t : t = 1, \dots, n\} = \{x_1, \dots, x_n\}$. It has n observations sampled at discrete time $1, \dots, n$. The notion will be abbreviated to $\{x_t\}$ when the length of n does not need to be specified.

In this chapter, we firstly discuss the basic and important concepts in time series analysis, stationarity and non-stationarity. A first step in a time series analysis is to plot the data and check for anonymous observations. Time series plot is also a visualisation tool for exploring apparent deterministic features such as trend and seasonal variation. If a time series shows trend and seasonal variation features, to progress with the analysis we need to decompose the time series, either additive decomposition or multiplicative decomposition ([Section 2.6](#)). The residuals obtained after decomposition are deseasonalised and detrended series, which is typically considered as a sequence of independent random variables. However, in many cases, the deseasonalised and detrended series still show serial correlations. Identifying such correlations can improve forecasting performance dramatically if the series is highly auto-correlated. The sample auto-correlations can be calculated, and the plot of the correlations is named as correlogram. Moreover, sample partial auto-correlations are sample auto-correlations given the information of correlations of previous lags, are useful for determining the order of autoregressive term ([Section 2.7](#)). Spectral analysis ([Section 2.7.6](#)) can be used before fitting time series models, and it allows us to discover underlying periodicities. We then introduce three univariate time series models: autoregressive moving average (ARMA) models which are stationary ([Section 2.8](#)); autoregressive integrated moving average (ARIMA) models which are variations on a random walk and so non-stationary in the variance ([Section 2.9](#)); and autoregressive fractionally integrated moving average (FARIMA) models which are stationary but have

long memory and are fractal processes (Beran et al. 1998) (Section 2.10).

The time series analysis involves two or more variables is more complicated than the univariate analysis, and become of increased interest in recent years. This is because multivariate time series analysis can be used to capture and explain the inter-dependencies and co-movements among the variables (Masum 2020). Moreover, it is important to investigate the co-integration between multiple time series because the ordinary regression method will give spurious results (Section 2.11). Unit root tests are useful tools for investigating co-integration (Section 2.12). If two or more time series variables are co-integrated, the pre-whitening procedure should be implemented in order to explore the relationship between them (Section 2.13), whether the variables are associated or Granger causal (Section 2.14). Two multivariate autoregressive models, vector autoregressive (VAR) models which are linear (Section 2.15), and threshold vector autoregressive (TVAR) models which are non-linear (Section 2.16), are discussed and used as tools for investigating Granger causality.

2.1 Stationarity

Stationarity plays a fundamental role in time series analysis, and is a property of many stochastic models. Two different types of stationarity are commonly defined, strict stationarity and weak stationarity. A time series variable $\{x_t\}$ is strictly stationary if the joint statistical distribution of x_{t_1}, \dots, x_{t_n} is the same as the joint distribution of $x_{t_1+m}, \dots, x_{t_n+m}$ for all t_1, \dots, t_n , and m . Mathematically,

$$F(x_{t_1+m}, \dots, x_{t_n+m}) = F(x_{t_1}, \dots, x_{t_n}). \quad (2.1)$$

In particular, $E[x_t]$ is constant and the auto-covariance $\text{cov}(x_t, x_s)$ only depends on lag $k = |t - s|$. A time series variable $\{x_t\}$ is weakly stationary if

- $E[x_t^2] < \infty$ for all t
- $E[x_t] = \mu$ for all t
- $\text{cov}(x_t, x_s)$ only depends on $t - s$ for all t and s .

The condition $E[x_t^2] < \infty$ for all t is the condition for variance to exist. Weak stationarity does not imply strict stationarity, and strict stationarity does not necessarily imply weak stationarity. As the condition $E[x_t^2] < \infty$ for all t is not assumed in the definition of strict stationarity. That is, if a time series $\{x_t\}$ is strictly stationary and has finite second moment ($E[x_t^2] < \infty$ for all t), then $\{x_t\}$ is a weakly stationary process.

2.2 Ergodicity and Stability

A stationary time series model is ergodic in the mean if the time average of any realisation tends to the mean. Ergodicity in higher moments is defined in a similar manner. An alternative definition of ergodicity is that the evolution of the process is independent of the initial value. In practice, a climatological time series is considered as a single realisation of an underlying ergodic process. An ergodic weakly stationary process is stable because any realisation will return to the mean if innovations are set to 0. In particular, in the case of linear autoregressive processes, stationarity is identical to stability. A crucial part of the estimation procedure is to determine whether the model is stable or, equivalently, whether the parameters are within the appropriate parameter space for stationarity (Tjøstheim 1994). Stationarity justifies simulation and resampling procedures which can be used for inference (e.g., Cline & Huay-min 2001).

2.3 White Noise

A basic stationary process is white noise. A discrete white noise (DWN), say $\{w_t\}$, is a time series with all the observations are independent and identically distributed with a mean 0 and variance σ^2 . A DWN process is strictly stationary. Additionally, if the variable follows a normal distribution, the series is called Gaussian white noise. Suppose we have a white noise process $\{w_t\}$ with mean 0 and variance 1, the time series plot (Figure 2.1) does not show any decreasing or increasing with time, and the variance is constant over time.

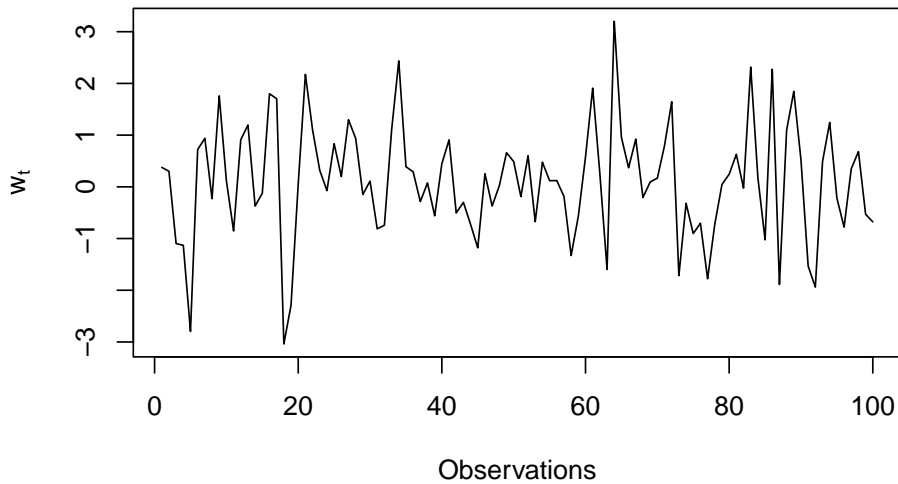


Figure 2.1: Time series plot of a realisation of DWN process with mean 0 and variance 1.

The residuals obtained after fitting an appropriate time series model should resemble a realisation of a white noise process, this can be examined using a correlogram which is discussed in detail in [Section 2.7.2](#).

2.4 Non-stationarity

A feature of stationary processes is that the mean is constant. However, we could also have time series that are not constant in means due to the presence of trends or seasonal patterns, these time series are classified as non-stationary processes. A well known example of a non-stationary time series is the monthly totals of international air travel passengers from January 1949 to December 1960 ([Figure 2.2](#)). The data are available in R and the data are stored as a time series.

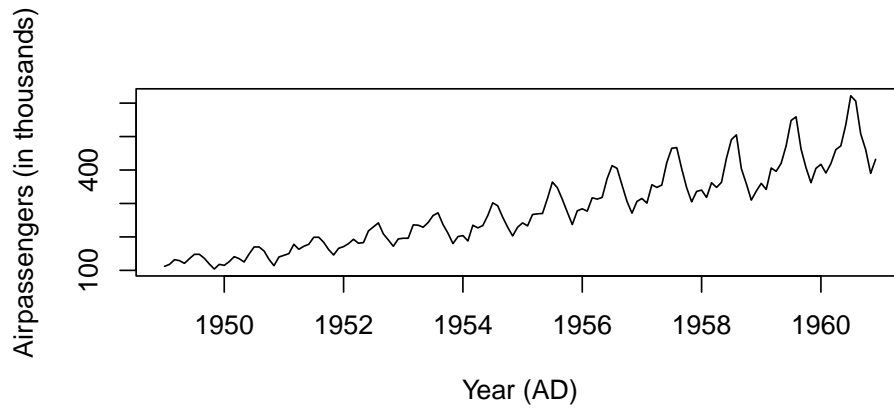


Figure 2.2: Time series plot of the Airpassenger data.

The time series appears to be periodic but increasing variation which is known as the seasonal variation, that is the series is increasing from the beginning of the year and reach a peak around June, then decreasing from June till the end of the year. In other words, the series share the same seasonal pattern for each individual year, but with non-constant variance as the seasonal variation is increasing over time. Taking log transformation of the series will give constant variance (Figure 2.3).

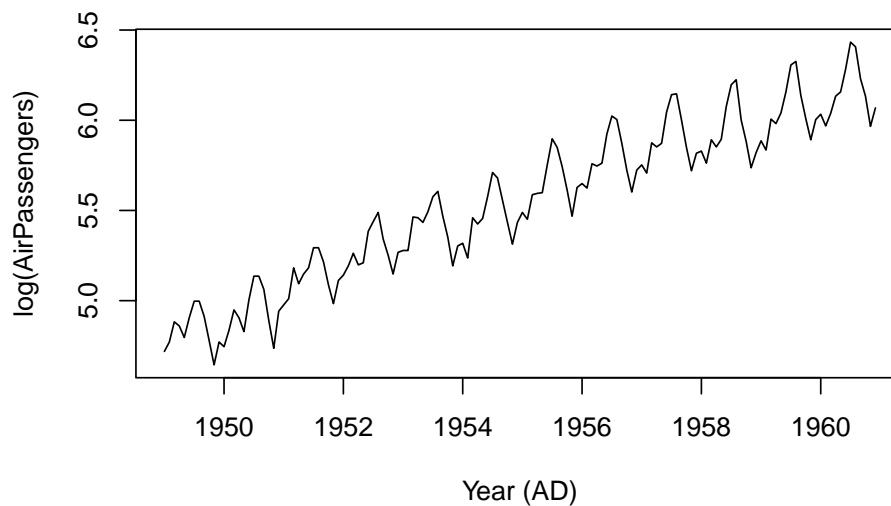


Figure 2.3: Time series plot of the log transformed Airpassenger data.

There is also a clear trend as the total number of international airline passengers increase with time. Decomposition of such time series provides an initial analysis that aims to identify seasonal variation in the mean together with an empirical low frequency component which represents a trend. The decomposition method is discussed in details in [Section 2.6](#).

2.5 Random Walk

A time series can be classified as non-stationary not only because of the non-constant mean value but also because of the non-constant variance. The random walk is a basic non-stationary process that is not constant in variance. The data with stochastic trends are well modelled by random walk processes, particularly in finance. A pure random walk is a time series that the current value depends on the previous value plus a random error term. Mathematically, a time series $\{x_t\}$ is a pure random walk if

$$x_t = x_{t-1} + w_t, \tag{2.2}$$

and given $x_0 = \mu$, where $\{w_t\}$ is a discrete white noise. It is stationary in the mean because $E[x_t] = \mu$ but $\text{var}(x_t) = t\sigma^2$, that is, the variance increases as time t increases. The time series plot of a simulated random walk ([Figure 2.4](#)) shows the series is increasing due to the non-constant variance. It is important that the increase in series cannot be classified as a deterministic trend, unit root tests can be used as a tool to determine whether a non-stationary process should be detrended or differencing to render the stationary. One of the crucial features of a random walk process is that it need not return to a fixed level, and this is a characteristic of financial and economic variables.

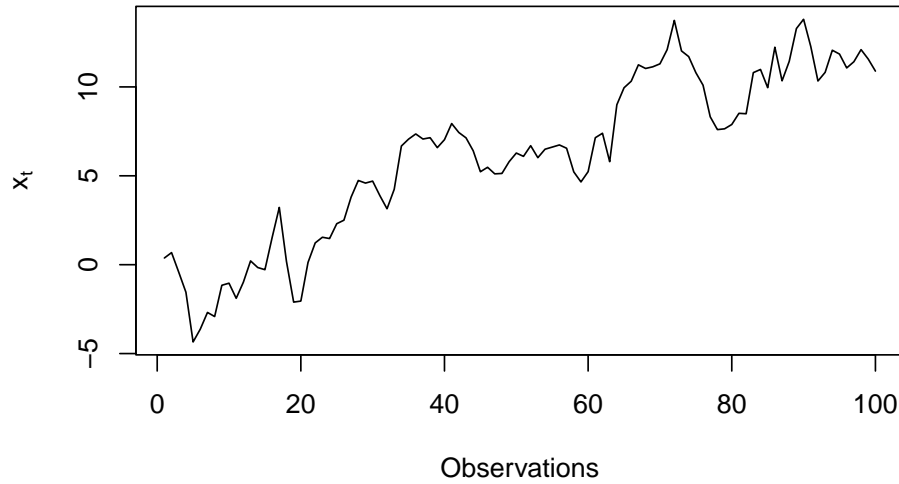


Figure 2.4: Time series plot of the simulated pure random walk.

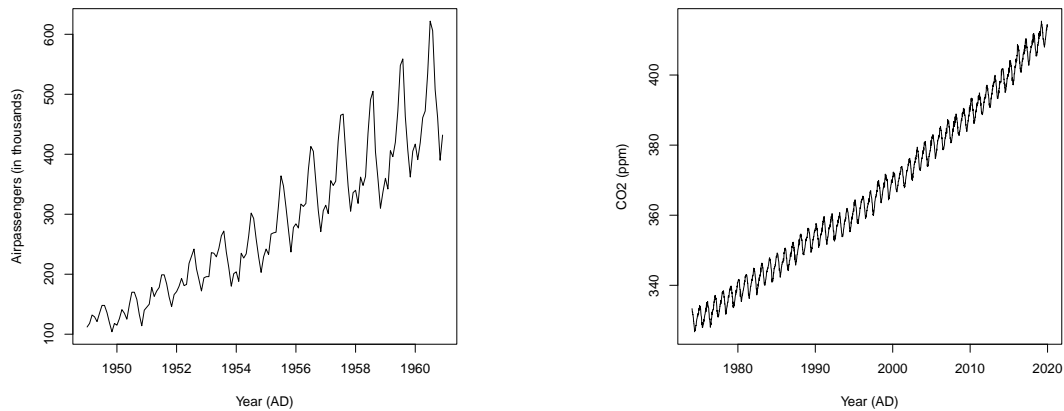
2.6 Decomposition of Series

Decomposition is an important technique for time series analysis as it gives us brief knowledge of the data and hence helps us to better analyse the data (Chourasia 2020). Decomposition is a process that can split the data into three components: trend, seasonal variation, and random error (noise). Based on the dominance of trend and seasonal effects in time series models, two models that are based on these components are additive decomposition model (Equation 2.3) and multiplicative decomposition model (Equation 2.4), defined below, respectively.

$$x_t = m_t + s_t + z_t, \quad (2.3)$$

$$x_t = m_t * s_t * z_t, \quad (2.4)$$

where $\{x_t\}$ is the observed series, $\{m_t\}$ is the trend, $\{s_t\}$ is the seasonal effects, and $\{z_t\}$ is an error term. The random error, $\{z_t\}$, is a sequence of correlated variables with mean zero in general. The main difference between these two models is, for the multiplicative decomposition model the seasonal effect tends to increase as the trend increases (Figure 2.5a), whereas additive decomposition model does not (Figure 2.5b).



(a) Monthly totals of international airline passengers, 1949 to 1960

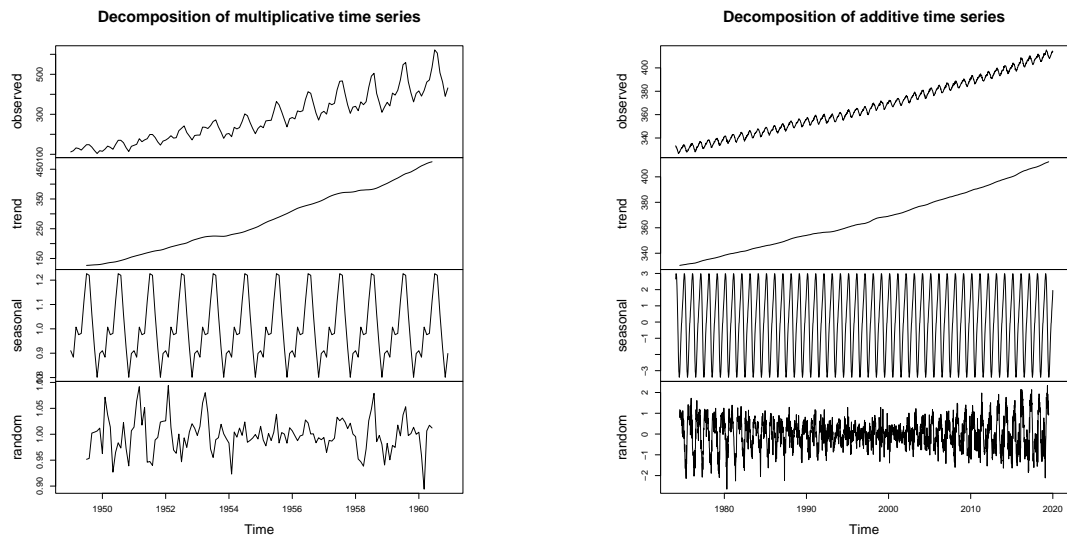
(b) Weekly CO_2 from May 1974 to January 2020, Mauna Loa.

Figure 2.5: Time series plots showing trends and seasonal variations.

A multiplicative model has non-constant variance as the seasonal variation increases over time, we can take the log transformation on the time series variable $\{x_t\}$ in order to have an additive model, as an example we showed in [Section 2.4](#) (The Airpassenger data). That is,

$$\log(x_t) = m_t + s_t + z_t. \quad (2.5)$$

The `decompose` function in R allows us to specify the type of the model, whether it is additive or multiplicative. As well as the original series, the `decompose` function also returns trend, seasonal and random components ([Figure 2.6](#)).



(a) Monthly totals of international airline passengers, 1949 to 1960

(b) Weekly CO₂ from May 1974 to January 2020, Mauna Loa.

Figure 2.6: Decomposition of the series.

There are clear increasing trends and seasonal variation for both of the series. The random component returned should be an estimate of a realisation of the random error, $\{z_t\}$, as it is obtained by subtracting estimated trend and seasonal components from the observations. We can examine whether the residuals can reasonably be considered as a realisation of a white noise process by plotting correlogram ([Section 2.7](#)).

2.7 Correlation

Once we decomposed a time series, the random component can be obtained and should resemble as a realisation of a white noise process. However, the random component cannot be modelled by independent random variables in many situations, particularly in finance, as consecutive values of the random component are correlated. We can improve the forecasting performance if we identify such correlations, especially when the correlations are high.

2.7.1 Auto-correlation

The auto-correlation plays an important role in the time series analysis as it helps us to explore patterns in the data, and hence select the best prediction model (Georgiou 2019). Correlation is a measure of the relationship between two variables, whereas auto-correlation measures the correlation of a time series variable with itself at different times. The lag k auto-covariance function (acvf) for a weak stationary time series model is defined as

$$\gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)]. \quad (2.6)$$

Thus, the lag k auto-correlation function (acf) is defined as

$$\rho_k = \frac{\gamma_k}{\sigma^2}. \quad (2.7)$$

By definition, $\rho_0 = 1$ as $\gamma_0 = \text{var}(x_t) = \sigma^2$.

The population acf is only defined for stationary process, whereas sample acf is defined for both of the stationary and non-stationary process. The sample auto-covariance function, c_k , can be calculated from a time series by its sample equivalent

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}), \quad (2.8)$$

where n is the number of observations. Hence, the sample auto-correlation function, r_k , is defined as

$$r_k = \frac{c_k}{c_0}. \quad (2.9)$$

One of the assumptions of the acf is that the observations are equally spaced. Variograms relax the assumption of equal spacing, and it describes the covariance structure over space, or time, as a function of distance h between points, and is defined as

$$2\gamma(h) = E[(Z(u) - Z(u+h))^2] \quad (2.10)$$

for all locations u , where $Z(s)$ is the value of the variable of interest at location s .

The linear interpolation and cubic spline interpolation are two methods that are commonly used to deal with unequal spaced data. The case study on multi-scale analysis of CO₂ uses variograms to choose between two interpolation methods.

2.7.2 Correlogram

The sample acf (Equation 2.9) can be calculated in R using `acf` function, this function also returns a plot called correlogram that shows the correlations against lag k . The blue dashed lines returned in the correlogram (e.g. Figure 2.7) are approximate limits within which 95% of r_k would lie if the time series is discrete white noise. The limits in R are calculated as:

$$-\frac{1}{n} \pm \frac{2}{\sqrt{n}}, \quad (2.11)$$

where n is the number of observations. Therefore, we expect 95% of r_k of a white noise process lie in the blue dashed lines, except the correlation at lag 0, which equals to 1 by definition. Here, we use Airpassenger data as an example (Figure 2.5a). There is a clear increasing trend as well as periodicity with increasing variance in the series, thus it has the property of non-stationarity. As log transformation can eliminate the increasing variance in the periodicity, we then plot the correlograms of the original series and the log transformed series (Figure 2.7). There is no apparent difference between the two correlograms, except the correlogram of the log transformed series decays slower than the original series.

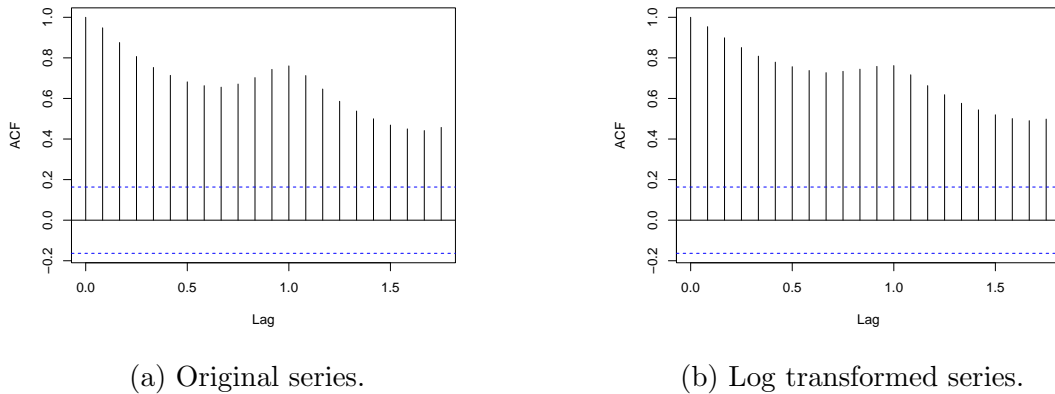


Figure 2.7: Time series plots of the AirPassenger data.

The y-axis shows the auto-correlation (r_k) at each lag (x-axis). It is noticeable that the x-axis has decimal values. This is because the default index in R is year, and we have monthly data in this case. For example, lag 0.5 represents 6 months (half year), it is also the sixth lag from the lag 0. The correlograms (Figure 2.7) decay slowly from lag 1, this is the indication of the presence of a trend. Additionally, there is a spike at a lag of 12 months which suggests the presence of seasonality. Even though the correlogram is a way to explore the presence of trends and seasonal patterns, we do not necessarily rely on it.

The main application of the correlogram is to detect the auto-correlations in the time series after we have removed or reduced any existing trends and seasonal patterns. The Airpassenger data has been detrended and deseasonalised in R using `decompose` function, the correlogram of the residuals is shown in [Figure 2.8](#) (either use multiplicative model for original series or use additive model for log transformed series).

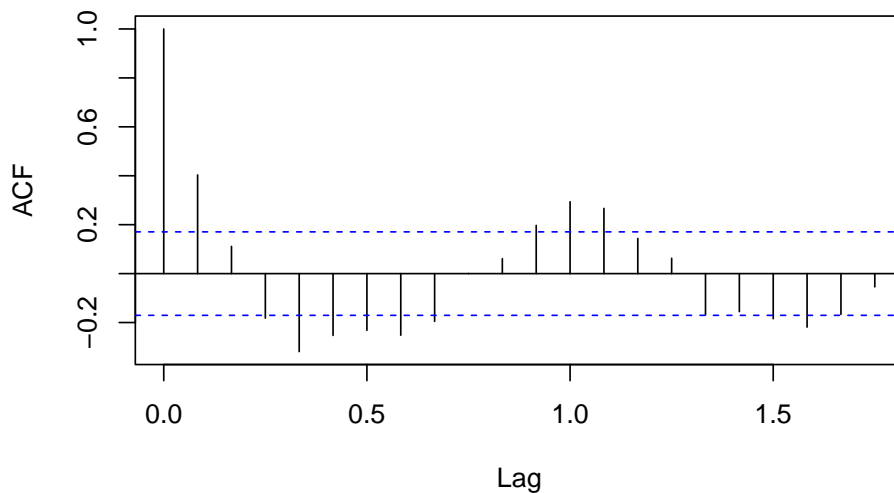


Figure 2.8: Correlogram of the detrended and deseasonalised Airpassenger data.

The correlogram ([Figure 2.8](#)) shows damped cosine shape, which suggests an AR(2) process (see [Section 2.8](#)) or the presence of seasonal variation. The latter explanation is unreasonable because we applied decomposition method which allows the monthly variations. The standard deviation of the original series is 109.42, and it decreased to 0.03 for the detrended and deseasonalised series. The reduction in the standard deviation suggests the seasonality has been removed effectively. Thus, this series might be reasonable to be modelled as an AR(2) process.

2.7.3 Partial auto-correlation

General speaking, a partial auto-correlation is a conditional auto-correlation, that is the correlation between two time points for a time series variable, say x_t and x_{t+k} , conditional on the set of observations between these two time points t and $t+k$. Mathematically, the lag k partial auto-correlation (pacf) is defined as:

$$\rho'_k = \frac{\text{cov}(x_t, x_{t+k} | x_{t+1}, \dots, x_{t+k-1})}{\sqrt{\text{var}(x_t | x_{t+1}, \dots, x_{t+k-1}) \text{var}(x_{t+k} | x_{t+1}, \dots, x_{t+k-1})}}. \quad (2.12)$$

By definition, the lag 1 partial auto-correlation ρ'_1 equals to the lag 1 auto-correlation ρ_1 . For example, we expect the partial auto-correlation of all the lags except lag 1 equal to zero for an AR(1) process. In general, for an AR(p) process, the partial auto-correlation will be zero for all lags greater than p . Therefore, it could be useful to calculate the estimated partial auto-correlation in order to determine the order of a suitable AR process for a time series.

2.7.4 Partial correlogram

The sample partial auto-correlation can be calculate in R using `pacf` function, this function also returns a plot called partial correlogram that shows the pacf against lag k . The partial correlogram of the detrended and deseasonalised Airpassenger data (Figure 2.6a) has been obtained as an example (Figure 2.9).

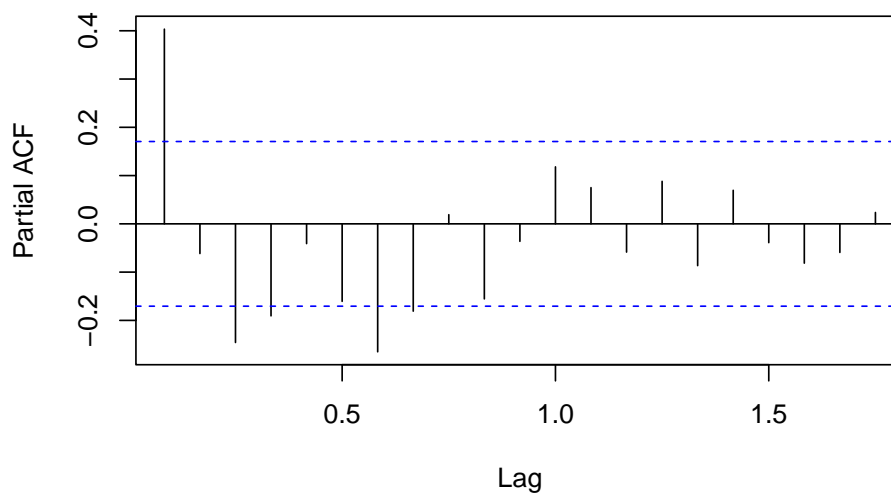


Figure 2.9: Partial correlogram of the detrended and deseasonalised Airpassenger data.

2.7.5 Cross correlation

Cross correlation is a standard method of estimating to the degree to which two series are correlated. The lag k sample cross covariance function (ccvf), $c_k(x, y)$, is defined as

$$c_k(x, y) = \frac{1}{n} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x})(y_t - \bar{y}). \quad (2.13)$$

Thus, the lag k sample cross correlation function (ccf), $r_k(x, y)$, is defined as

$$r_k(x, y) = \frac{c_k(x, y)}{\sqrt{c_0(x, x)c_0(y, y)}}. \quad (2.14)$$

It is important to note that the cross correlation function is not symmetric, and the time series variable $\{x_t\}$ is lagging $\{y_t\}$ by k . If the spikes in the ccf between $\{x_t\}$ and $\{y_t\}$ at negative lags, then the past values of $\{x_t\}$ are associated with $\{y_t\}$, and the inclusion of the past values of $\{x_t\}$ help to predict future values of $\{y_t\}$. In contrast, if the spikes in the ccf between $\{y_t\}$ and $\{x_t\}$ at negative lags, then the inclusion of the past values of $\{y_t\}$ help to predict future values of $\{x_t\}$. The ccf function in R gives a plot called cross correlogram, and the lag k value returned by `ccf(x, y)` estimates the correlation between x_{t+k} and y_t .

2.7.6 Spectrum

The spectrum is formally defined as the Fourier transform of the population auto-correlation. But, this is not an intuitive description. The spectrum can be estimated by: fitting a finite Fourier series, with term at 1 cycle per record length up to 0.5 cycle per sampling interval; calculating the squared amplitude of each contribution; plotting the squared amplitude against frequency; smoothing the plot by averaging neighbouring ordinates. This is known as smoothed spectrum. A spectrum of the deseasonalised and detrended Airpassenger data is obtained in [Figure 2.10](#).

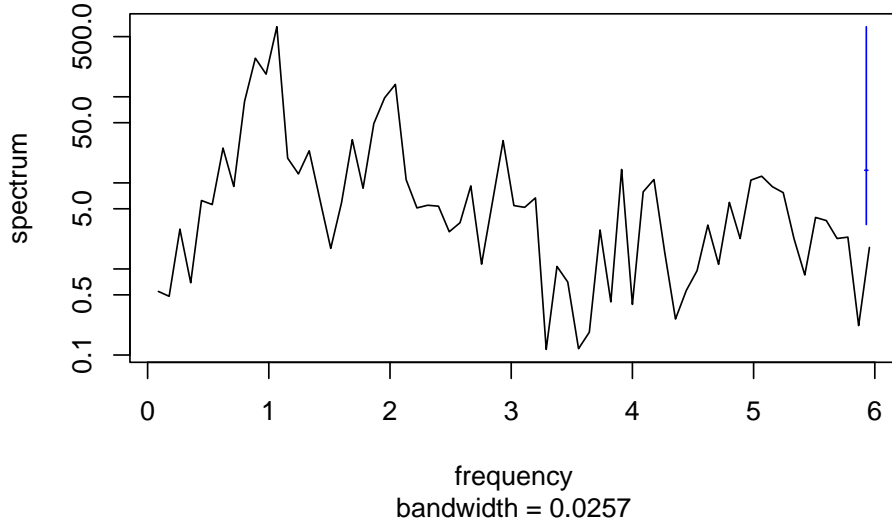


Figure 2.10: Spectrum of the detrended and deseasonalised Airpassenger data.

2.8 Autoregressive Moving Average (ARMA) Model

An autoregressive (AR) model of order p , denoted $AR(p)$, is a linear regression of the current value of a variable, $\{x_t\}$ say, on p past values. Mathematically, an $AR(p)$ model has the form:

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + w_t, \quad (2.15)$$

where $\{w_t\}$ is a white noise process. A moving average (MA) model of order q , denoted $MA(q)$, is a linear combination of q past errors. A $MA(q)$ model has the form:

$$x_t = \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q}, \quad (2.16)$$

where $\{w_t\}$ is a white noise process. These can be generalised to an ARMA process of order (p, q) , denoted $ARMA(p, q)$,

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q}, \quad (2.17)$$

where $\{w_t\}$ is a white noise process. The backward shift operator \mathbf{B} is defined as $\mathbf{B}x_t = x_{t-1}$, and the Equation 2.17 can be rewritten in terms of the backward shift operator \mathbf{B} ,

$$\Phi(\mathbf{B})x_t = \Theta(\mathbf{B})w_t, \quad (2.18)$$

where Φ and Θ are polynomials of orders p and q , respectively.

Ex 2.8.1. Let a time series variable $\{x_t\}$ to be a realisation of an ARMA(2, 2) process

$$x_t = 0.5x_{t-1} - 0.25x_{t-2} + w_t + 0.5w_{t-1} - 0.3w_{t-2},$$

where $\{w_t\}$ is a white noise process. Using backward shift operator \mathbf{B} , we yields

$$(1 - 0.5\mathbf{B} + 0.25\mathbf{B}^2)x_t = (1 + 0.5\mathbf{B} - 0.3\mathbf{B}^2)w_t.$$

Thus, we have $\Phi(\mathbf{B}) = 1 - 0.5\mathbf{B} + 0.25\mathbf{B}^2$, and $\Theta(\mathbf{B}) = 1 + 0.5\mathbf{B} - 0.3\mathbf{B}^2$.

An ARMA(p, q) model is strictly stationary, and hence stable, if all the roots of the polynomial

$$(1 - \phi_1 z - \dots - \phi_p z^p) \tag{2.19}$$

lie outside the unit circle of the complex plane, where z is a complex variable. In particular, AR(1) is stationary for $|\phi_1| < 1$, and AR(2) is stationary for $|\phi_2| < 1$ and $\phi_2 \pm \phi_1 < 1$. For any AR(p) processes with p greater than 2, we can solve the polynomial (Equation 2.19), the process is stationary and hence stable, if absolute values of all the roots are greater than 1.

Ex 2.8.2. Suppose we have a time series $\{x_t\}$ follows an ARMA(2, 1) process, that is

$$x_t = 0.3x_{t-1} + 0.8x_{t-2} + w_t + 0.7w_{t-1}.$$

In this case, $\phi_1 = 0.3$ and $\phi_2 = 0.8$. The series $\{x_t\}$ is said to be stationary if it satisfies two conditions: $|\phi_2| < 1$, and $\phi_2 \pm \phi_1 < 1$. Thus, the time series variable $\{x_t\}$ is non-stationary as $\phi_2 + \phi_1 = 0.8 + 0.3 = 1.1$ is greater than 1.

Fitting an ARMA model in R using `arma` function needs us to pre-specify the appropriate values of p and q , which can be determined by `pacf` and `acf`, respectively. For example, partial auto-correlation function has a cut-off after lag p suggests the autoregressive term of order p . However, we cannot totally rely on this. Brockwell et al. (1991) recommended using Akaike information criterion (AIC) to choose between different ARMA(p, q) models.

2.9 Autoregressive Integrated Moving Average (ARIMA) Model

Many time series are non-stationary in the mean due to periodicity or trends. Differencing can be used to remove a stochastic trend, as in a random walk, or deterministic trend,

as in the case of a linear trend. For a random walk process $x_t = x_{t-1} + w_t$, first-order differencing ($d = 1$) will return the white noise process $\{w_t\}$ ($x_t - x_{t-1} = w_t$), and thus a stationary process. The crucial feature of an unstable process is that it does not return to some fixed level, a well known example being the random walk as they are non-stationary in variance. Conversely, a stationary moving average process is obtained if we first-order difference a time series $x_t = \alpha + \beta t + w_t$, that is $x_t - x_{t-1} = \beta + w_t - w_{t-1}$. Consequently, we yields an MA(1) process rather than the white noise process $\{w_t\}$. Therefore, we should be decide discreetly whether it is sensible to remove a deterministic trend using differencing (Cowpewartwait & Metcalfe 2009).

