UNIVERSITY OF ZULULAND

A DATA ADAPTIVE APPROACH TO THE ANALYSIS OF SOUTH AFRICAN CLIMATE VARIABILITY AND ITS AGRICULTURAL IMPACTS

By

WILLARD ZVAREVASHE

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy (Statistics)

in the

Department of Mathematical Sciences

Faculty of Sciences and Agriculture

Supervisor: Dr. Syamala Krishnannair

Co-Supervisor: Prof. Venkataraman Sivakumar

2020

DECLARATION

I, Willard ZVAREVASHE (Student No: 201719050), solemnly declare that the thesis entitled "A data adaptive approach to the analysis of South African climate variability and its agricultural impact " submitted by the undersigned was carried out under the supervision of Dr. Syamala Krishnannair in the Department of Mathematical Sciences, University of Zululand and Prof. Venkataraman Sivakumar , Discipline of Physics, University of KwaZulu Natal. This study represents original work by the author and has not been submitted in any form to another university for any degree. The materials used are acknowledged in the text.

Candidate: Willard Zvarevashe

Signature:....................................

Date:.................................

Supervisor: Dr Syamala Krishnannair

Signature:....................................

Date:.............................

Co-Supervisor: Prof Venkataraman Sivakumar

Signature:....................................

Date:.............................

"Logic will take you from A to B. Imagination will take you everywhere"

Albert Einstein

ABSTRACT

Climate variability, which is mostly physically identified through extreme rainfall and temperature, has a significant impact on the daily livelihoods of people and other species. The temperature has been increasing at an average of 0.65ºC per decade since 1900, and this may be attributed to natural and human factors. Climate variability has more adverse effects on Africa due to lack of resources, failure to adapt, and high dependence on the rain for agricultural purposes. Although Africa is the most affected continent, there are few studies on the climate variability impact on the continent. Therefore, there is a need to carry out more studies using historical data to model the climate variability impact. The climate has become more variable recently, as can be observed by the increase in the occurrence of floods and droughts. Due to this high volatility, linear statistical methods are inconclusive. Some of the non-linear methods, such as Fourier and wavelet transform-based methods make several assumptions. In this study, a data-adaptive method, ensemble empirical mode decomposition (EEMD), is used to decompose the climate data into multiple time series called intrinsic mode functions (IMF). IMFs represents the variability of the time series at different time scales. EEMD is a noise assisted method and there is a challenge on the amplitude of noise to be added. Therefore, in this study, a method to find the optimal amplitude of the noise is proposed. To determine the physical meaning of the IMFs synchronisation is employed. Synchronisation is a method that can be used to find the correlations between oscillating systems. Rainfall is synchronised with temperature, El

Niño-Southern Oscillation (ENSO) and Quasi-Biennial Oscillation (QBO) to investigate their variability impact on rainfall at different time scales.

It has been shown that models, coupled with decomposition, perform better than single models; hence, for the first time, the decomposed data (IMFs) are modelled using generalised extreme value distribution (GEVD). GEVD was found in previous studies to be the best suitable for many hydrological and meteorological problems. However, many studies assume stationarity of the data. In this study, the diagnostic plots (quantile-quantile plot, probability-probability plot and probability density plot) reveals that the IMFs GEVD model provided a better fit. The goodness-of-fit test using Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) test shows that IMFs GEVD models are more robust, compared to the original rainfall time series GEVD model.

EEMD is used to investigate the climate variability impact on agriculture using grapevines and sugarcane as case studies. The IMFs from the Normalised Difference Vegetative Index (NDVI) are synchronised to rainfall and temperature. NDVI is used because it was found in previous studies to be positively correlated to the yield in many crops and has been used to gauge the expected yield for crops. The results show that an increase of temperature for the period under study does not have noticeable impact on grapevines' NDVI. However, this is not the case with sugarcane where it is shown that temperature variability has an impact on the NDVI at seasonal, and 26 months (approximately 2 years) periods. This study is the first to use the decomposition of NDVI to investigate the climate variability impact on plants. Furthermore, few statistical studies do not look at the growth phase of crops and plantations but use yield as a basis for the analysis of climate variability impact. The proposed model may be used for short and medium term planning to predict the future NDVI and expected yields.

For my beloved uncle and torch bearer the late Julias Mafoko Zvarevashe.

ACKNOWLEDGMENTS

I would like to acknowledge my wife Yvonne and daughter Tehila (she always told me to study) for their morale support. To my new born son Atida this one is for you. My journey in academia would not have been possible without the financial support , grooming and guidance of my parents Mr and Mrs Zvarevashe. To my relatives Prof and Mrs Mudzi, my friend and brother-in-law Marko T. and many other relatives and friends thank you for everything. Department of Mathematical Sciences you provided encouragement and support. My supervisors Dr S. Krishnannair and Prof. V. Sivakumar thank you for believing in me and for your guidance. I would like to acknowledge Aderonke Olaitan Adesina for editing this thesis.

I would like to acknowledge University of Zululand (Research office) for funding of two international conferences that attended. Lastly and not the least this journey of the research would be very difficult if it was not the support from National research Fund (NRF) scholarship.

Willard Zvarevashe, KwaDlangezwa

Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS AND ACRONYMS

AIC	Akaike information criterion
AO	Annual Oscillation
A-D	Anderson-Darling
ARMA	Autoregressive-moving average
BIC	Bayesian information criterion
DMI	Dipole Mode Index
DSLP	Darwin sea level pressure
EEMD	Ensemble Empirical Mode Decomposition
ENSO	El-Niño Southern Oscillation
EOF	Empirical orthogonal function
EEMD	Ensemble Empirical Mode Decomposition
FFT	Fast Fourier transform
GDP	Gross Domestic Product
GEVD	Generalised extreme value distribution
KNMI	Royal Netherlands Meteorological Institute
IMF	Intrinsic mode function
IPCC	Intergovernmental Panel on Climate Change
IOD	Indian Ocean Dipole
MAPE	Mean absolute percentage error
NDVI	Normalised Difference Vegetative Index
p.d.f.	Probability density function
PP	Probability-Probability plot
POT	Peaks-Over-Threshold

RMSE	Root Mean Square Error
SAO	Semi-annual Oscillation
SST	Sea surface temperatures
SOI	Seasonal Oscillation Index
QBO	Quasi Biennial oscillations
QQ	Quantile-Quantile plot
QTO	Quasi Triangular Oscillation
SAWS	South African Weather Service
SSA	Singular spectrum analysis
SSN	11-year solar cycle
UN	United Nations
VI	Vegetation Index

LIST OF NOTATION AND SPECIAL SYMBOLS

$c_t(i)$	i^{th} intrinsic mode function
$G_{\mu,\sigma,\xi}(x)$	Generalised extreme value distribution function
m_t	mean of the envelope
ϵ_j	white noise
IO	Index of orthogonality
r_t	residual
x_t	time series
μ	location of GEVD
σ	scale of GEVD
ξ	shape of GEVD
Q_p	quantile function

RESEARCH OUTPUT

A list of publication from this thesis is given below.

Journal Publications and Peer Reviewed Conference Proceedings

1. Zvarevashe, W, Krishnannair, S. & Sivakumar, V., (2019). Analysis of Rainfall and Temperature Data Using Ensemble Empirical Mode Decomposition. *Data Science Journal*, 18: 46, pp. 1–9. DOI: https://doi.org/10.5334/dsj-2019-046.

2. Zvarevashe, W., Krishnannair, S. & Sivakumar, V. (2019). Use of Satellite-Based NDVI Time Series to Detect Sugarcane Response to Climate Variability, Proc. *of 35th Annual conference of South African Society for Atmospheric Science*s, 8-9 October 2019, Vanderbijl Park, South Africa, ISBN 978-0-6398442-0-6.

3. Zvarevashe, W., Krishnannair, S., & Sivakumar, V. (2018). Analysis of Austral Summer and Winter Rainfall Variability in South Africa Using Ensemble Empirical Mode Decomposition. *IFAC-PapersOnLine*, 51(5),

132-137. DOI:10.1016/j.ifacol.2018.06.223.

4. Zvarevashe, W., Krishnannair, S. & Sivakumar, V. (2018). Annual and Seasonal Spatial Analysis of Rainfall Variability in Western Cape, South Africa, Proc. *of 34th Annual conference of South African Society for Atmospheric Science*s, 21-22 September 2018, Durban South Africa, ISBN 978-0-620-80825-5.

5. Zvarevashe, W. Krishnannair, S. & Sivakumar, V. (2017). Time Series Analysis of Rainfall Data at Cape Point Using Empirical Mode Decomposition, Proc. *of 33rd Annual conference of South African Society for Atmospheric Science*s, 21-22 September 2017, Polokwane, South Africa, ISBN 978-0-620-77401-7.

Conference Presentations

1. "Annual and Seasonal Spatial Analysis of Rainfall Variability in Western Cape, South Africa." 35th Annual conference of South African Society for Atmospheric Sciences, 8-9 October 2019, Vanderbijl Park, **South Africa**.

2. "Spatial Analysis of Winter Rainfall in South Africa." 62nd International Statistical Institute World Statistics Congress, 18-23 August 2019, Kuala Lumpur, **Malaysia**.

3. "Time Series Analysis of Rainfall Variability in South Africa Using Ensemble Empirical Mode Decomposition." The 43rd 2019 South African Numerical and Applied Mathematics (SANUM), 27-29 March 2019. University of Pretoria (UP), **South Africa.**

4. “Annual and Seasonal Spatial Analysis of Rainfall Variability in Western Cape, South Africa.” 34th Annual conference of South African Society for Atmospheric Sciences, 20-21 September 2018. University of KwaZulu Natal (UKZN), Durban, **South Africa.**

5. “Spatial and Temporal Analysis of Rainfall in Western Cape South Africa.” Southern Africa Mathematical Science Association Annual Conference, 19-22 November 2018. Botswana International University of Science and Technology (BIUST), Palapye, **Botswana.**

6. “Analysis of rainfall and temperature data using ensemble empirical mode decomposition.” International Data Week, 5-8 November 2018, Gaborone, **Botswana**

7. “Spatial Analysis of Winter rainfall variability in Western Cape, South Africa.” 60th South African Statistical Association conference, 26-29 November 2018, Johannesburg, **South Africa.**

8. “Analysis of Austral Summer and Winter Rainfall Variability in South Africa Using Ensemble Empirical Mode Decomposition.” IFAC Workshop on Integrated Assessment Model for Environmental Systems, 11-12 May 2018, Brescia, **Italy.**

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Introduction

Climate change refers to a "change in the state of the climate that can be identified by changes in the mean and/or the variability of its properties, and that persists for an extended period, typically decades or longer" (IPCC, 2014). This may be attributed to natural processes such as solar cycles, volcanic eruptions and persistent anthropogenic changes or external forcing. United Nations Framework Convention on Climate Change (UNFCCC), in its Article 1, defines climate change as: "a change of climate which is attributed directly or indirectly to human activity that alters the composition of the global atmosphere and which is in addition to natural climate variability observed over comparable time periods " (UNFCC, 1992).

Climate variability describes the changing interactions between the various components of Earth's climate system, when those changes go beyond individual weather events. This is defined by IPCC as "variations in the mean state and other statistics (such as standard deviations, the occurrence of extremes, etc.) of the climate, on all spatial and temporal scales beyond that of individual weather events" (IPCC, 2014). The variations over a short term, when they are averaged they are called "internal climate variability" and over the longer term they are called "climate change"(IPCC, 2014). The study will mainly focus on the climate variability impact using South Africa as a case study since the study is done over a short period.

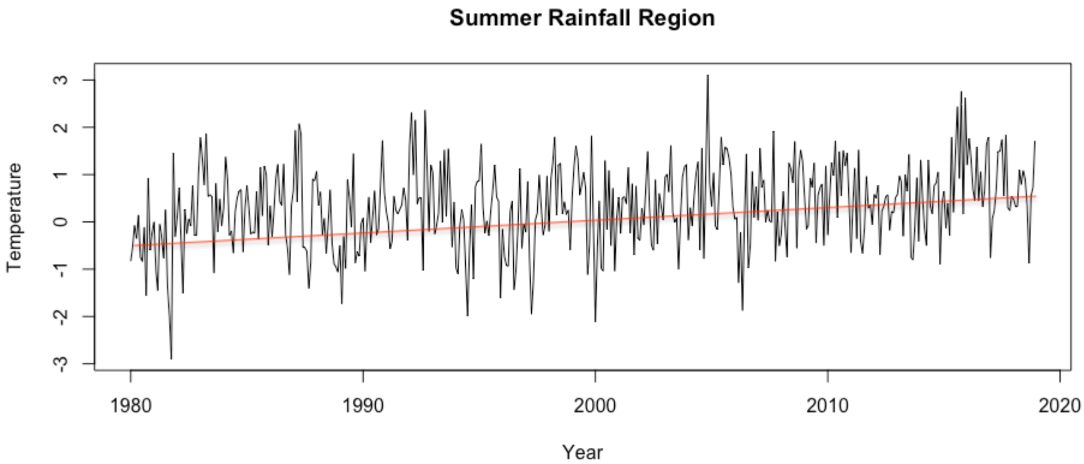
Despite the known knowledge that climate variability will have greater impact on third world countries there has been limited research on the impact it will cause. There is inadequate knowledge on socioeconomic impact, incomplete impacts modelling and lack of cross-sectorial integration on impacts and adaptation assessments (Ziervogel et al., 2014). Due to climate variability the climate variables become nonlinear and non-stationary such that use of linear methods present inconclusive results due to the high sensitivity of model results to model parameters (Molla et al., 2005). In this thesis a data adaptive methods to analyse the climate variability impact on South African weather and agriculture is done using methods that do not assume linearity or stationarity of the climate variables.