Autoregressive integrated moving average (ARIMA) model, $\text{ARIMA}(p, d, q)$, allows for a non-stationary series to be differenced d times before fitting an $\text{ARMA}(p, q)$ model. So, the general definition is as follows:

$$\begin{aligned} (1 - \phi_1 \mathbf{B} - \dots - \phi_p \mathbf{B}^p)(1 - \mathbf{B})^d x_t &= (1 + \theta_1 \mathbf{B} + \dots + \theta_q \mathbf{B}^q) w_t, \\ \Phi(\mathbf{B})(1 - \mathbf{B})^d x_t &= \Theta(\mathbf{B}) w_t, \end{aligned} \quad (2.20)$$

where $\{w_t\}$ is a white noise process.

Typically a range of ARIMA models can provide plausible fits to data from a dynamic system. The AIC is one option for choosing between them, but it tends to favour models with a large number of parameters, and the AIC may be appropriate for forecasting applications.

2.10 Autoregressive Fractionally Integrated Moving Average (FARIMA) Model

ARIMA models are popular and widely used in many disciplines. However, Granger (1980) showed that the aggregation of equations that involved lagged dependent variables, can lead to a long memory property. Thus, the use of ARIMA models for analysis are not appropriate. Granger & Joyeux (1980) introduced the concept of fractional differencing and proposed autoregressive fractionally integrated moving average (FARIMA) model, to capture the long memory property in time series. Long memory processes have marked auto-correlations at high lags. Correlograms of such series show more slowly decays than exponential decays, and usually modelled as power-like decays (Joshi 2016).

Ex 2.10.1. The correlogram of the FARIMA(1, 0.4, 0) is obtained as an example (Figure 2.11). We can observe that the auto-correlation at lag 23 is still highly significant.

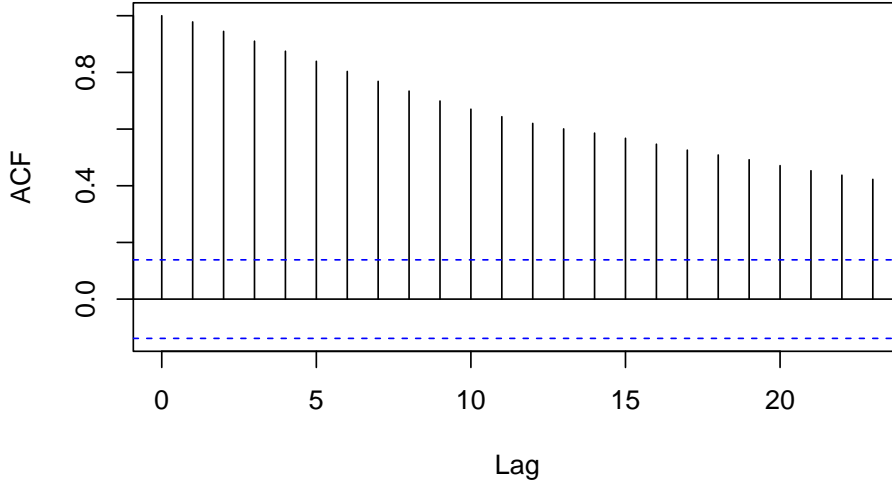


Figure 2.11: Correlogram of the FARIMA(1, 0.4, 0).

Mathematically, FARIMA models have the form

$$\Phi(\mathbf{B})(1 - \mathbf{B})^d x_t = \Theta(\mathbf{B})w_t, \quad (2.21)$$

for $-0.5 < d < 0.5$, and $\{w_t\}$ is a DWN process with mean 0 and variance σ_w^2 .

FARIMA models are useful for capturing both short-range and long-range dependence, whereas ARIMA models are considered to model the short-memory time series only. The range $0 < d < 0.5$ gives a long memory process, and it lies between a stationary AR(1) model and a non-stationary random walk (Beran et al. 1998). We are interested in this model because it is close to being a random walk yet still stays bounded, and returns to a mean value if there are no error inputs. Formally, the $(1 - \mathbf{B})^d$ term in the FARIMA model is expanded with the generalised binomial expansion:

$$1 - d\mathbf{B} + \frac{d(d-1)}{2!}\mathbf{B}^2 - \frac{d(d-1)(d-2)}{3!}\mathbf{B}^3 + \dots, \quad (2.22)$$

as far as \mathbf{B}^L , where L is some cut-off point typically around 30.

Ex 2.10.2. If we have $d = 0.45$ and $L = 30$, then

$$(1 - \mathbf{B})^{0.45} x_t \approx x_t - 0.45x_{t-1} - 0.124x_{t-2} - \dots - 0.002x_{t-30}.$$

For the value of d is between -0.5 and 0 , we say the time series has intermediate memory or long-range negative dependence. Moreover, a FARIMA($p, 0, q$) has short-memory property ($d = 0$), and corresponds to a stationary and invertible ARMA(p, q) model. As Granger & Joyeux (1980) shown that the auto-correlation function of a long memory process follows that

$$\rho_k = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(k+1-d)} \quad (2.23)$$

$$\approx \frac{\Gamma(1-d)}{\Gamma(d)} k^{2d-1}, \quad (2.24)$$

where lag k is a positive value to represent the number of lag, and $\Gamma(\cdot)$ is a Gamma function. Therefore, for any value of $d \geq 0.5$, the process is non-stationary as it has infinite variance. For any value of d such that $-0.5 < d < 0.5$, a FARIMA($p, d, 0$) model can be treated as an AR process with order $p + L$, where L is a cut-off point, which has form:

$$\Phi(\mathbf{B})(1 - \mathbf{B})^d x_t = w_t. \quad (2.25)$$

A long memory process $\{x_t\}$ is said to be stationary and invertible if all the roots of $\Phi(\mathbf{B})$ lie outside the unit circle and $-0.5 < d < 0.5$.

Ex 2.10.3. Suppose we have a time series $\{x_t\}$ follows a FARIMA(4, 0.3, 0) process, that is

$$(1 - 0.2\mathbf{B} - 0.4\mathbf{B}^2 - 0.12\mathbf{B}^3 + 0.65\mathbf{B}^4)(1 - \mathbf{B})^{0.34} x_t = w_t \quad (2.26)$$

To find the roots of the polynomial $1 - 0.2z - 0.4z^2 - 0.12z^3 + 0.65z^4$, we use `polyroot` function, the absolute values of all the roots are greater than 1, they are 1.10, 1.12, 1.12, and 1.10. Therefore, the series $\{x_t\}$ is a stationary process, and hence stable.

2.11 Cointegration

Before the introduction of the concept of cointegration, researchers normally used linear regression to explore the relationship between non-stationary time series, and also they

sought to identify direct causation between predictor variables and response. However, Granger et al. (1974) showed linear regression approach could produce spurious claims of causality as the multiple time series may share a common stochastic trend. In other words, the results of linear regression models may indicate that variables are associated to each other even though there is no causal relationship between the time series variables. More formally, two non-stationary time series $\{x_t\}$ and $\{y_t\}$ are said to be cointegrated if a linear combination of these two time series is stationary.

Ex 2.11.1. Given two time series variables, $\{x_t\}$ and $\{y_t\}$, we have

$$x_t = \mu_t + u_t \tag{2.27}$$

$$y_t = \mu_t + w_t, \tag{2.28}$$

where $\{\mu_t\}$ is a random walk process, and $\{u_t\}$ and $\{w_t\}$ are two independent white noise processes with mean zero. Both of the series $\{x_t\}$ and $\{y_t\}$ are non-stationary as they contain the random walk component. However, the linear combination of $\{x_t\}$ and $\{y_t\}$, $x_t - y_t$, is stationary as $u_t - w_t$ is a stationary white noise process. Therefore, the series $\{x_t\}$ and $\{y_t\}$ are cointegrated and they share a common stochastic trend.

2.12 Unit Root Tests

Unit root tests can be used to determine whether a non-stationary process should be detrended or differencing to render the stationarity. Furthermore, if two time series are cointegrated, then they share the common stochastic trend. Unit root tests are used for checking the presence of stochastic trend in univariate time series. If both of them show evidence to reject the hypothesis of presence of unit root, then the two series are not cointegrated. Otherwise, cointegration tests should be performed. That is, testing for unit root is a first step in the cointegration modelling. Three unit root tests are discussed, the Augmented Dickey-Fuller (ADF) test, the Phillips-Perron (PP) test, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The ADF and the PP test a null hypothesis of a unit root, which is non-stationary in variance. Conversely, the KPSS tests a null hypothesis of stationary in mean or variance. Hassler & Wolters (1994) showed that the ADF test performs poorly for fractionally integrated time series. Whereas the PP test is consistent under non-stationary fractionally integrated processes (Lee & Shie 2004). Lee & Schmidt (1996) showed that the KPSS test is consistent against stationary long memory alternatives. In other words, the KPSS test can be used to detect long memory, with short memory as the null hypothesis (e.g., Kokoszka & Young 2016). It is

a good idea to use unit root tests with different null hypotheses, as we cannot accept the null hypothesis even we have no evidence to reject the null hypothesis when p-value is greater than the expected significance level.

2.12.1 Augmented Dickey-Fuller (ADF) test

Dickey & Fuller (1979) invented the Dickey-Fuller (DF) test, for testing unit root with the null hypothesis of the presence of unit root ($\rho - 1 = 0$), in the series $y_t - y_{t-1} = (\rho - 1)y_{t-1} + u_t$. However, as it only considers AR(1) models, and thus has extremely low power on testing unit root. Therefore, an Augmented Dickey-Fuller (ADF) test was proposed in which allows for a more general process, that is higher-order autoregressive process (Dickey & Fuller 1979). The model for the ADF test is as follows:

$$\Delta y_t = \beta y_{t-1} + \sum_{k=1}^K \alpha_k \Delta y_{t-k} + w_t,$$

where $\Delta = 1 - \mathbf{B}$ is the first difference operator, K is the lag order of the autoregressive process, and β and α_k are constants for $k = 1, \dots, K$. The summation term in the equation above is to correct for the presence of serial correlation.

The null hypothesis of the ADF test is the presence of unit root, that is $H_0: \beta = 0$ against the alternative which is $H_a: \beta < 0$. We compare a t-statistic with the values of this Dickey-Fuller distribution. If the t-statistic less than the critical value, then we have evidence to reject the null hypothesis of presence of unit root.

There are two functions in R that can perform the ADF test: `adf.test` in the `aTSA` package (Qiu 2015), and `ur.df` in the `urca` package (Pfaff 2008a). We need to pre-specify the lag order K . The default value for K in the `adf.test` function is defined as `trunc((length(y) - 1)1/3)`. Whereas, the `ur.df` function does allow us to select the lag according to AIC or BIC.

2.12.2 Phillips-Perron (PP) test

There is another unit root test that has the null hypothesis of the presence of a unit root, that is the Phillips-Perron (PP) test, which is developed by Phillips & Perron (1988). The main difference between the ADF and the PP tests are how they deal with serial correlations and heteroskedasticity in the errors. In particular, the ADF test uses a parametric autoregression to approximate the ARMA structure of the errors in the test

regression, whereas the PP test uses a corrected form of t-test in order to correct for the presence of serial correlation and heteroskedasticity in the error term (Zivot & Wang 2007). The model is as follows:

$$y_t = c + \delta t + \alpha y_{t-1} + \epsilon_t.$$

The null hypothesis of this test is that $\alpha = 1$, against the alternative which is $\alpha < 1$. The advantages of the PP test over the ADF test are that the PP test is robust to general forms of heteroskedasticity in the error term, and we do not need to specify a lag length for the test regression. If the t-statistic less than the critical value, then we have evidence to reject the null hypothesis of the presence of a unit root. There are two functions in R that can perform the PP test: `pp.test` and `ur.pp`.

2.12.3 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test proposed by Kwiatkowski et al. (1992) was motivated by the ADF test, as the ADF test with the null hypothesis of the presence of a unit root has low power in many applications (Kokoszka & Young 2016). Thus, Kwiatkowski et al. (1992) proposed a test with null hypothesis to be trend stationary against the hypothesis of the presence of a unit root. Rejection of the null hypothesis can be viewed as a convincing evidence that the series has a unit root. The KPSS test can also be used to detect long memory, with the short memory as the null hypothesis (e.g., Lee & Schmidt 1996, Giraitis et al. 2003).

Let $\{y_t\}$ be the observed time series for which we wish to test stationarity. We assume that we can decompose the series into the sum of a deterministic trend, a random walk, and a stationary error (Kwiatkowski et al. 1992):

$$y_t = \zeta t + \mu_t + \epsilon_t.$$

Here μ_t is a random walk:

$$\mu_t = \mu_{t-1} + u_t,$$

where the $\{u_t\}$ are independent and identically distributed $N(0, \sigma_u^2)$. The initial value μ_0 is treated as fixed and serves the role of an intercept. The stationarity hypothesis is simply $\sigma_u^2 = 0$. Since ϵ_t is assumed to be stationary, under the null hypothesis $\{y_t\}$ is trend-stationary (Tang et al., 2019). The statistics they used are the one-sided Lagrange Multiplier (LM) statistic and the Local Best Invariant (LBI) test statistic for the hypothesis $\sigma_u^2 = 0$, under the stronger assumptions that the u_t are normal and that the ϵ_t are independent and identically distributed $N(0, \sigma_\epsilon^2)$. Because the parameter value specified by the null hypothesis is on the boundary of the parameter space, they were interested in

a one-sided LM test rather than a two-sided test (Kwiatkowski et al. 1992).

Let e_t , $t = 1, 2, \dots, T$, be the residuals from the regression of $\{y_t\}$ on an intercept and time trend. Let $\hat{\sigma}_\epsilon^2$ be the estimate of the error variance from this regression (the sum of squared residuals, divided by T). Define the partial sum process of the residuals:

$$S_t = \sum_{i=1}^t \epsilon_i, \quad t = 1, 2, \dots, T.$$

Then the LM (and LBI) statistic is

$$\text{LM} = \sum_{t=1}^T \frac{S_t^2}{\hat{\sigma}_\epsilon^2},$$

and high values (compare with critical values) are taken as evidence against the null hypothesis. The function `kpss.test` and `ur.kpss` in R can be used.

2.13 Pre-whitening

The pre-whitening is a pre-processing procedure used in time series analysis to eliminate or reduce common trends in two time series variables, and this method has been applied numerously in geophysics. The correlation between two time series variables may arise from their common trends not from the statistical coincidence between fluctuations (Verstraete 2018), and hence has effect on the cross correlogram.

Ex 2.13.1. Consider two time series variables $\{x_t\}$ and $\{y_t\}$ having the following formulas:

$$x_t = x_{t-1} + \alpha_1(x_{t-1} - x_{t-2}) + u_t, \quad (2.29)$$

$$y_t = \beta_1 x_{t-3} + \beta_2 x_{t-4} + w_t, \quad (2.30)$$

where $\{u_t\}$ and $\{w_t\}$ are two white noises. Based on [Equation 2.29](#) and [Equation 2.30](#), we would like to expect the cross correlogram between the pre-whitened $\{x_t\}$ and $\{y_t\}$ to show significant lags at lag 3 and 4. However, it is difficult to identify from their cross correlogram ([Figure 2.12](#)). We could only tell that the time series variable $\{x_t\}$ might predict $\{y_t\}$, as the highest peak is on the left of the lag 0. Therefore, pre-whitening procedure should be applied in order to examine the lagged regression relationship between the two time series variables.

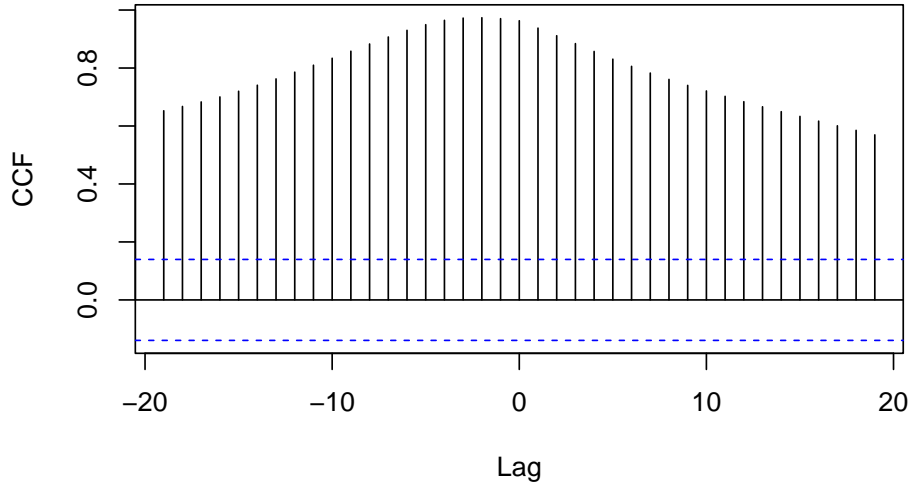


Figure 2.12: Cross correlogram between the variable $\{x_t\}$ and the $\{y_t\}$.

The main idea of the pre-whitening process is that the cross correlogram between $\{w_t\}$ (white noise) and $\{z_t\}$, where $\{z_t\}$ is a linear combination of lags of the $\{w_t\}$, are easily derived and identifiable. The steps along with mathematical explanations using $\{x_t\}$ and $\{y_t\}$ (Equation 2.29 and Equation 2.30) as the two time series variables, are as follows:

Step 1: Take the residuals from the fitted $\{x_t\}$ model. Ideally, we can have the residuals to be

$$\begin{aligned} u_t &= x_t - x_{t-1} - \alpha_1 x_{t-1} + \alpha_1 x_{t-2}, \\ u_t &= x_t - \alpha_2 x_{t-1} + \alpha_1 x_{t-2}, \end{aligned} \tag{2.31}$$

where $\alpha_2 = 1 + \alpha_1$. It is unrealistic to find the exact model for the variable $\{x_t\}$ in real life. Therefore, we want the residuals as close as to the white noise. In general, only autoregressive terms are considered (with differencing if needed) when fitting the $\{x_t\}$ than the moving average terms, for the purpose of simplicity.

Step 2: Filter the variable $\{y_t\}$ using the model of the $\{x_t\}$, we yields

$$y_{\text{filtered}} = y_t - \alpha_2 y_{t-1} + \alpha_1 y_{t-2}. \tag{2.32}$$

Substitute the [Equation 2.30](#) into the equation above, we have

$$\begin{aligned} y_{\text{filtered}} &= (\beta_1 x_{t-3} + \beta_2 x_{t-4} + w_t) - \alpha_2 (\beta_1 x_{t-4} + \beta_2 x_{t-5} + w_{t-1}) + \alpha_1 (\beta_1 x_{t-5} + \beta_2 x_{t-6} + w_{t-2}) \\ &= \beta_1 x_{t-3} + (\beta_2 - \alpha_2 \beta_1) x_{t-4} + (\alpha_1 \beta_1 - \alpha_2 \beta_2) x_{t-5} + \alpha_1 \beta_2 x_{t-6} + w_t - \alpha_2 w_{t-1} + \alpha_1 w_{t-2}. \end{aligned} \quad (2.33)$$

We can observe that the u_{t-3} is a linear combination of the x_{t-3} , x_{t-4} and the x_{t-5} . Moreover, the u_{t-4} is a linear combination of the x_{t-4} , x_{t-5} and the x_{t-6} . Mathematically,

$$u_{t-3} = x_{t-3} - \alpha_2 x_{t-4} + \alpha_1 x_{t-5}, \quad (2.34)$$

$$u_{t-4} = x_{t-4} - \alpha_2 x_{t-5} + \alpha_1 x_{t-6}. \quad (2.35)$$

We can then rewrite the [Equation 2.33](#) as a linear combination of the u_{t-3} and the u_{t-4} , yields

$$y_{\text{filtered}} = \beta_1 u_{t-3} + \beta_2 u_{t-4} + w_t - \alpha_2 w_{t-1} + \alpha_1 w_{t-2}. \quad (2.36)$$

Step 3: In this step, we generate the cross correlogram of the pre-whitened x_t (u_t) and the filtered y_t (y_{filtered}).

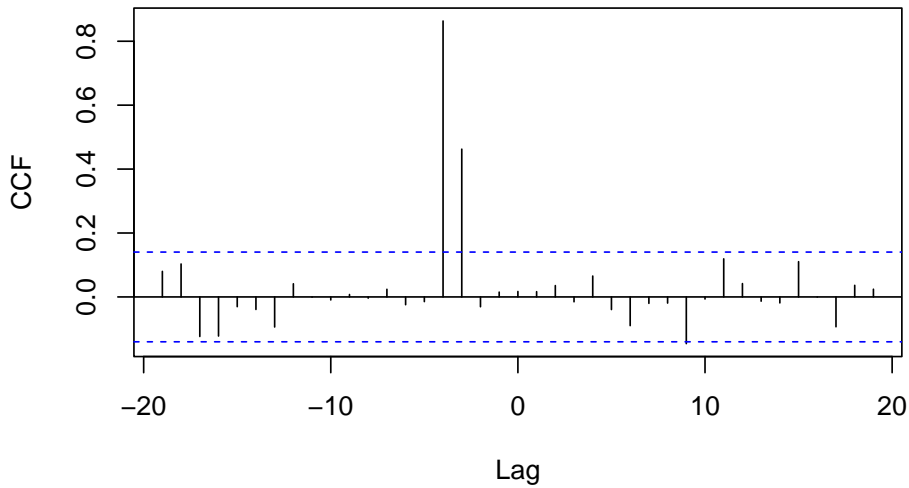


Figure 2.13: Cross correlogram of $\{u_t\}$ and $\{y_{\text{filtered}}\}$.

There are only two significant lags in the cross correlogram of pre-whitened series ([Figure 2.13](#)), lag 3 and lag 4 on the left hand side of the lag 0, which is what we expected.

2.14 Granger Causality

A working definition of causality is that A causes B if a specific physical change in A results in some corresponding specific change in B. According to this definition, causality can be inferred from controlled experiments. But, controlled experiments are not always practical, ethical, or feasible. In particular, controlled experiments are unlikely to be possible in macro-economics or climatology. Until 19th century, causality was closely connected to the statistical inference and correlation measures. However, correlation in itself is a measure of association and does not provide any direct evidence of causation. In some cases, causation can be inferred from physical reasoning. Physical reasoning relies on a time lag, since cause must occur before the effect, although the time lag can be small compared to sampling interval. However, a cause precedes its effect does not necessarily imply causality, it is one of the conditions that needs to be satisfied in order to claim for a causal relationship. The other conditions are the presence of covariation, and non-spuriousness. Spurious causations occur when two variables are both depend on a third variable (Oppewal 2010).

The requirement for the effect to not precede the cause led to further developments of statistical techniques, such as multiple correlation and regression were frequently associated with causal inference. The direction of regression respects the direction of causality. There are various approaches to causality, but the most influential one in economics is proposed by Granger (1969). Granger causality is an approach without direct reference to background economic theory, and it was developed to apply to dynamic time series models (Hoover 2006). Eichler (2013) reviews four approaches for defining causality. The first is average causal effect (ACE) of an intervention setting X_t to x^* on the response $Y_{t'}$, with $t' > t$ is

$$\text{ACE}_{x^*} = \text{E}[Y_{t'}|X_t = x^*] - \text{E}[Y_{t'}]. \quad (2.37)$$

The definition of non-causality with respect to interventions is

$$\text{F}(Y_{t'} \leq y|X_t = x^*) = \text{F}(Y_{t'} \leq y), \quad (2.38)$$

for all $y \in \mathbb{R}$ and all $t' > t$. The second is direct structural causation, which is closely related to causal effects of interventions. Neither can be detected empirically without a controlled experiment. The third is Granger causality which is considered in detail below. The fourth is a variation on Granger causality, known as Sims causality. The definition is given in terms of Sims non-causality. The process X is Sims non-causal for process

Y with respect to $V = (X, Y, Z)$ if $P(Y_{t'} \leq y | t' > t)$ is independent of $X_t | X^{t-1}, Y^t, Z^t$ for all t , where, for example, X^{t-1} is the history of X up to time $t - 1$, and Z includes all confounders for the dependence of X on Y . Sims causality is slightly stronger than Granger causality because of the inclusion of Z .

In time series forecasting, we normally use univariate data to predict its further values using its past values. Sometimes, we need multiple series for better predictions by introducing other time series variables that have causal relations with existing variable. The concept of Granger causality was established by Granger (1969) for stationary time series. The general ideas of the Granger causality are that a cause occurs before its effect, and the including of the past values of a variable help to predict the future values of another variable. That is, a time series $\{x_t\}$ is Granger causal for another time series say, $\{y_t\}$, if inclusion of past values of $\{x_t\}$ improves prediction for future $\{y_t\}$ compared to including the past values of $\{y_t\}$ alone. In other words, $\{y_t\}$ is not only the function of its own lags but also a function of the lagged $\{x_t\}$. It is important to note that a cause has unique information about the future values of its effect.

The `grangertest` function in the `lmtest` package in R (Zeileis & Hothorn 2002) can be implemented to investigate Granger causality. This function only works for bivariate series, using a Wald test comparing between two models: $\{y_t\}$ regressed on the past values of $\{y_t\}$ and the past values of $\{x_t\}$, and $\{y_t\}$ regressed on its past values only.

2.15 Vector Autoregressive (VAR) Model

Vector autoregressive (VAR) models are multivariate time series models, denoted by $\text{VAR}(p)$ where p is the order of autoregressive terms. Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{nt})$ be a vector with each entry to be a time series variable, $\text{VAR}(p)$ model has the form

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{nt} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} \alpha_{11}^1 & \dots & \alpha_{1n}^1 \\ \vdots & \dots & \vdots \\ \alpha_{n1}^1 & \dots & \alpha_{nn}^1 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ \vdots \\ y_{nt-1} \end{pmatrix} + \dots + \begin{pmatrix} \alpha_{11}^p & \dots & \alpha_{1n}^p \\ \vdots & \dots & \vdots \\ \alpha_{n1}^p & \dots & \alpha_{nn}^p \end{pmatrix} \begin{pmatrix} y_{1t-p} \\ \vdots \\ y_{nt-p} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \vdots \\ \epsilon_{nt} \end{pmatrix}. \quad (2.39)$$

Alternatively, it has the vector form

$$\mathbf{y}_t = \boldsymbol{\mu} + \alpha_1 \mathbf{y}_{t-1} + \dots + \alpha_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t, \quad (2.40)$$

where α_i are $n \times n$ coefficient matrices for $i = 1, \dots, p$. The vector $\boldsymbol{\mu}$ represents the constants and each entry in the $\boldsymbol{\epsilon}_t$ is a white noise process. The dimension of a $\text{VAR}(p)$

model depends on the number of time series variables in the model, an $AR(p)$ model is a special case of a $VAR(p)$ model with dimension of 1.

The VAR models are first known in economics by Sims (1980), these models are used to capture the evolution and the inter-dependencies between multiple time series. Each variable is regressed on its own past values, and also the past values of all the other variables in the model. A number of studies have shown that the VAR models are useful for describing the dynamic behaviour of the data and also provide better forecasting performance (Singh 2018). Moreover, the ease of implementing these models makes them become competitive against other multivariate linear time series models, such as VARMA models.

If the main purpose of a study is to analyse the relationships between time series variables, then we might lose some information on any long-term relationship by detrending and deseasonalising to induce stationarity. In contrast, if we interested in hypothesis tests, either singly or jointly, to evaluate the statistical significance of the coefficients, then it is important that all the components in the VAR model are stationary (Chris 2014). The `VAR` function in `vars` package in R allows us to fit the VAR model with trend and/or seasonal variation (Pfaff 2008b).

2.15.1 Stationarity of the VAR models

Representing the [Equation 2.40](#) in terms of lag operators, we yields,

$$I\mathbf{y}_t = \boldsymbol{\mu} + \alpha_1 L^1 \mathbf{y}_t + \cdots + \alpha_p L^p \mathbf{y}_t + \boldsymbol{\epsilon}_t, \quad (2.41)$$

where I is the identity matrix, and $L^j \mathbf{y}_t = \mathbf{y}_{t-j}$ for $j = 1, \dots, p$. Rearrange the equation, we have:

$$\begin{aligned} I\mathbf{y}_t - \alpha_1 L^1 \mathbf{y}_t - \cdots - \alpha_p L^p \mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \\ (I - \alpha_1 L^1 - \cdots - \alpha_p L^p) \mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \\ \Phi(L) \mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \end{aligned} \quad (2.42)$$

where $\Phi(L) = I - \alpha_1 L^1 - \cdots - \alpha_p L^p$. Hence,

$$\mathbf{y}_t = \Phi(L)^{-1}(\boldsymbol{\mu} + \boldsymbol{\epsilon}_t). \quad (2.43)$$

A series is said to be stationary if the modulus of the roots of the [Equation 2.44](#) are greater than 1.

$$|I - \alpha_1 z^1 - \cdots - \alpha_p z^p| = 0, \quad (2.44)$$

where $|\cdot|$ is the determinant of a matrix, and z is a complex variable.

Ex 2.15.1. Given two time series variables, $\{x_t\}$ and $\{y_t\}$, follow a VAR(1) model:

$$\begin{aligned}x_t &= 2.4x_{t-1} + 1.3y_{t-1}, \\y_t &= 0.4x_{t-1} + 0.6y_{t-1}.\end{aligned}$$

Rewritten in the vector form,

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 2.4 & 1.3 \\ 0.4 & 0.6 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}.$$

Now, we would like to solve the equation

$$\left| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 2.4 & 1.3 \\ 0.4 & 0.6 \end{pmatrix} z \right| = 0.$$

Hence,

$$\begin{aligned}\left| \begin{pmatrix} 1 - 2.4z & -1.3z \\ -0.4z & 1 - 0.6z \end{pmatrix} \right| &= 0, \\(1 - 0.6z - 2.4z + 2.4 \times 0.6z^2) - 1.3 \times 0.4z^2 &= 0, \\1 - 3z - 1.96z^2 &= 0.\end{aligned}$$

There are two roots for this polynomial, 0.28 and 1.82, which suggest the series are non-stationary as one of the roots is smaller than 1.

2.16 Threshold Vector Autoregressive (TVAR) Model

The threshold vector autoregressive (TVAR) model is a non-linear time series model characterized by two or more regimes that follow a VAR structure and where the switching among them is regulated by a latent variable (Niglio & Vitale 2015). The general idea of a TVAR model is splitting the time series into different regimes / intervals, and for each individual regime we fit a linear VAR model. Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{nt})$ be a vector with each entry to be a time series variable, a TVAR model with order p , denoted by TVAR(p), has the vector form

$$\mathbf{y}_t = \boldsymbol{\mu} + \alpha_1 \mathbf{y}_{t-1} + \dots + \alpha_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t \quad \text{if } r_{j-1} < z_{t-d} < r_j, \quad (2.45)$$

for $j = 1, \dots, s$, where s represents the number of regimes. The matrices α_i are $n \times n$ coefficient matrices for $i = 1, \dots, p$. The vector $\boldsymbol{\mu}$ represents the constants and each entry in the $\boldsymbol{\epsilon}_t$ is a white noise process. Also, z_t is the threshold variable, the threshold variable

can either be some external variables or response variable itself, and d is the threshold lag which normally set to one. The values r_{j-1} and r_j are the threshold values for the regime j .

The `TVAR.LRtest` is a function in R (Narzo et al. 2020) that used to test linearity in the multivariate systems by using likelihood ratio tests comparing the covariance matrix of linear VAR, TVAR(1), and TVAR(2) (Lo & Zivot 2001). The default settings of this test will return the likelihood ratio tests for covariance matrices of linear VAR compared with TVAR(1), and linear VAR compared with TVAR(2).

Chapter 3

Influence of climate indicators on rainfall in South Australia

In this chapter, we start with an investigation of the inter-relations between four climate indicators, the Southern Oscillation Index (SOI), the Indian Ocean Dipole (IOD), the Pacific Decadal Oscillation (PDO), and the Southern Annular Mode (SAM). We then explore the relationships between these four climate indicators and South Australian monthly rainfall using VAR and TVAR models.

3.1 Introduction

Monitoring and understanding rainfall variability is an important area of research as almost every industry relies on water as an input to the nation's economy, particularly agriculture and hydroelectric power plants. Luo et al. (2005) showed that the change in rainfall is the most influential factor for the wheat yield, among rainfall, temperature and CO₂ concentration, in the medium to low rainfall areas. In particular, arid areas are more sensitive to the rainfall change. This study also predicted that future climate change will reduce South Australian wheat yield. Moreover, Australia's climate has impact on water supply and consumption. In 2007 - 2008, 90% of water was used for irrigation. In the following year, the Murray-Darling Basin encountered strong demand for water transfers due to the continuing low rainfall levels. This resulted in about 32,000 water trades during the year, totaling \$2.2 billion and almost 4,000 giganlitres. It is now widely recognized that taking too much water out of Australia's rivers and groundwater systems can have detrimental economic and environmental consequences. These include reduced:

agricultural production; plant populations; and number of native animals (Pink 2007). Rainfall has caused drought and flood events over the past decades, including a prolonged multi-year drought from 1995 to 2009 known as the “Millennium drought” (Ummerhofer et al. 2009, Van Dijk et al. 2013). Therefore, accurate prediction of water availability is immensely beneficial for making informed policy, planning and management decisions, and can assist with the more sustainable operation of water resource systems (Bagirov et al. 2017).

Variability in rainfall is increasing over time and additionally varies from region to region, especially in Australia (Bagirov et al. 2017). The variability is due to Australia covering a large range of climate zones. Also, Australia is surrounded by tropical and subtropical oceans, and its climate is sensitive to large-scale ocean-atmosphere interactions (Forootan et al. 2016). The Southern Oscillation Index (SOI) is recognized as the principal tropical influence in the Pacific Ocean, and has relatively large impact in eastern Australia (e.g., Pui et al. 2012, Cai et al. 2011). The Indian Ocean Dipole (IOD) is recognized as a principal influence that contributes to rainfall variability both directly and in interaction with the SOI, in the tropical Indian Ocean (Cai et al. 2011). Risbey et al. (2009) found that the impact of the SOI is substantially reduced if the covariance with the IOD is removed. The IOD has been shown to have impact mainly in southern Australia with anomalously wet conditions during negative phase and dry conditions during positive phase (e.g., Nicholls 1989, Ashok et al. 2003, Cai et al. 2009, Min et al. 2013). Cai et al. (2011) and Chowdhury & Beecham (2013) showed the significant correlations between the SOI, the IOD and monthly rainfall in South Australia during spring and winter. However, a strong concurrent relationships does not necessarily lead to strong lagged relationships (Schepen et al. 2012).

As well as the two mainly studied climate indicators, SOI and IOD, other climate indicators are attracting researchers’ attention. The Southern Annular Mode (SAM) is recognized as the dominant mode of climate variability over the Southern Hemisphere extra-tropics (Min et al. 2013), with wet conditions during positive phase over much of Australia, and warm conditions during negative phase over the southwest and southeast coasts of the continent (e.g., Gillett et al. 2006, Cai & Cowan 2006, Hendon et al. 2007). The Pacific Decadal Oscillation (PDO) is shown to affect rainfall with positive phase inducing warmer conditions over much of the eastern Australia, particularly Queensland (e.g., Mantua & Hare 2002, Power et al. 2006, Pezza et al. 2007, Kiem & Franks 2004). Climate indicators are correlated with each other, which needs to be take into account when interpreting their impacts as they may affect the Granger causal relationships. Min et al. (2013) showed a positive relationship between the SOI and the IOD in winter and spring, and the SAM has a negative correlation with the SOI in summer. Moreover, there is evidence of positive correlation between IOD and SAM in summer, at the 0.05 significance level. Inter-relations between climate indicators are still a point of contention, and

required further investigation (Min et al. 2013). Thus, one of the objective of this study is to explore the associations and Granger causations between the SOI, the IOD, the PDO, and the SAM.