In the next sections of this chapter the background of the study, the statement of the problem, research aim, objectives and the significance of the study; the contribution of the study and the thesis layout, are all presented.

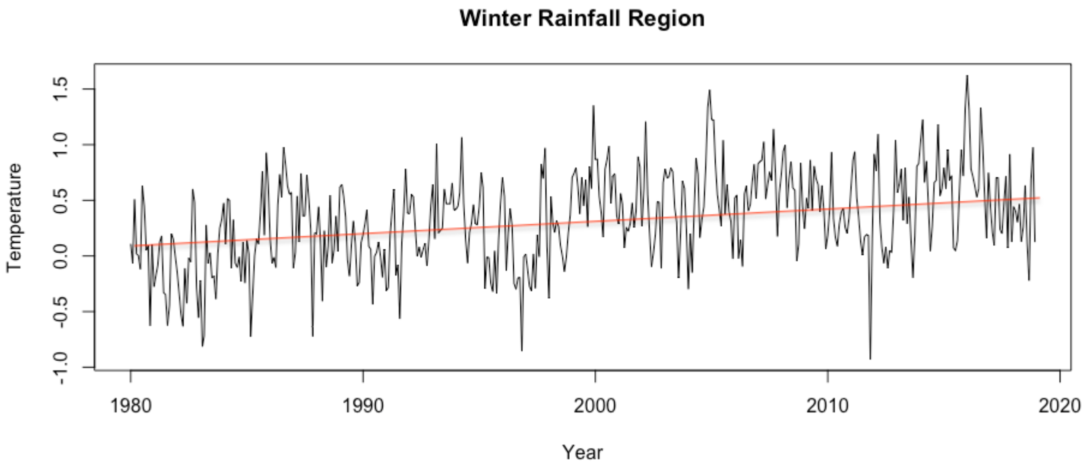
1.2 Background

The increase in economic activity and population growth has resulted in corresponding increase in greenhouse gas emissions, which has in turn, led high concentration of carbon dioxide (CO_2), methane and nitrogen oxide (WHO, 2015). This and other natural factors has resulted in increase of temperatures, at an average of 0.65ºC per decade (Ziervogel et al., 2014; Maúre et al., 2018). Over the past five decades, temperature in South Africa has increased by over one and half times more than the global observed increase in temperature increases (Ziervogel et al., 2014). Historical temperature studies have revealed a positive trend in the average temperatures for the whole of South Africa, with the interior areas recording higher percentage increase compared to coastal regions (Kruger & Shongwe, 2004). The season with highest mean temperature changes is autumn (March to May) and the least is spring (September to November) for the period 1960 to 2003. Fig. 1.1 shows the average temperature trend for South Africa from 1980 to 2018 for two regions, one in the summer rainfall region and the other in the winter rainfall region. The average temperature data extracted from National Center for Environmental Information was used to plot the graphs (Zhao & Li, 2015). The plots show that there has been an increasing trend of the temperature in both regions over the years from 1980 to 2018.

The effects of climate change have been felt in water resources, food security, hydropower and human health (Aggarwal et al., 2006; Kanga et al., 2009). It has the potential to amplify problems of human security on food, health, economic insecurity, ecosystem and water resources. Africa is the most vulnerable continent in the world, due to many factors such as over reliance on rain-fed crop production (98% in sub-Saharan Africa) and poverty, limited capacity to adapt, and low research (Niang et al., 2014). Climate change has immediate impact as a result of extreme weather patterns (climate variability) and longer term impacts such as shortage of food and increase in infectious disease (WHO, 2015). World Health Organisation (WHO) reported that there were between 80-120 deaths per million, as a direct result of climate change in Southern Africa as compared to 0-2 in China and United States of America (WHO, 2002). Given the adverse effect of climate variability on human livelihoods, it becomes necessary to examine the extreme weather patterns and its impact (Gbetibouo & Hassan, 2005).



(a) Time Series plot of temperature for summer rainfall region



(b) Time Series plot of temperature for winter rainfall region

Figure 1.1: Time series plot of standardised temperature trend in South Africa from 1980 to 2018 for the two selected regions which receives summer and winter rainfall.

The weather patterns in Africa are highly volatile, with large spatial variability over different regions and South Africa has not been spared from this variability (Jury, 2013). Rainfall trend analysis, on different spatial and temporal scales, has been of great concern, mainly because of the climate variability (Longobardi & Villani, 2010). In Southern Africa, some regions are receiving less water whilst others are receiving more rainfall (Jury, 2013). In this study, trend analysis and spatial variability maps are presented for Western Cape province. This is done in order to select a study area with similar rainfall patterns. The greatest restriction on the meteorological studies is the availability of data over a very long period (MacKellar et al., 2014). Fortunately, South Africa has the advantage of having a good station network, when compared with the rest of Africa, which makes it possible to carry out studies based on the actual data collected from the stations (Hughes & Balling Jr, 1996).

South Africa experiences different types of rainfall throughout the country, with most of the country receiving austral summer rainfall (November to March) and the south western parts of the country receiving austral winter rainfall (May to August) (Dieppois et al., 2016). The map in Fig. 1.2 shows the different rainfall seasons in South Africa (Schulze, 1997). The northern part of South Africa receives summer rainfall. The summer rainfall is further divided into early, mid, late and very late summer. South western areas of the country, which covers parts of Western Cape and parts Eastern Cape provinces receive winter rainfall. A very small portion found on the southern coastal area receives rainfall throughout the year. In recent years, the country has been experiencing draughts and floods. All of these add to the unpredictability of the weather patterns (Malherbe et al., 2016). A study, based on the period 1960 to 2010, conducted by MacKellar et al. (2014) indicated that there were drier conditions along the southern coastal regions and an increase of rain days in the west coast of the Western Cape province. Another study carried out for the period 1921 to 2015

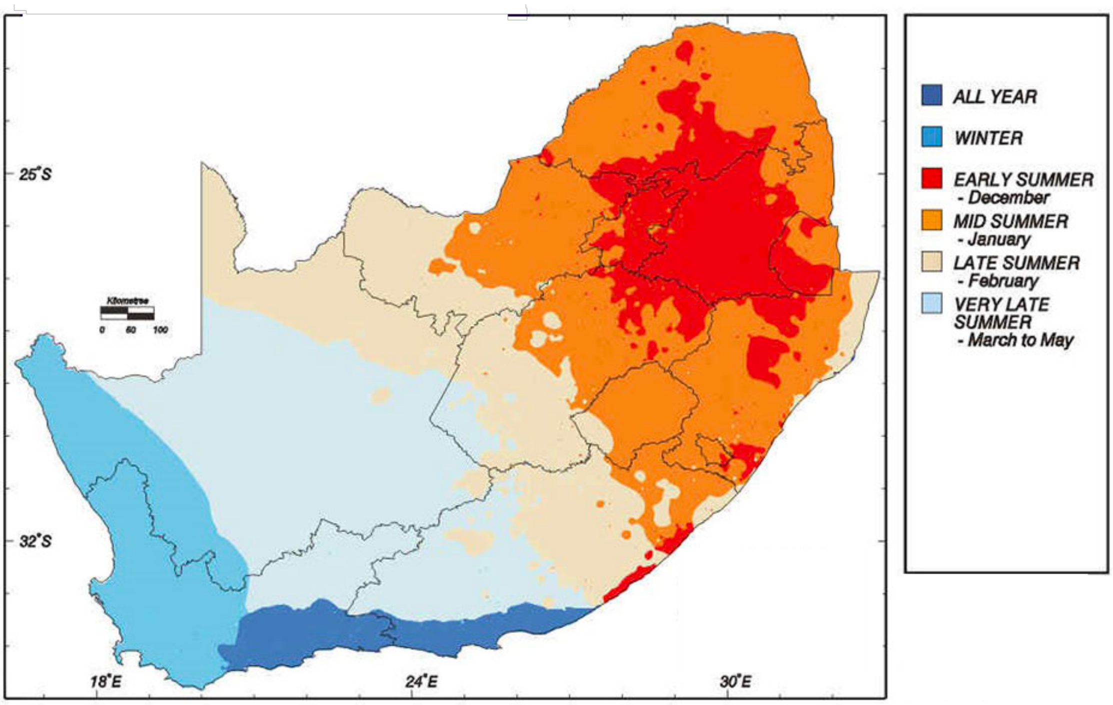


Figure 1.2: Rainfall Seasons in South Africa.
(Source: Schulze (1997))

by Kruger & Nxumalo (2017)showed that there is rainfall increase of more than 2.5mm per decade for the June-July-August season for the regions located in the Western Cape province.

A research by Kruger (2006) showed that few regions in South Africa had noticeable changes in the precipitation for the period from 1910 to 2004. According to the study, in most areas, there were no considerable changes in the average amount of rainfall per year. However, it was observed that there was a lot of fluctuations in the rainfall which is characterised by high inter-annual variability (50% to 200% of the mean) and long-term fluctuations (75% to 150% of the mean) for five-year moving average. A longer 21-year moving average did not indicate linear trends, either upwards or downwards (Kane, 2009; Malherbe et al., 2016). Given the high variability of rainfall, it is therefore, important that statistical method must be developed or applied, that can predict the patterns with minimal error.

The primary driver of South Africa's inter-annual climate variability has been the Quasi Biennial oscillations (QBO), El-Niño Southern Oscillation (ENSO), Indian Ocean Dipole (IOD) and 11-year solar cycle (SSN). Both summer and winter rainfall have been shown to be influenced by ENSO (Fauchereau et al., 2009; Kane, 2009; Philippon et al., 2012). El Niño describes the warming of sea surface temperature that occurs periodically, typically concentrated in the central-east equatorial Pacific (Lehodey et al., 1997). La Niña is the term adopted to describe the opposite side of the fluctuations. These climate variables are non-stationary and have nonlinear effect on the rainfall patterns (Deressa et al., 2005; Schulte, 2016). A stochastic process is said to be stationary if the parameters mean and variance do not change over time (Molla et al., 2005).

Linear statistical models, such as Auto Regressive Integrated Moving Average (ARIMA), Markov Chains and linear regression have been applied to climate data: however, they could not predict with high percentage of certainty due to the extreme climatic conditions (Molla et al., 2005; Tettey et al., 2017; Unnikrishnan & Jothiprakash, 2018). Fast Fourier Transform (FFT) method and other Fourier transform-based methods have been used to analyse climate data. However, they assume that the data is linear and the data must be strictly periodic, which is not the case with climate data. Wavelet Analysis was adopted to overcome the shortcomings of FFT; however, it sometimes fails to distinctly define local frequency changes (Coughlin & Tung, 2005; Schulte, 2016). There are abrupt climate shifts that occur many times, which have ecological and economic impact, making it is essential to understand and quantify these nonlinear events (Schulte, 2016).

In the second assessment report of the IPCC the issue of more research on observed historical trends in climate extremes was raised (IPCC, 2014). Studies with historical data would give more credibility to the models, as they would be verified against available data that could be used for future projections. Since then there has been a lot of research with focus on extreme weather (Nicholls et al., 1996; Hartmann et al., 2013). Generalised logistic probability distribution function was used to analyse the extreme annual maximum daily rainfall in Nelspruit, South Africa and its environs (Masereka et al., 2018). It was shown that the return period of floods in the area was 10 years, while the occurrence probability of flood disaster event (>100mm) at least once in 1, 2, 3, 4 and 5 years was 0.10, 0.19, 0.27, 0.34 and 0.41, respectively. Another study on extreme temperatures in South Africa used the generalised extreme value distribution (GEVD) for r largest order statistics which was fitted to the average maximum daily temperature for the non-winter season (Nemukula & Sigauke, 2015).

Ensemble empirical mode decomposition (EEMD) is a data-adaptive method that is best suitable for data that is oscillatory and maybe nonlinear and non-stationary characteristics such as climate data (Huang et al., 1998; Liu et al., 2019b). The time series is decomposed into different time scales called intrinsic mode functions (IMFs), which can reveal intrinsic changes in the climate system. EEMD will be used to decompose the rainfall and other times series in this study.

Generalised extreme value distribution (GEVD) has been found to be best suitable for hydrological and meteorological problems (Maposa et al., 2014; Chikobvu & Chifurira, 2015; Boudrissa et al., 2017; Namitha & Ravikumar, 2018). Additionally, it has been applied in modelling the rainfall pattern in South Africa (Bhagwandin, 2017). Previous studies have shown that methods coupled with decomposition perform better than single methods (Zhao & Chen, 2015; Zhu et al., 2016; Wang et al., 2018). Methods that have been coupled with EMD/EEMD are; ARIMA, artificial neural network, radial basis function, support vector machine, least square support vector machines, Auto Regressive and generalised S-transform. In this study, the rainfall time series will be decomposed using EEMD and modelled using GEVD.

Climate variability has a direct impact on the agricultural production; therefore an understanding of the trend and periods is very important for future planning and sustainability (Linnenluecke et al., 2018). The agriculture sector has not been spared from rainfall and temperature variability. Several previous studies used experimental analyses, agro-ecological models and process-based dynamic crop growth models and a few used historical data statistical analysis to investigate the climate variability impact (Kumar & Sharma, 2014; Jones et al., 2014; Zhao & Li, 2015). Furthermore, a number of the statistical methods use yield or production as a basis of analysis of the impact of climate variability and do not look at the growth phase of the plants (Linnenluecke et al., 2018). Therefore, in this thesis, Normalised Difference Vegetative Index (NDVI), which shows the crop progression and has positive correlation with expected yield, EEMD and synchronisation is used to analyse the climate variability impact on grapevines and sugarcane.

1.3 Statement of the Problem

South Africa's economy is vulnerable to the variability of weather and climate, as it heavily depends on agriculture. The severe impact of climate variability is due to minimal resources, which are not adequate to deal with conditions that come with extreme weather conditions. Hence, there is need to model monthly and seasonal rainfall and temperature patterns, with the aim of determining the climate variability impact and the extreme patterns. A number of existing statistical models make several assumptions; such that inconclusive results are found. Additionally, there are few statistical studies that have been done for South African climate variability. Therefore, there is a need to use methods that are suitable for South African weather conditions. The analysis can then be used for better preparedness and planning.

1.4 Aim and Objectives

The aim of the study is to develop a modelling framework that can be used to analyse the climate variability of South Africa and its agricultural impacts.

The aim of the research will be achieved through the following objectives:

i. To carry out a spatial variability analysis of rainfall for selected regions in South Africa;

ii. To find the optimal amplitude of noise to be added in the EEMD method to minimise the redundancy of IMFs;

iii. To investigate the quasi-biennial and El-Niño Southern Oscillations' impact on rainfall and temperature at different time scales;

iv. To fit generalised extreme value distribution to the generated decomposed monthly rainfall data and;

v. To find the climate variability impact of rainfall and temperature on agriculture, at different time scales, using data adaptive methods and synchronisation.

1.5 Significance of the Study

The climate variability in South Africa directly and indirectly affect human livelihoods and ecosystem through droughts, temperature changes, water supply and agricultural production (Dieppois et al., 2016). Despite the devastating impact of climate variability, there is still limited quantitative research on its impact on the agricultural sector and the economy. This study provides the first statistical analysis of decomposed rainfall and temperature, using ensemble empirical mode decomposition and generalised extreme value distributions, in South Africa. The decomposed data provides a more robust GEVD model. It aims to enhance the understanding of rainfall and temperature patterns, in order to reduce the impact of extreme weather patterns on the agricultural sector. Knowledge of rainfall patterns can give an early warning of droughts or floods in the coming season, for better preparedness, particularly in terms of crops to be planted, water harvesting and disaster management. In this study, an optimal amplitude of noise to be used in ensemble empirical mode decomposition is also proposed.

1.6 Contributions

The main contribution of the thesis is to apply statistical techniques to analyse and model the mean rainfall and temperature for selected regions in South Africa. The specific contributions are:

i. Presentation of the rainfall spatial patterns for Western Cape at annual, seasonal and monthly level, which may be used for planning and management.

ii. Offering, for the first time, to the best of our knowledge, the use of synchronisation to investigate the impacts of ENSO and QBO on rainfall at different time scales.

iii. Modelling of the decomposed rainfall time series (IMFs), using GEVD, thereby having a model that combines EEMD and GEVD to provide a better fit for a selected region in Western Cape. These findings may not only be important to Western Cape, but also to other areas, as it can be generalised and applied to other areas with similar climatic conditions.

iv. Investigation of rainfall and temperature variability impact on the sugarcane's and grapevines' NDVI, using EEMD and synchronisation.

v. Modelling of the environmental factors impact on sugarcane and grapevines.

vi. Future perspectives.

1.7 Thesis Layout

The rest of the thesis is organised as follows: **Chapter** 2 reviews the linear, transform based and data adaptive climate modelling methods. The strengths and shortfalls of the modelling methods is done. A review of previous studies, which used the empirical mode decomposition and generalised extreme value distribution is done.

A discussion of the methodology adopted for this thesis is done in **Chapter** 3. The methods include spatial analysis, empirical mode decomposition, ensemble empirical mode decomposition, and generalised extreme value distribution. Furthermore, the methods for model diagnostic, selection and performance are also discussed.

Results and discussion are divided into two chapters with **Chapter** 4[1234] containing the spatial analysis and time series decomposition. The first section of the chapter discusses the results from the spatial analysis of the study area. Results of the decomposition of the climate data, using the empirical mode decomposition and ensemble empirical mode decomposition, are provided in the sections that followed. The climate variability impact, at different time scales, is shown. A proposed method to determine the amplitude of noise to be added, and the modelling of the decomposed data results, using generalised extreme value distribution are also presented.

Chapter 5[5] contains the second part of the results and discussion, with the main focus on the climate variability on agriculture. Using sugarcane and vineyards as case studies, rainfall and temperature variabilities are investigated at different time scales.

Chapter 6 provides the summary and recommendations of the study.

The main statistical package used throughout this thesis is R. Other software that were used are MATLAB and ArcGIS.

[1]Analysis of Rainfall and Temperature Data Using Ensemble Empirical Mode Decomposition. *Data Science Journal*, 18: 46, pp. 1–9. DOI: https://doi.org/10.5334/dsj-2019-046.

[2]Analysis of Austral Summer and Winter Rainfall Variability in South Africa Using Ensemble Empirical Mode Decomposition. *IFAC-PapersOnLine*, 51(5), 132-137. DOI: 10.1016/j.ifacol.2018.06.223.

[3]Annual and Seasonal Spatial Analysis of Rainfall Variability in Western Cape, South Africa, *Proc. of 34th Annual conference of South African Society for Atmospheric Science*s, Durban, (South Africa), ISBN 978-0-620-77401-7, 21 -22 September 2018.

[4]Time Series Analysis of Rainfall Data at Cape Point Using Empirical Mode Decomposition, Proc. *of 33rd Annual conference of South African Society for Atmospheric Science*s, 21-22 September 2017, Polokwane, South Africa, ISBN 978-0-620-77401-7.

[5]Use of Satellite-Based NDVI Time Series to Detect Sugarcane Response to Climate Variability, *Proc. of 35th Annual conference of South African Society for Atmospheric Science*s, Vanderbijl Park , (South Africa), ISBN 978-0-620-77401-7, 8-9 October 2019.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The interaction between Earth's system results in the variability of rainfall and temperature. Begue et al. (2010) found the following climate variables to be dominant in influencing the surface temperature of South Africa: Annual Oscillation (AO), Semi-annual Oscillation (SAO), Quasi-biennial oscillations (QBO), El-Niño Southern Oscillation (ENSO), Indian Ocean Dipole (IOD) and 11-year solar cycle (SSN). Trend run method, a linear regression-based method, was used to find the important climate oscillations for Southern Africa. The IOD had minimal effect on the temperature change. The IOD is commonly measured by Dipole Mode Index (DMI), which is the Tropical Pacific sea surface temperatures (SST) anomaly difference between the western and eastern tropical Indian Ocean. Rainfall variability in South Africa was linked to SST and SO anomalies (Kane, 2009; Malherbe et al., 2016). QBO and Quasi-triangular oscillation (QTO) accounted for 30%-50% variability in the rainfall pattern. QBO is 2-3 year oscillation period and QTO is 3-4 year oscillation period.

Changes in temperature and rainfall patterns have a direct impact on crop production especially if the crops are rain-fed, which is the case in most parts of Africa (Niang et al., 2014). The increase of temperature that has been observed over the past decades has a direct impact on both animals and human beings through shortage of water resources, increased health risks, food security and others (Kanga et al., 2009). Recently, there has been increase in extreme rainfall patterns, with corresponding increase in the frequency of droughts (Malherbe et al., 2016; Masereka et al., 2018). Previous studies have revealed that this increase of extreme rainfall patterns is closely related to the increasing temperatures in Southern Africa (Maúre et al., 2018; Nangombe et al., 2018; Nikulin et al., 2018). Given the direct impact of increasing temperature and extreme rainfall patterns it is therefore important to use historical data to forecast future events (IPCC, 2014). The models developed using historical data can easily be evaluated using existing data, and the outcome can be used for future projection (Hartmann et al., 2013). The weather patterns have become very unpredictable, nonlinear and non-stationary, such that existing standard methods may fail to analyse and forecast future events (Molla et al., 2005; Liu et al., 2019b). Some of the methods that have been applied to climate variability modelling are ;Autoregressive-moving average, linear regression, Markov Chains, Fourier based methods, Wavelet analysis and Singular spectrum analysis. These methods can either make assumptions, or pre-process the data. A discussion of these methods is done in sections 2.2 and 2.3 below.

2.2 Linear Methods for Climate Variability Modelling

A number of climate modelling methods make assumptions about climate data, linearise or difference the data, so that a statistical method can be applied. Examples of these methods are autoregressive-moving average (ARMA), linear regression and Markov Chains. The unit root and fractional unit root/integration processes methods were applied in the study of rainfall and temperature in South Africa and 5 other African countries (Gil-Alana et al., 2018). However, the study did not fully consider the nonlinearity of the rainfall and temperature data. Trend run - a multi-linear regression based-model, was used to model the impact of atmospheric force on temperature (Begue et al., 2010). The model showed the influence of annual and semi-annual oscillation, QBO, ENSO and 11-year solar cycle (SSN) on Reunion temperature. However, the model did not consider the nonlinear and non-stationary characteristics of the climatic variables.

A non-parametric method, Markov Chains, has been applied in many fields and rainfall and temperature prediction are some of them. A monthly rainfall assessment for a 30-year period for 5 cities in Ghana was done using Markov chains (Tettey et al., 2017). The probability of rain and the general trend for the 5 cities was established. Using Markov chains a prediction of annual rainfall was also done in Nigeria and 4-6 states of rainfall were estimated (Yusuf, 2014; O Ibeje et al., 2018). However, the Markov chains do not take into consideration the full chain of the prior states and they sometime take long to converge (Mossel & Vigoda, 2006).

Stochastic modelling has gained attention in hydraulic time series modelling. Some of the common stochastic models are: Auto Regressive (AR), Moving Average (MA), Auto Regressive Moving Average (ARMA) and Auto Regressive Integrated Moving Average (Unnikrishnan & Jothiprakash, 2018). ARMA was used to model the Darwin sea level pressure (DSLP)(Trenberth & Hoar, 1996). DSLP is used to monitor the El-Niño-La-Niña events. For the period 1882-1995, the ARIMA model (3,1) was found. The research provided preliminary evidence that greenhouse emissions were causing global warming. Harrison & Larkin (1997) used ARMA, student's t-test and Bootstrap/Monte Carlo to model the strong El-Niño event of 1990-1995. The authors provided evidence of climate change and that it was not an entirely natural occurrence. However, the ARMA model for the DSLP assumes stationarity and linear dynamics underlying ENSO. In the study, they predicted that the 1990-1995 events would occur once every 150-200 years at 95% confidence interval. The research was inconclusive as it did not find enough evidence to link greenhouse emissions to the 1990-1995 occurrence.

Rainfall is unpredictable and almost random; therefore, it would be difficult, using linear models to make predictions to any reasonable degree of accuracy (Kane, 2009; Malherbe et al., 2016). Linear methods, such as ARIMA use smoothing and differencing to generate stationary data, and in the process, some information on the trends of the data is lost; particularly in the case of highly non-stationary climatic data. In order to achieve linearity; transformation can be done before finding the ARIMA model. This was done in the study on the effects of air temperature and precipitation on streamflow in glacier mountain, China (Liu et al., 2015). The study managed to find that, there was an effect of climate change but the extent and periods could not be quantified. The stochastic and linear models assume that the time series is stationary and follows a normal distribution, whereas climatic data in most cases is nonlinear and non-stationary. Therefore, it is difficult to predict peak values (Unnikrishnan & Jothiprakash, 2018). The results found from linear and stochastic models are usually inconclusive (Huang et al., 1999; Molla et al., 2005).