Risbey et al. (2009) found that any single climate indicator accounts for less than 20% of South Australia monthly rainfall variability. Rasel et al. (2015) conducted an extended study based on the research of Risbey et al. (2009), to investigate the relationship between the lagged SOI and the lagged SAM and South Australian monthly rainfall, and the combined influence of these lagged climate indicators on rainfall, using multiple regression models. Rasel et al. (2015) suggested that the inclusion of the interaction between the SOI and the SAM increase the monthly rainfall predictability. The square root of R^2 , which is the correlation between fitted values and observations, increased from 0.31 to 0.41 for North Adelaide and 0.34 to 0.45 for Mount Bold Reservoir. However, they did not deseasonalise and/or detrend the monthly rainfall and climate indicator time series. Cleverly et al. (2016) also found that the variability in annual rainfall in Australia is due to periods of synchronisation amongst the SOI, the IOD and the SAM, rather than any single climate indicator. Therefore, the second goal of this study is to explore the relationship between monthly rainfall in South Australia and four climate indicators: the SOI, the IOD, the PDO, and the SAM together with their interactions and their quadratic terms, after we understand the inter-relationship between these four climate indicators.

3.2 Data

3.2.1 Rainfall

The monthly rainfall data at four different Bureau stations ([Figure 3.1](#)): Adelaide Airport (station number: 023034); Murray Bridge (station number: 024521); Pooraka (station number: 023026) and North Adelaide (station number: 023011), were acquired from the Australian Bureau of Meteorology (BoM). Murray Bridge is relatively distant from the other three weather stations.

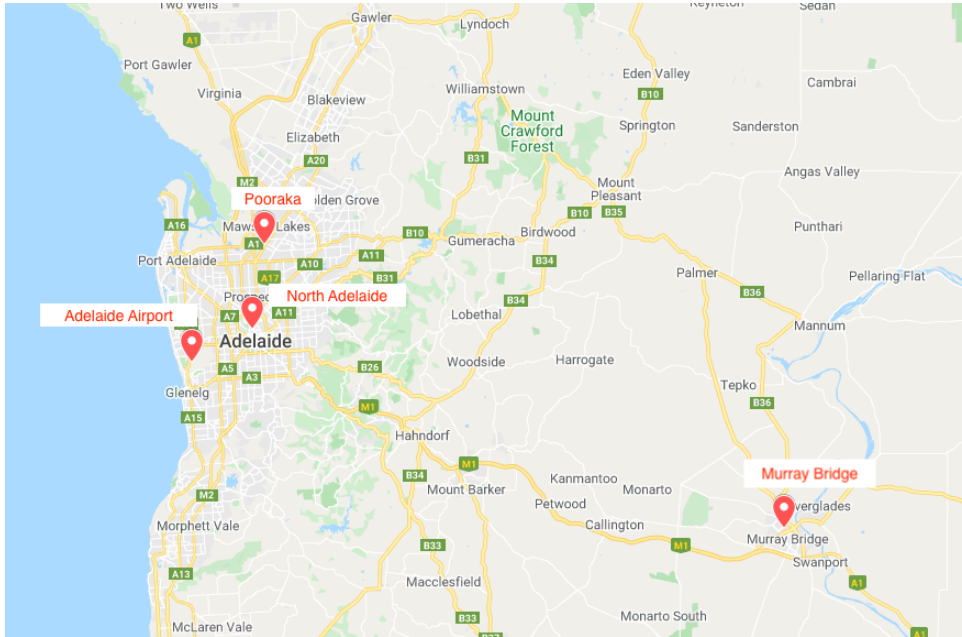
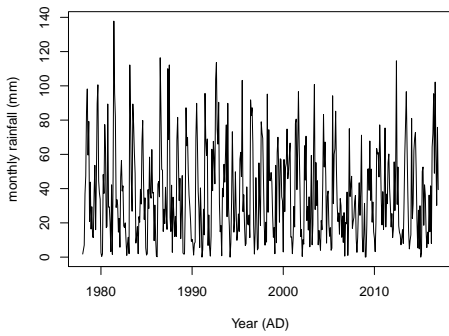
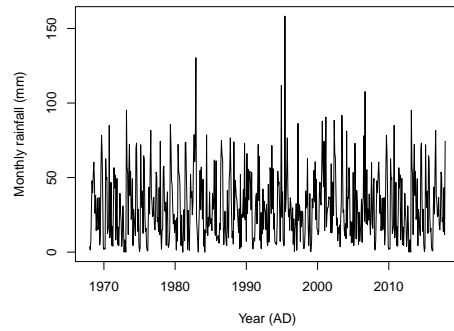


Figure 3.1: Locations of the four weather stations.

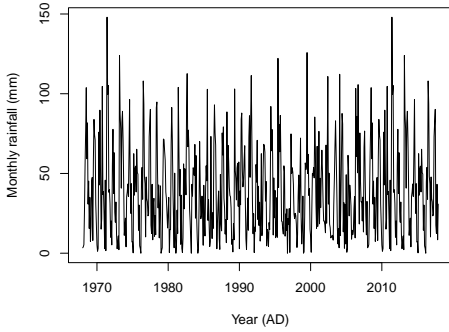
The BoM defined rainfall as all forms of precipitation that reach the ground, such as rain, drizzle, hail and snow. The monthly rainfall is the total of all available daily rainfall for the month, where daily rainfall measurements are nominally made at 9 am local time and record the total for the previous 24 hours (Bureau of Meteorology 2021*a*). The four monthly rainfall time series have a common time period from January 1978 to December 2017. There is one missing value for the Adelaide Airport station and the Murray Bridge station, and 2 and 12 missing values for the North Adelaide and the Pooraka stations, respectively. All missing values were imputed by linear interpolation between the corresponding months in the preceding year and the corresponding months in the following year.



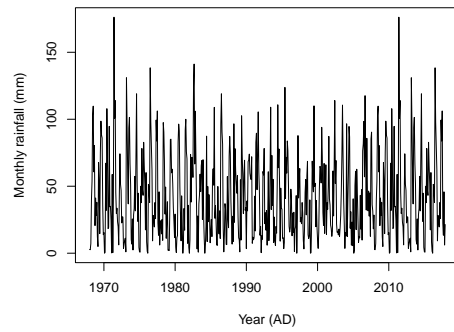
(a) Adelaide Airport.



(b) Murray Bridge.



(c) Pooraka.



(d) North Adelaide.

Figure 3.2: Time series plots of the monthly rainfall at the four weather stations.

3.2.2 Climate indicators

We consider four climate indicators, the SOI, the IOD, the PDO, and the SAM. The monthly IOD and the monthly PDO were provided by the National Oceanic and Atmospheric Administration in the U.S.A. (NOAA). The monthly SOI and the monthly SAM data were acquired from the BoM. The analysis period covers 480 months from January 1978 to December 2017. In this context, an anomaly is the difference between a particular monthly value and its expected value.

Southern Oscillation Index (SOI)

The Southern Oscillation Index (SOI) is defined as the standardised anomaly of the Mean Sea Level Pressure (MSLP) difference between Tahiti and Darwin (Equation 3.1).

$$\text{SOI} = 10 \times \frac{P_{\text{diff}} - E[P_{\text{diff}}]}{\text{Sd}(P_{\text{diff}})}, \quad (3.1)$$

where P_{diff} is the monthly average Tahiti MSLP minus the monthly average Darwin MSLP. The expected value of P_{diff} ($E[P_{\text{diff}}]$) is the long term average of P_{diff} for a particular month, and $\text{Sd}(P_{\text{diff}})$ is the long term standard deviation of P_{diff} for a particular month. Note that the multiplication by 10 is so that the SOI value can be quoted as an integer. This climate indicator is best represented by monthly average, because daily or weekly SOI values can fluctuate markedly due to short-lived, day-to-day weather patterns, particularly if a tropical cyclone is present. Sustained positive SOI values above about +7 indicate a La Niña event while sustained negative values below about -7 indicate an El Niño (Henderson 2012). Noted that, the rainfall across Australia is reduced during El Niño events. Figure 3.3 shows the time series of the SOI from January 1978 to December 2017, where the red dashed line is to separate La Niña event and neutral phase, and the blue dashed line is to separate neutral phase and El Niño event.

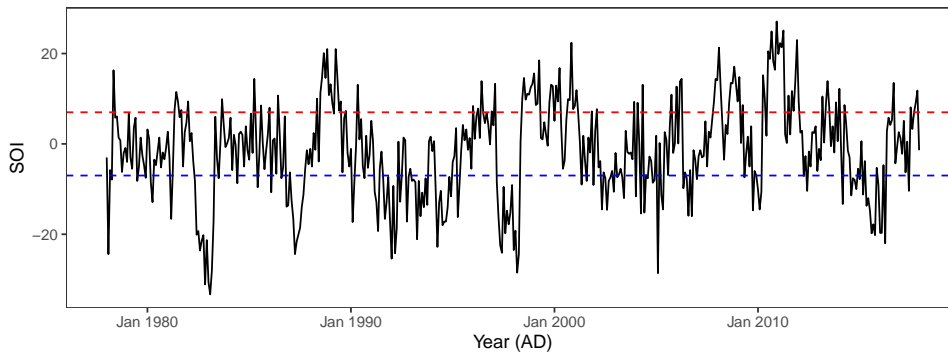


Figure 3.3: Time series plot of the SOI.

Indian Ocean Dipole (IOD)

The monthly Indian Ocean Dipole (IOD) dataset was provided by the NOAA and was dated from January 1978 to December 2017. IOD events are driven by changes in the tropical Indian Ocean. Sustained changes in the difference between normal sea surface temperatures in the tropical western and eastern Indian Ocean are what characterise IOD events. The IOD is commonly measured by an index (sometimes referred to as the

Dipole Mode Index, or DMI) that is the difference between sea surface temperature (SST) anomalies in two regions of the tropical Indian Ocean (Figure 3.4): 1. IOD west: 50°E to 70°E and 10°S to 10°N. 2. IOD east: 90°E to 110°E and 10°S to 0°S. The IOD is calculated as the SST in region 1 minus the SST in region 2 (Bureau of Meteorology 2012).

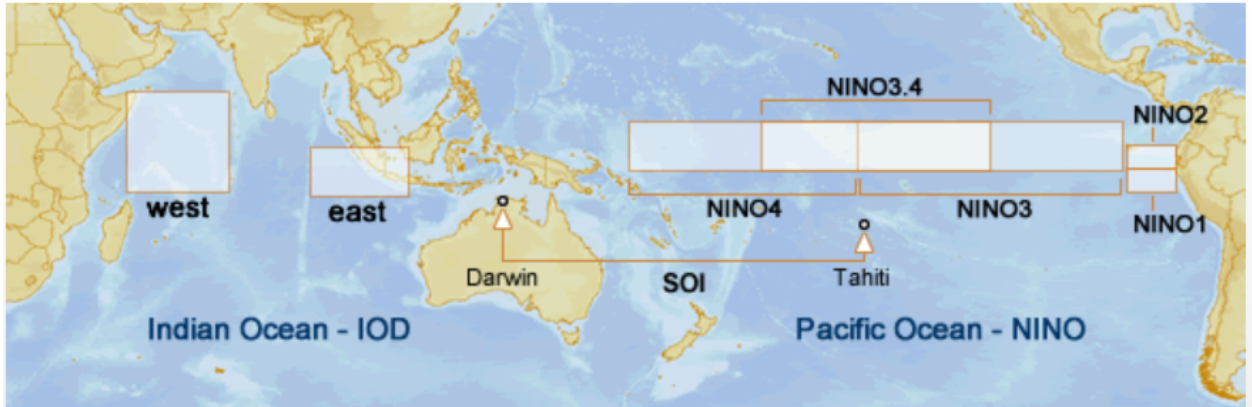


Figure 3.4: The Indian Ocean Dipole (Bureau of Meteorology 2021b).

The IOD has three phases: neutral, positive and negative. For monitoring the IOD, Australian climatologists consider sustained values above 0.4°C as typical of a positive IOD, and values below -0.4°C as typical of a negative IOD (Figure 3.5). During a positive IOD period, west parts of Australia tend to have more rain and storms. In contrast, east parts of Australia tend to have less rain and storms. A negative IOD typically results in above-average winter-spring rainfall over parts of southern Australia.

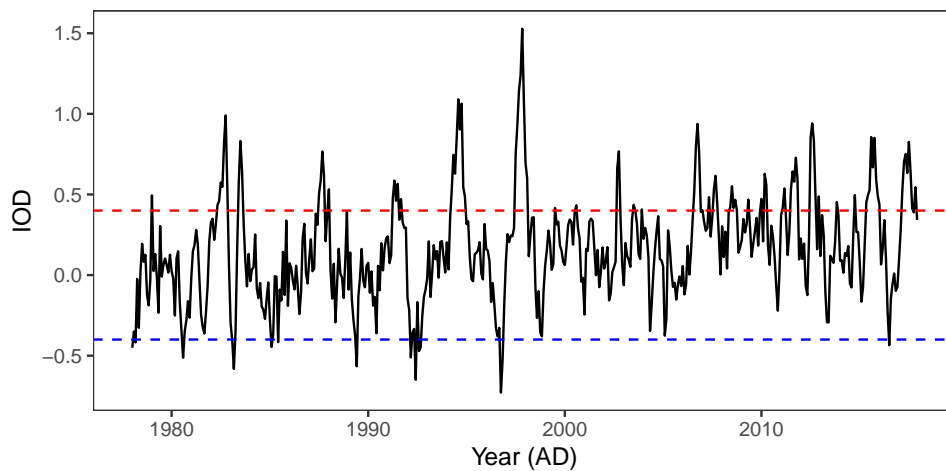


Figure 3.5: Time series plot of the IOD.

Pacific Decadal Oscillation (PDO)

The Pacific Decadal Oscillation (PDO) dataset was acquired from the NOAA and dated from January 1978 to December 2017 (Figure 3.6). It is a pattern of Pacific climate variability similar to the SOI in character, but which varies over a much longer time scale. The PDO is defined as monthly SST anomalies over the Pacific Ocean north of 20°N minus the monthly global average anomaly. The purpose of subtraction of the global average anomaly is to account for the global warming (Mantua & Hare 2002). The PDO has positive and negative phases. The PDO has positive phase when sea level temperatures are below average over the North Pacific, and negative phase when sea level temperatures are above average. The PDO can influence weather events and global average temperatures. Positive phases of the PDO tend to be associated with periods of more rapid global warming, while negative phases can be linked to times of slower warming.

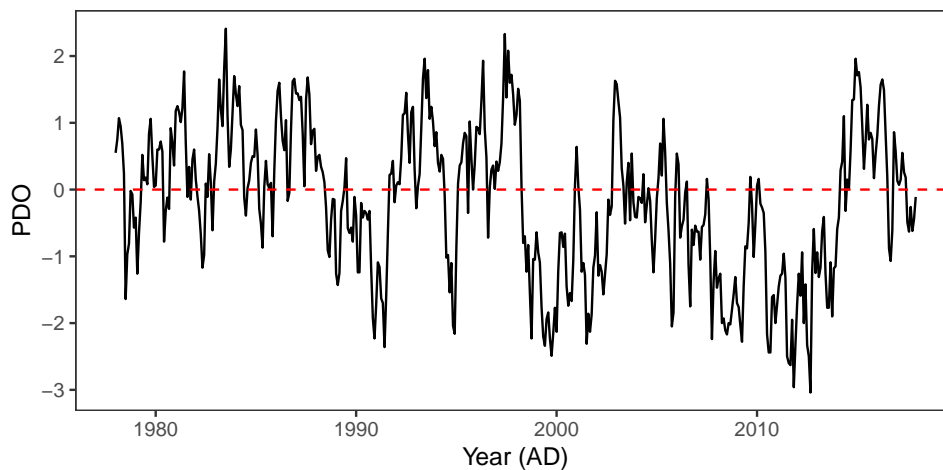


Figure 3.6: Time series plot of the PDO.

Southern Annular Mode (SAM)

The monthly Southern Annular Mode (SAM) dataset was acquired from the BoM and dated from January 1978 to December 2017 (Figure 3.8). The SAM is defined as the zonal mean sea level pressure at 40°S minus the zonal sea level pressure at 65°S. The zonal mean sea level pressures are calculated using records from 6 stations at roughly 40°S and 6 stations at roughly 65°S (Figure 3.7).

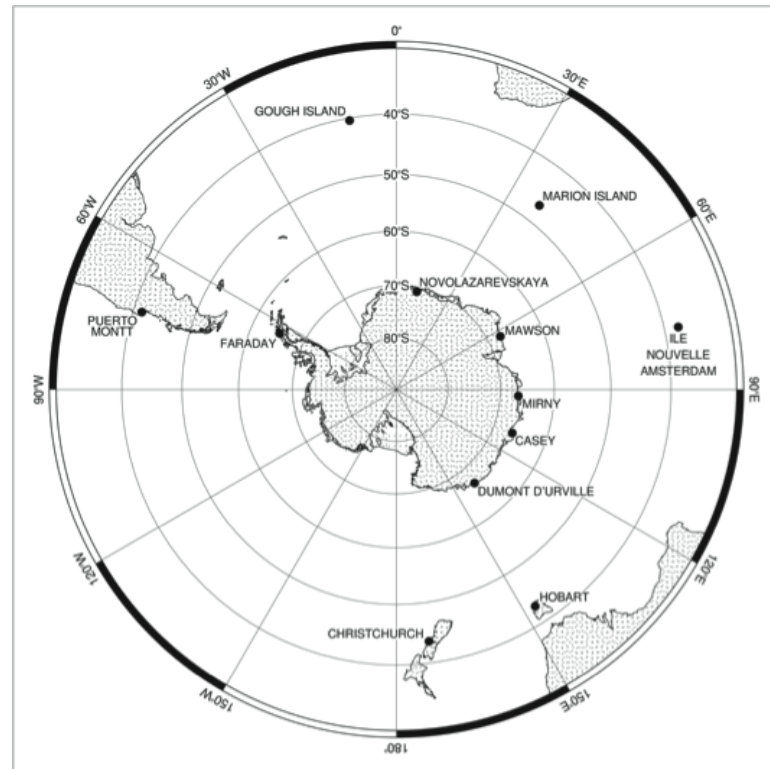


Figure 3.7: Locations of 12 stations used to calculate the Southern Annular Mode (Marshall 2021).

It describes the north-south movement of the westerly wind belt that circles Antarctica, dominating the middle to higher latitudes of the southern hemisphere. The SAM varies between three phases. Positive, neutral and negative. These phases last around two weeks but can sometimes persist longer (Bureau of Meteorology 2021c). The SAM has positive phase if the belt of strong westerly winds contracts towards Antarctica. Conversely, the SAM has negative phase if the belt of strong westerly winds contracts towards the equator. Normally, values above approximately 0.4 are classified as positive SAM events, and values below -0.4 are negative SAM events. Moreover, effect of the SAM on Australian weather patterns varies during the year.

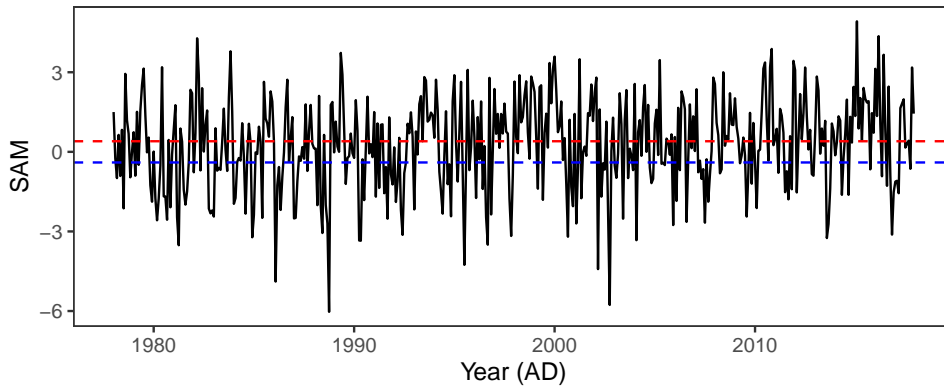


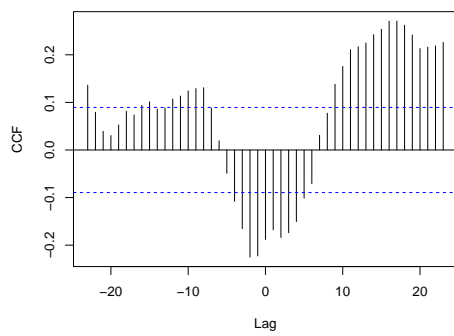
Figure 3.8: Time series plot of the SAM.

3.3 Analysis

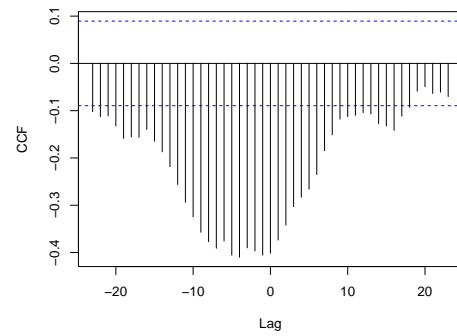
We begin by examining the inter-relations between the four climate indicators (SOI, IOD, PDO, SAM). We then fit VAR(2) and VAR(5) models to the original series, deseasonalised and detrended series, and pre-whitened series, for the investigation of Granger causality. Also, linear regression models are used for the exploration of any associations by including contemporaneous values of the climate indicators.

3.3.1 Relationships between climate indicators

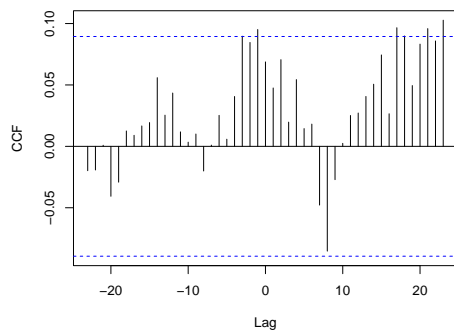
The cross correlograms of the original time series ([Figure 3.9](#)) include any trend or seasonal effects, although these are not dominant. The most striking relationship appears to be negative association between the SOI and the PDO, with the SOI leading the PDO by 4 months. It appears that the SOI is negatively associated with the IOD and leads the IOD by 1 month. There is also a negative association between the IOD and the PDO, with the IOD apparently leading by a year. There are no other clear associations.



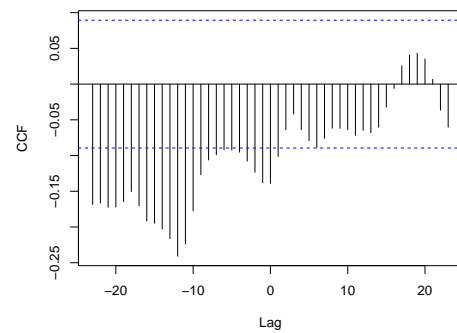
(a) SOI and IOD.



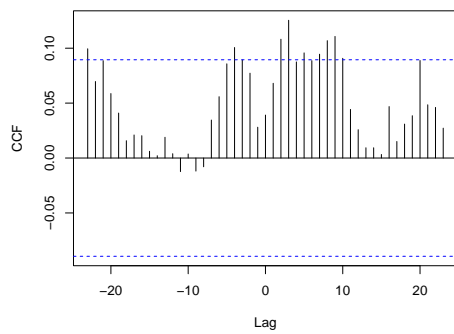
(b) SOI and PDO.



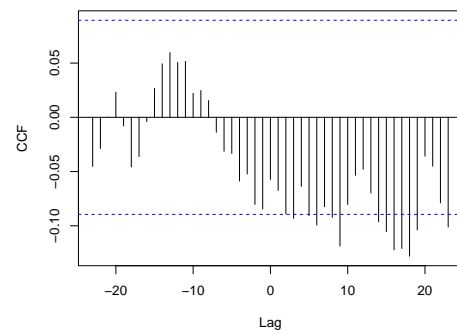
(c) SOI and SAM.



(d) IOD and PDO.



(e) IOD and SAM.



(f) PDO and SAM.

Figure 3.9: Cross correlograms of the four climate indicators.

The significance of seasonal patterns and trends are examined by fitting linear regressions on time and harmonic terms. The p-values of the estimated coefficients would be affected

due to the presence of the autoregressive terms of the time series. To allow for this, we include the lag 1 and lag 2 terms of the variables in the linear regression models. The linear regression model for each individual climate indicator has the form:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 \cos(2\pi t/12) + \beta_4 \sin(2\pi t/12) + \beta_5 (t - \bar{t}) + \beta_6 (t - \bar{t})^2, \quad (3.2)$$

where $\{y_t\}$ represents each individual climate indicator. The sine and cosine terms represent monthly variations, and $t - \bar{t}$ is the mean adjusted time. The term $(t - \bar{t})^2$ is also included in order to capture any quadratic trends. The semi-annual sinusoidal curve is included, however, the adjusted R^2 increases. Therefore, we have not included it in the model. The results (Table 3.1) indicate that there is no significant trend in the SOI, and no significant quadratic trends in any of the series, at the 0.05 significance level. There is evidence of slight linear trends for the IOD, the PDO and the SAM with coefficients of 0.00067, -0.0025 and 0.0019, respectively. There is evidence of seasonality in the IOD and the PDO but not in the SAM. The SOI is a deseasonalised measure but the monthly means and standard deviations are calculated over longer periods than considered in the analysis, and there is weak evidence of slight seasonality.

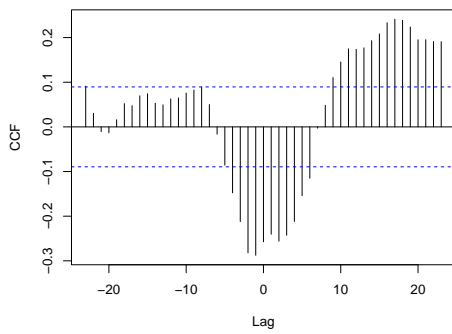
Table 3.1: Statistical significances of the lagged terms, trends and the seasonal patterns with adjusted R^2 .

Climate indicator	Intercept	Lag 1	Lag 2	Cosine	Sine	$t - \bar{t}$	$(t - \bar{t})^2$	Adjusted R^2
SOI	.41	.00	.00	.95	.11	.29	.82	.47
IOD	.01	.00	.28	.03	.22	.02	.99	.62
PDO	.12	.00	.16	.00	.39	.05	.34	.77
SAM	.11	.00	.83	.65	.44	.01	.70	.05

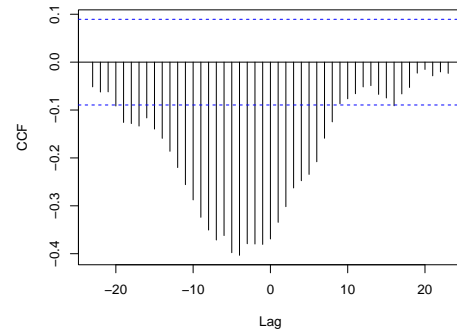
This linear regression model (Equation 3.2) provides a good fit to the IOD as well as the PDO series, as the values of the adjusted R^2 are 0.62 and 0.77, respectively. However, this model only explains 5% of the variation of the time series variable SAM. Deseasonalised and detrended series, \tilde{y}_t , are defined by

$$\tilde{y}_t = y_t - \hat{\beta}_0 - \hat{\beta}_3 \cos(2\pi t/12) - \hat{\beta}_4 \sin(2\pi t/12) - \hat{\beta}_5 (t - \bar{t}), \quad (3.3)$$

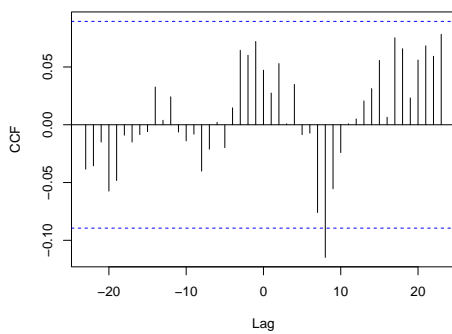
where $\hat{\beta}_i$ is the estimated coefficient returned from the linear regression model, using ordinary least squares.



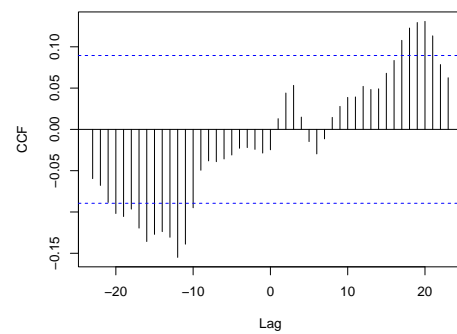
(a) SOI and IOD.



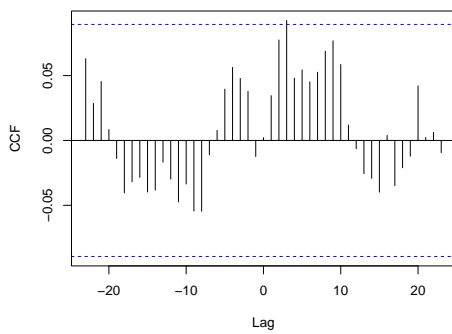
(b) SOI and PDO.



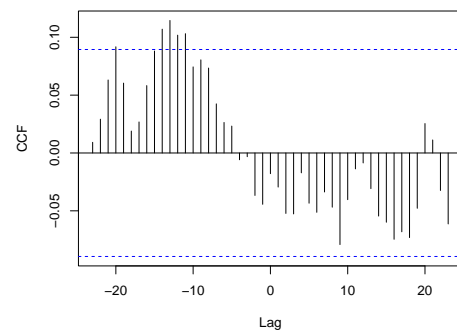
(c) SOI and SAM.



(d) IOD and PDO.



(e) IOD and SAM.

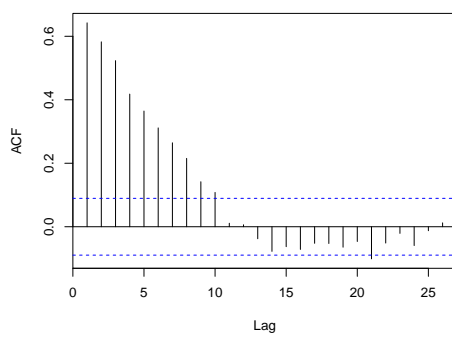


(f) PDO and SAM.

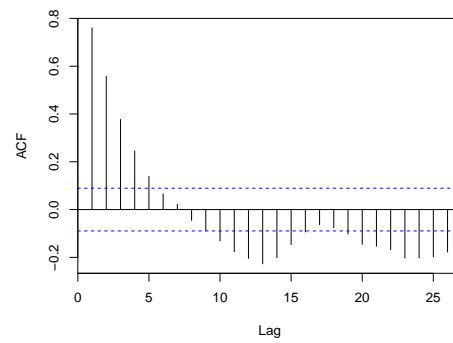
Figure 3.10: Cross correlograms of the deseasonalised and detrended climate indicators.

We could observe that there is little difference in the cross correlograms between climate indicators after removing any estimated trends and seasonal variation. In particu-

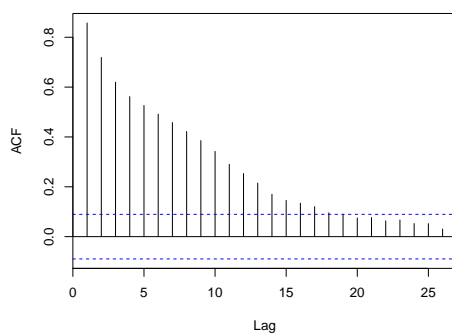
lar, the correlograms of the deseasonalised and detrended series (Figure 3.11) still show auto-correlations at higher lags except for the SAM. Therefore, pre-whitening method is implemented to facilitate interpretation.



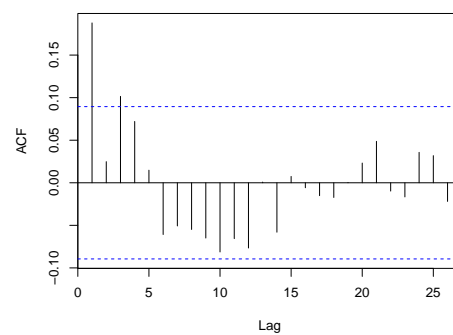
(a) SOI.



(b) IOD.



(c) PDO.



(d) SAM.

Figure 3.11: Correlograms of the deseasonalised and detrended series.

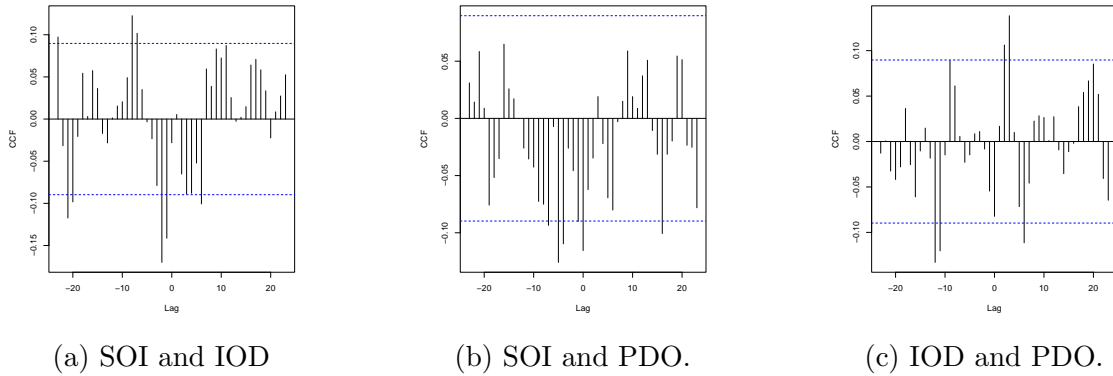


Figure 3.12: Cross correlograms of the pre-whitened series.

The inter-relations between climate indicators are explored by fitting linear regression models. We only present the analysis of the SOI, the IOD, and the PDO below. The detailed results of the SAM are not presented here, because the SAM does not appear to have any relationship with the other three climate indicators. In particular, there are no striking patterns in the cross correlograms between the SAM and other climate indicators. We followed the procedures in [Section 2.13](#), and obtained cross correlograms of the pre-whitened series in [Figure 3.12](#). That is, we fitted AR(p) model to the SOI, with fitted model as follows:

$$\text{SOI}_t = 0.42 * \text{SOI}_{t-1} + 0.23 * \text{SOI}_{t-2} + 0.13 * \text{SOI}_{t-3}.$$

Then we fitted the autoregressive model with the same coefficients to the IOD, the PDO, and the SAM. The cross correlogram ([Figure 3.12a](#)) suggests that the SOI is Granger causal for the IOD at lag 1 and lag 2, and Granger causality is demonstrated by the result of fitting a linear regression with the lagged pre-whitened SOI as predictor variables and the pre-whitened IOD as the response variable (p-value = 0.00). The lagged terms of the response variable are not included in the regression model as we used pre-whitened series for the analysis. This is because that the pre-whitening procedure reduces or removes the serial correlation. Additionally, there is no association suggested by the linear regression model with pre-whitened IOD as response variable and pre-whitened SOI at lag 0 as predictor variable (p-value = 0.64).

Table 3.2: The estimated coefficients and their standard errors of the fitted regression model for the pre-whitened IOD on the lagged pre-whitened SOI.

	$\text{SOI}_{\text{prewhiten}_{t-1}}$	$\text{SOI}_{\text{prewhiten}_{t-2}}$
$\text{IOD}_{\text{prewhiten}_t}$	✓ -0.0039 (0.0013)	✓ -0.0048 (0.0013)

The red check mark means that term is statistically significant at the 0.05 level.
Adjusted $R^2 = 0.045$.

Inspection of the cross correlogram of the pre-whitened series ([Figure 3.12b](#)) suggests that the SOI might be associated with the PDO and also Granger causal for the PDO at lag 4 and 5, with lag 7 to be marginally significant at the 0.05 level. To explore the association, a linear regression model has been fitted to the pre-whitened PDO and the pre-whitened SOI at lag 0. The result shows a statistically significant association between the two variables (p-value = 0.012). The Granger causal relationship is investigated further by fitting a linear regression with the pre-whitened SOI from lag 1 up to lag 7 with the pre-whitened PDO as response variable ([Table 3.3](#)).