2.3 Multiple Scale Methods for Climate Variability Modelling

Multiple scale modelling methods either use existing functions or decompose the data to analyse it at different time scales. Given the characteristics of climate, the transform based methods, such as Fourier transform and Wavelet transform, are very common. Maximum Entropy method and Fourier transform was used to detect all possible peaks in South African rainfall data for different regions (Kane, 2009). In the study, 2-3 year period (QBO) with 30% - 50% variance contribution and 3-4 year period (QTO), 11-year sunspot cycle, 20, 35 and 65 years, all with less than 10% variance contribution were identified. However, traditional Fourier transform-based techniques tend to spread the signal energy into multiple bins, which sometimes leads to misinterpretations of the data. Moreover, the underlying conditions that go with Fourier transform is that, the system must be linear and stationary. Furthermore the data must be strictly periodic and requires large data samples (Coughlin & Tung, 2005).

Wavelet analysis was developed in the 1980s, and ever since it has been used extensively in signal processing, climate modelling and other disciplines. It can identify the fluctuations that are multi-scale in nature. The idea behind wavelet analysis is to convolve a time series with a function satisfying certain conditions. Such functions are called wavelets, of which the most widely used is the Morlet wavelet, a sinusoid damped by a Gaussian envelope (Torrence & Compo, 1998). It is capable of localising the variability of a time series in both time and frequency domains (Torrence & Compo, 1998; Varikoden et al., 2016). Malherbe et al.

(2016) used wavelet analysis to identify statistically significant regional climate variation with which draughts is associated. They explored the occurrence of draughts over South Africa, from 1920 to 2014, using Standardised Precipitation Index. However, Wavelet transform requires the choice of a mother wavelet, and hence it is not data adaptive. Moreover there is also a problem of shift variance. If the starting point is varied, it produces completely different results (Yiou et al., 1996; Coughlin & Tung, 2005; Tang et al., 2012). Given the unpredictability of the climate, data transform based methods may sometimes fail to predict extreme weather conditions. Therefore, data driven methods, which are fully driven by the data itself, may be more suitable.

Empirical orthogonal function (EOF) analysis is a decomposition of a signal or data set in terms of orthogonal basis functions which are determined from the data (Zhang & Moore, 2015). It is often used to study the spatial patterns of climate variability and how they change with time. EOF depends on spatial decompositions to filter noisy data. Similar to what is done in Fourier analysis and wavelet analysis, the original climate data is projected on orthogonal basis in EOF analysis. However, the orthogonal basis is derived by computing the eigenvectors of a spatially weighted anomaly covariance matrix, and the corresponding eigenvalues provide a measure of the percentage variance explained by each pattern (Zhang & Moore, 2015).

EOF was applied to a study to find the relationship between rainfall and El Niño in Oman (Varikoden et al., 2016). In most of the years when there was El Niño there was increase in rainfall. Given the unpredictability of the climate, data transform based methods may sometimes fail to predict extreme weather conditions. Therefore, data driven methods, which are fully driven by the data itself, may be more suitable. Moreover, they could not attribute the variabilities found to other atmospheric or oceanic variables (Varikoden et al., 2016). Climate oscillations for a longer period usually have smaller variance that is often grossly misrepresented or totally missing in the leading EOFs (Mak, 1995; Schulte, 2016). Another well-known shortcoming of EOF analysis is that the eigenvectors are constrained by their mutual orthogonality and the maximisation of variance over the entire analysis domain (Hurrell, 2015). Therefore, they do not represent the actual physical modes of the climate system. Moreover, the loading values of EOFs do not reflect the local behavior of the data: values of the same sign at two different spatial points in an EOF do not imply that those two points are highly correlated. This implies that the pattern structure of any EOF must be interpreted with care.

Pattern recognition of a time series becomes very important in cases where there is need to study the influence of other factors. Therefore, decomposing the time series into a number of components that are interpretable can help in modelling the time series in a better way (Unnikrishnan & Jothiprakash, 2018). Singular spectrum analysis (SSA) is a fully data-driven method of analysis that is more efficient than Fourier transform based methods in detecting periodic feature in a time series (Wilks, 2011). The periodic and quasi-periodic feature in SSA is represented by pairs of eigenfunctions and principal component. SSA is data adaptive and can extract intermittent periodicities that are not sinusoidal in nature (Wilks, 2011). SSA was utilised in forecasting the daily rainfall time series in Maharashtra, India (Unnikrishnan & Jothiprakash, 2018). This was done by decomposing the time series into trend, periodic, noise and cyclic components, which were forecasted separately and added together. SSA was used to forecast annual precipitation, monthly runoff and hourly water temperature time series (Marques et al., 2006). The method was also reported to be very successful in extracting various components of the time series with prior knowledge of the time series' trend (Alexandrov, 2008). However, SSA sometimes fails to fully separate the signals found in the time series (Lu & Saniie, 2016). Therefore, another data adaptive method, empirical mode decomposition (EMD) turn to perform better in terms of extraction of signals.

EMD was introduced by Huang et al. (1998) and does not make assumptions about linearity and stationarity. The intrinsic mode functions (IMFs) derived from the data are usually easy to interpret and relevant to the underlying physical system. EMD is a nonlinear technique that adaptively represent non-stationary time series data in different time scales (Huang et al., 1998; Molla et al., 2005, 2007). It only requires that the data must consist of simple intrinsic mode of oscillations (Huang et al., 1998).

2.3.1 Review of Relevant Studies on EMD

EMD method has had many applications since its inception. The technique has found application in ocean wave data analysis, analysis of seismological data, diagnosis of the rate of flactuations of heartbeats, and signal image processing (Huang et al., 1999; Zhang et al., 2003; Balocchi et al., 2004; Linderhed, 2002). Chiew et al. (2005) applied it to annual stream flow to find important oscillations in the data. Bootstrapping was used to test the significant of the identified trends. Twenty unimpaired catchments from different parts of the world were used and some important observations were made on the effect of climatic oscillations in the stream flow. Two-dimensional EMD, which was applied to image compression, was first introduced by Linderhed (2002). In two-dimensional EMD instead of decomposing data into IMFs it is decomposed into intrinsic mode surfaces (IMSs). It was also applied to Bethlehem (South Africa) rainfall nowcasting (present weather conditions and forecast of those immediately expected) to separate high frequency components from the low frequency ones in the data. This was done by removing the noisy high frequency components, which do not exhibit a strong temporal correlation and add little structural information to nowcasting algorithms (Sinclair & Pegram, 2005). The natural variations in the climate were identified using EMD, which was cumbersome and difficult to identify using traditional Fourier techniques. The climate trend was broken down into only 5 modes, with 4 of them having positive correlations to the known climate parameters. These were AO, QBO, ENSO, SSN and the forth unknown one (Coughlin & Tung, 2005). Climate variability was also noticeable in the rainfall data in India between 1989 to 2004. EMD method showed that global warming was contributing to more frequent long lasting draughts and floods in India (Molla et al., 2007).

EMD has a challenge of “mode mixing”, where a single IMF contains multiple components at different frequencies, which is not representative of any signal. To overcome mode mixing Huang et al. (1999) used intermittence test, which is based on scales subjectively selected. With this subjective selection, the EMD ceases to be totally adaptive and works when there are clearly separable and definable time scales in the data. Therefore, it does not work very well as majority of signals have scales that are not clearly separable. To overcome these challenges, a data assisted method, Ensemble Empirical Mode Decomposition (EEMD), which consists of adding noise to the data before carrying out the EMD algorithm, was developed (Wu & Huang, 2009). The noise will be cancelled out as it takes advantage of the zero mean of the noise, once it has served its function.

2.3.2 Review of Relevant Studies on EEMD

EEMD has been applied widely in hydrological and atmospheric studies. Chiew et al. (2005) applied EMD to annual stream flow to find important oscillations in the data. Twenty unimpaired catchments from different parts of the world were used and 3, 6–7, 11–15 and 20–25 year oscillations in the stream flows were identified. In another study, the climate variability was observed in the case study of East and Central Africa, which was demonstrated by mapping the coupling between precipitation variables, inter-annual vegetation changes and the ENSO and Indian Ocean Dipole (IOD). EEMD was adapted because of its ability to break down normalised vegetation index (NDVI) series into multiple time scales components and its generic ability to be used with other time series analysis tools (Hawinkel et al., 2015, 2016). EEMD has been applied to rainfall analysis in a number of studies. In Australia, using 44 stations for a selected region, cycles of between 3-20 years were identified (Srikanthan et al., 2011). A follow up study found the footprints of dominant such as SOI, IOD and Tasman Sea Index, in the oscillations identified (Peel et al., 2011). The Pacific Decadal Oscillation (PDO) is known to have an influence over East China's annual and summer rainfall. Using EEMD, the monthly rainfall influence of PDO was identified (Yang et al., 2016). The intrinsic oscillations in the land surface temperature of Wuhan, China were revealed by using EEMD to decompose the data into annual, inter-annual, noise and trend (Liu et al., 2019b). In South Korea, several climate drivers have been shown to have an impact on the precipitation; however, many studies did not take into consideration the inherent cycles in the long term precipitation (Kim et al., 2018). The EEMD is a data adaptive method that also considers the cycles in any data variable. Using EEMD, cross correlation and multiple linear regression on the influence of ENSO, QBO, Arctic Oscillation, Atlantic Meridional Mode and others, was shown at monthly level (Kim et al., 2018). Some notable applications of EEMD include finding the effects of temperature and precipitation trends in Plateau droughts (Sun & Ma, 2015) and identifying the variability of monthly precipitation in Iran (Alizadeh et al., 2019).

Although EEMD has found a lot of applications in many fields, there are issues with the amplitude of the noise to be added to the original data. If the added noise is too small, relative to the the original data, there will not be any improvement on the mode mixing. In South Korea, several climate drivers have been shown to have an impact on the precipitation; however, many studies did not take into consideration the inherent cycles in the long term precipitation (Tabrizi et al., 2014).

2.3.3 EMD/EEMD Hybrid Methods

Recently, EMD and EEMD have been used, as a hybrid method with other known methods. Some of the them are; EMD-Artificial Neural Network (ANN), EMD-Radial Basis Function (RBF), EMD-Support Vector Machine (SVM), EMD-Least Square Support Vector Machines (LS-SVM), EMD-Relevant Vector Machine (RVM) methods, EMD-Auto Regressive, and EEMD-generalised S-transform (Huang et al., 2014; Zhao & Chen, 2015; Zhang et al., 2015; Liu et al., 2019a). The prediction accuracy of these hybrid forecasting methods has been improved, depending on the area of application.

2.3.3.1 EMD-Support Vector Machine

A model that combines EMD and support vector machine was proposed by Huang et al. (2014) to predict the streamflow in a certain river in China. The streamflow time series is decomposed into IMFs; and residual, using EMD and IMF1, is subtracted from the original data. IMF1 is known to contain a lot of noise captured in the time series. A comparison of the performance of other models, using root mean square, mean absolute error and mean absolute percentage error showed that EMD-SVM was more superior. Another study used EMD-SVM and discrete wavelet transform-SVM and this improved the predictability of monthly streamflow as shown by the improvement of RMSE, MAPE and MAE (Zhu et al., 2016). The study, concluded that the models coupled with decomposition, performed better than those without. Similar conclusion was made for a study to forecast monthly rainfall at a weather station in China (Ouyang et al., 2016).

2.3.3.2 EEMD-Auto-Regressive

A hybrid of EEMD and Auto-regressive (EEMD-AR) model was proposed by Zhao & Chen (2015) and tested for river runoff forecasting for hydraulic stations. The method consist of firstly, decomposing the runoff data into IMFs and a trend term, using EEMD to make the time series stationary. Secondly, these IMFs are forecasted using an optimum auto-regressive (AR) model and a quadratic polynomial equation is used for the residual. The AR model of order p that was used is given by:

$$x_t = \mu + \varphi_1(x_{t-1} - \mu) + \varphi_2(x_{t-2} - \mu) + \cdots + \varphi_p(x_{t-p} - \mu) + \epsilon_t$$

where ϵ_t is the residual term and $\varphi_1, \varphi_2, \ldots, \varphi_p$ are the parameters of the AR(p) model. Lastly, the predicted results from the AR models and the quadratic model are summed together to formulate an ensemble forecast for the original runoff time series. The EEMD-AR model was found to perform better than the AR and EMD-AR model (Zhao & Chen, 2015).

2.3.3.3 EMD-ARIMA

EMD-Autoregressive Integrated Moving Average (EMD-ARIMA) was proposed and applied in streamflow prediction (Wang et al., 2018). The non-stationary monthly streamflow time series, was decomposed into IMFs and taking advantage of the stationarity of the IMFs, these were modelled using ARIMA. The individual forecast from these models is then summed up together to predict the overall expected streamflow.

In this study, the rainfall data will be decomposed into a number of IMFs and modelled using GEVD. GEVD has been found to be best suitable for hydrological and meteorological problems, and EMD/EEMD is best suitable for nonlinear and non-stationary data. Therefore, motivated by the performance of the hybrid methods some of them reviewed in section 2.3.3, in this thesis a combination of EEMD and GEVD model will be proposed. Additionally, it has been shown that models coupled with decomposition perform better than the single models. In the next section, a discussion of the generalised extreme value distribution is done.