Table 3.3: The estimated coefficients and their standard errors of the fitted regression model for the pre-whitened PDO on the lagged pre-whitened SOI.

	$\text{PDO}_{\text{prewhiten}_t}$
$\text{SOI}_{\text{prewhiten}_{t-1}}$	-0.0064 (0.0035)
$\text{SOI}_{\text{prewhiten}_{t-2}}$	-0.0011 (0.0035)
$\text{SOI}_{\text{prewhiten}_{t-3}}$	-0.0017 (0.0035)
$\text{SOI}_{\text{prewhiten}_{t-4}}$	✓ -0.0081 (0.0035)
$\text{SOI}_{\text{prewhiten}_{t-5}}$	✓ -0.011 (0.0035)
$\text{SOI}_{\text{prewhiten}_{t-6}}$	-0.00038 (0.0035)
$\text{SOI}_{\text{prewhiten}_{t-7}}$	✓ -0.0069 (0.0035)

The red check mark means that term is statistically significant at the 0.05 level.
Adjusted $R^2 = 0.033$

The cross correlogram of the pre-whitened series ([Figure 3.12c](#)) suggests that the PDO is Granger causal for the IOD at lag 3 with positive correlation, this might due to the association and causation between the SOI and the PDO. This conclusion is further demonstrated

by fitting a linear regression with the pre-whitened PDO from lag 1 up to lag 3, and fitting a linear regression with both of the pre-whitened PDO and the pre-whitened SOI from lag 1 to lag 3 (Equation 3.4), with the pre-whitened IOD as the response variable (Equation 3.5). Equations are as follows:

$$\text{IOD}_{\text{prewhiten}_t} = \alpha_0 + \mathbf{c}_1 \text{PDO}_{\text{prewhiten}_{t-[1:3]}}, \quad (3.4)$$

$$\text{IOD}_{\text{prewhiten}_t} = \beta_0 + \mathbf{c}_2 \text{PDO}_{\text{prewhiten}_{t-[1:3]}} + \mathbf{c}_3 \text{SOI}_{\text{prewhiten}_{t-[1:3]}}, \quad (3.5)$$

where the term $\text{PDO}_{\text{prewhiten}_{t-[1:3]}}$ represents the pre-whitened PDO at lag 1, 2, and 3, same for the pre-whitened SOI. The vector \mathbf{c}_i , for $i = 1, 2, 3$, is the vector of coefficients.

Table 3.4: The estimated coefficients and their standard errors of the fitted regression models for the pre-whitened IOD on the pre-whitened PDO and the pre-whitened SOI.

	$\text{IOD}_{\text{prewhiten}_t}$	$\text{IOD}_{\text{prewhiten}_t}$
$\text{PDO}_{\text{prewhiten}_{t-1}}$	-0.046 (0.019)	-0.014 (0.018)
$\text{PDO}_{\text{prewhiten}_{t-2}}$	0.023 (0.020)	0.017 (0.020)
$\text{PDO}_{\text{prewhiten}_{t-3}}$	✓ 0.041 (0.019)	0.035 (0.018)
$\text{SOI}_{\text{prewhiten}_{t-1}}$		✓ -0.0041 (0.0013)
$\text{SOI}_{\text{prewhiten}_{t-2}}$		✓ -0.0045 (0.0013)
$\text{SOI}_{\text{prewhiten}_{t-3}}$		-0.0018 (0.0013)

The red check mark means that term is statistically significant at the 0.05 level.

Adjusted $R^2 = 0.016$ for Equation 3.4.

Adjusted $R^2 = 0.060$ for Equation 3.5.

Based on the results above, there is no evidence that the PDO is Granger causal for the IOD, as the inclusion of the lagged pre-whitened SOI in the linear regression model leads to the non-significance of the pre-whitened PDO at lag 3, at the 0.05 level. However, the SOI is Granger causal for the IOD, with the SOI leading the IOD by 1 and 2 months. The contemporaneous association between the IOD and the PDO is investigated by fitting linear regression model on the pre-whitened IOD with the pre-whitened PDO as the predictor. The result does not show any significant association, at the 0.05 level (p-value = 0.072). Also, there is no evidence of association between the IOD and the SOI (p-value = 0.54).

The cross correlogram between the pre-whitened IOD and the pre-whitened PDO suggests the negative Granger causation between the IOD and the PDO, with the IOD leading the PDO by a year (Figure 3.12c). To investigate further, two linear regression models are fitted to the pre-whitened PDO,

$$\text{PDO}_{\text{prewhiten}_t} = \alpha_0 + \mathbf{c}_1 \text{IOD}_{\text{prewhiten}_{t-[1:12]}}, \quad (3.6)$$

$$\text{PDO}_{\text{prewhiten}_t} = \beta_0 + \mathbf{c}_2 \text{IOD}_{\text{prewhiten}_{t-[1:12]}} + \mathbf{c}_3 \text{SOI}_{\text{prewhiten}_{t-[1:12]}}, \quad (3.7)$$

where the term $\text{IOD}_{\text{prewhiten}_{t-[1:12]}}$ represents the pre-whitened IOD at lag 1, ..., 12, same for the pre-whitened SOI. The vector \mathbf{c}_i , for $i = 1, 2, 3$, is the vector of coefficients.

The pre-whitened IOD at lag 12 in the regression model (Equation 3.6) is not statistically significant at the 0.05 level (p-value = 0.054). However, as we introduced the lagged pre-whitened SOI in the model (Equation 3.6), the pre-whitened IOD at lag 11 and 12 are both significant at the 0.05 level. It is noticeable that the pre-whitened SOI at lag 8 is statistically significant (p-value = 0.04), which is not suggested by the cross correlogram.

Table 3.5: The estimated coefficients and their standard errors of the fitted regression models for the pre-whitened PDO on the lagged pre-whitened IOD and the SOI, and conditional parameters are chosen.

	$\text{PDO}_{\text{prewhiten}_t}$	$\text{PDO}_{\text{prewhiten}_t}$
$\text{IOD}_{\text{prewhiten}_{t-11}}$	-0.25 (0.15)	✓ -0.35 (0.15)
$\text{IOD}_{\text{prewhiten}_{t-12}}$	-0.27 (0.14)	✓ -0.31 (0.14)
$\text{SOI}_{\text{prewhiten}_{t-4}}$		✓ -0.010 (0.0037)
$\text{SOI}_{\text{prewhiten}_{t-5}}$		✓ -0.013 (0.0037)
$\text{SOI}_{\text{prewhiten}_{t-7}}$		✓ -0.0093 (0.0036)
$\text{SOI}_{\text{prewhiten}_{t-8}}$		✓ -0.0086 (0.0036)

The red check mark means that term is statistically significant at the 0.05 level.

Adjusted $R^2 = 0.010$ for Equation 3.6.

Adjusted $R^2 = 0.063$ for Equation 3.7.

To sum up, there exists contemporaneous association between the SOI and the PDO. There is some evidence that the SOI is Granger causal for both the PDO and the IOD. However, these effects are slight. The Granger causation between the IOD and the PDO, with the IOD leading the PDO by 11 and 12 months, is statistically significant if we take

into account of the effects of the SOI. Generally, a multivariate regression analysis has two substantial advantages over regressions based on just one of the predictor variables. The first is that the multivariate model is likely to account for more of the variance of the response, the second is that the multivariate analysis allows for the investigation of interactions. In the case of correlated predictor variables, the usual situation with survey data, different models may have a similar predictive performance, and it is necessary to qualify the effect of one predictor by stating the other predictors in the model. The Granger causality test in R is considered in a pairwise fashion, but given the association and Granger causality between climatic indicators it seems more appropriate to describe the Granger causality of climatic indicators in terms of the other predictor variables in the model.

Granger causality for multiple time series is usually examined by fitting VAR models. Here, we fit the VAR(2) model to the original series of four climate indicators, and the deseasonalised and detrended series of four climate indicators, shown in [Table 3.6](#) and [Table 3.7](#), respectively. The values in the brackets are the standard errors of the estimated coefficients. Despite the autoregressive terms of each climate indicator, we have the SOI is Granger causal for the IOD at lag 1, and the SOI is Granger causal for the PDO at lag 1, with fairly small estimated coefficients. Moreover, the IOD is Granger causal for the SAM at lag 2. However, the Granger causal relationship between the IOD and the SAM is eliminated if we deseasonalised and detrended the series ([Table 3.7](#)).

Table 3.6: The estimated coefficients and their standard errors of the fitted VAR(2) model for the four climate indicators, along with standard deviations (sd) of residuals and adjusted R^2 .

	SOI _t	IOD _t	PDO _t	SAM _t
SOI _{t-1}	✓0.43 (0.04)	✓-0.003 (0.00)	✓-0.01 (0.00)	0.01 (0.01)
SOI _{t-2}	✓0.25 (0.05)	-0.001 (0.00)	-0.003 (0.003)	0.004 (0.01)
IOD _{t-1}	0.56 (1.82)	✓0.79 (0.05)	-0.15 (0.13)	-0.48 (0.40)
IOD _{t-2}	-3.10 (1.79)	-0.05 (0.05)	0.03 (0.12)	✓0.83 (0.39)
PDO _{t-1}	-1.06 (0.67)	-0.02 (0.02)	✓0.91 (0.05)	-0.08 (0.15)
PDO _{t-2}	0.03 (0.66)	0.01 (0.02)	-0.09 (0.05)	0.04 (0.14)
SAM _{t-1}	-0.08 (0.21)	0.01 (0.01)	-0.01 (0.01)	✓0.20 (0.05)
SAM _{t-2}	0.20 (0.21)	0.01 (0.01)	-0.01 (0.01)	0.00 (0.05)
Sd of residuals	7.76	0.20	0.54	1.70
Adjusted R^2	0.47	0.62	0.76	0.05

The red check mark means that coefficient is statistically significant at the 0.05 level.

Table 3.7: The estimated coefficients and their standard errors of the fitted VAR(2) model for the four deseasonalised and detrended climate indicators, along with standard deviations (sd) of residuals and adjusted R^2 .

	SOI_{ds_t}	IOD_{dsdt_t}	PDO_{dsdt_t}	SAM_{dt_t}
$SOI_{ds_{t-1}}$	✓0.43 (0.04)	✓-0.003 (0.001)	✓-0.01 (0.003)	0.01 (0.01)
$SOI_{ds_{t-2}}$	✓0.26 (0.05)	-0.001 (0.001)	-0.004 (0.003)	0.002 (0.01)
$IOD_{ds_{t-1}}$	0.18 (1.84)	✓0.76 (0.05)	-0.11 (0.13)	-0.54 (0.40)
$IOD_{ds_{t-2}}$	-3.29 (1.81)	-0.06 (0.05)	-0.004 (0.12)	0.73 (0.39)
$PDO_{ds_{t-1}}$	-0.93 (0.68)	-0.01 (0.02)	✓0.88 (0.05)	-0.09 (0.15)
$PDO_{ds_{t-2}}$	0.10 (0.67)	0.01 (0.02)	-0.07 (0.05)	0.07 (0.15)
$SAM_{dt_{t-1}}$	-0.09 (0.21)	0.01 (0.01)	-0.004 (0.01)	✓0.19 (0.05)
$SAM_{dt_{t-2}}$	0.19 (0.21)	0.01 (0.01)	-0.01 (0.01)	-0.01 (0.05)
Sd of residuals	7.74	0.20	0.53	1.69
Adjusted R^2	0.47	0.59	0.74	0.03

The red check mark means that coefficient is statistically significant at the 0.05 level.

As we have concluded that the SOI is Granger causal for the PDO at lag 4 and 5. Therefore, we also fit VAR(5) model to the deseasonalised and detrended climate indicators, to see how is that different compared with the VAR(2) model. As the high dimensionality of the VAR(5) model, we have not listed all the predictors but significant ones at the 0.05 level (Table 3.8). The difference between the fitted VAR(2) model to the deseasonalised and detrended series and the fitted VAR(5) model is the Granger causation between the SOI and the PDO at lag 4. The VAR(2) model suggests that the SOI is Granger causal for the PDO at lag 1, whereas the VAR(5) model suggests that the SOI is Granger causal for the PDO at lag 4.

Table 3.8: The estimated coefficients and their standard errors of the fitted VAR(5) model for the deseasonalised and detrended series, along with standard deviations (sd) of residuals and adjusted R^2 .

	SOI_{ds_t}	IOD_{dsdt_t}	PDO_{dsdt_t}	SAM_{dt_t}
$SOI_{ds_{t-1}}$	✓ 0.41 (0.05)	✓ -0.004 (0.001)	-0.004 (0.003)	0.01 (0.01)
$IOD_{dsdt_{t-1}}$	0.35 (1.87)	✓ 0.75 (0.05)	-0.08 (0.13)	-0.45 (0.41)
$PDO_{dsdt_{t-1}}$	-0.79 (0.69)	-0.01 (0.02)	✓ 0.87 (0.05)	-0.09 (0.15)
$SAM_{dt_{t-1}}$	-0.13 (0.21)	0.01 (0.01)	-0.004 (0.01)	✓ 0.19 (0.05)
$SOI_{ds_{t-2}}$	✓ 0.22 (0.05)	-0.002 (0.001)	-0.002 (0.004)	0.003 (0.01)
$SOI_{ds_{t-3}}$	✓ 0.17 (0.05)	0.001 (0.001)	-0.0005 (0.004)	0.01 (0.01)
$SOI_{ds_{t-4}}$	-0.05 (0.05)	0.001 (0.001)	✓ -0.01 (0.003)	-0.01 (0.01)
Sd of residuals	7.67	0.19	0.53	1.69
Adjusted R^2	0.48	0.60	0.74	0.03

The red check mark means that coefficient is statistically significant at the 0.05 level.

In a VAR(p) model for a climate indicator, the covariate climate indicators can only enter at lags 1, 2, up to p . A univariate regression model can include covariate indicators at lag 0 and allows for a comparison of contemporaneous association with Granger causality. However, contemporaneous association does not improve the predictive power of the covariate indicator. The results of univariate regressions for the IOD and the PDO on the SOI are given in Table 3.9 and Table 3.10, respectively, with only significant terms recorded at the 0.05 level. The results for linear regression models are consistent with the results given by the VAR(5) model, except the linear regression indicates the association between the SOI and the PDO.

Table 3.9: The statistically significant coefficients and their standard errors, at the 0.05 level, of the fitted regression model for IOD on SOI.

	$IOD_{dsdt_{t-1}}$	$SOI_{ds_{t-1}}$
IOD_{dsdt_t}	0.76 (0.05)	-0.004 (0.001)

Table 3.10: The statistically significant coefficients and their standard errors, at the 0.05 level, of the fitted regression model for PDO on SOI.

	$PDO_{dsdt_{t-1}}$	SOI_{ds_t}	$SOI_{ds_{t-4}}$
PDO_{dsdt_t}	0.87 (0.05)	-0.01 (0.003)	-0.01 (0.003)

In summary, there is some evidence that the SOI is Granger causal for the IOD at lag 1, with slight effects, as the estimated coefficients are smaller than -0.005. Moreover, the SOI is also Granger causal for the PDO at lag 4 based on the results of the fitted VAR(5) model. The results of the linear regression suggest the IOD is Granger causal for the PDO at lag 11 and lag 12. However, these are not picked up by the VAR(12) model (not presented in details). There exists contemporaneous association between the SOI and the PDO, and between the IOD and the PDO. However, these cannot be explored by VAR models and do not improve the prediction power.

3.3.2 Relationship between rainfall and climate indicators

We explore the relationship between the average monthly rainfall, over the four weather stations, and the climate indicators, as well as the relationship between the monthly rainfall and the climate indicators for each individual weather station. Then, we look for consistency between the two analyses, and also the differences, using VAR and TVAR models.

Average weather station average rainfall analysis

The average monthly rainfall is calculated as the average of the monthly rainfall at the four weather stations. The significance of seasonal variation and trend are examined by fitting linear regression on lagged average monthly rainfall, time, quadratic trend and harmonic terms (Equation 3.2). In this regression, the variable $\{y_t\}$ represents average monthly rainfall series.

Table 3.11: Statistical significances of the lagged terms, trends and the seasonal patterns with adjusted R^2 .

Response variable	Intercept	Lag 1	Lag 2	Cosine	Sine	$t - \bar{t}$	$(t - \bar{t})^2$	Adjusted R^2
Average monthly rainfall	.00	.00	.63	.00	.05	.42	.49	.32

The results show highly significant seasonal variation, and also the results indicate that there is no significant trend in the average monthly rainfall series, at the 0.05 significance level. Thus, the process of deseasonalisation needs to be performed before fitting the VAR models.

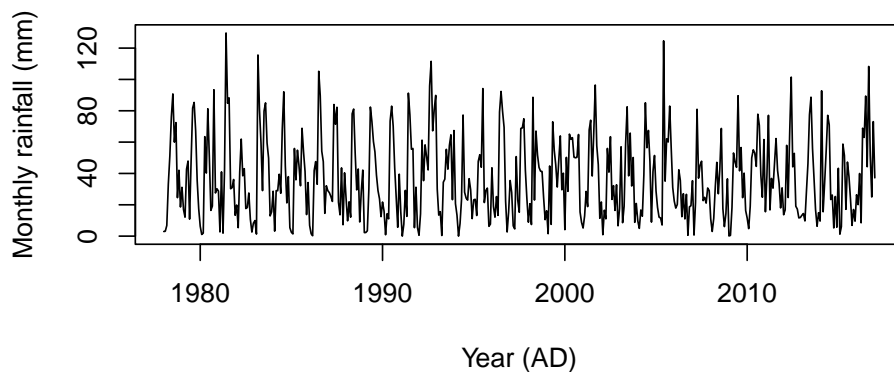


Figure 3.13: Time series plot of the average monthly rainfall.

The seasonality has been removed after we deseasonalised the data (Figure 3.14). The partial correlogram of the deseasonalised series suggests the series is a realisation of an AR(1) process.

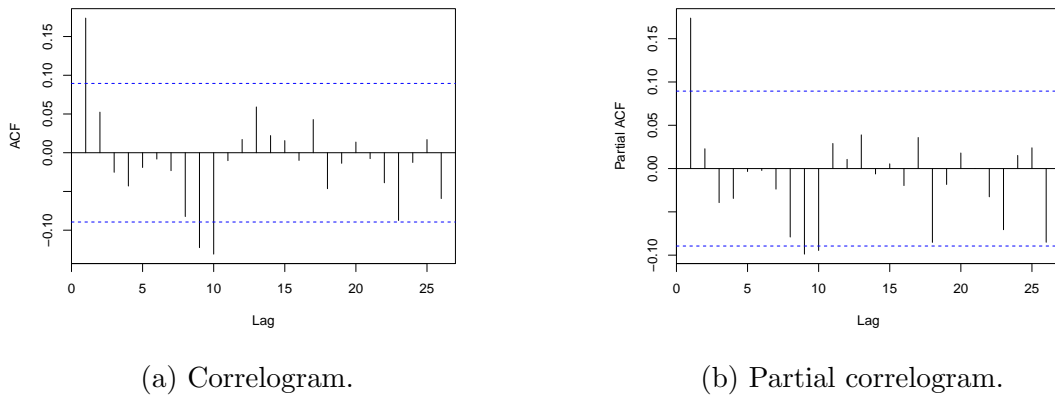
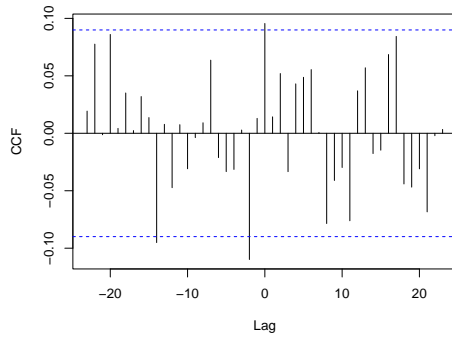
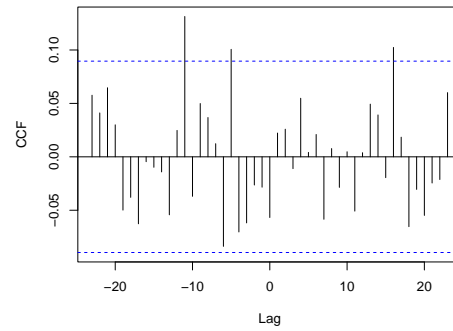


Figure 3.14: ACF and PACF of the deseasonalised average monthly rainfall.

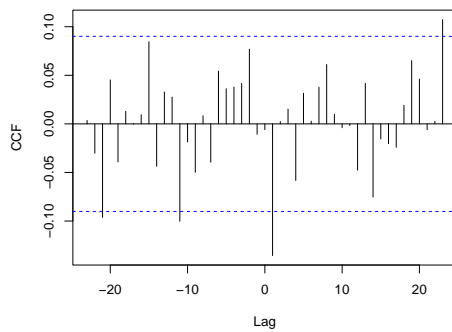
The cross correlations between the four pre-whitened climate indicators and the average monthly rainfall are shown in [Figure 3.15](#), using `prewhiten` function in R. The cross correlogram indicates significant negative association between the SOI and the average rainfall at lag 2, with the SOI leads the average rainfall by 2 months. The positive association between the IOD and the average rainfall at lag -11 suggests that the IOD leads the average rainfall by 11 months, which is physically unrealistic. The cross correlogram between the PDO and the average rainfall suggests that the average rainfall leads the PDO by 1 month, but this is not plausible. This might be due to the association between the average rainfall, and the association between the SOI and the PDO at lag 1. The SAM does not have any significant correlation with the average rainfall.



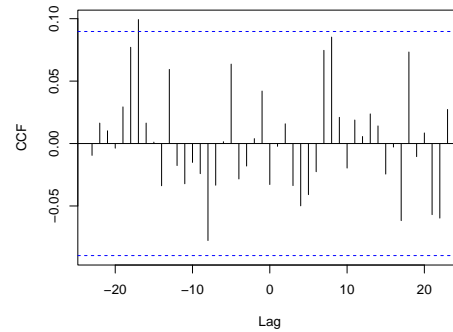
(a) SOI and average monthly rainfall.



(b) IOD and average monthly rainfall.



(c) PDO and average monthly rainfall.



(d) SAM and average monthly rainfall.

Figure 3.15: Cross correlograms between the climate indicators and average monthly rainfall.

Given that Hendon et al. (2007) found seasonal variation in the effect of the SAM on rainfall, we fitted a VAR model that included interactions between seasonal indicators and climatic indicators including the SAM. The inclusion of such interactions did not improve the fit of the model, and the detailed results are not presented. To deal with seasonality, we must either deseasonalise the rainfall data, or include months as indicators in the VAR models, or use temperature data as a proxy. In this study, all of these three methods were used, and they give consistent results. Hence, we only use deseasonalised data for presenting and comparing between different VAR models. We also deseasonalised and detrended climate indicators series.

As Rasel et al. (2015) pointed out that interactions between climate indicators may affect monthly rainfall. Therefore, we have included all the possible two-way interactions between the climate indicators, as well as their quadratic terms. Two VAR(1) and two VAR(2) models have been fitted with variables as follows:

- VAR(1) models:

- Model 1

lag 1: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, \text{rain}_{ave_{ds}}$.

- Model 2

lag 1: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, SOI_{ds}^2, IOD_{dsdt}^2, PDO_{dsdt}^2, SAM_{dt}^2, SOI_{ds} * IOD_{dsdt}, SOI_{ds} * PDO_{dsdt}, SOI_{ds} * SAM_{dt}, IOD_{dsdt} * PDO_{dsdt}, IOD_{dsdt} * SAM_{dt}, PDO_{dsdt} * SAM_{dt}, \text{rain}_{ave_{ds}}$.

- VAR(2) models:

- Model 3

lag 1,2: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, \text{rain}_{ave_{ds}}$.

- Model 4

lag 1,2: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, SOI_{ds}^2, IOD_{dsdt}^2, PDO_{dsdt}^2, SAM_{dt}^2, SOI_{ds} * IOD_{dsdt}, SOI_{ds} * PDO_{dsdt}, SOI_{ds} * SAM_{dt}, IOD_{dsdt} * PDO_{dsdt}, IOD_{dsdt} * SAM_{dt}, PDO_{dsdt} * SAM_{dt}, \text{rain}_{ave_{ds}}$.

The subscripts ds , dt , and $dsdt$ represent deseasonalised, detrended, and deseasonalised and detrended, respectively. The variable $\text{rain}_{ave_{ds}}$ represents the deseasonalised average monthly rainfall.

Table 3.12: Estimated coefficients and their associated standard errors of the four fitted VAR models, and conditional parameters are chosen for model 3 and model 4.

	VAR(1)		VAR(2)	
	Model 1	Model 2	Model 3	Model 4
$\text{rain}_{ave_{ds_{t-1}}}$	✓ 0.14 (0.05)	✓ 0.10 (0.05)	✓ 0.12 (0.05)	0.08 (0.05)
$\text{SOI}_{ds_{t-1}}$	0.13 (0.10)	✓ 0.22 (0.10)	✓ 0.26 (0.12)	✓ 0.33 (0.12)
$\text{IOD}_{dsdt_{t-1}}$	✓ -9.67 (3.33)	✓ -11.75 (3.63)	✓ -10.96 (4.98)	✓ -10.84 (5.08)
$\text{PDO}_{dsdt_{t-1}}$	1.41 (1.00)	1.49 (1.05)	-2.14 (1.84)	-1.87 (1.90)
$\text{SOI}_{ds_{t-1}}^2$		✓ 0.02 (0.01)		0.01 (0.01)
$\text{IOD}_{dsdt_{t-1}}^2$		✓ 15.59 (7.43)		7.62 (9.06)
$\text{PDO}_{dsdt_{t-1}}^2$		✓ 2.35 (0.87)		1.89 (1.08)
$\text{SOI}_{ds_{t-1}} * \text{IOD}_{dsdt_{t-1}}$		✓ 0.70 (0.34)		0.71 (0.38)
$\text{SOI}_{ds_{t-1}} * \text{PDO}_{dsdt_{t-1}}$		0.14 (0.13)		0.12 (0.14)
$\text{IOD}_{dsdt_{t-1}} * \text{PDO}_{dsdt_{t-1}}$		3.83 (3.40)		✓ 9.56 (4.49)
$\text{SOI}_{ds_{t-2}}$			-0.24 (0.12)	-0.19 (0.13)
$\text{IOD}_{dsdt_{t-2}}$			-0.80 (4.91)	-1.85 (5.08)
$\text{PDO}_{dsdt_{t-2}}$			✓ 3.61 (1.82)	1.89 (1.08)
adj- R^2	0.05	0.07	0.06	0.08
sd	21.14	20.88	21.05	20.85

The red check mark means that coefficient is statistically significant at the 0.05 level.

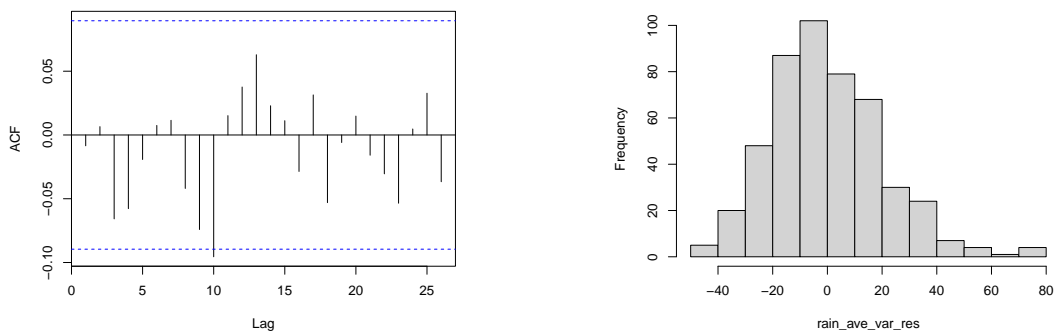
In terms of the standard deviation of residuals and hence adjusted R^2 , model 4 is the best fitted model. However, the difference between model 2 and model 4 is not statistically significant using F test, and model 2 is simpler than model 4 with 15 parameters less.

The effects of the SAM are not presented in Table 3.12 for all four models, as the SAM does not show any relationships with rainfall. Not all the parameters are presented in Table 3.12 for model 3 and model 4, as these two models have large number of parameters. The results show the average monthly rainfall series has an autoregressive term of order 1, whereas the lag 1 average monthly rainfall term in model 4 is significant at the 0.10 level

(p-value = 0.08). The IOD is Granger causal for the average monthly rainfall, and this is consistent for all fitted models. If the IOD is negative, we expect the amount of rainfall in the following month to be increased. The effect of the lagged SOI is statistically significant for all models except model 1. Moreover, the quadratic terms of climate indicators at lag 1 are statistically significant at the 0.05 level, but not for the SAM. The estimate coefficient for the quadratic term of the SOI is 0.02, who has the confidence interval (0.0014, 0.030), which is relatively small compare to the IOD² and the PDO². The interaction between the SOI and the IOD is marginally significant (p-value = 0.04). As both of the SOI and the IOD at lag 1 are highly significant, we do expect their interaction to be significant as well. The significant quadratic terms as well as the interaction between the lagged SOI and the lagged IOD in model 2 become non-significant in model 4 as we introduce lag 2 terms. This is because that the climate indicators are highly correlated. The SAM does not have any effect on the average rainfall.

As model 2 suggests, if the IOD is 0.4°C, which is the boundary of the positive phase, the average rainfall will decrease by 2.21 mm for the following month, that is 5.98%. Furthermore, if the SOI is -7, which is the boundary of an El Niño, the average rainfall will decrease by 2.52 mm, that is 6.83%. Given the value of the monthly IOD in December 2020 is 0.10°C, and the monthly SOI is 16.9, we expect the average rainfall increased by 9.59 mm, which is 25.99% for the following month, within South Australia.

The residuals have been obtained from model 2, that is VAR(1) model with interactions and quadratic terms. The correlogram of the residuals appears as a realisation of white noise process, and the histogram shows positively skewed.



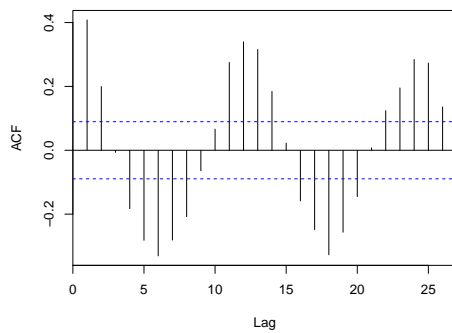
(a) Correlogram.

(b) Histogram.

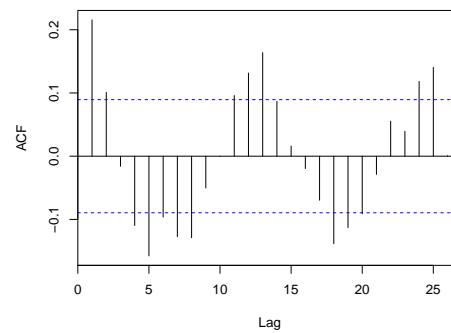
Figure 3.16: Plots of the residuals obtained from model 2.

Individual weather station analysis

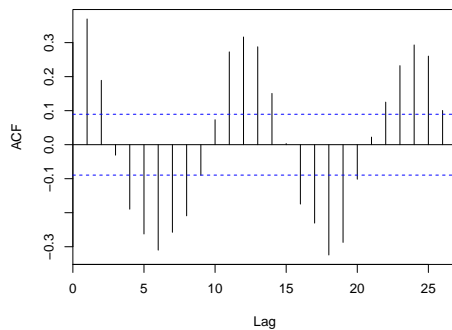
The correlograms show significant peaks at lags 6, 12, 18, 24, . . . , and the sinusoidal curves for the Adelaide Airport, Pooraka, and North Adelaide weather stations (Figure 3.17a, Figure 3.17c, Figure 3.17d), indicate the presence of seasonal variations. The correlogram for the Murray Bridge station (Figure 3.17b) shows spikes at lags 5, 13, 18, and 25 with sinusoidal curve, and the auto-correlations here are relatively smaller than the other three weather stations. Thus, the process of deseasonalisation needs to be performed before fitting the VAR models. Additionally, the KPSS test shows that the deterministic trend is not statistically significant at the 0.05 level, for each of the weather station.



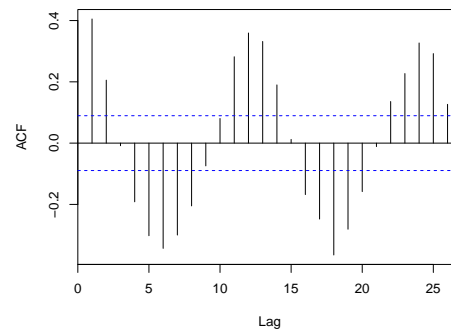
(a) Adelaide Airport.



(b) Murray Bridge.



(c) Pooraka.

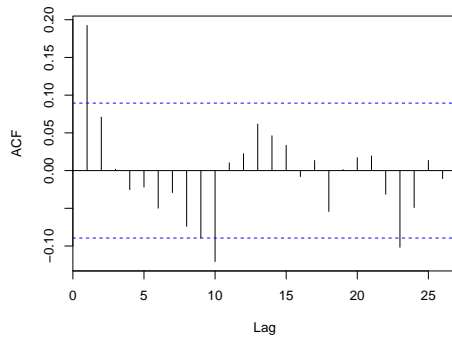


(d) North Adelaide.

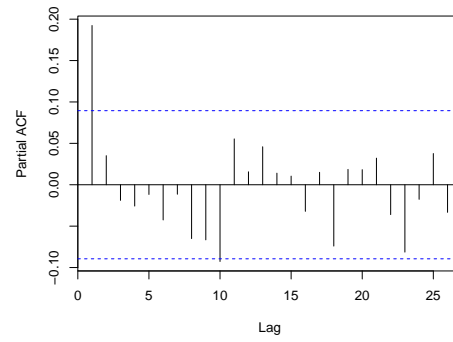
Figure 3.17: Correlograms of the monthly rainfall at four weather stations.

The seasonal patterns for all of the weather stations have been removed by fitting linear regressions on months as indicators. The correlograms of the residuals of the monthly rainfall (Figure 3.18) do not show any sinusoidal curve. The partial correlograms for the Adelaide Airport and the North Adelaide monthly rainfall show significance at lag 1, with

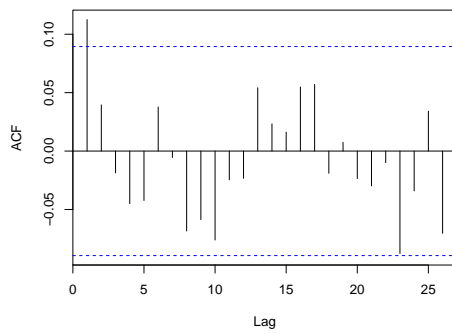
correlations greater than 0.15. Whereas, the partial correlations at lag 1 for the monthly rainfall at the other two weather stations are still statistically significant but relatively small in value. Therefore, AR(1) models are appropriate to model the monthly rainfall.



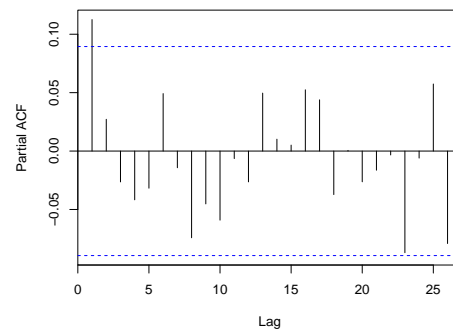
(a) Adelaide Airport.



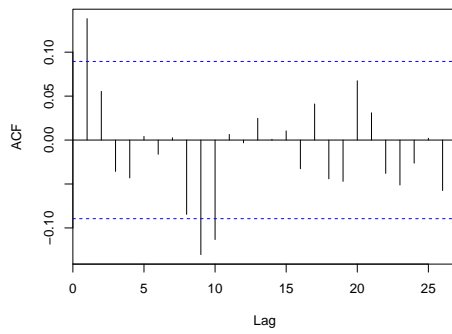
(b) Adelaide Airport.