2.4 Generalised Extreme Value Distribution

The generalised extreme value distribution (GEVD) consist of a family of continuous probability distributions that were developed within extreme value theory and combines the distributions Gumbel, Fréchet and Weibull, which is also known as type I, II and III extreme value distributions (Coles, 2001). The GEVD is given by:

$$G_{\sigma,\mu,\xi}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right]\right\} \tag{2.1}$$

defined on $\{x : 1+\xi(x-\mu)/\sigma > 0\}$ with $-\infty < \mu < \infty$ and $-\infty < \xi < \infty$, where μ,σ and ξ are location, scale and shape parameters respectively. The value of the shape parameter ξ distinguishes the type of extreme value distribution, that is

$$G(x) = \begin{cases} \text{Gumbel}, & \text{if } \xi{=}0 \\ \text{Fréchet}, & \text{if } \xi{>}0 \\ \text{Weibull}, & \text{if } \xi{<}0 \end{cases} \tag{2.2}$$

(Coles, 2001).

A number of studies in hydrology and climatology have revealed that GEVD is the best model to fit extreme weather conditions (Maposa et al., 2014; Chikobvu & Chifurira, 2015; Boudrissa et al., 2017; Namitha & Ravikumar, 2018). The modelling of maximum rainfall for Zimbabwe from 1901 to 2009, using Anderson Darling goodness-of-fit test revealed that the simpler Gumbel Distribution was the most suitable. Some of the key findings included the return period of extreme floods, such as 1193mm, with return period of 300 years (Chikobvu & Chifurira, 2015).

Nemukula & Sigauke (2015) fitted a Generalised Pareto Distribution (GPD) on average daily minimum temperature in South Africa. The study showed that a Weibull family of distribution is appropriate in modelling the upper tail of the distribution. However, the QQ plot revealed that there were a number of outliers. Univariate and multivariate extreme theory were used to model rainfall data in Western Cape province using 5 stations. The block maxima approach did not perform well in modelling the weather variables for the stations, in both the univariate and multivariate, when compared to the threshold excess and point process approaches. The study assumed that the rainfall data was stationary (Bhagwandin, 2017).

2.5 Climate Variability Impact on Agriculture

The agriculture sector has not been spared from rainfall variability and warming climate. Driven by the growing concern about climate variability impact on agricultural sector, there has been a growing interest on the subject. From 1990s onwards, there are a number of studies that have been carried out, which mainly focused on the impact of climate variables such as air temperature, precipitation and CO_2 on the agricultural yield (Linnenluecke et al., 2018). Regression modelling was employed to investigate the agriculture production's sensitivity to changes in climate in South Africa over different regions (Blignaut et al., 2009). The study, however, used a number of unrelated variables in their model. A number of studies used experimental analyses, agro-ecological models, and process-based dynamic crop growth models; while a few used statistical analysis (Kumar & Sharma, 2014; Jones et al., 2014; Zhao & Li, 2015; Linnenluecke et al., 2018). Another study by Le Roux et al. (2009) used an existing agricultural model to carry out seasonal maize simulations in South Africa. Therefore, there is need to carry out more statistical based analysis using historical data. These models can be evaluated using existing data and used to forecast future trend. Furthermore, a number of the statistical methods use yield or production as a basis of analysis of the impact of climate variability, and do not look at the growth phase of the sugarcane. With the advent of technology, crop monitoring has also advanced. One of the method that are used to monitor crops and other plantations is satellite imaging. A number of vegetation index have been derived from the satellite imaging and one of these is Normalised Difference Vegetative Index (NDVI). Using this index, any deficiency or excess of nutrients in the leaves can be diagnosed through spectral responses. The NDVI can show nitrogen status and chlorophyll at micro level (Lisboa et al., 2018).

A number of studies have used vegetation index to monitor crop growth (Tarnavsky et al., 2008; Peralta et al., 2016; Sanches et al., 2018). In these studies it has been found that the vegetation index is positively correlated to the yield of crops such as maize, soybean and sugarcane.

2.6 Summary

In this chapter, a literature review of various methods that have used to model climate was done. This included a review of the strengths and shortcomings of linear methods. Transform based methods, such as Wavelet analysis and Fourier analysis, were discussed. A discussion on data adaptive methods, such as singular spectrum analysis was done. The review of literature on EMD and EEMD was covered in detail. Literature has shown that a number of hybrid models that includes EEMD performed better in forecasting. Some of the models discussed were EEMD-SVM and EEMD-AR. EEMD is used to decompose the original time series into IMFs and the decomposed data will be modelled in this study using GEVD; therefore, a review of GEVD was done.

The literature reviewed in this chapter has revealed several gaps in theory and applications. Several authors have advocated for the improvement of the existing methods in literature so that there can be a reduction in the prediction error. There is also a need to evaluate new methods adopted in more developed countries in developing countries, such as South Africa, with different climatic conditions.

In the next chapter, the methodology employed on the study of the analysis South Africa climate variability and its agricultural impact is done.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In Chapter 2, review of literature related to climate variability modelling were discussed. In this chapter, a discussion of the methodology that will be directly followed or applied throughout the thesis, to analyse the South African climate variability and its agricultural impact is done. The climate data collected from weather stations have the challenge of missing data. A discussion of how missing data is dealt with in this study is explained in section 3.2. South Africa has highly spatial differences in rainfall patterns; therefore, spatial variability analysis will be done to identify areas with similar rainfall pattern, which is selected for this study. This method is discussed in section 3.3. The decomposition methods, empirical mode decomposition (EMD) and ensemble empirical mode decomposition (EEMD), to be used are discussed in sections 3.4 and 3.5, followed by the modelling of the decomposed data method. The methods for model diagnostics, selection and performance are also discussed. Most of the statistical analysis were performed using MATLAB and R, a software package that is useful as a language of programming and statistical modelling. The MATLAB code used to decompose the time series using EMD and EEMD is attached in appendix A.

3.2 Missing Data

The average monthly rainfall and temperature data used in this study were obtained from the South African Weather Service (SAWS), for the period 1980 to 2018. Rainfall and temperature data from weather stations has challenge of having missing data. Therefore, to cater for the missing data, multiple imputation by chained equations (MICE) method was used to impute missing data (van Buuren & Groothuis-Oudshoorn, 2011). It has been used to impute missing data in rainfall data in previous studies, and Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) revealed that it performed better for rainfall data than other methods (de Carvalho et al., 2017; Norazizi & Deni, 2019). MICE was chosen for this study because it does not assume normal distribution of the data and it assumes missing at random (MAR).

Let Y be a complete data set of a partially observed random sample, a multivariate distribution with p-variables $P(Y|\theta)$. Assume that the multivariate distribution of Y is completely specified by unknown vector parameters θ. The problem is how to get the multivariate distribution of θ. The MICE algorithm obtains the posterior distribution of θ by sampling iteratively from conditional distributions of the form

$$
\begin{gathered}
P(Y_1|Y_{-1}, \theta_1) \\
\vdots \\
P(Y_p|Y_{-p}, \theta_p).
\end{gathered}
$$

The parameters $\theta_1, \ldots, \theta_p$ are specific to the respective conditional densities and are not necessarily the product of the factorisation of the joint distribution $P(Y|\theta)$. Assuming that the missing data is random; that is, when the missing values pattern in a variable is predictable from the other variables of the data set. A simple sample of the observed marginal distribution is taken out and the t^{th} iteration of the chained equations is a sample of Gibbs, which successively withdraws

$$\begin{aligned}
\theta_1^{*(t)} &\sim P\left(\theta_1|Y_1^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}\right) \\
Y_1^{*(t)} &\sim P\left(\theta_1|Y_1^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)}\right) \\
&\vdots \\
\theta_p^{*(t)} &\sim P(\theta_p|Y_p^{\text{obs}}, Y_1^{t}, \ldots, Y_{p-1}^{(t-1)} \\
Y_p^{*(t)} &\sim P\left(\theta_p|Y_p^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_p^{*(t)}\right)
\end{aligned} \tag{3.1}$$

where $Y_j^{(t)} = \left(Y_j^{\text{obs}}, Y_j^{*(t)}\right)$ is the j^{th} imputed variable at iteration t and θ_p^* is the estimation of the parameter for the t^{th} iteration of the chained equations. Observe that previous imputations $Y_j^{*(t-1)}$ only enter $Y_j^{*(t)}$ through its relation with other variables, and not directly (van Buuren & Groothuis-Oudshoorn, 2011). The name, chained equations, refers to the fact that the MICE algorithm can be easily implemented as a concatenation of univariate procedures to fill out the missing data. The R package '*mice*' developed by van Buuren & Groothuis-Oudshoorn (2011) is used for imputation in this study.

3.3 Spatial Variability

Mann Kendall (MK) has been widely used in hydrological and meteorological problems to investigate the trend at annual, seasonal, monthly and daily level (Longobardi & Villani, 2010; Yanming et al., 2012; Ahmad et al., 2015; Nema et al., 2018). It was used to detect trends in the monthly, seasonal and annual precipitation over a certain region in Pakistan (Ahmad et al., 2015). MK was applied in a study in India over 27 districts for a study period of 115 years and found that both monthly and seasonal trend showed an overall decline in rainfall (Nema et al., 2018). This study also used coefficient of variation to investigate the variability over the districts. MK is widely used in hydrological studies because it is a non-parametric method and the data does not necessarily have to be normally distributed.

The South African weather patterns are highly volatile with large spatial variability over different regions (Jury, 2013). It is therefore, important to carry out a rainfall trend analysis, on different spatial and temporal scales, mainly because of the climate variability (Longobardi & Villani, 2010). In this study, trend analysis and spatial variability maps are presented for Western Cape province to select a study area with similar rainfall patterns. The spatial and trend analysis of rainfall over the areas under study will be carried out using Mann-Kendall (MK), coefficient of variation (CV%) and Kriging. These are explained in the following sub-sections.

3.3.1 Mann-Kendall and Coefficient of Variation

Mann-Kendall (MK) trend test is used to find any trend in the rainfall for the given period. The advantage of MK is that it does not have any underlying assumptions about the distribution of the data, except that it must be independent and identically distributed. It can indicate the temporal patterns in a time series (Kendall, 1980). The null hypothesis for MK is that, there is no monotonic trend in the time series. The test is carried out at 5% level of significance, and if the probability estimate (p-value) was less than 0.05, the trend was deemed to be significant.

Given the time series $x_t = x_1, x_2, x_3, \ldots, x_n$ then the MK statistic, S, is defined as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{sgn}(x_j - x_i) \tag{3.2}$$

where x_j and x_i are the data values in years j and i, respectively, with $j > i$, n is the total number of years and sgn() is the sign function. Let $x_j - x_i = \theta$ then

$$\mathrm{sgn}(\theta) = \begin{cases} 1 & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \\ -1 & \text{if } \theta < 0 \end{cases} \tag{3.3}$$

The standard normal variable Z is then formulated as

$$z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ S - 1 & \text{if } S < 0 \end{cases} \tag{3.4}$$

The coefficient of variation (CV%) is used to find the variability of rainfall. CV%, also known as "relative variability", equals the standard deviation divided by the mean (Lebel et al., 1987). It measures how a value is varied around the mean. Since there are un-sampled areas around the weather stations, ordinary Kriging interpolation method is then used to find the spatial distribution of the variability throughout the region selected (Robinson & Metternicht, 2003).

3.3.2 Kriging

Kriging is one of the widely used estimation techniques. It is named after the South African mining engineer, D.G. Krige, who helped pioneer the development of geo-statistical methods in the 1950s (Robinson & Metternicht, 2003). Kriging is a linear weighted average method. The weights used in kriging are based on the semivariogram model of spatial correlation. Kriging is a stochastic interpolation method that is used to map un-sampled locations with available data sets. It is also known as Wiener-Kolmogorov prediction (Robinson & Metternicht, 2003). It was chosen for this study, as it has been found to comparatively perform better than other interpolation methods, such as Spline and inverse distance weighting, when used to estimate the un-sampled areas in rainfall problems (Kim et al., 2011; Yang et al., 2015).

According to Bonaventura & Castruccio (2005) and Malvić & Balić (2009), Kriging is defined as follows: given a point P, the ordinary kriging estimator at Z_p based on the data $Z_i : i = 1, \ldots, n$ is defined as the linear unbiased estimator

$$Z_p = \sum_{I=1}^{n} \omega_i Z_i \tag{3.5}$$

of Z_p with minimum mean square prediction error. Where $\omega_i \in \mathbb{R}$ is the unknown weights corresponding to the influence of the variable Z_i in the computation of Z_p. The weights $\{\omega_i : i = 1, 2, \ldots, n\}$ are calculated from the set of n equations

$$\begin{aligned}
\omega_1\gamma(h_{11}) + \omega_2\gamma(h_{12}) + \cdots + \omega_n\gamma(h_{1n}) + \lambda &= \gamma(h_{1P}) \\
\omega_1\gamma(h_{21}) + \omega_2\gamma(h_{22}) + \cdots + \omega_n\gamma(h_{2n}) + \lambda &= \gamma(h_{2P}) \\
\vdots \quad &\vdots \quad \vdots \\
\omega_1\gamma(h_{n1}) + \omega_2\gamma(h_{n2}) + \cdots + \omega_n\gamma(h_{nn}) + \lambda &= \gamma(h_{nP})
\end{aligned} \tag{3.6}$$

where $\gamma(h_{ij})$ is the semivariogram at lag distance h_{ij} between two points (P_i, P_j). The semivariogram $\gamma(h_{iP})$ is the semivariogram at lag distance h_{ip} between control point P_i and the point P where attribute Z_p is being estimated.