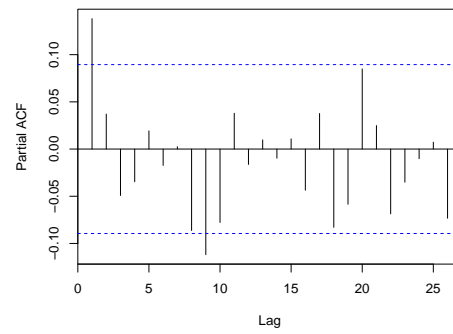
(c) Murray Bridge.



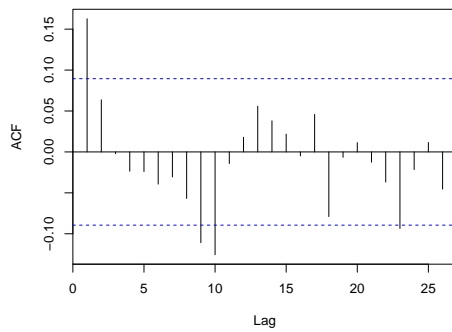
(d) Murray Bridge.



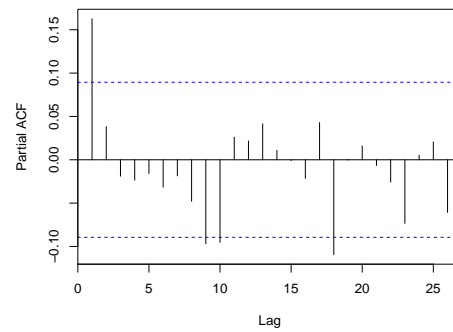
(e) Pooraka.



(f) Pooraka.



(g) North Adelaide.

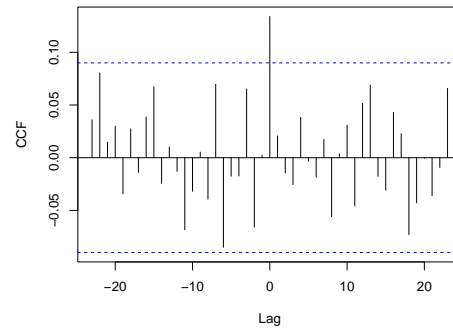
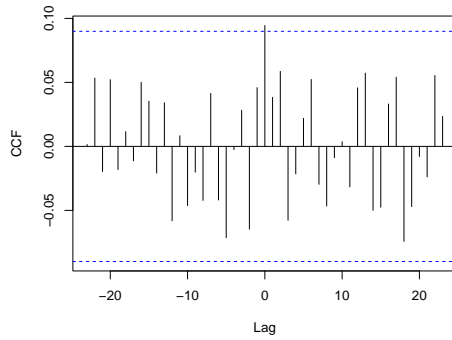


(h) North Adelaide.

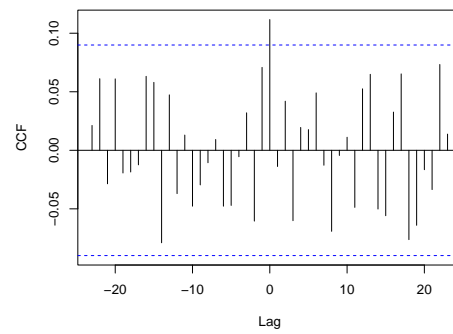
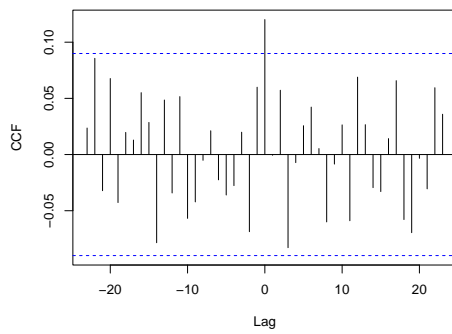
Figure 3.18: Correlograms and partial correlograms of the deseasonalised monthly rainfall at four weather stations.

VAR models

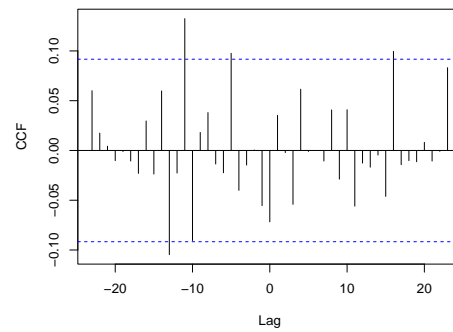
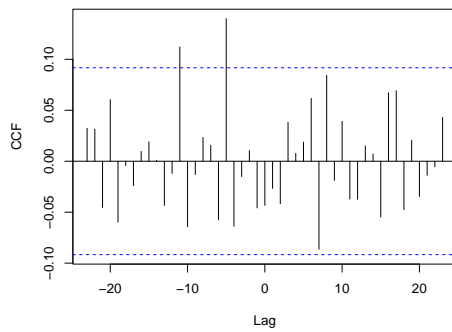
As a part of the exploratory data analysis, we looked at the cross correlograms between the individual pre-whitened climate indicator and the monthly rainfall using `prewhiten` function in R, at four weather stations (Figure 3.19, Figure 3.20). They can be used as an initial investigation on data so as to discover which climate indicators are associated with, or Granger causal for, rainfall. The cross correlograms suggest significant association between the SOI and monthly rainfall at all stations, with correlations smaller than 0.15. The IOD appears to be Granger causal for monthly rainfall at lag 5, but this is physically unrealistic. There is no evidence of any relationship between SAM and monthly rainfall at any of the weather stations. Figure 3.20a suggests that monthly rainfall is Granger causal for the PDO at lag 1, but this is not plausible. This might be due to the association between rainfall and the SOI, and the causality between the SOI and the PDO at lag 1. However, several climate indicators are Granger causal for rainfall, and given that the climate indicators are themselves associated and that some indicators are Granger causal for others, it is likely that a multivariate VAR model will be a better predictor of rainfall than a model restricted to just one single climate indicator.



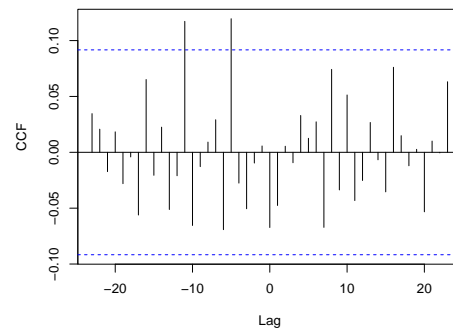
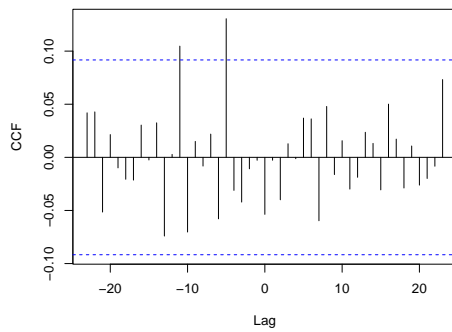
(a) SOI and Adelaide Airport monthly rainfall. (b) SOI and Murray Bridge monthly rainfall.



(c) SOI and Pooraka monthly rainfall. (d) SOI and North Adelaide monthly rainfall.

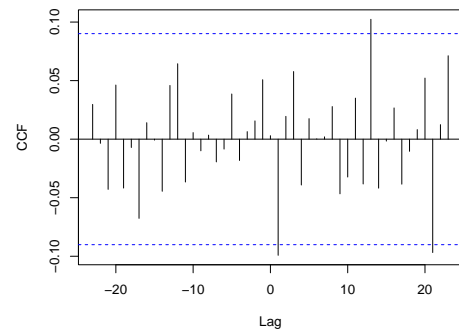
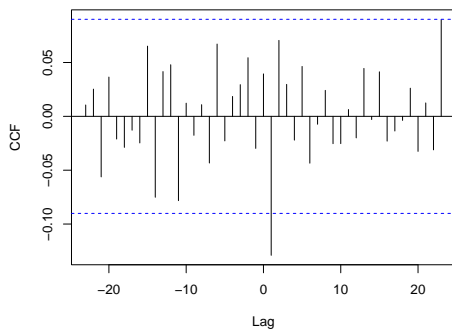


(e) IOD and Adelaide Airport monthly rainfall. (f) IOD and Murray Bridge monthly rainfall.

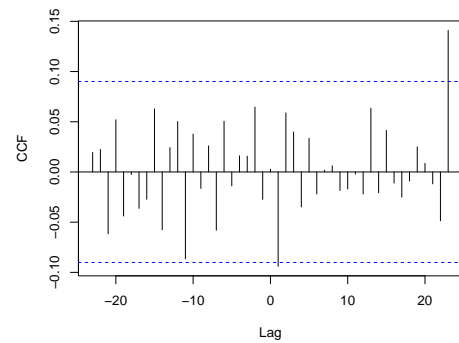
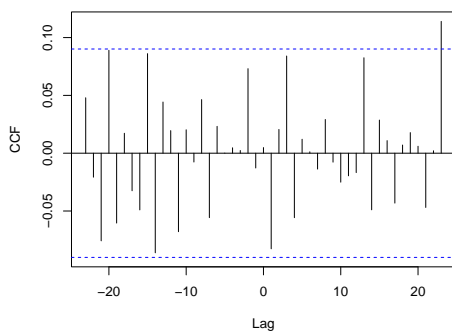


(g) IOD and Pooraka monthly rainfall. (h) IOD and North Adelaide monthly rainfall.

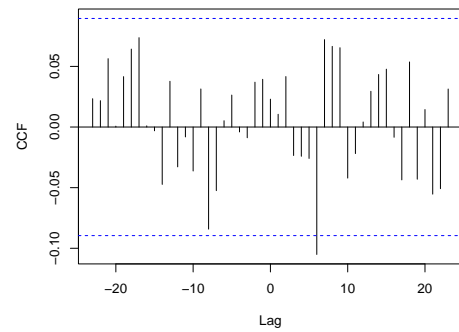
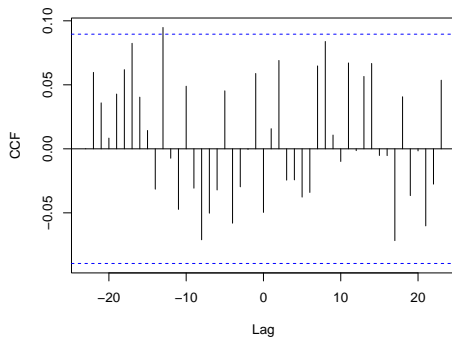
Figure 3.19: Cross correlograms between climate indicators and pre-whitened rainfall at four weather stations.



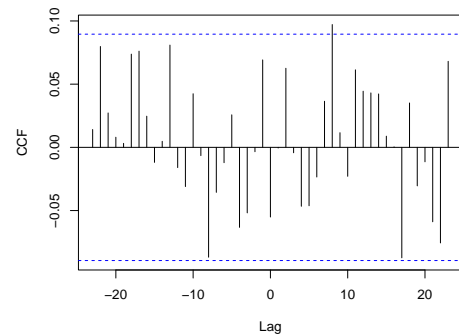
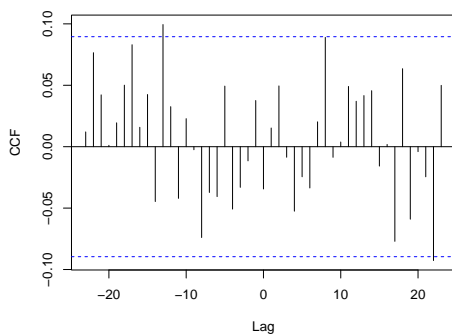
(a) PDO and Adelaide Airport monthly rainfall. (b) PDO and Murray Bridge monthly rainfall.



(c) PDO and Pooraka monthly rainfall. (d) PDO and North Adelaide monthly rainfall.



(e) SAM and Adelaide Airport monthly rainfall. (f) SAM and Murray Bridge monthly rainfall.



(g) SAM and Pooraka monthly rainfall. (h) SAM and North Adelaide monthly rainfall.

Figure 3.20: Cross correlograms between climate indicators and pre-whitened rainfall at four weather stations.

We fit the same four VAR models as in the average rainfall analysis. Recall these four models are:

- VAR(1) models:
 - Model 1
lag 1: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, rainfall_{ds}$.
 - Model 2
lag 1: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, SOI_{ds}^2, IOD_{dsdt}^2, PDO_{dsdt}^2, SAM_{dt}^2, SOI_{ds} * IOD_{dsdt}, SOI_{ds} * PDO_{dsdt}, SOI_{ds} * SAM_{dt}, IOD_{dsdt} * PDO_{dsdt}, IOD_{dsdt} * SAM_{dt}, PDO_{dsdt} * SAM_{dt}, rainfall_{ds}$.
- VAR(2) models:
 - Model 3
lag 1,2: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, rainfall_{ds}$.
 - Model 4
lag 1,2: $SOI_{ds}, IOD_{dsdt}, PDO_{dsdt}, SAM_{dt}, SOI_{ds}^2, IOD_{dsdt}^2, PDO_{dsdt}^2, SAM_{dt}^2, SOI_{ds} * IOD_{dsdt}, SOI_{ds} * PDO_{dsdt}, SOI_{ds} * SAM_{dt}, IOD_{dsdt} * PDO_{dsdt}, IOD_{dsdt} * SAM_{dt}, PDO_{dsdt} * SAM_{dt}, rainfall_{ds}$.

The dimension of the model is equivalent to the number of variables in the model. That means model 1 and model 3 have dimension of 5, whereas model 2 and 4 have dimension of 15.

The standard deviation of errors and adjusted R^2 for each model and for each weather station are shown in [Table 3.13](#). Model 2 is the one that returns the lowest standard deviation of errors and hence highest adjusted R^2 for the Murray Bridge station, whereas model 4 is the best for the other three weather stations. However, the differences between standard deviations of errors are smaller than 0.1, and the differences between adjusted R^2 s are smaller than 0.01, for those two models.

Table 3.13: The standard deviations of residuals and adjusted R^2 of fitted four models for four weather stations.

	VAR(1)				VAR(2)			
	Model 1		Model 2		Model 3		Model 4	
	sd	adj- R^2	sd	adj- R^2	sd	adj- R^2	sd	adj- R^2
Adelaide Airport	21.66	0.06	21.43	0.07	21.54	0.07	21.31	0.09
Murray Bridge	21.06	0.03	20.89	0.04	21.11	0.03	20.99	0.04
Pooraka	22.8	0.04	22.54	0.06	22.72	0.04	22.47	0.06
North Adelaide	24.76	0.05	24.44	0.07	24.61	0.06	24.35	0.08

Based on [Table 3.14](#), we can conclude that the IOD is Granger causal for monthly rainfall, with the IOD apparently leading by 1 month. This is the case for all four weather stations. The lagged rainfall at the Murray Bridge station is not statistically significant at the 0.05 level. The correlogram for the Murray Bridge station ([Figure 3.18c](#)) shows the lowest auto-correlation at lag 1 compared with other three stations. The SOI at lag 1 is only significant for Pooraka and North Adelaide station, in model 2. The quadratic term of SOI at lag 1 is statistically significant for all weather stations but Murray Bridge, at the 0.05 level. Moreover, the quadratic term of PDO is significant at all weather stations. The interaction between the SOI and the IOD at lag 1 is marginally significant at the 0.05 level, but not for the Murray Bridge station.

[Table 3.15](#) only showed a part of the predictor variables as model 3 and model 4 have large number of parameters. As the results returned from fitted VAR(1) models, we have strong evidence that the IOD Granger causal for monthly rainfall, and the effect of the lagged SOI is equivocal. However, the VAR(2) models suggest the Granger causal relations between the SOI and monthly rainfall at all weather stations except Murray Bridge. The effects of the IOD become non-significant, at the 0.05 level, at Pooraka and North Adelaide stations. This might due to the significance of the PDO at lag 2, as the SOI is Granger causal for IOD, and the SOI is associated with the PDO. Therefore, it is not unlikely to observe the causation between the PDO and the IOD at lag 2 and 3. The quadratic terms of the climate indicators become non-significant at the 0.05 level, except the SOI_{dst-1}^2 at Pooraka. Moreover, the interaction between the SOI and the IOD at lag 1 becomes non-significant, whereas the interaction between the IOD and the PDO becomes significant. This is because climate indicators are highly correlated with each other.

There is strong evidence that IOD is Granger causal for monthly rainfall, with negative relationship. That is, for example, if IOD is 0.4, which is the boundary of the positive phase, according to the model 2 at the Adelaide Airport station, the average rainfall would decrease by 4.69 mm, which is decreased by 13%. Moreover, the average rainfall will decrease by 5.15 mm (17%), 4.27 mm (11%), and 4.99 mm (12%), for Murray Bridge, Pooraka, and North Adelaide, respectively.

The results for the average monthly rainfall analysis are consistent with the analysis of each individual weather station. In particular, there is strong evidence of the IOD to be Granger causal for monthly rainfall, with the IOD leads rainfall by a month. Moreover, the IOD at lag 2 does not have any impact on monthly rainfall, according to the results of any fitted VAR(2) models. The SAM does not have any relationships between monthly rainfall, even take into account of the seasonal effects. Furthermore, the relationships between the interactions of climate indicators (except the SAM) and monthly rainfall are equivocal in both of the analyses. The main differences between the two analyses are, the Granger causations between: the SOI and monthly rainfall; the quadratic terms of

Table 3.14: Estimated coefficients and their associated standard errors of the two VAR(1) models for four weather stations, at the 0.05 significant level, with conditional parameters are chosen.

	Adelaide Airport		Murray Bridge		Pooraka		North Adelaide	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Rainfall $_{d_{st-1}}$	✓0.16 (0.05)	✓0.13 (0.05)	0.08 (0.05)	0.06 (0.05)	✓0.11 (0.05)	0.07 (0.05)	✓0.13 (0.05)	✓0.10 (0.05)
SOI $_{d_{st-1}}$	0.11 (0.10)	0.20 (0.11)	0.06 (0.10)	0.13 (0.11)	0.18 (0.11)	✓0.27 (0.11)	0.19 (0.12)	✓0.29 (0.12)
IOD $_{d_{st-1}}$	✓-9.61 (3.44)	✓-11.73 (3.76)	✓-10.09 (3.33)	✓-12.87 (3.65)	✓-9.00 (3.61)	✓-10.67 (3.94)	✓-10.77 (3.94)	✓-12.47 (4.29)
PDO $_{d_{st-1}}$	1.48 (1.04)	1.63 (1.09)	1.04 (1.01)	0.73 (1.06)	1.57 (1.09)	1.79 (1.14)	1.62 (1.19)	1.85 (1.24)
SAM $_{d_{st-1}}$	0.68 (0.58)	0.42 (0.60)	0.59 (0.56)	0.36 (0.58)	0.37 (0.61)	0.10 (0.63)	0.86 (0.66)	0.59 (0.68)
SOI $_{d_{st-1}}$		✓0.02 (0.01)		0.01 (0.01)		✓0.02 (0.01)		✓0.02 (0.01)
IOD $_{d_{st-1}}$		14.48 (7.70)		✓16.09 (7.51)		14.78 (8.07)		✓18.01 (8.77)
PDO $_{d_{st-1}}$		✓2.37 (0.90)		✓2.10 (0.88)		✓2.44 (0.95)		✓2.54 (1.03)
SAM $_{d_{st-1}}$		-0.37 (0.24)		-0.28 (0.23)		-0.37 (0.25)		-0.45 (0.27)
SOI $_{d_{st-1}}$ *		✓0.68 (0.35)		0.60 (0.34)		✓0.74 (0.37)		✓0.82 (0.40)
IOD $_{d_{st-1}}$								

The red check mark means that coefficient is statistically significant at the 0.05 level.

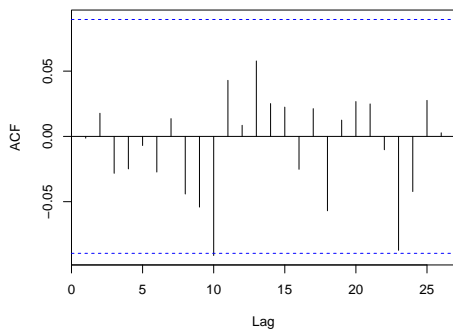
Table 3.15: Estimated coefficients and their associated standard errors of the two VAR(2) models for four weather stations, at the 0.05 significant level, with conditional parameters are chosen.

	Adelaide Airport		Murray Bridge		Pooraka		North Adelaide	
	Model 3	Model 4	Model 3	Model 4	Model 3	Model 4	Model 3	Model 4
Rainfall _{dst-1}	✓0.15 (0.05)	✓0.11 (0.05)	0.07 (0.05)	0.04 (0.05)	✓0.10 (0.05)	0.05 (0.05)	✓0.12 (0.05)	0.08 (0.05)
SOI _{dst-1}	✓0.26 (0.12)	✓0.32 (0.13)	0.14 (0.12)	0.22 (0.13)	✓0.31 (0.13)	✓0.37 (0.14)	✓0.35 (0.14)	✓0.45 (0.15)
IOD _{dsdtt-1}	✓-13.40 (5.13)	✓-13.27 (5.24)	✓-11.19 (5.02)	✓-11.62 (5.16)	-9.26 (5.41)	-8.81 (5.52)	-10.61 (5.87)	-10.11 (6.00)
SOI _{dst-2}	✓-0.26 (0.13)	-0.21 (0.13)	-0.16 (0.12)	-0.14 (0.13)	-0.24 (0.13)	-0.18 (0.14)	✓-0.30 (0.15)	-0.25 (0.15)
PDO _{dsdtt-2}	✓4.38 (1.88)	✓4.30 (1.91)	1.24 (1.84)	0.93 (1.88)	✓4.01 (1.98)	✓4.33 (2.01)	✓5.04 (2.14)	✓5.05 (2.18)
SAM _{dt-2}	-0.14 (0.59)	0.11 (0.62)	0.47 (0.58)	0.50 (0.61)	-0.27 (0.62)	-0.14 (0.65)	-0.30 (0.67)	-0.21 (0.71)
SOI _{dst-1} ²		0.01 (0.01)		0.01 (0.01)		✓0.02 (0.01)		0.02 (0.01)
IOD _{dsdtt-1} ²		5.35 (9.35)		11.61 (9.22)		6.39 (9.85)		7.29 (10.68)
PDO _{dsdtt-1} ²		1.88 (1.12)		1.79 (1.10)		1.72 (1.18)		2.14 (1.27)
SOI _{dst-1} *		0.72 (0.39)		0.71 (0.39)		0.62 (0.42)		0.84 (0.45)
IOD _{dsdtt-1} *		✓9.80 (4.63)		✓9.08 (4.56)		✓10.04 (4.89)		9.59 (5.29)
PDO _{dsdtt-1}								
IOD _{dsdtt-2} *		✓4.19 (2.02)		3.76 (2.00)		2.13 (2.14)		3.45 (2.31)
SAM _{dt-2}								

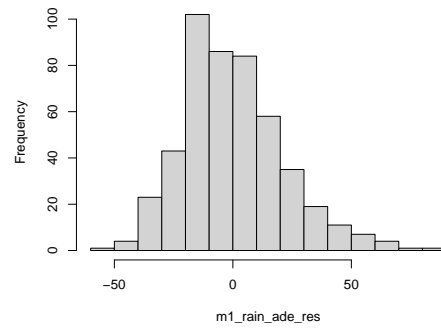
The red check mark means that coefficient is statistically significant at the 0.05 level.

the climate indicators and monthly rainfall; and lag 2 terms of the climate indicators, are equivocal in each individual weather station analysis.

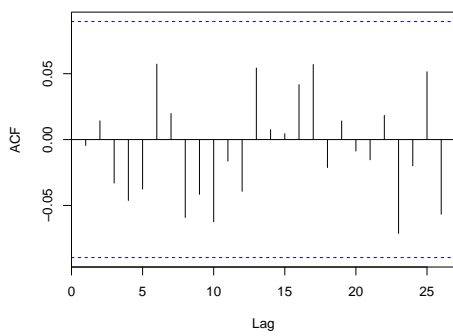
The residuals for each individual weather station have been obtained from fitted VAR(1) model with linear term of climate indicators only. The residuals of monthly rainfall at each weather station appear as a realisation of white noise process, and the histograms of residuals are likely to be positively skewed as monthly rainfall is itself positively skewed (Figure 3.21). Consequently, limits of prediction will not be accurate based on the assumption of normality. Additionally, p-values for coefficients will not be precisely correct, but the error will be slight. The cross correlograms between the residuals of the monthly rainfall show high correlations at lag 0 (Figure 3.22). It is noticeable that the associations of the monthly rainfall at the Murray Bridge and the other three weather stations are lower than 0.7, whereas the associations among other three weather stations are greater than 0.8. This can be explained by the fact that the distance between Murray Bridge and the other three stations is much greater than the distances between the other three stations. Also, Murray Bridge is separated from the other three weather stations by Adelaide Hills.



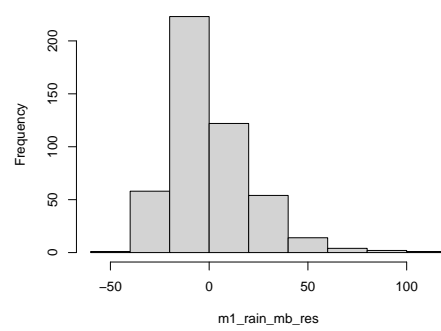
(a) Adelaide Airport.



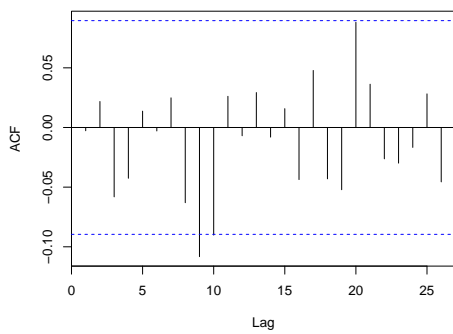
(b) Adelaide Airport.



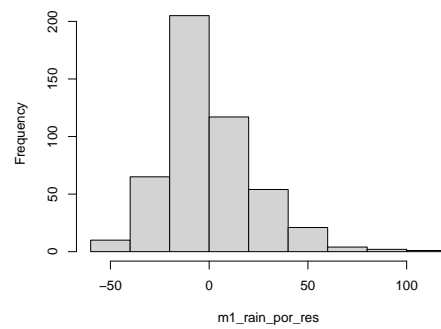
(c) Murray Bridge.



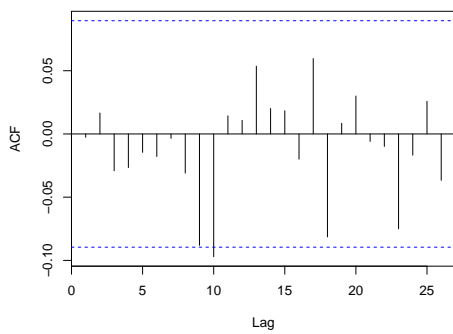
(d) Murray Bridge.



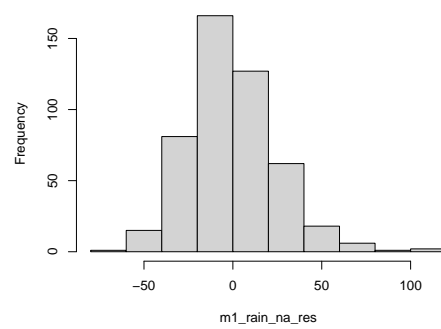
(e) Pooraka.



(f) Pooraka.

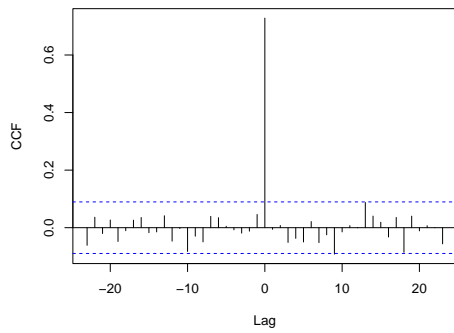


(g) North Adelaide.

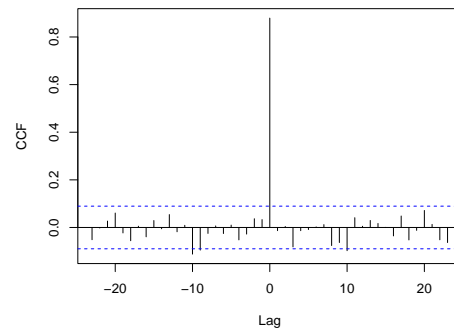


(h) North Adelaide.

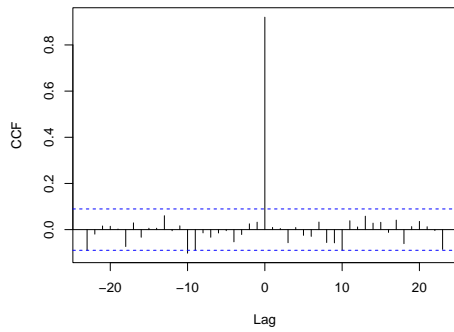
Figure 3.21: Correlograms and histograms of the residuals of monthly rainfall obtained from fitted VAR(1) with linear terms only, for four weather stations.



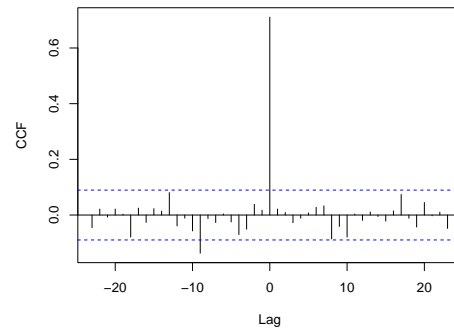
(a) Adelaide Airport and Murray Bridge.



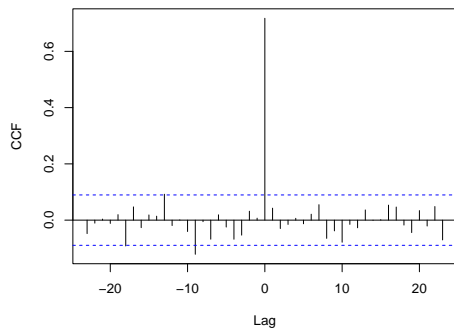
(b) Adelaide Airport and Pooraka.



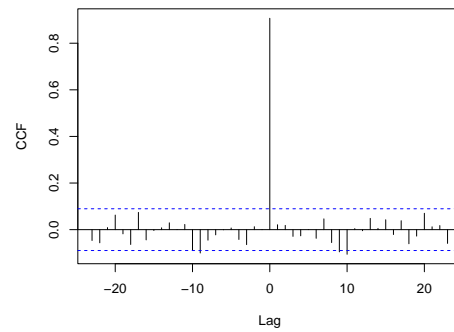
(c) Adelaide Airport and North Adelaide.



(d) Murray Bridge and Pooraka.



(e) Murray Bridge and North Adelaide.



(f) Pooraka and North Adelaide.

Figure 3.22: Cross correlograms between the residuals of the monthly rainfall obtained from model 1.

TVAR models

Linear models are sometimes preferred to non-linear models, as they provide close approximation to the time series and are simpler in calculation and interpretation. However, most of the data in real life show the property of non-linearity and sometimes non-linear models outperform linear models for forecasting. The non-linear time series model, threshold vector autoregressive (TVAR) models, are used in this study. Not only because the TVAR model is the non-linear extension of the VAR model, but also because the ease of implementation of the TVAR model. Most importantly, threshold are inherently directional (Lawrance 1991), and the TVAR model uses threshold space to capture several non-linear characteristics of a series (Mansor et al. 2018). Deseasonalised rainfall show significant directionality using peak-over-threshold (POT) test, at the 0.05 level (Mansor et al. 2019). The threshold value we used in the POT test is 80%, and h , the number of observations before and after the peak, is set to be 2.

We fit a 1 threshold TVAR(1) model to the deseasonalised and detrended series at each individual weather station, without (model A) and with (model B) any significant interactions that found in fitted VAR models (Table 3.14, Table 3.15). The two models are:

- Model A: $\text{Rainfall}_{\text{dsdt}}, \text{SOI}_{\text{ds}}, \text{IOD}_{\text{dsdt}}, \text{PDO}_{\text{dsdt}}, \text{SAM}_{\text{dt}}$.
- Model B: $\text{Rainfall}_{\text{dsdt}}, \text{SOI}_{\text{ds}}, \text{IOD}_{\text{dsdt}}, \text{PDO}_{\text{dsdt}}, \text{SAM}_{\text{dt}}, \text{SOI}_{\text{ds}} * \text{IOD}_{\text{dsdt}}, \text{IOD}_{\text{dsdt}} * \text{PDO}_{\text{dsdt}}$.

As we can observe from Table 3.16 that adjusted R-squares for all eight models are smaller than 0, that might due to the small sample size relative to the number of parameters, as we have 60 ((5 predictors + intercept) \times 2 regimes \times 5 equations / response variables) and 112 ((7 predictors + intercept) \times 2 regimes \times 7 equations / response variables) parameters in model A and model B, respectively. Hence, we fit a 1 threshold TVAR(1) model to the deseasonalised and detrended climate indicators: SOI_{ds} ; IOD_{dsdt} ; and PDO_{dsdt} , as SAM_{dt} does not show any impact on monthly rainfall according to fitted VAR(1) and VAR(2) models. In this case, we have adjusted R-squares fairly close to 0, and the estimated coefficients with their standard errors are shown in Table 3.17

Table 3.16: Threshold values, percentage of observations below thresholds, standard deviations of residuals and adjusted R^2 s for 1 threshold TVAR(1) model with and without interaction terms, for four weather stations.

		Threshold	Observations below the threshold (%)	Sd of residuals	Adjusted R^2
Adelaide Airport	Model A	21.54	85.2	21.30	-0.05
	Model B	-13.07	31.9	21.16	-0.18
Murray Bridge	Model A	-21.40	11.1	20.79	-0.08
	Model B	14.00	78.7	20.68	-0.22
Pooraka	Model A	-14.27	28.4	22.47	-0.07
	Model B	-14.63	27.3	22.28	-0.20
North Adelaide	Model A	23.96	82.7	24.39	-0.06
	Model B	-14.78	29.2	24.05	-0.17

The red check mark means that coefficient is statistically significant at the 0.05 level.

Resulting from fitting TVAR models, there is strong evidence the IOD is Granger causal for rainfall with a negative relationship, as the IOD is statistically significant for four weather stations either in the upper regimes or in the lower regimes. This is consistent with the results of fitted VAR(1) models. The SAM does not have any significant effects on rainfall at these four weather stations, and the PDO has significant effects in the upper regime (threshold value = 21.54) for Adelaide Airport weather station.

Table 3.17: Estimated coefficients and their standard errors for each regime for the fitted 1 threshold TVAR(1) model, along with threshold values, standard deviations of residuals and adjusted R^2 .

	Adelaide Airport	Murray Bridge	Pooraka	North Adelaide
Regime 1				
Rainfall $_{dst-1}$	0.06 (0.07)	0.91 (0.72)	0.21 (0.23)	0.40 (0.20)
SOI $_{dst-1}$	0.10 (0.12)	0.42 (0.32)	-0.14 (0.21)	0.39 (0.22)
IOD $_{dsdt_{t-1}}$	✓ -10.80 (3.91)	0.02 (10.33)	✓ -19.01 (7.10)	-3.98 (7.40)
PDO $_{dsdt_{t-1}}$	0.60 (1.14)	5.38 (2.74)	0.73 (1.93)	2.44 (2.08)
Regime 2				
Rainfall $_{dst-1}$	-0.01 (0.19)	✓ 0.11 (0.05)	✓ 0.18 (0.06)	✓ 0.21 (0.07)
SOI $_{dst-1}$	0.28 (0.26)	0.03 (0.11)	✓ 0.33 (0.13)	0.20 (0.15)
IOD $_{dsdt_{t-1}}$	-10.69 (7.50)	✓ -11.09 (3.54)	-5.98 (4.20)	✓ -12.27 (4.69)
PDO $_{dsdt_{t-1}}$	✓ 5.53 (2.57)	0.29 (1.09)	1.82 (1.33)	1.30 (1.45)
Threshold value	21.54	-21.40	-14.27	-14.48
Sd of residuals	21.41	20.84	22.48	24.46
Adjusted square	R- -0.01	-0.04	-0.02	-0.01

The red check mark means that coefficient is statistically significant at the 0.05 level.