The constant λ is the Lagrange multiplier for the weights where

$$\sum_{i=1}^{n} \omega_i = 1. \tag{3.7}$$

The resulting matrix for the equations (3.6) and (3.7) is given by

$$\begin{bmatrix} 1 & \varphi(h_{11}) & \varphi(h_{12}) & \cdots & \varphi(h_{1n}) \\ 1 & \varphi(h_{21}) & \varphi(h_{22}) & \cdots & \varphi(h_{2n}) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \varphi(h_{n1}) & \varphi(h_{n2}) & \cdots & \varphi(h_{nn}) \\ 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{bmatrix} = \begin{bmatrix} \varphi(h_{1P}) \\ \varphi(h_{2P}) \\ \vdots \\ \varphi(h_{nP}) \\ 1 \end{bmatrix}. \tag{3.8}$$

3.4 Empirical Mode Decomposition (EMD)

The climate variables, which include rainfall and temperature time series, will be decomposed into multiple time series using empirical mode decomposition and ensemble empirical mode decomposition. Empirical mode decomposition is a data adaptive method, which requires only that the data must consist of simple intrinsic mode of oscillations (Huang et al., 1998). The original data is decomposed into several time series, called intrinsic mode functions (IMFs), through a sifting process.

3.4.1 Intrinsic Mode Functions

The EMD methodology is based on sifting process, which identifies local extrema (maxima and minima) resulting in the formation of IMFs. The first IMF may contain the highest fluctuations and this is subtracted from the original data and subsequent IMFs are then derived from this new data (Peel et al., 2011). The IMFs and residual data approximate the original data when they are summed together. The conditions that must be satisfied by each IMF are that (i) in the whole data set, the number of extrema must be equal to the number of zero-crossings, or differ from it at most by one; and (ii) at any point, the mean value of the two envelopes determined by the local maxima and minima must be zero (Huang et al., 1998). The IMFs is the main conceptual innovation; it captures the local properties of the data which makes instantaneous frequency meaningful and eliminates the need to create false harmonics for representation of nonlinear and non-stationary data (Huang et al., 1998; Zhao & Chen, 2015).

3.4.2 Sifting Process

Given a time series defined by $x_t = (x_1, x_2, ..., x_n) \in \mathbb{R}$, where x_i is the value of time series at time i. The upper envelope, $x_t^{\max}$ and lower envelope, $x_t^{\min}$ of x_t, are defined as the cubic spline interpolation of the coordinates of its local maxima and minima respectively (New, 2013). Figure 3.1 shows the upper and lower envelopes of the given data. The red line is the average of the two.

To decompose the time series x_t into IMFs, we use the following algorithm:

(i) Construct the upper $x_t^{\max}$ and lower $x_t^{\min}$ envelopes by connecting all the maxima and minima of x_t, respectively.

(ii) Compute

$$\Delta x_t = x_t - \left[\frac{x_t^{\max} + x_t^{\min}}{2}\right] \tag{3.9}$$

(iii) Repeat steps (i) and (ii) for Δx_t until the resulting signal will possess the properties that the number of extrema is equal (or differ at most by one) to

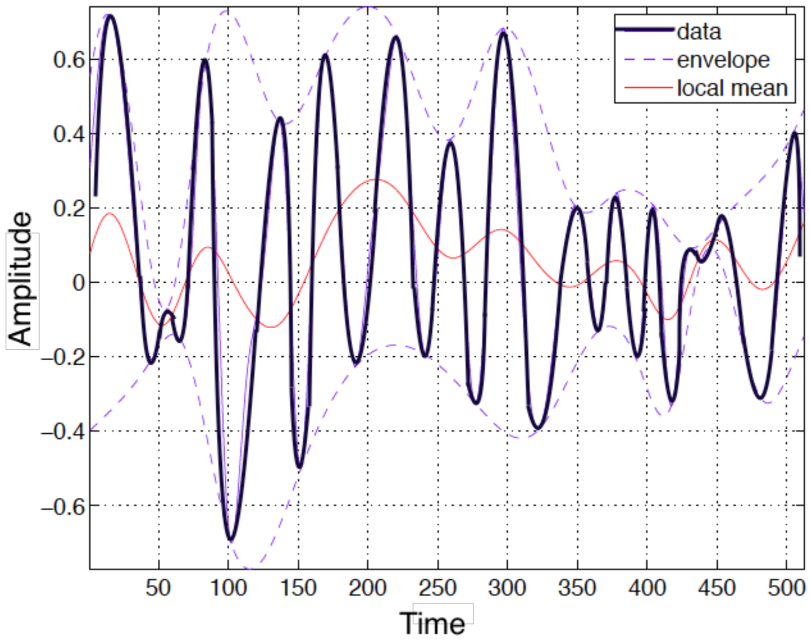


Figure 3.1: Upper and lower envelopes and local mean of a sample time series. (New, 2013)

the number of zero crossings, and the mean value between the upper and lower envelope is equal to zero at any point. Denote the resulting signal by $c_t(1)$, which is the first IMF.

(iv) Take the difference $x_t(1) = x_t - c_t(1)$ and repeat steps (i)–(iii) for it to obtain the second IMF $c_t(2)$ (Huang et al., 1998; Guo et al., 2016).

This decomposition continued until no other IMF can be identified. The summation of the IMFs and residual approximates the original data, that is for a set of IMFs $c_t(i)$

$$x_t \approx \sum_{i=1}^{n} c_t(i) + r_t, \tag{3.10}$$

where $c_t(i)$ is the i^{th} IMF time series and r_t is the residual from the decomposition. (Huang et al., 1998; Molla et al., 2005; Guo et al., 2016).

The sifting process must be carried out with care, to avoid over-sifting, where each IMF produces smooth amplitude with physical meaningful amplitude being sifted away; or under-sifting, where the IMFs fails to reveal all the riding waves.

3.4.2.1 Stop Criteria

Stop criteria that have been used are: Cauchy-type convergence, mean fluctuations thresholds, energy difference tracking, resolution factor, bandwidth and orthogonality criterion (Tabrizi et al., 2014).

- *Cauchy-type convergence (SD)* - Huang et al. (1998) suggested the use of a constant value that is derived from the standard deviation of the IMFs. The sifting process is stopped once the standard deviation reaches a pre-defined value (most instants being less than 0.2 or 0.3); sifting must stop, where SD is defined as

$$SD = \sum_{t=0}^{T} \frac{c_t(k-1) - c_t(k)}{c_t^2(k-1)} \tag{3.11}$$

where $c_t(k-1)$ is the $(k-1)^{\text{th}}$ and $c_t(k)$ is the k^{th} IMF.

- *Mean Fluctuations threshold (MFT)* - The stopping criterion is dependent on three threshold (θ_1, θ_2 and α) aimed at attaining small global fluctuations in the mean whilst taken into consideration local large excursions. For $(1-\alpha)$ fraction of data, sifting will be continued for $\sigma(t) < \theta_1$ and for remaining fraction $\sigma(t) < \theta_2$, where $\sigma(t) = \left|\frac{m_t}{a_t}\right|$ for $m_t = \left(x_t^{\max} + x_t^{\min}\right)/2$ and $a_t = \left(x_t^{\max} - x_t^{\min}\right)/2$ (Rilling et al., 2003).

- *Energy difference tracking (EDT)* - EDT considers the error of original data and the sum of decomposed data. When it reaches a certain minimum and the mean value of envelopes becomes small, the sifting process is stopped (Tabrizi et al., 2014).

- *Resolution factor (RF)* - When using RF, the sifting process is stopped when the ratio between the energy of the signal at the beginning of the sifting and the energy of average of the envelopes ascends to a predefined factor (resolution factor) (Tabrizi et al., 2014).

- *Bandwidths* - Bandwidths use MFT criterion to get a result that almost satisfies conditions (i) and (ii) in section 3.4.1. The sifting process is continued until the difference of variance of consecutive instantaneous frequencies is very small (Qiwei et al., 2007).

- *Orthogonality criterion (OC)* - The sifting process is stopped when the OC reaches a pre-defined value, where OC is given by:

$$OC = \left| \sum_{t=1}^{N} \frac{m_t \cdot x_t}{m_t \cdot x_t - m_t} \right| \tag{3.12}$$

where $m_t = \left(x_t^{\max} + x_t^{\min}\right)/2$ is the mean of the i^{th} envelope (Lin & Hongbing, 2009).

The number of IMFs after decomposition is close to $\log_2 N$, with N being the number of total data points (Wu & Huang, 2004). Most of the IMFs are normally distributed according to central limit theorem. The efficiency of decomposition needs to be measured and the elements need to be orthogonal to each other (Molla et al., 2007). Higher orthogonality corresponds to less amount of information leakage. Index of orthogonality (IO) is used, which is given by

$$IO = \frac{1}{T} \sum_{t} \frac{1}{x^2(t)} \left(\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} c_t\left(i\right) \cdot c_t\left(j\right) \right), \tag{3.13}$$

where i and j represents the i^{th} and j^{th} IMFs. Orthogonality between two IMFs is given by

$$IO_{(i,j)} = \frac{1}{T}\sum_{t}\frac{c_t(i) \cdot c_t(j)}{c_t(i) + c_t(j)}. \tag{3.14}$$

The orthogonality between two IMFs must be less than 0.100 for it to be acceptable (Chang et al., 1997; Molla et al., 2005).

In this study, the Cauchy-type of convergence is used because it was demonstrated that the IMFs obtained by the energy difference tracking method can reflect the intrinsic information included in the signal more clearly and because their index of orthogonal (IO) is smaller when compared to other methods (Junsheng et al., 2006).

3.4.3 Significance Test of the IMFs

An IMF can either contain true signal or random noise. To test this, Monte-Carlo method is used to perform a significant test (Wu & Huang, 2005). The true signal is determined by examining the distribution of the energy, with respect to period. The i^{th} IMF energy density component is defined as follows:

$$E_i = \frac{1}{N}\sum_{t=1}^{N}|c_t(i)|^2 \tag{3.15}$$

where N is the length of the IMF component and $c_t(i)$ denotes the i^{th} IMF component.

The relationship between energy density, E_i, and mean oscillation period, $\overline{T}_i$, is given by

$$\ln E_i + \ln \overline{T}_i = 0 \tag{3.16}$$

(Wu & Huang, 2009).

If the IMF energy with a certain mean period exceeds the upper limit of a certain confidence interval, it is assumed that the corresponding IMF contains statistically significant information at selected confidence level. The confidence interval (CI) is given by

$$CI = -x_t \pm k\sqrt{\frac{2}{N}}e^{w/2}, \tag{3.17}$$

where x_t is the rainfall time series, $w = \ln\left(\overline{T}_n\right)$ and k is a constant determined by percentiles of the standard normal distribution (Wu & Huang, 2014).

3.4.4 Synchronisation

The derived IMFs will need to be interpreted in terms of the physical meaning by correlating them against known systems. One method that can be used to do this is synchronisation. This is defined as the appearance of certain relations between their phases and frequencies whilst the amplitudes can remain non-correlated (Rosenblum et al., 2001). The phase synchronisation of chaotic systems is defined as the appearance of a certain relation between the phases of interacting systems or between the phase of a system and that of an external force. The amplitudes of the systems can remain chaotic and non-correlated. The phase synchronisation properties in chaotic systems are similar to those found in periodic noisy oscillators (Rosenblum et al., 2001). Therefore, this allows the use of the same framework in this study, where the climate variables rainfall, temperature, QBO and ENSO (represented by Niño 3.4) are taken as the periodic noisy oscillators. According to Pikovsky et al. (2001), synchronisation is formally defined as follows:

Definition 3.1. (Synchronisation) Construct the analytic signal $\zeta(t)$ which is a complex function of time defined by

$$\zeta(t) = x(t) + i\tilde{x}(t) = A(t)e^{\phi(t)}, \tag{3.18}$$

where $A(t)$ is the instantaneous amplitude and $\phi(t)$ is the instantaneous phase of the time series, $x(t)$ and the function $\tilde{x}(t)$ is the Hilbert transform of $x(t)$, where

$$\tilde{x}(t) = \frac{1}{\pi}\int\limits_{-\infty}^{\infty} \frac{x(\omega)}{t-\omega} d\omega. \tag{3.19}$$

To detect synchronisation between two signals, we calculate the phase difference or relative phase given by

$$\varphi_{a,b}^{12} = a\phi_1 - b\phi_2, \tag{3.20}$$

where ϕ_1 and ϕ_2 are the phases of the first and the second signals and a and b are integers. The $a : b$ phase synchronization is defined by the condition

$$\left|\varphi_{a,b}^{12} - \delta\right| < const, \tag{3.21}$$

where δ is some average phase shift. In this case, the relative phase difference fluctuates around a constant value.