3.4 Conclusion

There was no evidence of a trend, at the 0.05 significance level, in the station average monthly rainfall series or in the individual station monthly rainfall series. However, there is evidence of slight linear trends in all the climate indicators, except the SOI. All time series were detrended before investigating Granger causality, to avoid a common trend. As the common trend could lead to a spurious claim of Granger causal relationship.

South Australia has a Mediterranean climate with higher rainfall during the winter and hot dry summers. Seasonality was also present in all the climate indicators, except the SAM. It is necessary to deseasonalise time series before investigating Granger causality, because one deterministic seasonal patterns can be predicted from another and this is not an interesting Granger causal relationship.

There are no substantial relationships between the four climate indicators, but some features are statistically significant. The SOI is Granger causal for the IOD, with the SOI leading by one and two months, with a negative relationship. Additionally, the SOI is associated with the PDO, as well as leading the PDO by 4 months and 5 months, with negative relations.

The results for the average weather station average rainfall analysis show strong evidence that the IOD is Granger causal for monthly rainfall, with a negative relationship. Inclusion of a quadratic terms and interactions with the SOI and the PDO give a slight improvement in predictions. Including lag 2 terms gives a very slight, but not statistically significant, improvement in predictions. The SAM does not show any significant Granger causation with monthly rainfall, at the 0.05 level.

The results for each individual weather station are consistent with each other, and also consistent with the station average rainfall analysis. Based on lag one models there is strong evidence that the IOD is Granger causal for monthly rainfall, with positive IOD predicting lower rainfall. This conclusion is slightly more equivocal, at two of the four stations, if lag 2 models are used but lag 2 models only give a very slight improvement in forecasting ability. The increased complexity offered by threshold vector autoregressive models is not justified. There is no evidence of any relationship between the SAM, or its interactions with the other climatic indicators, and monthly rainfall.

By comparing the results for the average weather station with the results for individual weather stations, we can observe the Granger causation between the IOD and the monthly rainfall is consistent. Additionally, the PDO² in both of the analyses shows significant Granger causation with monthly rainfall, in fitted VAR(1) models, but the effect becomes non-significant with the inclusion of lag 2 terms. The difference between these two analyses

is that the Granger causations between: the SOI and monthly rainfall; the quadratic terms of the climate indicators and monthly rainfall; and lag 2 terms of the climate indicators and monthly rainfall, are equivocal in each individual weather station analysis.

The Granger causal relationship between the IOD and monthly rainfall is potentially useful for agribusiness. Rainfall in the next month can be forecast from rainfall in the current month, the IOD in the current month and seasonal variation. Forecasts with a longer lead time can be based on extrapolation of the IOD. The inclusion of the SOI and the PDO in the forecasting models could offer some small improvement in forecasts.

Chapter 4


Multi-scale analysis of CO₂

Given the importance of atmospheric level of CO₂, and also the availability of CO₂ data from ice core, following an article published in Nature, it is worth to fit time series models to CO₂ series. The main objective of this study is to look for consistent features of time series models, at different sampling intervals, from 0.5 million-year up to one week. This has been written up in the following paper.

Statement of Authorship

Title of Paper	Multi-scale CO ₂
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	This paper was submitted to Environmental Modelling & Assessment on 26 th December 2020, and currently under review.

Principal Author

Name of Principal Author (Candidate)	Xiaomeng Gu		
Contribution to the Paper	Xiaomeng Gu collected from online open sources and conducted the research under the supervision of Andrew V. Metcalfe. Xiaomeng Gu wrote the manuscript in consultation with Andrew V. Metcalfe and Gary F. V. Glonek.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	23/2/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Andrew V. Metcalfe		
Contribution to the Paper	Andrew V. Metcalfe conceived of the presented idea, and provided critical feedback and helped shaped the research, analysis and manuscript.		
Signature		Date	17 March 2021

Name of Co-Author	Gary F. V. Glonek		
Contribution to the Paper	Gary F. V. Glonek provided critical feedback and helped shape the research, analysis and manuscript.		
Signature		Date	15 April 2020

Please cut and paste additional co-author panels here as required.

Multi-scale CO₂

X. Gu^{*1}, A. V. Metcalfe^{†1}, and G. F. V. Glonek^{‡1}

¹School of Mathematical Sciences, University of Adelaide, SA 5005,
Australia

Abstract

Five time series of estimated atmospheric CO₂ with sampling intervals ranging from 0.5 million years to the relatively high frequency of one week are analysed. The yearly series shows a clear increasing trend since the beginning of the first Industrial Revolution around 1760. The weekly series shows a clear increasing trend and also seasonal variation. In both cases, the trend is fitted by a conceptual model that consists of a baseline value with an exponential trend superimposed. For the weekly series, the seasonal variation is modelled as an exponential of a sum of sine and cosine terms. The deviations from these deterministic models are treated as detrended and deseasonalised time series. Then, three sub-categories of autoregressive integrated moving average (ARIMA) models are fitted to the five time series: ARMA models which are stationary; FARIMA models which are stationary but have long memory and are fractal processes, and ARIMA models which are variations on a random walk and so non-stationary in the variance. The FARIMA and ARIMA models provide better fits to the data than the corresponding ARMA models. All the fitted models are close to the boundary of stability, and are consistent with claims that climate change due to an increase in atmospheric CO₂ may not quickly be reversed even if CO₂ emissions are stopped.

1 Introduction

One of the major challenges confronting society today is the increase of carbon dioxide emissions, which are considered the major cause of global warming. Global warming is one

*xiaomeng.gu@adelaide.edu.au

†andrew.metcalfe@adelaide.edu.au

‡gary.glonek@adelaide.edu.au

of the most severe environmental, social and economic threats that the world has faced in the recent century, and it is potentially the most difficult to avoid (Shirmohammadi et al. 2018). Global warming not only stresses ecosystems through various mechanisms, but also impacts other areas which can lead to positive feedback. For example, Australia's bushfires, intensified through global warming, are contributing to one of the biggest increases in the concentration of atmospheric CO₂ in the past 60 years (Green 2020).

Solomon et al. (2009) claim that the climate change that takes place due to increases in carbon dioxide concentration is largely irreversible for 1,000 years after emissions stop. Cessation of emission of carbon dioxide will decrease radiative forcing, but this decrease is largely offset by slower loss of heat to the ocean. A consequence is that global atmospheric temperatures are unlikely to drop significantly for at least 1,000 years.

We fit time series models to records of atmospheric CO₂ at very different time scales, starting with a record based on pedogenic carbonate assessments extending back 420 million years, with a sampling interval of 0.5 million years (Foster et al. 2017). At an intermediate level we consider ice core records extending back around a million years with sampling intervals of the order of 1,000 years. These are compared with relatively high frequency data from the past 2,000 and last 50 years sampled at yearly, and weekly intervals respectively. The objective is to compare the time series models and look for unifying features, with particular emphasis on stationarity. The results are discussed in the context of Solomon et al's (2009) claim that increases in carbon dioxide are largely irreversible.

2 Data

The main objective of this study is to compare time series models that are fitted to estimates of atmospheric CO₂ made over quite different sampling intervals, and in the case of weekly data direct measurements of CO₂. The time intervals are: 0.5 million years; around one thousand years, yearly; and weekly (Table 1).

Table 1: Five CO₂ time series – source, sampling interval, and length.

Source	Length	start date – end date	Sampling Interval	Number of observations
Compiled Data (Foster et al., 2017)	420 million years	420 million years ago - nominally up to geological present	0.5 million years	840
Vostok Ice Core (Petit et al. 2001, Siegenthaler et al. 2005)	~ 440 thousand years	438,986 yr Bp – 2,690 yr Bp in EDC3 chronology	Unequally spaced	372
Pre-Industry Revolution (Law Dome Ice Core (Macfarling Meure et al. 2006))	1,760 years	1 yr AD – 1760 yr AD	yearly	1,760
Post-Industry Revolution (Law Dome Ice Core (Macfarling Meure et al. 2006))	244 years	1761 yr AD – 2004 yr AD	yearly	244
Mauna Loa Atmospheric Measurement (McGee 2020)	46 years	May 1974 – February 2020	weekly	2,389

Foster et al. (2017) compiled the dataset using five independent techniques applied to data from 112 published studies covering the last 420 million years. In order to ensure the highest-quality compilation, they applied a set of criteria that left 1,241 observations to be analysed. Monte Carlo resampling and a LOESS fit were used to interpolate the series to a regular 0.5 million years interval. The data is available from the supplementary information from <https://www.nature.com/articles/ncomms14845>, with DOI: 10.1038/ncomms14845.

The Vostok ice core dataset was obtained from the National Oceanic and Atmospheric Administration (NOAA) on 10th of March 2020. It can be accessed via <https://www.ncdc.noaa.gov/paleo-search/study/17975>. This original publication is “Antarctic Ice Cores Revised 800KYr CO₂ Data”, with data set ID noaa-icecore-17975.

The original publication of Law Dome dataset is “Law Dome Ice Core 2000-Year CO₂, CH₄, and N₂O Data”, which was contributed by David Etheridge (Original receipt by WDC Paleo) and last updated on July 2010. This dataset was acquired from the National Oceanic and Atmospheric Administration (NOAA) on 25th of February 2020, it can be accessed via <https://www.ncdc.noaa.gov/paleo-search/study/9959>, with data set ID noaa-icecore-9959. This dataset has been split into two time intervals, 1 yr AD to 1760 yr AD, and 1761 yr AD to 2004 yr AD. This is because the dramatic increase due to the first Industrial Revolution, which started in 1760.

The weekly time series was acquired from the National Oceanic and Atmospheric Administration (NOAA) on 6th of March 2020, and it can be accessed via <https://www.co2.earth/co2-datasets>. These data are weekly carbon dioxide observations, measured as mole fraction in dry air, from the Mauna Loa Observatory in Hawaii since May 1974.

3 Method

3.1 Filtering deterministic models

There is a clear increasing trend in CO₂, that appears to be approximately exponential, since the beginning of the first Industrial Revolution, around 1760. The trend is estimated with a conceptual model that consists of an exponential increase in CO₂ superimposed on some baseline value (a),

$$y_t = a + be^{ct}. \quad (1)$$

The model is fitted by non-linear least squares, using the *nls* function in R, and the residuals are considered to be the detrended time series. There is also a marked seasonal pattern in the time series of weekly data from Mauna Loa. The seasonal pattern is modelled by modulating the trend with an exponential of a linear sum of cosine and sine functions,

$$y_t = (a + be^{ct}) \times (e^{\beta_0 + \beta_1 C_1 + \beta_2 S_1 + \dots + \beta_{2M-1} C_M + \beta_{2M} S_M}), \quad (2)$$

where $C_m = \cos(m2\pi t/p)$, $S_m = \sin(m2\pi t/p)$ for $m = 1, \dots, M$, The value of p is equal to the seasonal period, 365.25/7 in the case of weekly data, and M is selected as the smallest value such that neither of the coefficients of C_{M+1} nor S_{M+1} is statistically significant at a 0.05 level. The residuals (observations less fitted values) from the deterministic models are considered to be detrended and deseasonalised time series. There is no evidence of

any deterministic trend in the time series that pre-date the first Industrial Revolution, so they are assumed to be realisations of models that are stationary in the mean.

We then calculated autocorrelations, partial autocorrelations, and spectra, and applied various unit root tests that aim to detect non-stationarity in variance before fitting ARIMA and FARIMA models. We review these methods briefly in the following sections.

3.2 Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) (Akaike 1974) is the most commonly used criterion for model selection when maximum likelihood is used to estimate parameters. The model with a lowest AIC is considered as the best fitting model. However, AIC tends to favour models with quite large numbers of parameters, and we are looking for a succinct model to emulate the dynamics of the processes. Burnham & Anderson (2004) suggest that a simpler model may be preferable to a model with more parameters if the simpler model has an associated increase in AIC that does not exceed two. But, this advice can be rather restrictive if applied to long time series because a small reduction, of no practical significance, in the estimated standard deviation of the errors can be statistically significant. In general, a sensible choice of model, or comparison of models, depends on the context, and it is useful to consider both AIC and the estimated standard deviation of the errors (based on the unbiased estimator of the variance of the errors).

3.3 ARIMA model

An AR(p) model is a linear regression of the current value of a variable, $\{y_t\}$ say, on p past values. This can be generalised to an ARMA(p, q) model which includes a linear combination of q past errors. An ARMA(p, q) model is strictly stationary, and hence stable, if all the roots of the polynomial

$$(1 - \phi_1 z - \dots - \phi_p z^p)$$

lie outside the unit circle, where z is a complex variable. In particular AR(1) is stationary for $|\phi_1| < 1$, and AR(2) is stationary for $|\phi_2| < 1$ and $\phi_2 \pm \phi_1 < 1$. The ARIMA(p, d, q) model allows for a non-stationary series to be differenced d times before fitting an ARMA(p, q) model. So, the general definition is as follows:

$$(1 - \phi_1 \mathbf{B} - \dots - \phi_p \mathbf{B}^p)(1 - \mathbf{B})^d y_t = (1 + \theta_1 \mathbf{B} + \dots + \theta_q \mathbf{B}^q) \epsilon_t,$$

where \mathbf{B} is the backward shift operator, defined by $\mathbf{B}y_t = y_{t-1}$.

Typically a range of ARIMA models can provide plausible fits to data from a dynamic system. The AIC is one option for choosing between them, but it is entirely empirical and tends to favour models with a large number of parameters. The AIC may be appropriate for forecasting applications, but we prefer to focus on models with some conceptual basis for modelling CO₂. The differencing could be used to remove a deterministic trend, but in the context of a process that is stationary in the mean a d of 1 allows for an unstable process. The crucial feature of an unstable process is that it does not return to some fixed level, a well known example being the random walk. In mathematical terms, the defining characteristic of an unstable process is that the polynomial in \mathbf{B} acting on y_t has a unit root. Unless strongly indicated by model fit, we prefer AR(1) and AR(2) models, without moving average terms, because they have a conceptual interpretation in terms of elementary linear dynamic systems. The AR(2) process is a discrete time version of the second order differential equation that describes a linear oscillator subject to a sequence of random impulses. The derivatives are approximated by differences. The canonical example of a linear oscillator is a mass suspended by a spring with some damping arrangement attached (e.g. Thomson 1993, *Theory of Vibration With Applications* (4e), Prentice Hall)

3.4 FARIMA model

Some time series exhibit marked correlations at high lags, and they are referred to as long-memory processes (Beran et al. 1998). The time series is considered to have long memory if the acf decays more slowly than an exponential decay. This slower decay is usually modelled as a power-like decay (Joshi 2016). The FARIMA model allows for fractional differencing, generalising the ARIMA model's integer order of differencing to allow the d parameter to take on fractional values, $-0.5 < d < 0.5$ (Baum 2013). The range $0 < d < 0.5$ gives long memory processes, and it lies between a stationary AR(1) model and a non-stationary random walk (Beran et al. 1998). We are interested in this model because it is close to being a random walk yet still stays bounded, and returns to a mean value if there are no error inputs. Formally, the $(1 - \mathbf{B})^d$ term in the ARIMA model is expanded with the generalised binomial expansion:

$$1 - d\mathbf{B} + \frac{d(d-1)}{2!}\mathbf{B}^2 - \frac{d(d-1)(d-2)}{3!}\mathbf{B}^3 + \dots,$$

as far as \mathbf{B}^L , where L is some cut-off point typically around 30.

3.5 Unit root tests

3.5.1 Random Walk

The ARIMA(0,1,0) model is a random walk. It is generally defined as follows:

$$x_t = x_{t-1} + w_t,$$

where w_t is discrete white noise with mean 0 and variance σ_w^2 .

Using the backward shift operator \mathbf{B} , we can rewrite the defining equation as

$$\begin{aligned}(1 - \mathbf{B})x_t &= w_t \\ x_t &= (1 - \mathbf{B})^{-1}w_t \\ x_t &= (1 + \mathbf{B} + \mathbf{B}^2 + \dots)w_t \\ x_t &= w_t + w_{t-1} + w_{t-2} + \dots\end{aligned}$$

It is formally stationary in the mean, but non-stationary in the variance and so unbounded. The current value is the best predictor of future values and the model has no tendency to return to the initial value.

3.5.2 Augmented Dickey-Fuller (ADF) test

The model for the Augmented Dickey-Fuller test is as follows:

$$\Delta y_t = \beta y_{t-1} + \sum_{k=1}^K \gamma_k \Delta y_{t-k} + \epsilon_t,$$

where $\Delta = 1 - \mathbf{B}$ is the first difference operator, and in the R implementation K is defined as $\text{trunc}((\text{length}(y) - 1)^{(1/3)})$. The summation term in the equation above is to correct for the presence of serial correlation.

The null hypothesis of the Augmented Dickey-Fuller test is the presence of unit root, that is $H_0: \beta = 0$ against the alternative which is $H_a: \beta < 0$. Under the null hypothesis, y_t itself is non-stationary. Hence the ordinary central limit theorem does not apply to the least squares estimator, $\hat{\beta}$, of β , and its sampling distribution does not tend to a t-distribution as the length of the time series increases. Dickey and Fuller have tabulated the asymptotic distribution of the least squares estimator for β under the null hypothesis of it being a unit root. It turns out that we can compare a t-statistic with the values of this Dickey-Fuller distribution. If the t-statistic less than the critical value, then we reject the null hypothesis (EViews 2019).

3.5.3 Phillips-Perron (PP) test

The model is as follows:

$$y_t = c + \delta t + \alpha y_{t-1} + \epsilon_t.$$

The null hypothesis of this test is that $\alpha = 1$, against the alternative which is $\alpha < 1$. They used a corrected form of t-test in order to correct for the presence of serial correlation and heteroscedasticity in the error term (Zivot & Wang 2007). Again, there are two functions in R that can perform the PP test: *pp.test* and *ur.pp*.

3.5.4 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests

Let $\{y_t\}$ be the observed time series for which we wish to test stationarity. We assume that we can decompose the series into the sum of a deterministic trend, a random walk, and a stationary error (Kwiatkowski et al. 1992):

$$y_t = \zeta t + \gamma_t + \epsilon_t.$$

Here γ_t is a random walk:

$$\gamma_t = \gamma_{t-1} + u_t,$$

where the u_t are independent and identically distributed $N(0, \sigma_u^2)$. The initial value γ_0 is treated as fixed and serves the role of an intercept. The stationarity hypothesis is simply $\sigma_u^2 = 0$. Since ϵ_t is assumed to be stationary, under the null hypothesis $\{y_t\}$ is trend-stationary (Tang et al., 2019). The statistic they used is both the one-sided Lagrange Multiplier (LM) statistic and the Local Best Invariant (LBI) test statistic for the hypothesis $\sigma_u^2 = 0$, under the stronger assumptions that the u_t are normal and that the ϵ_t are independent and identically distributed $N(0, \sigma_\epsilon^2)$. Because the parameter value specified by the null hypothesis is on the boundary of the parameter space, they were interested in a one-sided LM test rather than a two-sided test (Kwiatkowski et al. 1992).

Let ϵ_t , $t = 1, 2, \dots, T$, be the residuals from the regression of y on an intercept and time trend. Let $\hat{\sigma}_\epsilon^2$ be the estimate of the error variance from this regression (the sum of squared residuals, divided by T). Define the partial sum process of the residuals:

$$S_t = \sum_{i=1}^t \epsilon_i, \quad t = 1, 2, \dots, T.$$

Then the LM (and LBI) statistic is

$$\text{LM} = \sum_{t=1}^T \frac{S_t^2}{\hat{\sigma}_\epsilon^2},$$

and high values are taken as evidence against the null hypothesis. The function `kpss.test` and `ur.kpss` in R can be used.

In summary, the ADF and PP test a null hypothesis of a unit root, which is non-stationary in variance. Conversely, the KPSS test a null hypothesis of stationarity.

4 Analysis

We fit $AR(p)$, $ARIMA(p, d, q)$ and $FARIMA(p, d, q)$ where both p and q are no greater than three, to all five time series. These five time series include million-year series; thousand-year series; pre Industrial Revolution yearly series; detrended post Industrial Revolution yearly series; and detrended and deseasonalised weekly series. To eliminate or reduce the trend and seasonal patterns, we fit exponential trends above a threshold to the time series after the Industrial Revolution, and a seasonal model to the weekly series, prior to model fitting. The residuals, defined as observations less fitted values, are considered to be a realisation of a process that is stationary in the mean.

4.1 Million-year time scales

The time series (Foster et al. 2017) is based on pedogenic carbonate assessments extending back 420 million years at 0.5 million years interval (Figure 1).

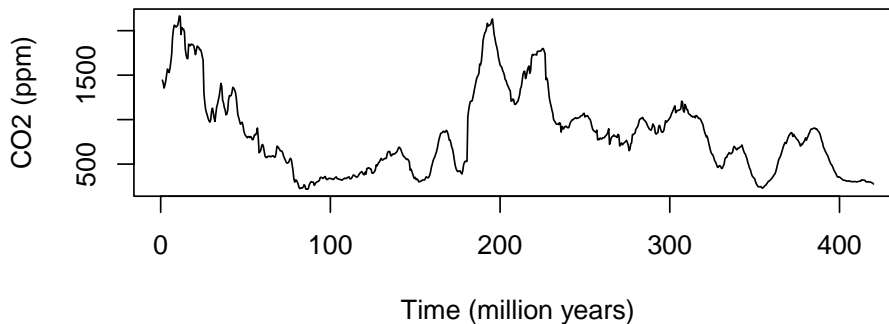
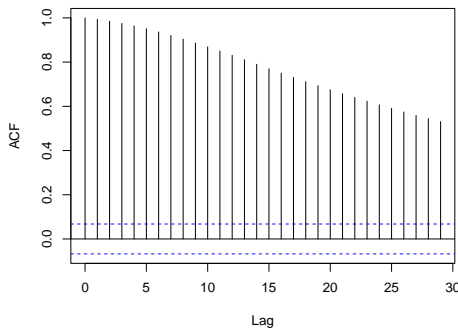


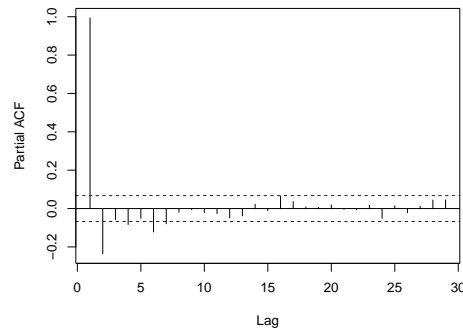
Figure 1: CO₂ from 420 million years ago until present at 0.5 million years interval.

At the 0.5 million year scale there are two peak values above 2000 ppm, around 400 and 200 million years ago. Current levels are around 400 ppm, but this is above the minimum value of 220 ppm that occurred around 300 million years ago. For comparison, the values around 300 and 80 million years ago are close to the value before the first Industrial Revolution (around 1760s), which is 270 ppm. The correlogram (Figure 2a) shows the series is highly auto-correlated, and the pacf (Figure 2b) shows the significant lags at 1

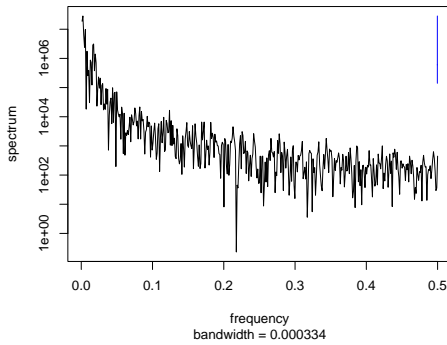
and 2. These suggest the series can be modelled reasonably as AR(2). But given the dominant autocorrelation at lag 1, an AR(1) model, with a coefficient close to 1, may be a reasonable approximation. Both periodogram and smoothed spectrum show that the time series is a realisation of a low frequency process. If the unit root tests are applied, neither ADF nor PP provide evidence against a unit root at the 0.05 level of significance and KPSS provides evidence against a null hypothesis of stationarity. Hence, all of those three tests are consistent with a hypothesis of a unit root. Furthermore, fitting a simple AR(1) model returns a coefficient of 0.9939, which is very close to 1. However, a random walk model is somewhat unsatisfactory for an environmental process that is physically bounded, and a stationary fractional differenced model is conceptionally preferable.



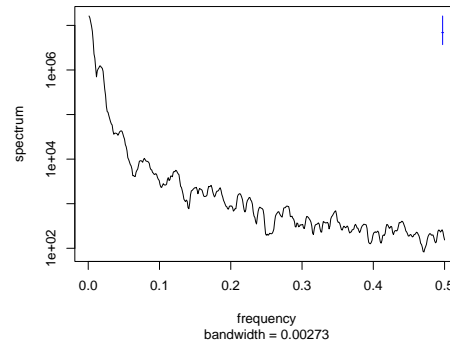
(a) Correlogram.



(b) Partial correlogram.



(c) Periodogram.

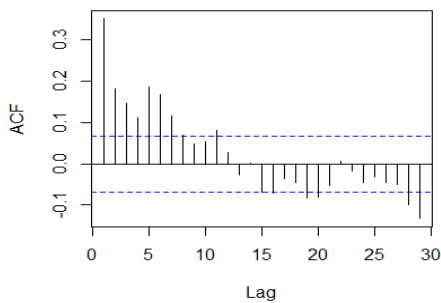


(d) Smoothed spectrum with span = 9.

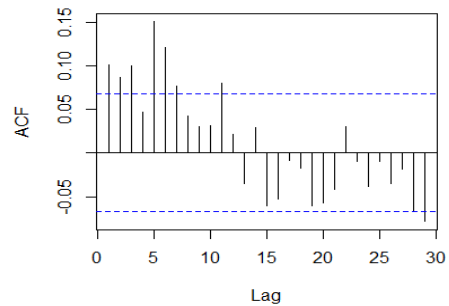
Figure 2: 420 million-year time series.

The results of fitting models are shown in [Table 2](#). The residuals from AR(1) and AR(2) models still show significant autocorrelations ([Figure 3a](#), [Figure 3b](#)). Based on Akaike information criterion, the ARIMA(3,1,2) would be chosen as best fitting model. A con-

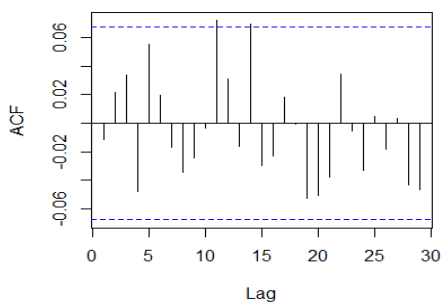
sequence of the differencing parameter, $d=1$, is that the process is unstable and so unbounded. Instability within the range of CO_2 in the time series record is quite plausible, and thresholds could be introduced to give a bounded non-linear variant to the model. The FARIMA models provide a better fit than the $\text{AR}(p)$ models, and as they are equivalent to a high order AR model they are also stationary. Even though $\text{FARIMA}(3, 0.330, 0)$ gives a lower AIC, as well as lower residual variance than $\text{FARIMA}(2, 0.241, 0)$, the third autoregressive term is non-significant. The third autoregressive term can be attributed to approximating derivatives with centred differences, and both models allow a reasonable physically interpretation of the process as a damped oscillator, which is modelled in terms of displacement and its first and second derivatives, with random inputs. The interpretation is of a borderline stable second order dynamic system with long memory.



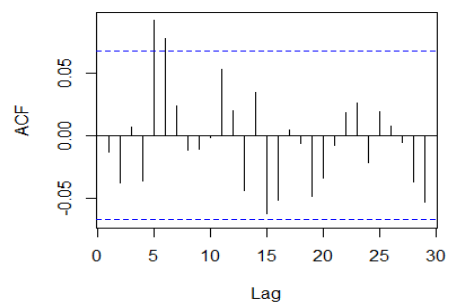
(a) AR(1) model.



(b) AR(2) model.



(c) ARIMA(3,1,2) model.



(d) FARIMA(2, 0.241, 0) model.

Figure 3: Correlograms of the residuals of the fitted models.

Although the difference in AICs between $\text{FARIMA}(2, 0.241, 0)$ and $\text{ARIMA}(3, 1, 2)$ is nearly 20, this can be attributed to the long time series, which allows identification of

Table 2: Estimated coefficients and their associated standard errors of AR, ARIMA, FARIMA models for the 420 million years data.

ARMA (p, q)	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	AIC	σ
(1,0)	0	0.996 (0.003)	—	—	—	—	8704.240	42.763
(2,0)	0	1.345 (0.032)	-0.350 (0.032)	—	—	—	8596.320	40.049
(1,0)	1	0.348 (0.032)	—	—	—	—	8580.240	40.125
(2,0)	1	0.325 (0.034)	0.066 (0.034)	—	—	—	8578.600	40.037
(1,0)	0.241 (0.024)	0.974 (0.008)	—	—	—	—	8598.784	40.270
(2,0)	0.241 (0.028)	1.065 (0.043)	-0.090 (0.041)	—	—	—	8575.806	39.675
(2,2)	0	1.886 (0.037)	-0.889 (0.036)	—	-0.588 (0.051)	-0.144 (0.038)	8576.080	39.522
(3,2)	1	2.093 (0.057)	-1.476 (0.115)	0.309 (0.050)	-1.791 (0.040)	0.934 (0.057)	8556.900	39.369

statistically significant terms which do not correspond to a substantial reduction in the variance of the residuals. The corresponding estimated standard deviations of the errors are 39.68 and 39.37 respectively.

The coefficients of the ARMA(2,2) are very close to the boundary of the stability region, that is $|\phi_2| = 0.889$, and $\phi_2 + \phi_1 = 0.997$. In contrast, if an unstable ARIMA(2,1,0) is fitted then the coefficients of the model for differences are far removed from the boundary of stability.

4.2 Thousand-year time scales

The Vostok CO₂ dataset has differing time intervals between observations (Figure 4) and the time scale is from past to present.

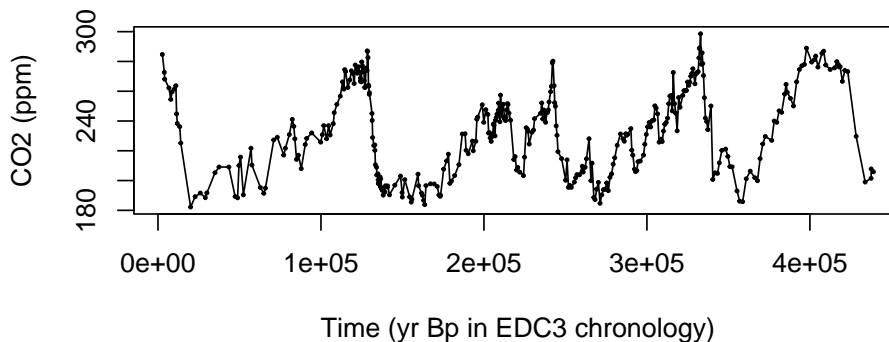


Figure 4: Time series plot of Vostok data from approximately 440,000 years ago until 2,690 years ago with unequally spaced time interval.

The usual approach to analysing time series with unequal time intervals is to interpolate to obtain equal time spacing. We considered linear interpolation and cubic spline interpolation to a time interval of 1,176 years corresponding to 372 points, which exactly equals the number of points in the original time series. The variogram describes the covariance structure over space, or time, as a function of distance h between points, and is defined as

$$2\gamma(h) = E[(Z(u) - Z(u + h))^2] \quad (3)$$

for all locations u , where $Z(s)$ is the value of the variable of interest at location s .

Estimation of the variogram does not assume equal spacing, so variograms of the original time series, the linear interpolated series, and the spline interpolated series are compared

in Figure 5. There is reasonable agreement between the original time series and the interpolated series. The difference between the variograms for the interpolated series is not statistically significant. The linear interpolation is slightly closer to the original time series overall, although the spline interpolation is closer at the beginning and the end.

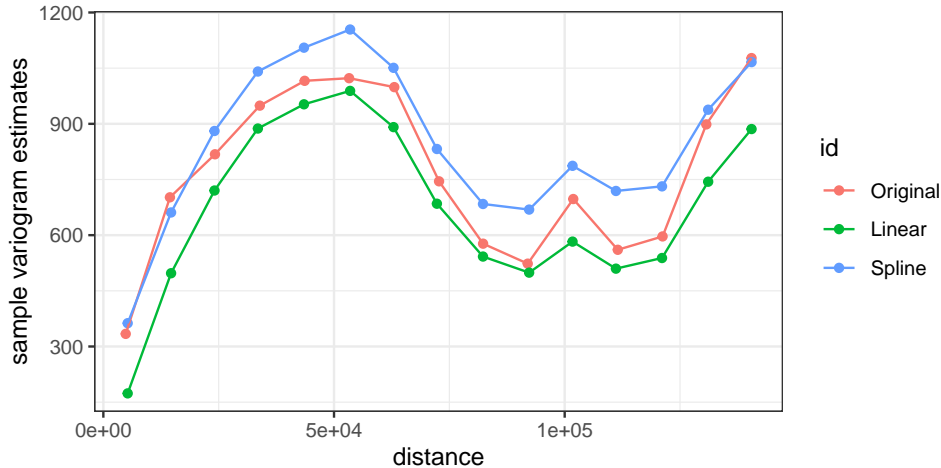


Figure 5: Variograms for raw data, linear interpolated series and series using spline interpolation.

As a further check of the interpolations we investigated frequency compositions. For all three time series we fitted regressions with cosine and sine terms with periods at multiples of one cycle per record length. That is regressions with $2M$ predictor terms of the form $\cos(m2\pi t/n)$ and $\sin(m2\pi t/n)$ for m from 1 up to M , and plotted the coefficients of determination (R-square) against M . The result is shown in Figure 6. The spline interpolation increases slowly relative to the other two series. The interpolated series have identical plots as we increase the number of harmonic terms, the cumulative periodograms, and are quite close to the regression results for the original series (not a precise periodogram because the orthogonality property is lost if the time intervals vary).

The conclusion is that both interpolation methods gave evenly spaced time series with a similar auto-correlation structure to the original time series. We chose to continue with the linear interpolation, because it is simpler and avoids a few unrealistic values at the beginning of the series where there are some relatively long time intervals between observations.

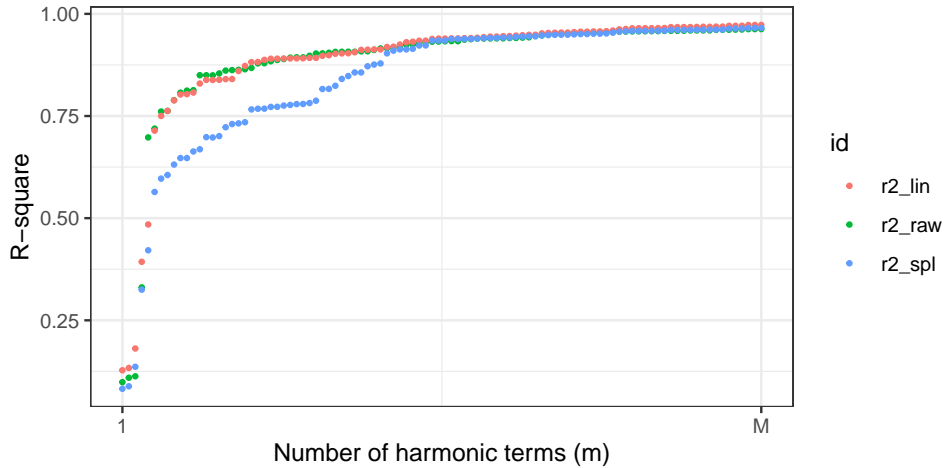


Figure 6: Values of R-square for raw data, linear interpolated series and series using cubic spline interpolation, with $M = 100$.