In this study, synchronisation is used to reveal the interaction between rainfall and other climatic drivers. The results of the phase synchronisation are presented in the form of histograms. Time series that are phase synchronised or locked exhibit a modal distribution with a prominent peak at a given phase difference, whereas unrelated time series are characterised by a uniform or diffuse distribution. R package '*synchrony*' is used; it measures phase synchrony between quasi-periodic times series (Cazelles & Stone, 2003).

3.5 Ensemble Empirical Mode Decomposition (EEMD)

EMD has a challenge of ”mode mixing”, where a single IMF contains multiple components at different frequencies; it is therefore, not representative of any signal. Wu & Huang (2009) proposed a data assisted method, called Ensemble Empirical Mode Decomposition (EEMD), which consists of adding noise to the data before carrying out the decomposition using EMD. The noise will be cancelled out, as it takes advantage of the zero mean of the noise once it has served its purpose.

EEMD is a noise-assisted method, which consists of adding noise to the data before carrying out the EMD algorithm. For a given time series x_t, noise is introduced to each data point , x_i. This noise is treated as random noise that can be encountered in the process. Therefore, the new time series y_t is denoted by

$$y_t = \varepsilon_j + x_t, \text{ for } j = 1, 2, ..., m \tag{3.22}$$

where ε_j is white noise added for m number of ensembles (Molla et al., 2005; Tabrizi et al., 2014).

An average of the IMFs found from the data with noise becomes the final IMF, that is,

$$c_t(j) = \frac{1}{m}\sum_{i=1}^{m} d_t(j) \tag{3.23}$$

where $d_t(j)$ is the IMF of the time series with added noise and $c_t(i)$ is the final i^{th} IMF for the original time series x_t. The average of residuals from y_t gives the final residual that is,

$$r_t(j) = \frac{1}{m}\sum_{i=1}^{m} r_t(j) \tag{3.24}$$

where $r_t(j)$ are residuals from the time series with added noise.

However, there are issues with the amplitude of the noise to be added to the original data. If the added noise is too small, relative to the original data there will not be any improvement in the mode mixing. On the other hand, if the amplitude is too high, it will lead to misinformation on the analysis of the results, due to redundancy in the IMFs and increased computational cost (Tabrizi et al., 2014). Hence in this study, an optimal noise to be added will be determined using variance contribution rate and threshold method to identify important IMFs. These methods are discussed in the next sub-sections.

3.5.1 Variable Contribution Rate

Wu & Huang (2009) proposed that the amplitude of noise to be added should be between 0.1 and 0.4 of the standard deviation. However, in this study an optimal amplitude of the noise to be added will be determined by variance contribution rate (VCR) and the threshold value (Ayenu-Prah & Attoh-Okine, 2010). VCR determines the significance of each IMF to the overall variability of the original signal (data) (Guo et al., 2016). VCR is calculated using the formula

$$VCR_i = \frac{\mathrm{var}(c_t(i))}{\sum_{i=1}^{n} \mathrm{var}(c_t(i)) + \mathrm{var}(r_t)} \cdot 100 \tag{3.25}$$

where $\mathrm{var}(c_t(i))$ is variance of i^{th} IMF and $\mathrm{var}(r_t)$ is variance of residual (Guo et al., 2016).

3.5.2 Threshold to determine a relevant IMF

The threshold method uses the idea that since the IMFs are supposed to be almost orthogonal components of the original signal, each of the IMFs may therefore, have a correlation with original signal (Huang et al., 1998). Therefore, this presupposes that irrelevant IMFs may have poor correlation with original signal. The method consist of finding the correlation coefficients between each IMF and the original signal, and determining a stringent threshold coefficient used to distinguish between a relevant IMF and irrelevant IMFs. A threshold λ is introduced where,

$$\lambda = \frac{\max(\rho_i)}{10 \cdot \max(\rho_i) - 3}, \qquad i = 1, 2, ..., n \tag{3.26}$$

where ρ_i is the correlation coefficient of the i^{th} IMF with the original signal, n is the total number of IMFs and $\max(\rho_i)$ is the maximum correlation coefficient observed.

3.6 Modelling of the IMFs

The decomposed data into IMFs will be modelled using GEVD. Several previous studies in hydrology and climatology revealed that GEVD is the best model to fit extreme weather conditions (Chikobvu & Chifurira, 2015; Maposa, 2016; Boudrissa et al., 2017; Namitha & Ravikumar, 2018; ?). This section covers the block maxima and generalised extreme value distribution in details.

3.6.1 Block Maxima and the Generalised Extreme Value Distribution

Suppose $X_1, ..., X_n$ is a sequence of independent and identically distributed (i.i.d.) random variables with distribution function F. Then, let

$$M_{k,m} = \max\{X_{(k-1)(m+1)}, ..., X_{k,m}\} \tag{3.27}$$

where $M_{k,m}$ is the set of mean monthly rainfall amounts over n-observation period. The block maxima approach in extreme value theorem, consists of dividing the observation period into non-overlapping periods of equal blocks and then restricts attention to the maximum observation in each period. That is, the observations, $n = k \times m$, are divided into k blocks of size m, where n is the total number of observations. "Then there exists, a non-degenerate function G, such that for a fixed block length $m \geq 1$ the variables $M_{k,m}$ are i.i.d. with distribution

$$P\left\{\left(\frac{M_{k,m} - b_m}{a_m}\right) \leq x\right\} = \lim_{m \to +\infty} F^m\left(a_m x + b_m\right) \to G_\xi(x) \text{ as } n, k \to +\infty \tag{3.28}$$

where G is the extreme value index given by

$$G_\xi(x) = \exp\left\{-\left[1 + \xi x^{-\frac{1}{\xi}}\right]\right\},\ \xi \in \mathbb{R} \tag{3.29}$$

" (Maposa, 2016, p. 47).

If the size of the blocks approaches infinity, then $G_\xi(x)$ approximates the GEVD which is given by

$$G_{\sigma,\mu,\xi}(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right]\right\} \tag{3.30}$$

where μ,σ and ξ are location, scale and shape parameters respectively (Dombry, 2015). The value of the shape parameter ξ determines the type of extreme value distribution (Coles, 2001; Maposa, 2016). There are two important theorems that concern the GEVD; non-degeneracy theorem and max-stability theorem.

Theorem 3.2. *(Non-degeneracy) If there exist sequences of constants $a_m > 0$ and b_m such that*

$$P\left\{\left(\frac{M_{k,m} - b_m}{a_m}\right) \leq x\right\} \rightarrow G(x), \text{ as } n \rightarrow \infty \tag{3.31}$$

for a non-degenerate distribution function G, then G is a member of the GEV family in (3.30)*.*

The proof of Theorem (3.2) is given in Coles (2001).

Theorem 3.3. *(Max-stability) A distribution is max-stable if, and only if, it is a generalised extreme value distribution in* (3.30)*.*

The proof of Theorem 3.3 is given in Coles (2001).

The GEVD takes different distributions, depending on the values of the location (μ), scale (σ) and shape (ξ). The distributions are classified into 3 types given below.

(i) Gumbel Distribution

If $\xi = 0$ (interpreted as $\xi \rightarrow 0$), the GEVD becomes a Gumbel distribution. It has two main parameters μ and σ. The Gumbel distribution comprises distributions such as exponential, normal, logistic and gamma distributions. The Gumbel Distribution is defined as:

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, -\infty < x < \infty \tag{3.32}$$

(ii) Fréchet distribution

If $\xi > 0$, the GEVD becomes a Fréchet distribution which has a heavy upper tail. The Fréchet family of distributions includes Pareto, log-gamma, Student t and Cauchy distributions. The Fréchet Distribution is defined as:

$$G(x) = \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right\}, x > 0 \tag{3.33}$$

for $x > \mu$, $\alpha > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$.

(iii) The Weibull Distribution

If $\xi < 0$, the GEVD becomes a Weibull distribution, which has a short tail for $0 < \alpha < 1$. The Weibull distribution family includes the Uniform and Beta distributions. Weibull distribution is defined as:

$$G(x) = 1 - \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right\}, x > 0 \tag{3.34}$$

for $x > \mu$, $\alpha > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$.

3.6.2 Estimation Procedure of GEVD Parameters

There are a number of methods that can be used to estimate the GEVD parameters, μ, σ and ξ, these are; L-moments, method of moments (MOM), least squares estimators (LSE) methods and maximum likelihood estimators (MLE). In this study, MLE is used to estimate the parameters because it is flexible and consistent (Maposa, 2016).

Maximum Likelihood Estimator

Maximum Likelihood Estimators are the solution of an equation system formed by setting the derivatives of L_k with respect to each parameter (μ, σ and ξ) to zero. Consider the distribution of $M_{k,m}$ given in equation (3.27) for large k and $m \geq 1$, which is approximately a GEVD. The log-likelihood function for the GEVD parameters when $\xi \neq 0$ is:

$$L_k(\mu, \sigma, \xi) = \frac{1}{k} \sum_{i=1}^{k} \ell_{(\mu,\sigma,\xi)}(M_{k,m}). \tag{3.35}$$

For stationary GEVD, the derivative of the log-likelihood, L_k (equation (3.35)), with respect to each parameter $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$ is a maximum likelihood estimator (MLE), if L_k has a local maximum at $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$ (Dombry, 2015). Therefore, MLEs are obtained from the solution of the likelihood equations

$$\begin{aligned} \tfrac{\partial L_k}{\partial \mu} &= 0 \\ \tfrac{\partial L_k}{\partial \sigma} &= 0 \\ \tfrac{\partial L_k}{\partial \xi} &= 0. \end{aligned} \tag{3.36}$$

The major findings concerning the existence of consistent of MLEs by Dombry (2015) are summarised in the following theorem.

Theorem 3.4. *(Existence of consistent) Suppose $F \in D(G_\xi)$ with $\xi > -1$ and assume that*

$$\lim_{n \to +\infty} \frac{m(n)}{\log n} = +\infty. \tag{3.37}$$

Then there exists a sequence of estimators $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$ and a random integer $N \geq 1$ such that

$$\mathbb{P}\left[(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k) \ is\ a\ MLE\ for\ all\ n \geq N\right] = 1 \tag{3.38}$$

and

$$\hat{\xi} \to \xi,\ \frac{\hat{\mu}_n - b_n}{a_m} \to 0 \ and\ \frac{\hat{\sigma}_n}{a_m} \to 1 \ as\ n \to +\infty \tag{3.39}$$

(Dombry, 2015, p. 423).

The likelihood in (3.36) has no solution for $\xi < 1$ and no consistent MLE exists when $\xi < 1$ (Dombry, 2015).

3.6.3 Return Level Estimates

Estimates of the extreme events are of particular interest in meteorological and climate modelling because these tend to have direct effect on human livelihood. The extreme quantiles give an estimate of the level the process is expected to exceed once, on average, in a given number of years (T) with a probability p. The quantiles are calculated by inverting (3.30) to give:

$$Q_p = \begin{cases} \mu + \frac{\sigma}{\xi}\left[(-\ln(1-p))^{-\xi} - 1\right], & \xi \neq 0 \\ \mu - \sigma \ln\left[-\ln(1-p)\right], & \xi = 0 \end{cases} \tag{3.40}$$

where $p = 1/T$ and Q_p is the quantity of the return level associated with time T year return period (**?**).

The confidence intervals for the return level are obtained using the profile likelihood method. The 100(1 - α)% confidence interval is given by:

$$\left\{\theta : 2\log\left(\frac{L(\hat{\mu}, \hat{\sigma}, \hat{\xi})}{L_m(\theta)}\right) < \chi^2_{1,1-\alpha}\right\}, \tag{3.41}$$

where $\chi^2_{1,1-\alpha}$ is $100(1-\alpha)\%$ quantile of the chi-square distribution with 1 degree of freedom, $L_m(\theta)$ is obtained by pre-parameterising the log-likelihood in terms of (Q_p, σ, ξ) and $L_m(Q_p)$ is the profile likelihood given by:

$$L_m(Q_p) = \max_{\sigma,\xi} L(Q_p, \sigma, \xi) \tag{3.42}$$

(**?**). The return level estimates for the rainfall and IMFs GEVD models is calculated in this study.

3.6.4 Goodness-of-fit Test

To test for the best model, two goodness-of-fit (GoF) tests methods, Kolmogorov-Smirnov (K-S) test and Anderson-Darling (A-D), were tested at 5% level of significance. Two tests methods are selected to verify and to be sure of the results obtained. Let $x_1, x_2, \ldots, x_n$ be a sample of n monthly average rainfall observed and suppose the c.d.f. of the random variable x_t is $G(.)$, then the K-S and A-D tests are as presented in the following sub-sections.