4.2.1 Linear interpolated series

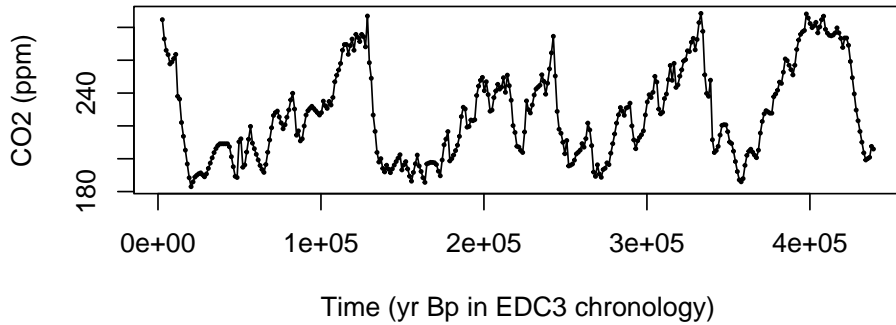
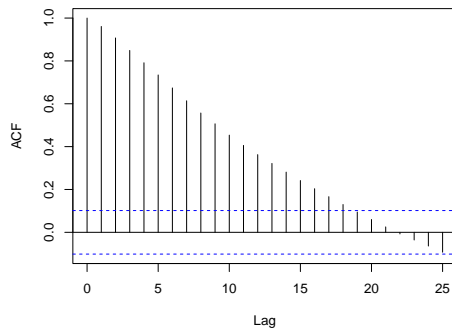
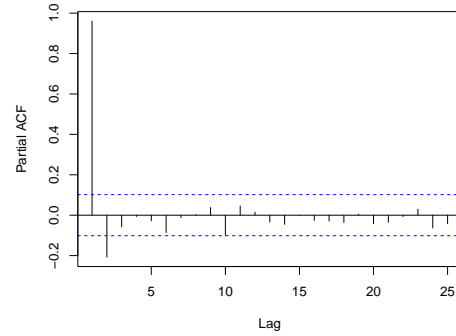


Figure 7: Time series plot for linear interpolated thousand-year series, from 440,000 yr Bp to 2,690 yr Bp.

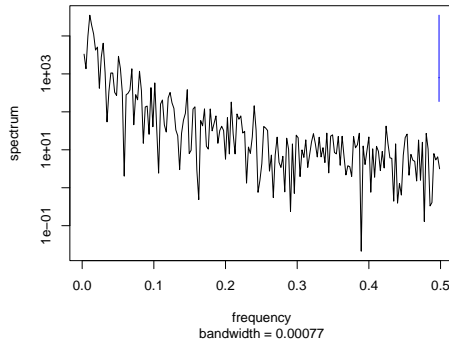
The correlogram (Figure 8a) shows a slow decay which might be compatible with a non-stationary time series model. Whilst non-stationary models might be considered physically unrealistic for the process over 420 million years, they can provide reasonable local approximation over sub-periods of, here, four hundred thousand years. The partial correlogram (Figure 8b) shows significant lags at lag 1 and lag 2, which suggests the time series model has two autoregressive terms. The spectrum indicates a low frequency process which is indicative of a process with a unit root or a fractionally differenced process.



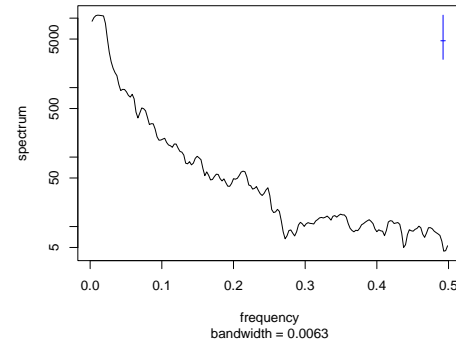
(a) Correlogram.



(b) Partial correlogram.



(c) Periodogram.



(d) Smoothed spectrum with span = 9.

Figure 8: Linear interpolated thousand-year time series.

The results of fitting various models to the carbon dioxide time series with an interval of 1,176 years are shown in [Table 3](#) below. The FARIMA(2, 0.064, 0) process is not statistically significantly worse than the ARIMA(3, 1, 2), and in practical terms they provide equally good fits (Burnham & Anderson 2004). The FARIMA model is preferred because it is simpler and it is stationary. It is also consistent with the FARIMA model fitted to the 420 million years series and a fractal process.

We also investigated directionality in the thousand-year time series (Mansor et al. 2020), especially whether there is a difference between the time from a threshold to a peak and the time from the peak back to the threshold. The threshold was taken as the 0.80 quantile of the marginal distribution of the time series. A peak was defined as the maximum value of the time series over the period from an upwards crossing of the threshold to the next downwards crossing. The statistic is calculated from all peaks for which there are four time steps following the up crossing and before the down crossing. The mean of the

Table 3: Estimated coefficients and their associated standard errors of AR, ARIMA, FARIMA models for the thousand-year data.

ARMA (p, q)	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	AIC	σ
(1,0)	0	0.971 (0.013)	-	-	-	-	2520.770	7.099
(2,0)	0	1.240 (0.050)	-0.279 (0.050)	-	-	-	2493.030	6.820
(1,0)	1	0.259 (0.050)	-	-	-	-	2488.590	6.904
(2,0)	1	0.253 (0.052)	0.021 (0.052)	-	-	-	2490.430	6.903
(1,0)	0.219 (0.017)	0.903 (0.022)	-	-	-	-	2499.382	6.903
(2,0)	0.064 (0.014)	1.163 (0.052)	-0.218 (0.052)	-	-	-	2488.849	6.791
(2,2)	0	1.951 (0.021)	-0.956 (0.020)	-	-0.760 (0.052)	-0.198 (0.050)	2489.510	6.747
(3,2)	1	-0.065 (0.052)	-0.872 (0.025)	0.224 (0.051)	0.327 (0.009)	1.000 (0.016)	2488.380	6.793

variable over the four time steps following the peak is subtracted from the mean over the four time steps before the peak. These differences are assumed to be independent and are analysed by a one-sample t-test for paired comparisons. The value of the test statistic is -2.07. So there is evidence that the rise to a peak is more rapid than the decline following it. This provides some support for Solomon et al's (2009) claim that climate change due to an increase in atmospheric CO₂ may not quickly reversed even if CO₂ emissions are ceased.

4.3 Yearly ice core data

The concentrations of CO₂ were inferred by using spline fits to the Law Dome firn (Meure et al. 2006), ice core records and the Cape Grim record (Figure 9).

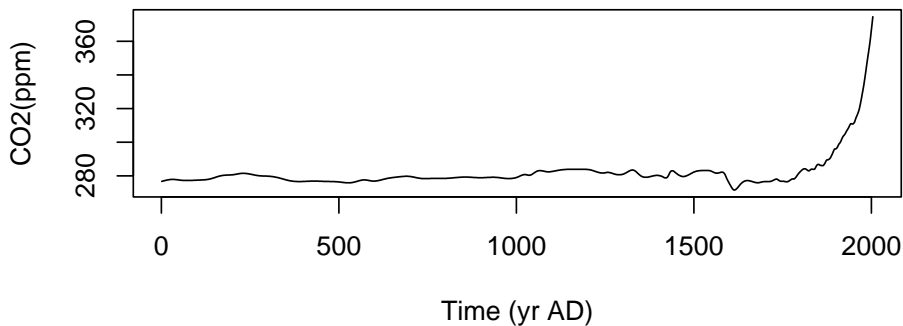
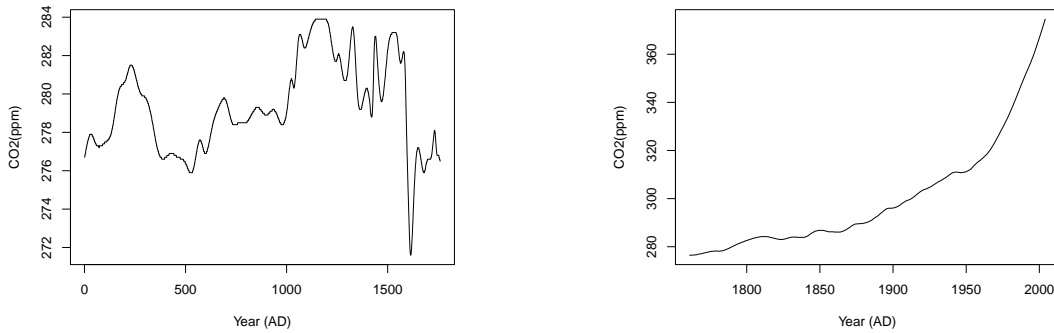


Figure 9: Time series plot of ice core data from 1 yr AD to 2004 yr AD at one year interval.

There is a dramatic increase from around 1800 yr AD, that can be attributed to the Industrial Revolution. We analyse the series in two parts: before the Industrial Revolution (1 yr AD – 1760 yr AD) and after the Industrial Revolution (1761 yr AD – 2004 yr AD). There is noticeable variation in the pre-Industrial Revolution series (Figure 10a), but the range for CO₂ is relatively small. In contrast, there seems to be an exponential increase for the series after 1760s (Figure 10b).



(a) Pre-Industrial Revolution series.

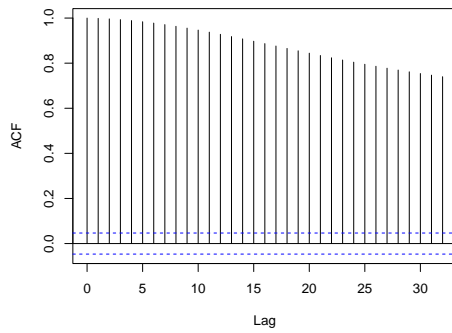
(b) Post-Industrial Revolution series.

Figure 10: Time series plots of the yearly series, from 1 yr AD to 2004 yr AD.

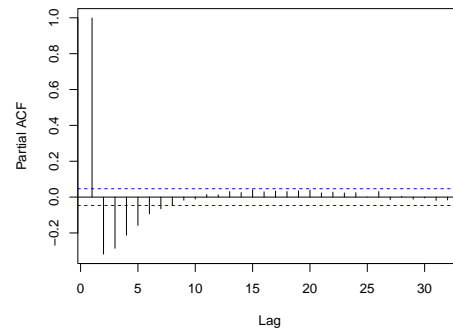
The ADF, PP and KPSS tests in R are applied to the pre-Industrial Revolution series only, as there is a clear trend in the post-Industrial Revolution series. There is no evidence to reject the null hypothesis that there is a unit root at the 0.05 significance level, based on the ADF and PP tests. Furthermore, the KPSS test shows evidence to reject the null hypothesis of the stationarity of the series.

4.3.1 Pre-Industrial Revolution (IR) period

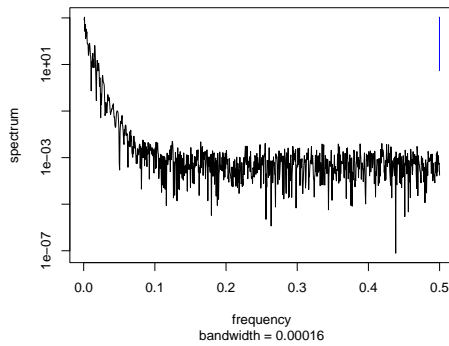
The correlogram decays very slowly down from unity, which is one of the characteristics of a random walk. The partial correlogram shows a positive partial autocorrelation at lag 1, and negative autocorrelations from lag 2 up to lag 5, and this might suggest a moving average term. The spectrum shows the property of low frequency and no dominant peak.



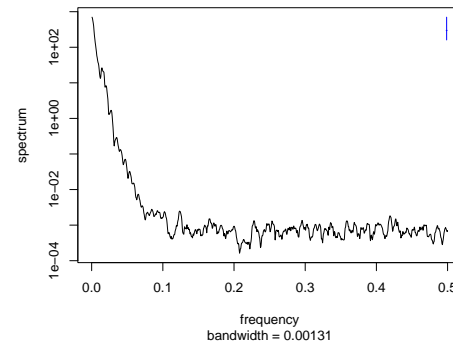
(a) Correlogram.



(b) Partial correlogram.



(c) Periodogram.



(d) Smoothed spectrum with span= 9.

Figure 11: Pre-IR yearly time series.

Table 4 shows the results for different models. There are no standard errors for the coefficients of FARIMA(1, 0.5, 0) because R is unable to compute correlation matrices for those two models. The autoregressive coefficient returned by AR(1) model suggest our series resembles a realisation of a random walk, this conclusion is consistent with the differencing operator d returned by the FARIMA(2, 0.499, 0). The ARIMA(3,1,2) again is the model that has the lowest AIC value, and although the unbounded property, a consequence of its being unstable, is not physically reasonable over a long period, it is an adequate approximation over a relatively short period. The FARIMA(2, 0.499, 0) is a succinct and more physically realistic model.

The appearance of stochastic trends, that are modelled by a random walk, may be accounted for as segments of a relatively low frequency oscillation in the underlying continuous process. The FARIMA($p, d, 0$) models are consistent with the underlying continuous process being fractal.

Table 4: Estimated coefficients and their associated standard errors of AR, ARIMA, FARIMA models for the pre-IR yearly data.

ARMA (p, q)	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	AIC	σ
(1,0)	0	0.999 (0.001)	-	-	-	-	-3621.480	0.086
(2,0)	0	1.695 (0.017)	-0.697 (0.017)	-	-	-	-4786.530	0.062
(1,0)	1	0.6956 (0.017)	-	-	-	-	-4792.300	0.062
(2,0)	1	0.296 (0.020)	0.574 (0.020)	-	-	-	-5492.000	0.051
(1,0)	0.5	0.985	-	-	-	-	-5139.295	0.056
(2,0)	0.499 (0.001)	0.703 (0.020)	0.287 (0.020)	-	-	-	-5300.493	0.053
(2,2)	0	1.958 (0.007)	0.959 (0.007)	-	-0.982 (0.024)	0.475 (0.021)	-5834.61	0.046
(3,2)	1	1.476 (0.032)	-0.237 (0.050)	-0.267 (0.030)	-1.481 (0.026)	0.683 (0.021)	-5902.590	0.045

4.3.2 Post-Industrial Revolution (IR) period

There is a clear exponential growth in atmospheric carbon dioxide concentration after the first Industrial Revolution, and this is fitted using the conceptual model of Equation 1. The detrended series, \tilde{y}_t , are defined by

$$\tilde{y}_t = y_t - 279.10 - 10.92 \times e^{0.02t} \quad (4)$$

A time series plot of the residuals which are the detrended series is shown in Figure 12.

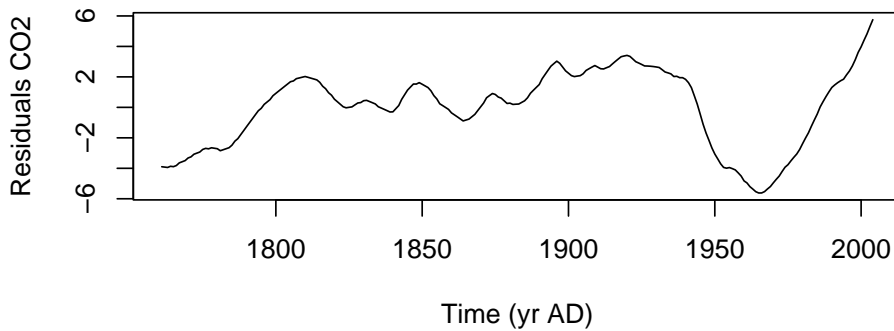
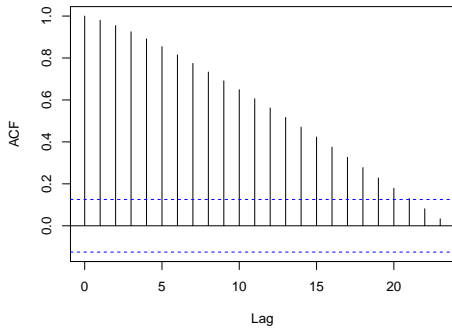


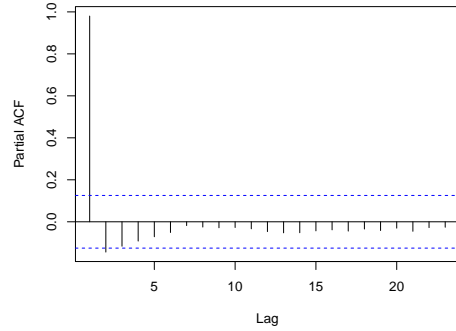
Figure 12: Time series plot of the detrended series for the post-IR yearly data.

Even after removing a trend, the ADF and PP unit root tests do not provide evidence to reject the null hypothesis of a unit root at the 0.05 significance level. Furthermore, the KPSS test provides evidence to reject a null hypothesis of stationarity at the 0.1 significance level.

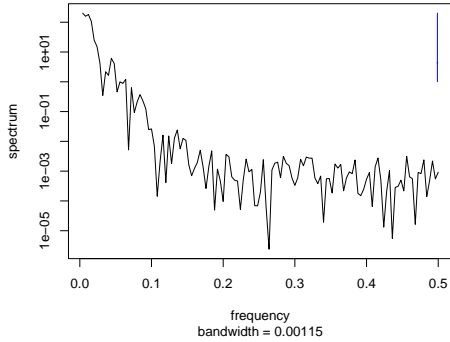
The correlogram shows a linear decay which also suggests non-stationarity (Figure 13a). The partial correlogram is significant at lag 1, suggesting an AR(1) model. Based on the results of three unit root tests, we might conclude that the series is a realisation of a random walk. Also, the estimated coefficient returned by the AR(1) model is 0.9988, which is very close to 1 and consistent with the unit root tests. Both the raw periodogram and smoothed spectrum show the dominance of low frequency variation, again consistent with a random walk.



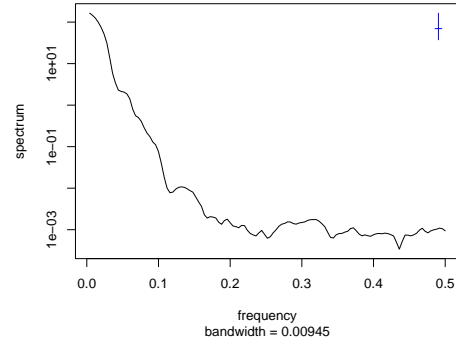
(a) Correlogram.



(b) Partial correlogram.



(c) Periodogram.



(d) Smoothed spectrum with span= 9.

Figure 13: Detrended post-IR yearly data.

Table 5 gives the results of fitting models for the post-IR yearly time series. We observe that for this series, $ARIMA(3,1,2)$, is again the model that has the lowest AIC. However, several other models have standard deviations of residuals that are close to 0.082. In particular, $AR(2)$ with a standard deviation of residuals of 0.087. The estimated values of the coefficients of the $AR(2)$ are close to the boundary of stability, and so realisations from the fitted $AR(2)$ trend to be more oscillatory than the observed time series. The fitted $FARIMA(1, 0.497, 0)$ model has a somewhat higher standard deviation of residuals at 0.105, but realisations from the fitted $FARIMA(1, 0.497, 0)$ show stochastic trends and are generally qualitatively more similar to the observed series. Again, the fitted $FARIMA(1, 0.497, 0)$ is close to the boundary of stability, and it is a plausible conceptual model that is consistent with a fractal process.

Table 5: Estimated coefficients and their associated standard errors of AR, ARIMA, FARIMA models for the post-IR yearly data.

ARMA (p, q)	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	AIC	σ
(1,0)	0	0.999 (0.002)	–	–	–	–	-40.040	0.218
(2,0)	0	1.921 (0.025)	-0.927 (0.025)	–	–	–	-482.590	0.087
(1,0)	1	0.919 (0.025)	–	–	–	–	-486.270	0.088
(2,0)	1	0.800 (0.063)	0.131 (0.064)	–	–	–	-488.430	0.087
(1,0)	0.497 (0.004)	0.993 (0.006)	–	–	–	–	-392.877	0.105
(2,0)	0.000	1.923	-0.929	–	–	–	-488.153	0.088
(2,2)	0	1.920 (0.027)	-0.925 (0.027)	–	-0.246 (0.070)	0.343 (0.073)	-502.560	0.083
(3,2)	1	1.447 (0.141)	-0.635 (0.214)	0.107 (0.135)	-0.696 (0.120)	0.550 (0.119)	-515.460	0.082

4.4 Weekly Mauna Loa Data

These data provide weekly carbon dioxide that measured as mole fraction in dry air (direct measure) since May of 1974, in Mauna Loa Observatory, Hawaii (Figure 14). It is acceptable to use Mauna Loa as a proxy for global CO₂ levels. This is because CO₂ mixes well throughout the atmosphere. This is supported by the statistically non-significant difference in trend in Mauna Loa CO₂ levels (1.64 ppm per year) and global CO₂ levels (1.66 ppm per year) (Cook 2019).

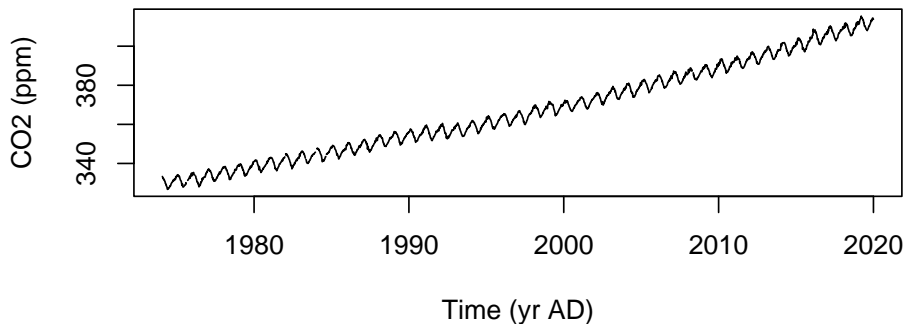


Figure 14: Time series plot of weekly data from May 1974 to January 2020.

There are 20 observations that have values equal to -999.99, which indicates they are not available (NA). The reasons for them to be unavailable have not been stated. These missing values are imputed using linear interpolation, which is easily implemented as our dataset is equally spaced.

From Figure 14, we can observe the seasonal patterns, and much of this can be explained by the role of plants in the carbon cycle (EPA 2019). Plants use CO₂ as well as sunlight and water, in order to make food or other substances that they need to grow, and they release O₂ as a by-product. This process is called photosynthesis, and this can draw down the CO₂ (Monroe 2013). Another process that is part of the carbon cycle is called respiration, this is the process which plants decompose and produce CO₂.

As there is an apparent trend as well as periodicity, hence the first thing that we need to do before any analysis is to remove the trend and seasonality. This can be done by fitting the model of subsection 3.1 with $M = 4$ (Equation 2). The detrended and deseasonalised series, \tilde{y}_t , is defined by

$$\tilde{y}_t = y_t - \left\{ (251.40 + 79.18e^{0.0003t}) \times e^{0.0073\cos(2\pi t/p) - 0.003\sin(2\pi t/p) + 0.001\cos(4\pi t/p) + 0.002\sin(4\pi t/p) - 0.0002\sin(6\pi t/p) + 0.0002\cos(8\pi t/p)} \right\}, \quad (5)$$

where p is equal to the seasonal period, $365.25/7$ in the case of weekly data. There is no clear trend or seasonal pattern of the residuals after detrended and deseasonalised (Figure 15).

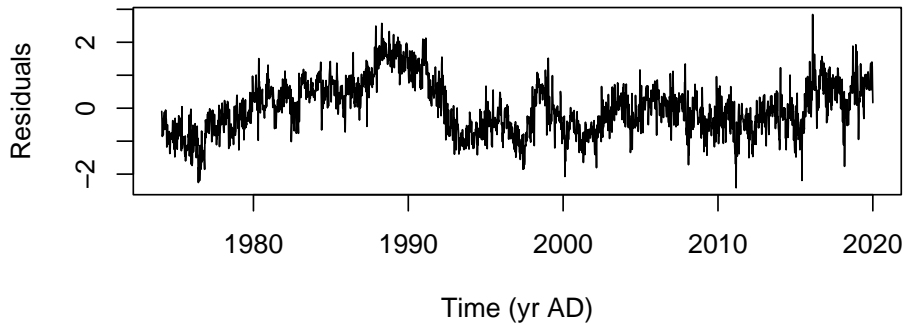
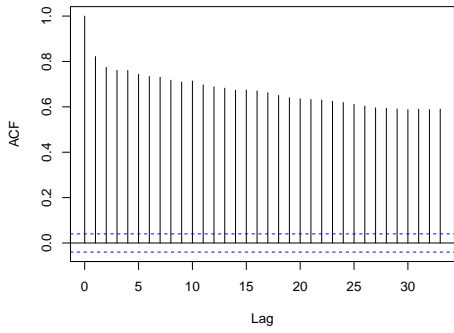


Figure 15: Time series plot of the detrended and deseasonalised weekly data.

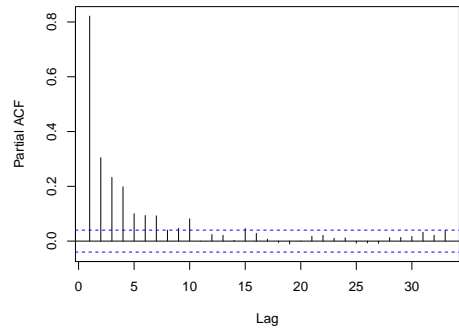
The correlogram (Figure 16a) shows highly significant correlations between the observations, and the partial correlogram (Figure 16b) shows statistically significant at lag 1, 2, 3 and 4, at the 0.05 level. The partial correlations for lag 5, 6, 7 and 10 are relatively small. Both the periodogram and smoothed spectrum (Figure 16c, Figure 16d) show the property of low frequency.

The ADF and PP tests on the residuals show the evidence to reject the null hypothesis of the presence of unit root, at the 0.05 significance level. However, the result returned by the KPSS test is contradicted with the ADF and PP tests, it shows the evidence to reject the null hypothesis of stationarity. If an AR(1) model is fitted to the residuals, the estimated coefficient should close to 1, that is 0.821.

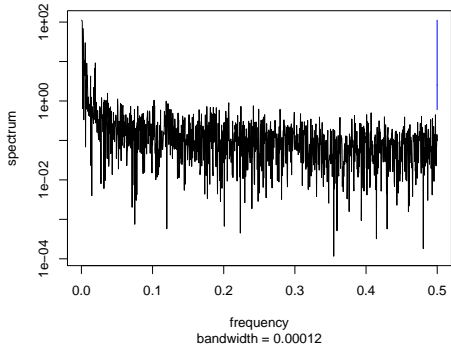
Finally, in Table 6 below, we looked at the weekly data. The ARIMA(2,0,1), which is equivalent to the ARMA(2,1), has the lowest AIC with standard deviation to be 0.401. Even though the difference in AICs between FARIMA(2, 0.497, 0) and ARMA(2, 1) is approximately 30, this can be attributed to the long time series, which allows identification of statistically significant terms which do not correspond to a substantial reduction in the variance of the residuals. They have standard deviation of residuals to be 0.403 and 0.401, respectively. Moreover, the FARIMA(2, 0.497, 0) is more conceptually reasonable, and it also gives a reasonable physically interpretation.



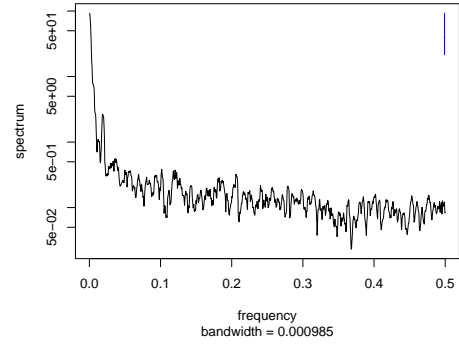
(a) Correlogram.



(b) Partial correlogram.



(c) Periodogram.



(d) Smoothed spectrum with span= 9.

Figure 16: Detrended and deseasonalised weekly data.

Table 6: Estimated coefficients and their associated standard errors of AR, ARIMA, FARIMA models for the weekly data.

ARMA (p, q)	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	AIC	σ
(1,0)	0	0.821 (0.012)	—	—	—	—	2974.820	0.450
(2,0)	0	0.571 (0.020)	0.305 (0.020)	—	—	—	2743.49	0.429
(1,0)	1	-0.367 (0.019)	—	—	—	—	2848.820	0.439
(2,0)	1	-0.466 (0.020)	-0.269 (0.020)	—	—	—	2672.14	0.423
(1,0)	0.494 (0.005)	-0.077 (0.021)	—	—	—	—	2468.456	0.405
(2,0)	0.497 (0.003)	-0.087 (0.021)	-0.073 (0.021)	—	—	—	2458.725	0.403
(1,1)	1	0.227 (0.028)	-0.827 (0.017)	—	—	—	2432.82	0.403
(2,1)	0	1.208 (0.029)	-0.216 (0.028)	—	-0.813 (0.019)	—	2426.25	0.401

5 Conclusion

Table 7: Preferred models for CO₂ time series at different resolutions.

Scale (time interval)	Preferred model without unit root (sd)	Preferred model with unit root (sd)
0.5 million years	FARIMA(2, 0.241, 0) (39.675)	ARIMA(3, 1, 2) (39.369)
1,176 thousand years	FARIMA(2, 0.0644, 0) (6.791)	—
Yearly (Pre-IR)	FARIMA(2, 0.499, 0) (0.053)	ARIMA(3, 1, 2) (0.045)
Yearly (Post-IR)	FARIMA(1, 0.499, 0) (0.105)	ARIMA(3, 1, 2) (0.082)
Weekly	FARIMA(2, 0.497, 0) (0.403)	ARIMA(1, 1, 1) (0.403)

The preferred models in [Table 7](#) are not necessarily those with lowest AIC, but are those we judge to be the most suitable given a preference for relatively simple conceptual models. In particular, autoregressive models without moving average terms, $AR(p)$, have a conceptual interpretation as a discrete approximation to a dynamic system driven by a sequence of independent shocks, are special cases of multiple regression and so relatively simple, and the pacf can provide a clear indication of a suitable order. All autoregressive models fitted to the time series were stationary and so stable, but close to the stability limit. Allowing moving average terms, $ARMA(p, q)$, did generally improve the fit in terms of AIC with the estimates of autoregressive coefficients remaining close to the stability limit. The values of coefficients of moving average terms do not affect stability. Realisations of time series from stable $AR(p)$, or $ARMA(p, q)$, models that are close to the stability boundary will eventually return to some mean value if inputs are set to zero, but return to the mean will be slow. In contrast, $ARIMA(p, 1, q)$ models have a unit root and are not second order stationary, and are unstable because realisations do not return to a mean value if inputs are set to zero. This qualitative difference between $ARMA(p, q)$ and $ARIMA(p, d, q)$ models can be critical, for example in the context of control systems, but in the context of modelling CO₂ concentration all the models are consistent with claims that anthropogenic emissions will not easily be reversed. The lack of any underlying mean for $ARIMA(p, d, q)$ models is conceptually appealing, but a less reasonable corollary is that they are unbounded. However, this limitation can be removed by introducing thresholds above, and below, which the process is modelled as stable. The $FARIMA(p, d, 0)$

models come between the ARMA(p, q) and ARIMA($p, 1, q$) models, but are still stable. The FARIMA($p, d, 0$) models account for a slow decay in the acf and indicate a fractal process. The notion of a fractal process fits nicely with the fact that weekly, and even yearly, sampling of an underlying continuous process with much lower natural frequency is likely to identify low frequency variation as a trend.

In [Table 7](#), a preferred model including unit root is only shown if it has lower estimated standard deviation of errors than the model without unit root. Also, the models after IR are for errors after the conceptual trend model, with seasonal terms in the case of weekly data, is fitted. In summary, both the FARIMA and ARIMA models provide a better fit to the time series than ARMA models with the same p and q . Up until the first Industrial Revolution there is little to choose between them although the ARIMA have slightly lower estimated standard deviations of errors at two of the three scales. Formally, there is no evidence against a hypothesis of a unit root and there is evidence against a hypothesis of stationarity at the 0.1 level of significance, for all except the weekly series. These findings support a choice of ARIMA models over the corresponding ARMA models. But unit root tests do not allow for moving average terms, and ARMA models with $q = 2$ are close to the corresponding ARIMA models in term of AIC. However, moving average terms do not have much conceptual interpretation in this context.

The FARIMA models suggest fractal processes. The ARIMA models are not stable, which is not physically plausible at the 0.5 million years scale but may be a realistic approximation over the shorter periods. The best fitting ARMA models are close to the boundary for stability. Whichever models are chosen, the borderline stability is consistent with a general claim that a reduction in atmospheric CO₂ following a cessation of CO₂ emissions will be a protracted process. In particular, the evidence that the rise to a peak is more rapid than the decline from it (found in the thousand-year time interval series) provides some support for the claim by (Solomon et al. 2009).

Following the Industrial Revolution the anthropomorphic effects appear to dominate the natural sequence of events. After fitting conceptual models for the trends, the prediction errors are well modelled as either FARIMA or ARIMA models. The former are preferred as superimposing unstable models on an exponential increase is not a convincing conceptual model. It is interesting that the FARIMA models still provide a reasonable fit as this is consistent with a fractal process at scales from half a million years down to one week.

Except for the weekly data from Mauna Loa, we have modelled time series of estimates of CO₂ rather than direct measurements. In all cases the CO₂ concentration is, nominally at least, aggregated over the sampling interval. It is possible that autocorrelations at short lags may be affected by the method of estimation. But, this cannot account for

the evidence of a lack of stability in the underlying process. The borderline stability has serious consequences because a unit root corresponds to a process that does not return to some average value. If this is realistic, ceasing CO₂ emissions will not necessary lead to a return to lower levels but it will help prevent levels increasing. Empirical evidence based only on time series models might not be convincing in itself, but when it augments physical explanation (e.g. Solomon et al., 2009) the forecast should be taken seriously.

Acknowledgements

Xiaomeng Gu thanks the Australian Research Council for a scholarship to pursue a Master of Philosophy.

Author Declarations

Funding

This study was funded by the Australian Research Council for a scholarship for the Master of Philosophy.

Conflicts of interest

Not applicable.

Availability of data and material

The Australian Bureau of Meteorology and the National Oceanic and Atmospheric Administration provided data for analysis.

Code availability

All the calculations are done in open source R.

Authors' contributions

A. V. Metcalfe conceived of the presented idea. X. Gu conducted the analysis under the supervision of A. V. Metcalfe and G. F. V. Glonek. X. Gu wrote the manuscript in consultation with A. V. Metcalfe and G. F. V. Glonek. All authors provided critical feedback and helped shape the research, analysis and manuscript.

References

- Akaike, H. (1974), ‘A new look at the statistical model identification’, *IEEE transactions on automatic control* **19**(6), 716–723.
- Baum, C. F. (2013), ‘Arfima (long memory) models’. <https://fmwww.bc.edu/ec-c/s2013/327/EC327.S2013.nn5.slides.pdf>.
- Beran, J., Bhansali, R. J. & Ocker, D. (1998), ‘On unified model selection for stationary and nonstationary short-and long-memory autoregressive processes’, *Biometrika* **85**(4), 921–934.
- Burnham, K. P. & Anderson, D. R. (2004), ‘Multimodel inference: understanding aic and bic in model selection’, *Sociological methods & research* **33**(2), 261–304.
- Cook, J. (2019), ‘How reliable are co2 measurements?’. <https://www.skepticalscience.com/co2-measurements-uncertainty.htm>.
- EPA, U. (2019), ‘Carbon through the seasons’.
URL: <https://archive.epa.gov/climatechange/kids/documents/carbon-through-the-seasons.pdf>
- EViews (2019), ‘Unit root testing’.
URL: http://www.eviews.com/help/helpintro.html#page/content/advtimeser-Unit_Root_Testing.html
- Foster, G. L., Royer, D. L. & Lunt, D. J. (2017), ‘Future climate forcing potentially without precedent in the last 420 million years’, *Nature Communications* **8**, 14845.
- Green, M. (2020), ‘Australia bushfires contribute to big rise in global co2 levels: Uk’s met office’. <https://www.reuters.com/article/us-climate-change-australia-emissions/australia-bushfires-contribute-to-big-rise-in-global-co2-levels-uks-met-office-idUSKBN1ZN0BT>.
- Joshi, P. (2016), ‘What is long memory in time series analysis’.
<https://prateekvjoshi.com/2016/08/27/what-is-long-memory-in-time-series-analysis/>.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., Shin, Y. et al. (1992), ‘Testing the null hypothesis of stationarity against the alternative of a unit root’, *Journal of econometrics* **54**(1-3), 159–178.
- Macfarling Meure, C., Etheridge, D., Trudinger, C., Steele, P., Langenfelds, R., Van Ommen, T., Smith, A. & Elkins, J. (2006), ‘Law dome co2, ch4 and n2o ice core records extended to 2000 years bp’, *Geophysical Research Letters* **33**(14).

- Mansor, M. M., Green, D. A. & Metcalfe, A. V. (2020), ‘Detecting directionality in time series’, *The American Statistician* **74**(3), 258–266.
- McGee, M. (2020), ‘Co2 datasets’. <https://www.co2.earth/co2-datasets>.
- Meure, C. M., Etheridge, D., Trudinger, C., Steele, P., Langenfelds, R., van Ommen, T., Smith, A. & Elkins, J. (2006), ‘Law dome co2, ch4 and n2o ice core records extended to 2000 years bp’.
- Monroe, R. (2013), ‘Why does atmospheric co2 peak in may?’. <https://sioweb.ucsd.edu/programs/keelingcurve/2013/06/04/why-does-atmospheric-co2-peak-in-may/>.
- Petit, J., Jouzel, J., Raynaud, D., Barkov, N., Barnola, J., Basile, I., Bender, M., Chappellaz, J., Davis, J., Delaygue, G. et al. (2001), ‘Vostok ice core data for 420,000 years’, *IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series* **76**, 2001.
- Shirmohammadi, R., Soltanieh, M. & Romeo, L. M. (2018), ‘Thermoeconomic analysis and optimization of post-combustion co2 recovery unit utilizing absorption refrigeration system for a natural-gas-fired power plant’, *Environmental Progress & Sustainable Energy* **37**(3), 1075–1084.
- Siegenthaler, U., Stocker, T. F., Monnin, E., Lüthi, D., Schwander, J., Stauffer, B., Raynaud, D., Barnola, J.-M., Fischer, H., Masson-Delmotte, V. et al. (2005), ‘Stable carbon cycle–climate relationship during the late pleistocene’, *Science* **310**(5752), 1313–1317.
- Solomon, S., Plattner, G.-K., Knutti, R. & Friedlingstein, P. (2009), ‘Irreversible climate change due to carbon dioxide emissions’, *Proceedings of the National Academy of Sciences* **106**(6), 1704–1709.
- Zivot, E. & Wang, J. (2007), *Modeling financial time series with S-Plus®*, Vol. 191, Springer Science & Business Media.

Chapter 5

Conclusion

In the case of just two time series, it may be relatively straight forward to demonstrate a Granger causal relationship, and cross-correlation analysis is useful for establishing a Granger causal relationship. The situation becomes far more complicated when there are several inter-related time series, and VAR models provide more insight than cross-correlation analyses which are individually bivariate. In time series analysis, there will typically be several models that give a similar fit in terms of mean squared error yet have somewhat different physical interpretation. Since there is no true model, various selection procedures can be promoted. Using AIC to choose a best model tends to prefer complex models and ignores physical plausibility. However, it seems reasonable to require some physical plausibility. But, this may lead to acceptance of whatever physical theory the researcher chooses to propose. We investigated the practical consequences of this dilemma in two cases: investigation of Granger causal for monthly rainfall, and multi-scale analysis of atmospheric CO₂.

In the case of the investigation of Granger causality for monthly rainfall, there were many models: VAR(1) and VAR(2) models without interactions between climate indicators; VAR(1) and VAR(2) models with interactions between climate indicators; and 1 threshold VAR(1) model with linear terms. However, the impact of the IOD is consistent, that is the IOD is Granger causal for South Australian monthly rainfall by 1 month with a negative relationship. The Granger causality of the SOI and the PDO for rainfall depended on the model fitted and was less substantial than the Granger causality of the IOD for rainfall. In contrast, the effect, if any, of the SAM is equivocal.

A realistic physically based climatic model that allows for climatic indicators and predicts rainfall would be highly complex and non-linear. So, non-linear empirical regression models which include higher order, than two predictor, interactions might be anticipated to

provide a better approximation. However, the non-linear time series models, TVAR(p), do not provide any improvement or additional information compared to the linear VAR(p) models. The understanding of Granger causations should help to improve predictions of monthly rainfall in South Australia. It has potential benefits for making informed policy, planning and management decisions, and can assist with the more sustainable operation of water resource systems. Investigation of causality through Granger causality is essentially about improved predictions rather than a direct demonstration of causality. Integrating rainfall prediction models with agricultural decision making models would be an extension of the work presented in this thesis.

The case of atmospheric CO₂ involved just one variable, but it was not straightforward to choose between ARMA, ARIMA, and FARIMA models. Overall, the ARIMA(3, 1, 2) model gave the minimum residual mean square at the most time scales. But, it has the drawbacks of introducing moving average terms and of introducing unit roots. Unit root models may provide good short term forecasts, but the long term physical interpretation of unbounded stochastic trends is not realistic. The FARIMA model is stable about an overall mean and this seems far more plausible. The FARIMA model is an example of a fractal, which would be repeated at the different time scales, making a very nice model. Rather than attempt to justify a best class of model, it seems more useful to emphasise that all the models are borderline stable. Borderline stability is consistent with warnings that increases in atmospheric CO₂ will take considerable time to reverse, even if emissions are drastically reduced. Time series models have the limitation that they are relatively simple empirical approximations to complex processes. Forecasts are based on extrapolations of current trends and cannot allow for unknowable cataclysmic changes.

Appendix A

R code

A.1 Section 2.3 White Noise

Simulation of a discrete white noise with mean 0 and variance 1.

```
# set seed
set.seed(2020)

# generate 100 observations with mean 0 and variance 1
wt <- rnorm(100)

# time series plot
ts.plot(wt, xlab = "Observations", ylab = expression(w[t]))
```

A.2 Section 2.6 Decomposition of Series

Decompose additive and multiplicative models

```
# multiplicative model
decompair <- plot(decompose(AirPassengers, type = "multiplicative"))

# additive model
decompco2 <- plot(decompose(co2.ts, type = "additive"))
```

A.3 Section 2.13 Pre-whitening

Simulating two time series variables $\{x_t\}$ and $\{y_t\}$ based on [Equation 2.29](#) and [Equation 2.30](#)

```
# set seed
set.seed(2020)

# simulate x which is ARIMA(1,1,0), n = 200
x = arima.sim(list(order = c(1,1,0), ar = 0.7), n = 200)

# simulate y use lagged x
z = ts.intersect(x, lag(x,-3), lag(x,-4))
y = 15+0.8*z[,2]+1.5*z[,3]

# ccf plot between x and y
ccf(z[,1],y,na.action = na.omit, ylab = "CCF", main = "")
```

The pre-whitening procedure can be implemented using `prewhiten` function in TSA package in R.

```
# instal the TSA package from R website
install.packages(
  "https://cran.r-project.org/src/contrib/Archive/TSA/TSA_1.2.1.tar.gz",
  repos = NULL, type = "source")

# load the library
library(TSA)

# pre-whiten the series
prewhiten(z[,1] , y)
```

We can also implement pre-whitening using the algorithm that discussed in [Section 2.13](#), and these two methods give identical result.

```
# fit ARIMA(1,1,0) to x
arx <- arima(x, order = c(1,1,0))

# get the residuals from the ARIMA model
x_res <- arx$residuals

# filter the y variable using the coefficients
#returned from the fitted ARIMA model for x
```

```
yfilter <- filter(y, filter = c(1, -1.6323, 0.6323), sides = 1)

# ccf plot
ccf(x_res, yfilter, na.action = na.omit, main = "", ylab = "CCF")
```

A.4 Chapter 3

A.4.1 Linear interpolation

We used linear interpolation method to deal with missing values in monthly rainfall data. Here, we used monthly rainfall data at the Adelaide Airport weather station as an example. There was one missing value found in August in 2014. We extracted all the observations in August, and applied linear interpolation.

```
# read in data
rain_ade<-read.csv("./Data/adelaide_rainfall.csv")

# get the observations from 1978 to 2017
lev1<-which(rain_ade$Year==1978)
lev2<-which(rain_ade$Year==2017)
rain_ade<-rain_ade[lev1:lev2,1:ncol(rain_ade)]

# get the observations in August from 1978 to 2017
yr <- rain_ade$Year
aug <- rain_ade$Aug

# linear interpolation
aug2 <- approx(yr, aug, n=length(aug))
rain_ade$Aug <- aug2$y
rain_ade <- rain_ade[,2:ncol(rain_ade)]

# store the series as a vector
rain_ade<-as.vector(t(rain_ade))
```

A.4.2 Deseasonalise and detrend the series

```
##### deseasonalise and detrend rainfall data
# months as indicators
```

```

month.f <- rep(seq(1,12),2017-1978+1)
month.f <- factor(month.f)

# fit linear regression on time and months as indicators
rain_ade_res <- residuals(lm(rain_ade ~ month.f + time(rain_ade)))

##### deseasonalise and detrend climate indicators
# adjusted mean of time
# it is same for all the series as they share the common period
t <- time(SOI) - mean(SOI)

# monthly cycle
p <- 12
c <- cos(2*pi*t/p)
s <- sin(2*pi*t/p)

# fit regression on harmonic terms
m1 <- lm(SOI ~ c + s)
# deseasonalised SOI
soids <- m1$residuals

# deseasonalised and detrended IOD
m2 <- lm(IOD ~ c + s + t)
ioddsdt <- m2$residuals

# deseasonalised and detrended PDO
m3 <- lm(PDO ~ c + s + t)
pdodsdt <- m3$residuals

# detrended SAM
m4 <- lm(SAM ~ t)
samdt <- m4$residuals

```

A.4.3 Fitting VAR models

```

# load the vars package
pacman::p_load(vars)

# combine the series with only linear terms
df <- cbind(rain_ade_res, soids, ioddsdt, pdodsdt, samdt)

# fit a VAR(1) model

```

```
var1 <- vars::VAR(df, p = 1)
```

A.4.4 Fitting TVAR models

```
# load the tsDyn package
pacman::p_load(tsDyn)

# fit a 1 threshold TVAR(1) model with only linear terms
adetvar <- tsDyn::TVAR(df, lag=1, nthresh = 1)
```

A.5 Chapter 4

A.5.1 Filtering deterministic models

We fitted exponential trend to the post-Industrial Revolution, as the series shows clearly increasing exponential trend.

```
t <- time(co2.post)
t_centered <- t - mean(t)

# get starting points
theta.0 <- min(co2.post) * 0.5
model.0 <- lm(log(co2.post - theta.0) ~ t_centered)
alpha.0 <- exp(coef(model.0)[1])
beta.0 <- coef(model.0)[2]
start <- list(alpha = alpha.0, beta = beta.0, theta = theta.0)

# fit nls
model <- nls(co2.post ~ alpha * exp(beta * t_centered) + theta,
start = start)
model_ts <- ts(residuals(model), start = 1761,
end = 2004, frequency = 1)
```

There is also a marked seasonal pattern in the time series of weekly data from Mauna Loa. The seasonal pattern is modelled by modulating the trend with an exponential of a linear sum of cosine and sine functions.

```
t <- time(co2)
```

```

p <- 365.25/7

# get starting points
theta.0 <- min(co2) * 0.5
model.0 <- lm(log(co2 - theta.0) ~ t + cos(2*pi*t/p)+sin(2*pi*t/p)
              + cos(4*pi*t/p)+sin(4*pi*t/p)
              + cos(6*pi*t/p)+sin(6*pi*t/p)
              +cos(8*pi*t/p)+sin(8*pi*t/p))

alpha.0 <- exp(coef(model.0)[1])
beta.0 <- coef(model.0)[2]
c1 <- coef(model.0)[3]
s1 <- coef(model.0)[4]
c2 <- coef(model.0)[5]
s2 <- coef(model.0)[6]
c3 <- coef(model.0)[7]
s3 <- coef(model.0)[8]
c4 <- coef(model.0)[9]
s4 <- coef(model.0)[10]
start <- list(alpha = alpha.0, beta = beta.0, theta = theta.0,
              c1 = c1, s1 = s1, c2 = c2, s2 = s2,
              c3 = c3, s3 = s3, c4 = c4, s4 = s4)

# fit nls
model <- nls(co2 ~ (alpha * exp(beta * t) + theta) * exp(c1*cos(2*pi*t/p)
              + s1*sin(2*pi*t/p)
              + c2*cos(4*pi*t/p)+s2*sin(4*pi*t/p)
              + c3*cos(6*pi*t/p)+s3*sin(6*pi*t/p)
              + c4*cos(8*pi*t/p)+s4*sin(8*pi*t/p)) ,start = start)

```

A.5.2 Fitting univariate time series models

We fitted $ARMA(p, q)$, $ARIMA(p, d, q)$, and $FARIMA(p, d, q)$ to five time series, with p and q no greater than 3. Additionally, the differencing parameter d for $ARIMA$ models set to be no greater than 2. For loops are used and the models are listed from lowest AIC to highest AIC values. The R code below is an example of fitting $ARIMA$ models

for million year data. For ARMA models, we can restrict the value of d to be zero.

```
# create container
order.matrix <- matrix(0, nrow = 3, ncol = 3 * 1 * 3)
aic.vec <- numeric(3 * 1 * 3)
sig.vec <- numeric(3 * 1 * 3)
k <- 1

# fir ARIMA(p, d, q) with p and q no greater than 3, and d = 1
for (p in 1:3) for (d in 1) for (q in 1:3) {
  tryCatch({order.matrix[, k] <- c(p,d,q)
    aic.vec[k] <- AIC(arima(co2, order=c(p,d,q)))
    sig.vec[k] <- arima(co2, order=c(p,d,q))$sigma
    k <- k+1 },
  # return error
  error=function(e){cat("ERROR_":", conditionMessage(e), "\n")})
}

# reorder the list to from lowest AIC to highest
ind <- order(aic.vec, decreasing = FALSE)
aic.vec <- aic.vec[ind]
order.matrix <- order.matrix[, ind]
order.matrix <- t(order.matrix)

# return p, d, q, AIC, and standard deviation of residuals
result <- cbind(order.matrix, aic.vec, sig.vec)
colnames(result) <- c("p", "d", "q", "AIC", "sigma")
result
```

For FARIMA models, we do not need to specify the value of d , it will returned by the function `fracdiff` in the `fracdiff` package in R.

```
# create container
order.matrix <- matrix(0, nrow = 2, ncol = 3 * 4)
aic.vec <- numeric(3 * 4)
sig.vec <- numeric(3 * 4)
k <- 1

# fit FARIMA(p, d, q) with p from 1 up to 3, and q from 0 up to 3
for (nar in 1:3) for (nma in 0:3) {
  if (has_warning(fracdiff(co2, nar=nar, nma=nma)) == TRUE) {k=k+1
  } else{
    order.matrix[, k] <- c(nar, nma)
    aic.vec[k] <- AIC(fracdiff(co2, nar=nar, nma=nma))
  }
}
```

```
sig.vec[k] <- fracdiff(co2, nar=nar,nma=nma)$sigma
k <- k+1
}}

# reorder the list from lowest AIC to highest
ind <- order(aic.vec, decreasing = FALSE)
aic.vec <- aic.vec[ind]
sig.vec <- sig.vec[ind]
order.matrix <- order.matrix[, ind]
order.matrix <- t(order.matrix)

# return the results
# remove the ones with errors
result <- cbind(order.matrix, aic.vec, sig.vec)
re <- which (result[,3] == 0)
result <- result[-c(1:tail(re,n=1)),]
colnames(result) <- c("nar", "nma", "AIC", "Sigma")
```

Bibliography

- Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.
- Aleem, A. & Lahiani, A. (2014), 'A threshold vector autoregression model of exchange rate pass-through in Mexico', *Research in International Business and Finance* **30**, 24–33.
- Aoki, M. (1987), 'State space modeling of time series'.
- Armstrong, J. S., Green, K. C., Jones Jr, R. J. & Wright, M. J. (2010), 'Predicting elections from politicians's faces', *International Journal of Public Opinion Research* **22**(4), 511–522.
- Ashok, K., Guan, Z. & Yamagata, T. (2003), 'Influence of the Indian Ocean dipole on the Australian winter rainfall', *Geophysical Research Letters* **30**(15).
- Athanasopoulos, G. & Vahid, F. (2008), 'Varma versus VAR for macroeconomic forecasting', *Journal of Business & Economic Statistics* **26**(2), 237–252.
- Bagirov, A. M., Mahmood, A. & Barton, A. (2017), 'Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach', *Atmospheric Research* **188**, 20–29.
- Bartlett, M. S. (1948), 'Smoothing periodograms from time-series with continuous spectra', *Nature* **161**(4096), 686–687.
- Bartlett, M. S. (1950), 'Periodogram analysis and continuous spectra', *Biometrika* **37**(1/2), 1–16.
- Beran, J., Bhansali, R. J. & Ocker, D. (1998), 'On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes', *Biometrika* **85**(4), 921–934.
- Beveridge, W. H. (1922), 'Wheat prices and rainfall in western Europe', *Journal of the Royal Statistical Society* **85**(3), 412–475.
- Blackman, R. B. & Tukey, J. W. (1958), 'The measurement of power spectra from the point of view of communications engineering—part I', *Bell System Technical Journal* **37**(1), 185–282.
- Box, G. E. & Jenkins, G. M. (1970), *Time series analysis forecasting and control*, Technical report, WISCONSIN UNIV MADISON DEPT OF STATISTICS.

- Brillinger, D. (1966), Computation and interpretation of kth order spectra, *in* ‘Proc. of an Advanced Seminar on Spectral Analysis of Time Series’.
- Brockwell, P. J., Davis, R. A. & Fienberg, S. E. (1991), *Time series: theory and methods: theory and methods*, Springer Science & Business Media.
- Bureau of Meteorology (2012), ‘About enso and iod indices’.
URL: <http://www.bom.gov.au/climate/enso/indices/about.shtml>
- Bureau of Meteorology (2021a). <http://www.bom.gov.au/climate/cdo/about/about-rain-data.shtml>.
- Bureau of Meteorology (2021b), ‘About enso and iod indices’.
<http://www.bom.gov.au/climate/enso/indices/about.shtml>.
- Bureau of Meteorology (2021c), ‘The southern annular mode (sam)’.
URL: <http://www.bom.gov.au/climate/enso/history/ln-2010-12/SAM-what.shtml>
- Burlando, P., Rosso, R., Cadavid, L. G. & Salas, J. D. (1993), ‘Forecasting of short-term rainfall using arma models’, *Journal of Hydrology* **144**(1-4), 193–211.
- Cai, W. & Cowan, T. (2006), ‘Sam and regional rainfall in ipcc ar4 models: Can anthropogenic forcing account for southwest western australian winter rainfall reduction?’, *Geophysical Research Letters* **33**(24).
- Cai, W., Cowan, T. & Sullivan, A. (2009), ‘Recent unprecedented skewness towards positive indian ocean dipole occurrences and its impact on australian rainfall’, *Geophysical Research Letters* **36**(11).
- Cai, W., Van Rensch, P., Cowan, T. & Hendon, H. H. (2011), ‘Teleconnection pathways of enso and the iod and the mechanisms for impacts on australian rainfall’, *Journal of Climate* **24**(15), 3910–3923.
- Chan, J. C. (2020), ‘Large bayesian vars: A flexible kronecker error covariance structure’, *Journal of Business & Economic Statistics* **38**(1), 68–79.
- Chourasia, A. (2020), ‘Decomposition in time series data’. <https://medium.com/analytics-vidhya/decomposition-in-time-series-data-b20764946d63>.
- Chowdhury, R. K. & Beecham, S. (2013), ‘Influence of soi, dmi and niño3. 4 on south australian rainfall’, *Stochastic environmental research and risk assessment* **27**(8), 1909–1920.
- Chris, B. (2014), ‘Introductory econometrics for finance, 3rd editio’.

- Cleverly, J., Eamus, D., Luo, Q., Coupe, N. R., Kljun, N., Ma, X., Ewenz, C., Li, L., Yu, Q. & Huete, A. (2016), 'The importance of interacting climate modes on australia's contribution to global carbon cycle extremes', *Scientific reports* **6**(1), 1–10.
- Cline, D. B. & Huay-min, H. P. (2001), 'Stability of nonlinear time series: What does noise have to do with it?', *Lecture Notes-Monograph Series* pp. 151–170.
- Cowpertwait, P. S. & Metcalfe, A. V. (2009), *Introductory time series with R*, Springer Science & Business Media.
- Daniell, P. J. (1946), 'Discussion on symposium on autocorrelation in time series', *Journal of the Royal Statistical Society* **8**(1).
- Dickey, D. A. & Fuller, W. A. (1979), 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American statistical association* **74**(366a), 427–431.
- Diebold, F. X., Husted, S. & Rush, M. (1991), 'Real exchange rates under the gold standard', *Journal of Political Economy* **99**(6), 1252–1271.
- Eichler, M. (2013), 'Causal inference with multiple time series: principles and problems', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**(1997), 20110613.
- Engle, R. F. & Granger, C. W. (1987), 'Co-integration and error correction: representation, estimation, and testing', *Econometrica: journal of the Econometric Society* pp. 251–276.
- Forootan, E., Awange, J., Schumacher, M., Anyah, R., Van Dijk, A., Kusche, J. et al. (2016), 'Quantifying the impacts of enso and iod on rain gauge and remotely sensed precipitation products over australia', *Remote sensing of Environment* **172**, 50–66.
- Gelper, S., Wilms, I. & Croux, C. (2016), 'Identifying demand effects in a large network of product categories', *Journal of Retailing* **92**(1), 25–39.
- Georgiou, A. D. (2019), 'Autocorrelation in time series data - dzone ai'. <https://dzone.com/articles/autocorrelation-in-time-series-data>.
- Gil-Alana, L. A. & Toro, J. (2002), 'Estimation and testing of arfima models in the real exchange rate', *International Journal of Finance & Economics* **7**(4), 279–292.
- Gillett, N. P., Kell, T. D. & Jones, P. (2006), 'Regional climate impacts of the southern annular mode', *Geophysical Research Letters* **33**(23).

- Giraitis, L., Kokoszka, P., Leipus, R. & Teyssière, G. (2003), ‘Rescaled variance and related tests for long memory in volatility and levels’, *Journal of econometrics* **112**(2), 265–294.
- Granger, C. W. (1969), ‘Investigating causal relations by econometric models and cross-spectral methods’, *Econometrica: journal of the Econometric Society* pp. 424–438.
- Granger, C. W. (1980), ‘Long memory relationships and the aggregation of dynamic models’, *Journal of econometrics* **14**(2), 227–238.
- Granger, C. W. (1981), ‘Some properties of time series data and their use in econometric model specification’, *Journal of econometrics* **16**(1), 121–130.
- Granger, C. W. & Joyeux, R. (1980), ‘An introduction to long-memory time series models and fractional differencing’, *Journal of time series analysis* **1**(1), 15–29.
- Granger, C. W., Newbold, P. & Economidis, J. (1974), ‘Spurious regressions in econometrics’, *Baltagi, Badi H. A Companion of Theoretical Econometrics* pp. 557–61.
- Graves, T., Gramacy, R., Watkins, N. & Franzke, C. (2017), ‘A brief history of long memory: Hurst, mandelbrot and the road to arfima, 1951–1980’, *Entropy* **19**(9), 437.
- Gray, H. L., Kelley, G. D. & McIntire, D. (1978), ‘A new approach to arma modeling’, *Communications in Statistics-Simulation and Computation* **7**(1), 1–77.
- Hannan, E. J. (1980), ‘The estimation of the order of an arma process’, *The Annals of Statistics* pp. 1071–1081.
- Hassler, U. & Wolters, J. (1994), ‘On the power of unit root tests against fractional alternatives’, *Economics letters* **45**(1), 1–5.
- Henderson, S. (2012), ‘weather’.
URL: <https://ussromantics.com/tag/weather/>
- Hendon, H. H., Thompson, D. W. & Wheeler, M. C. (2007), ‘Australian rainfall and surface temperature variations associated with the southern hemisphere annular mode’, *Journal of Climate* **20**(11), 2452–2467.
- Hoover, K. D. (2006), ‘Causality in economics and econometrics’, *Available at SSRN 930739*.
- Hosking, J. R. (1984), ‘Modeling persistence in hydrological time series using fractional differencing’, *Water resources research* **20**(12), 1898–1908.
- Hurst, H. E. (1951), ‘Long-term storage capacity of reservoirs’, *Transactions of the American society of civil engineers* **116**(1), 770–799.

- Johnson, E. R. & Bras, R. L. (1980), 'Multivariate short-term rainfall prediction', *Water Resources Research* **16**(1), 173–185.
- Jones, R. H. (1980), 'Maximum likelihood fitting of arma models to time series with missing observations', *Technometrics* **22**(3), 389–395.
- Joshi, P. (2016), 'What is long memory in time series analysis'.
<https://prateekvjoshi.com/2016/08/27/what-is-long-memory-in-time-series-analysis/>.
- Kalman, R. E. & Bucy, R. S. (1961), 'New results in linear filtering and prediction theory'.
- Kendall, M. G. (1945), 'On the analysis of oscillatory time-series', *Journal of the Royal Statistical Society* **108**(1/2), 93–141.
- Keng, C. Y., Shan, F. P., Shimizu, K., Imoto, T., Lateh, H. & Peng, K. S. (2017), Application of vector autoregressive model for rainfall and groundwater level analysis, in 'AIP Conference Proceedings', Vol. 1870, AIP Publishing LLC, p. 060013.
- Kiem, A. S. & Franks, S. W. (2004), 'Multi-decadal variability of drought risk, eastern australia', *Hydrological Processes* **18**(11), 2039–2050.
- Kirch, C., Muhsal, B. & Ombao, H. (2015), 'Detection of changes in multivariate time series with application to eeg data', *Journal of the American Statistical Association* **110**(511), 1197–1216.
- Kitagawa, G. & Gersch, W. (1996), *Smoothness priors analysis of time series*, Vol. 116, Springer Science & Business Media.
- Kohn, R. & Ansley, C. F. (1989), 'A fast algorithm for signal extraction, influence and cross-validation in state space models', *Biometrika* **76**(1), 65–79.
- Kokoszka, P. & Young, G. (2016), 'Kpss test for functional time series', *Statistics* **50**(5), 957–973.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., Shin, Y. et al. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root', *Journal of econometrics* **54**(1-3), 159–178.
- Lawrance, A. (1991), 'Directionality and reversibility in time series', *International Statistical Review/Revue Internationale de Statistique* pp. 67–79.
- Lee, C. & Shie, F.-S. (2004), 'Fractional integration and the phillips-perron test', *Academia Economic Papers* **32**(2), 273–312.
- Lee, D. & Schmidt, P. (1996), 'On the power of the kpss test of stationarity against fractionally-integrated alternatives', *Journal of econometrics* **73**(1), 285–302.

- Lo, M. C. & Zivot, E. (2001), ‘Threshold cointegration and nonlinear adjustment to the law of one price’, *Macroeconomic Dynamics* **5**(4), 533.
- Luo, Q., Bellotti, W., Williams, M. & Bryan, B. (2005), ‘Potential impact of climate change on wheat yield in south australia’, *Agricultural and Forest Meteorology* **132**(3-4), 273–285.
- Makridakis, S. (1976), ‘A survey of time series’, *International Statistical Review/Revue Internationale de Statistique* pp. 29–70.
- Mansor, M. M., Green, D. A. & Metcalfe, A. V. (2018), ‘Directionality and volatility in high-frequency time series’, *High Frequency* **1**(2), 70–86.
- Mansor, M. M., Green, D. A. & Metcalfe, A. V. (2019), ‘Detecting directionality in time series’, *The American Statistician* .
- Mantua, N. J. & Hare, S. R. (2002), ‘The pacific decadal oscillation’, *Journal of oceanography* **58**(1), 35–44.
- Marshall, G. (2021), ‘The climate data guide: Marshall southern annular mode (sam) index (station-based).’.
URL: <https://climatedataguide.ucar.edu/climate-data/marshall-southern-annular-mode-sam-index-station-based>
- Masum, M. (2020), ‘Multivariate time series forecasting’.
<https://towardsdatascience.com/multivariate-time-series-forecasting-456ace675971>.
- Min, S.-K., Cai, W. & Whetton, P. (2013), ‘Influence of climate variability on seasonal extremes over australia’, *Journal of Geophysical Research: Atmospheres* **118**(2), 643–654.
- Narzo, A. F. D., Aznarte, J. L., Stigler, M. & Tsung-wu, H. (2020), *tsDyn: Nonlinear Time Series Models with Regime Switching*. R package version 10-1.2.
URL: <https://CRAN.R-project.org/package=tsDyn>
- Nicholls, N. (1989), ‘Sea surface temperatures and australian winter rainfall’, *Journal of Climate* **2**(9), 965–973.
- Nielsen, A. (2019), *Practical time series analysis: prediction with statistics and machine learning*, ” O’Reilly Media, Inc.”.
- Nielsen, A. (2021), ‘Practical time series analysis’.
URL: <https://www.oreilly.com/library/view/practical-time-series/9781492041641/ch01.html>

- Niglio, M. & Vitale, C. D. (2015), ‘Threshold vector arma models’, *Communications in Statistics-Theory and Methods* **44**(14), 2911–2923.
- Oppewal, H. (2010), ‘Concept of causality and conditions for causality’, *Wiley International Encyclopedia of Marketing* .
- Pezza, A. B., Simmonds, I. & Renwick, J. A. (2007), ‘Southern hemisphere cyclones and anticyclones: Recent trends and links with decadal variability in the pacific ocean’, *International Journal of Climatology: A Journal of the Royal Meteorological Society* **27**(11), 1403–1419.
- Pfaff, B. (2008a), *Analysis of Integrated and Cointegrated Time Series with R*, second edn, Springer, New York. ISBN 0-387-27960-1.
URL: <http://www.pfaffikus.de>
- Pfaff, B. (2008b), ‘Var, svar and svec models: Implementation within R package vars’, *Journal of Statistical Software* **27**(4).
URL: <http://www.jstatsoft.org/v27/i04/>
- Phillips, P. C. & Perron, P. (1988), ‘Testing for a unit root in time series regression’, *Biometrika* **75**(2), 335–346.
- Pink, B. (2007), ‘Australia’s environment issues and trends 2007’.
<http://www.abs.gov.au>.
- Plackett, R. L. (1950), ‘Some theorems in least squares’, *Biometrika* **37**(1/2), 149–157.
- Power, S., Haylock, M., Colman, R. & Wang, X. (2006), ‘The predictability of interdecadal changes in enso activity and enso teleconnections’, *Journal of Climate* **19**(19), 4755–4771.
- Pui, A., Sharma, A., Santoso, A. & Westra, S. (2012), ‘Impact of the el niño–southern oscillation, indian ocean dipole, and southern annular mode on daily to subdaily rainfall characteristics in east australia’, *Monthly Weather Review* **140**(5), 1665–1682.
- Qiu, D. (2015), *aTSA: Alternative Time Series Analysis*. R package version 3.1.2.
URL: <https://CRAN.R-project.org/package=aTSA>
- Quenouille, M. H. (1957), *The analysis of multiple time-series*, Vol. 1, Griffin London.
- R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: <https://www.R-project.org/>

- Rasel, H., Imteaz, M. A. & Mekanik, F. (2015), Evaluating the effects of lagged enso and sam as potential predictors for long-term rainfall forecasting, *in* 'Water Resources and Environment: Proceedings of the 2015 International Conference on Water Resources and Environment (Beijing, 25-28 July 2015)', CRC Press, p. 125.
- Risbey, J. S., Pook, M. J., McIntosh, P. C., Wheeler, M. C. & Hendon, H. H. (2009), 'On the remote drivers of rainfall variability in australia', *Monthly Weather Review* **137**(10), 3233–3253.
- Schepen, A., Wang, Q. & Robertson, D. (2012), 'Evidence for using lagged climate indices to forecast australian seasonal rainfall', *Journal of Climate* **25**(4), 1230–1246.
- Schuster, A. (1898), 'On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena', *Terrestrial Magnetism* **3**(1), 13–41.
- Schuster, A. (1906), 'Ii. on the periodicities of sunspots', *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **206**(402-412), 69–100.
- Senin, P. V. (2009), 'Literature review on time series indexing'.
- Sims, C. A. (1980), 'Macroeconomics and reality', *Econometrica: journal of the Econometric Society* pp. 1–48.
- Singh, A. (2018), 'A multivariate time series guide to forecasting and modeling (with python codes)'.
- Tao, M., Wang, Y., Yao, Q. & Zou, J. (2011), 'Large volatility matrix inference via combining low-frequency and high-frequency approaches', *Journal of the American Statistical Association* **106**(495), 1025–1040.
- Tjøstheim, D. (1994), 'Non-linear time series: a selective review', *Scandinavian Journal of Statistics* pp. 97–130.
- Tong, H. (1978), 'On a threshold model'.
- Tong, H. (1983), 'Threshold models in non-linear time series analysis. lecture notes in statistics, no. 21'.
- Tong, H. & Lim, K. (1980), 'Threshold autoregression, limit cycles and cyclical data', *Journal of the Royal Statistical Society. Series B (Methodological)* **42**(3), 245–292.
- Toth, E., Brath, A. & Montanari, A. (2000), 'Comparison of short-term rainfall prediction models for real-time flood forecasting', *Journal of hydrology* **239**(1-4), 132–147.

- Tsay, R. S. (2000), ‘Time series and forecasting: Brief history and future research’, *Journal of the American Statistical Association* **95**(450), 638–643.
- Tsay, R. S. (2013), *Multivariate time series analysis: with R and financial applications*, John Wiley & Sons.
- Tsay, R. S. & Tiao, G. C. (1984), ‘Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary arma models’, *Journal of the American Statistical Association* **79**(385), 84–96.
- Ummenhofer, C. C., England, M. H., McIntosh, P. C., Meyers, G. A., Pook, M. J., Risbey, J. S., Gupta, A. S. & Taschetto, A. S. (2009), ‘What causes southeast australia’s worst droughts?’, *Geophysical Research Letters* **36**(4).
- Van Dijk, A. I., Beck, H. E., Crosbie, R. S., de Jeu, R. A., Liu, Y. Y., Podger, G. M., Timbal, B. & Viney, N. R. (2013), ‘The millennium drought in southeast australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society’, *Water Resources Research* **49**(2), 1040–1057.
- Verstraete, M. (2018), ‘Is it necessary to pre-whiten time series of climate variables for checking the presence of trend?’.
- Wang, D., Zheng, Y., Lian, H. & Li, G. (2020), ‘High-dimensional vector autoregressive time series modeling via tensor decomposition’, *Journal of the American Statistical Association* pp. 1–42.
- Wecker, W. E. & Ansley, C. F. (1983), ‘The signal extraction approach to nonlinear regression and spline smoothing’, *Journal of the American Statistical Association* **78**(381), 81–89.
- Wilms, I., Basu, S., Bien, J. & Matteson, D. S. (2017), ‘Sparse identification and estimation of large-scale vector autoregressive moving averages’, *arXiv preprint arXiv:1707.09208* .
- Wold, H. (1938), A study in the analysis of stationary time series, PhD thesis, Almqvist & Wiksell.
- Yule, G. U. (1927), ‘VII. on a method of investigating periodicities disturbed series, with special reference to wolfer’s sunspot numbers’, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **226**(636-646), 267–298.
- Zeileis, A. & Hothorn, T. (2002), ‘Diagnostic checking in regression relationships’, *R News* **2**(3), 7–10.
URL: <https://CRAN.R-project.org/doc/Rnews/>

Zheng, J. (2013), 'Effects of us monetary policy shocks during financial crises-a threshold vector autoregression approach'.

Zivot, E. & Wang, J. (2007), *Modeling financial time series with S-Plus®*, Vol. 191, Springer Science & Business Media.