Kolmogorov-Smirnov test

The K-S test is non-parametric test that quantifies the vertical distance between the empirical c.d.f., $G_n(x)$, and the theoretical c.d.f., $G(x)$. Let this distance be denoted by denoted by $D_{\max}$, then the test statistic of the K-S test is given by

$$D_{\max} = \max_{x} |G_n(x)\text{-}G(x)| \tag{3.43}$$

The hypothesis for the GoF under the K-S test procedure is

$$\begin{aligned} H_0 &: G(x) = G_0(x;\theta) \\ H_1 &: G(x) \neq G_0(x;\theta) \end{aligned} \tag{3.44}$$

where G_0 is a specified (hypothesised) distribution and θ is a vector of unknown parameters.

If $D_{\max}$ calculated is greater than the tabulated value of $D_{0.05} = 1.36/\sqrt{n}$, then the null hypothesis for this test is rejected at 5% level of significance (Sukla et al., 2014; Maposa, 2016).

Anderson-Darling test

The A-D test is used to assess whether a sample comes from specified distribution. In this thesis, it will be used to compare the fit of observed c.d.f. and theoretical (expected) c.d.f. (Sukla et al., 2014; Maposa, 2016). The A-D statistic is defined as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln G(x) + \ln(1 - G_n(x))\right] \tag{3.45}$$

where $G_n(x)$ is the empirical c.d.f., $G(x)$ is the theoretical c.d.f. for ordered data x and n is the sample size.

The hypothesis for GoF under the A-D test procedure is

H_0: The data follow a specified distribution, versus, H_1: The data do not follow the specified distribution. This can be written as

$$\begin{aligned} H_0 &: G(x) = G_0(x;\theta) \\ H_1 &: G(x) \neq G_0(x;\theta) \end{aligned} \tag{3.46}$$

where G_0 is a hypothesised distribution and θ is a vector of unknown parameters.

If $G(x)$ is very different from the hypothesised distribution $G_0(x;\theta)$, that is, if A^2 calculated is greater than the critical value then the null hypothesis for A-D test is rejected at 5% level of significance, (Sukla et al., 2014; Maposa, 2016).

The A-D and K-S tests are used to test and compare the GoF for the proposed decomposed and non-decomposed models.

3.6.5 Diagnostic Plots

The diagnostic plots that are commonly used are probability density function (p.d.f.) plot, probability probability (PP) plots and quantile-quantile (QQ) plots, among others (Berning, 2010). A more detailed discussion of QQ-plots is made in Berning (2010). "A QQ-plot is a graphical tool that is used to assess the GoF, that is, whether the underlying distribution G is the appropriate distribution for random variable x_t, whose observations are $x_1, x_2, \ldots, x_n$" (Maposa, 2016, p.116). Berning (2010) defines QQ-plot as follows:

Definition 3.5. (QQ-plot). A QQ-plot is a graphical tool used to visually assess the goodness-of-fit of a distribution. A plot of $(Q(i/(n+1)); x_{i,n})$, $i = 1, 2, ..., n$, should be approximately linear if $x_1, x_2, \ldots, x_n$ are from a distribution with quantile function Q.

"One of the main advantage of QQ-plot is that the location, μ, and scale, σ, parameters of the model do not need to be known in advance. A normal QQ-plot can be used to estimate the location and scale parameters" (Maposa, 2016, p. 117). The normal QQ-plot is defined by Berning (2010) as follows:

Definition 3.6. (Normal QQ-plot). Consider observations $x_1, x_2, \ldots, x_n$. Then a plot of $(\Phi^{-1}(i/(n+1)); x_{i,n})$, $i = 1, 2, ..., n$ is a normal QQ-plot of the data, where $\Phi\,(.)$ is the standard normal distribution function. If the observation $x_1, x_2, \ldots, x_n$ are from a normal distribution with mean, μ , and standard deviation, σ, then the slope should be approximately linear with gradient σ and intercept μ.

The QQ plot is very good at comparing the tails of the distribution therefore, which makes it a better graphical technique to use (Farrell & Rogers-Stewart, 2006). Another graphical method that can used is the PP-plot. The definition of PP-plot is as follows:

Definition 3.7. (PP-plot). A PP-plot is defined as a graphical tool used to visually assess the GoF of a specific distribution. It is based on a plot of the empirical c.d.f. values against the theoretical c.d.f. values. Just like the QQ-plot, the PP-plot is used to assess how well a specified distribution fits the observed data. The PP-plot should be approximately linear, if the specified theoretical distribution is the appropriate model for the random variable $X = \{x_1, x_2, \ldots, x_n\}$ (Maposa, 2016, p. 117).

PP-plot uses cumulative probabilities in plotting the graph, whereas the QQ-plot uses the quantiles. A p.d.f. can be defined for both discrete and continuous random variables. In this study, the focus is on the continuous random variables, therefore, a definition of p.d.f. for the continuous random variables is given as follows:

Definition 3.8. (p.d.f. Plot). Consider a continuous random variable X. Then the p.d.f. of X is a function $f(x)$, such that for any two data points (numbers) a and b $(a < b)$,

$$P(a \leq x \leq b) = \int_a^b f(x)dx \tag{3.47}$$

That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of a density function. The name usually given to the graph of $f(x)$ is the density curve (Maposa, 2016, p. 118).

The p.d.f. plot in this study will be plotted superimposed on the histogram, to show the symmetry (or asymmetry) and kurtosis of the empirical distribution. The QQ-plot, PP-plot and p.d.f. plot is used in this study to assess the GoF, that is, whether the underlying distribution G is the appropriate distribution for random variable x_t.

3.6.6 Model Selection

In searching for the best model, there should be an objective method that can be used to select the best amongst various candidates. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are widely used for model selection. Two methods are used to verify and to be sure about the model selected. Given $\hat{l}$ a maximised likelihood from a model that contains p parameters for sample of size n. AIC is defined as

$$AIC = -\frac{2\hat{l}}{n} + \frac{2p}{n} \tag{3.48}$$

and BIC is defined as

$$BIC = -2\hat{l} + p \ln n \tag{3.49}$$

(Claeskens & Hjort, 2008).

The negative log-likelihood (NLL) is defined as follows.

Definition 3.9. (Negative log-likelihood). The sample vector $\mathbf{X}$ is taken to be a vector of independently drawn samples from a distribution $\hat{F}$. That is

$$\mathbf{X} \leftarrow F. \tag{3.50}$$

The sample vector $\hat{F}$ is used to find the probability distribution as an approximation to the true distribution F. The likelihood that the samples were drawn from some given distribution $\hat{F}$, is defined as:

$$L\left(\mathbf{X}|\hat{F}\right) = \prod_{i=1}^{|k|} \hat{F}x_i. \tag{3.51}$$

Taking negative logarithm of both sides will result in an expression called NLL given by

$$- \ln \left[L \left(\mathbf{X} | \hat{F} \right) \right] = - \sum_{i=1}^{|k|} \ln \left(\hat{F}_{\bullet} x_i \right) \qquad (3.52)$$

(Bosman & Thierens, 2000).

The AIC, BIC and NLL are often used to select the better model based on the minimum of the chosen criterion (Coles, 2001). In this study, they will be used to compare the models of decomposed (IMFs) and non-decomposed rainfall data.

3.6.7 Sums of Random Variables

EEMD is a lossless decomposition method, such that the original data can be approximated by summing all the IMFs and residual. In this study, the derived IMFs will be modelled using GEVD, and estimates of the return level is added together to give the final return level for the rainfall. Addition of two random variables from same distribution gives a random variable with similar distribution. That is, given two random variables $X \sim G$ and $Y \sim G$ then the sum $X + Y = Z \sim G$, where G is a certain probability distribution (Leon-Garcia, 2017). This is defined formally by the following definitions, theorem and proof.

Definition 3.10. (Sum of discrete random variables). Let X and Y be two independent integer-values random variables, with distribution functions $m_1(x)$ and $m_2(x)$ respectively. Then the convolution of $m_1(x)$ and $m_2(x)$ is the distribution function $m_3 = m_1 * m_2$ given by

$$m_3(j) = \sum_k m_1(k).m_2(j-k),$$

for $j = \ldots, -2. -1, 0, 1, 2, 3, \ldots$. The function $m_3(x)$ is the distribution function of the random variable $Z = X + Y$ (Leon-Garcia, 2017).

The definition holds true for the sum of n independent random variables.

Definition 3.11. (Convolution). If $f(x)$ and $g(y)$, are two integrable real-valued functions, then the convolution of f and g is the real-valued function $f*g : \mathbb{R} \to \mathbb{R}$ is defined as

$$\begin{aligned} f * g &= \int_{-\infty}^{\infty} f(y)g(z-y)dy \\ &= \int_{-\infty}^{\infty} f(z-x)g(x)dx \\ &= g * f \end{aligned}$$

Theorem 3.12. *(Sum of Continuous random variables). Let X and Y be two continuous random variables with density functions $f_X(x)$ and $f_Y(y)$, respectively defined for all x. Then the sum $Z = X + Y$ is a random variable with density function $f_Z(z)$, where f_Z is the convolution of f_X and f_Y (Leon-Garcia, 2017).*

Proof. Let X and Y be independent random variables with probability density functions $f_X(x)$ and $f_Y(y)$. The cumulative distribution function $F_Z(z)$ of the random variable $Z = X + Y$ can be given as follows

$$\begin{aligned} F_Z(z) &= P(X + Y \leq z) \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f_X(x)f_Y(y)dy \right] dx \\ &= \int_{-\infty}^{\infty} f_X(x)F_Y(z-x)dx \end{aligned}$$

where F_Y is the c.d.f of Y. Differentiating $F_Z(z)$, we obtain the pdf $f_Z(z)$ of Z

as

$$\begin{aligned} f_Z(z) &= \frac{d}{dz}F_Z(z) \\ &= \int_{-\infty}^{\infty} f_X(x)\left(\frac{d}{dz}F_Y(z-x)\right)dx \\ &= \int\limits_{-\infty}^{\infty} f_X(x)f_Y(z-x)dy. \end{aligned}$$

(Leon-Garcia, 2017). □

3.7 Modeling of Climate Variability Impact on Agriculture

The use of vegetation index to monitor crop progression and yield has increased over the years, since it has been found to be positively correlated to yield for crops such as maize, soybean and sugarcane (Tarnavsky et al., 2008; Peralta et al., 2016; Sanches et al., 2018). One of the most commonly used index is Normalised Difference Vegetative Index (NDVI). NDVI is a graphical indicator, that can be used to analyse remote sensing measurements, and it assesses the target on the level of green vegetation by measuring it from value -1 to 1. It is one of the most commonly used index to measure the vegetation greenness. Using this index, the stress on the leaves can be diagnosed through spectral responses. The NDVI can show nitrogen status and chlorophyll, at micro level. Therefore, any limiting factors in the environment can be identified through the physiological stress in the leaves because of the variation in the concentration of nutrients (Lisboa et al., 2018). The NDVI is given by

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{3.53}$$

where NIR is reflection in the near-infrared spectrum and RED - reflection in the red range of the spectrum. This index takes values between -1.0 to 1.0, basically representing greens, with negative values mainly formed from clouds, water and snow; and values close to zero are primarily formed from rocks and bare soil. An NDVI value of between 0.2 to 0.3 represents shrubs and meadows, whilst large values of above 0.5 indicate crop fields and tropical forests.

The NDVI is decomposed into intrinsic mode functions (IMFs) and a residual. The IMFs represents the original data, at different time scales and the residual shows the general trend of the NDVI. In order to investigate the variability of the NDVI at different time scales, the index will be decomposed into different time scales using EEMD. The IMFs will be synchronised to temperature and precipitation, in order to identify the relationship between the time series.

The environmental factors: vapour pressure, soil moisture and sunlight together with temperature and rainfall, which were found to have considerable influence over the growth of plants and crops, are modelled using multiple linear regression.

3.7.1 Multiple Linear Regression Model of Environmental Factors

A multiple linear regression model is proposed to find and predict the influence of air temperature, rainfall, vapour pressure, soil moisture and sunlight on NDVI. Sunlight, vapour pressure and soil moisture have all been found to have considerable influence on plant growth (Zhao & Li, 2015; Miguez et al., 2018).

Probabilistic models that include more than one independent variables are called multiple regression. The model can be written as:

$$y_t \;\; = \beta_0 \;\; \sum_{i=1}^{n} \beta_i x_{i,t} + \epsilon_t, \qquad t = 1, 2, \ldots, n \tag{3.54}$$

Where y is the dependent variable NDVI, ϵ_t is error term, β_i is a coefficient and $x_{i,t}$ is the independent variables temperature, rainfall, vapour pressure, soil moisture. Parameters β_i and probabilistic error term ϵ_t are unknown. The value of the coefficient β_i determines the contribution of the independent variables $x_{i,t}$, if the other independent variables are held constant. The estimated values of $\beta_0, \beta_1, \ldots, \beta_n$ are given by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$ and the dependent variable is estimated by

$$\hat{y}_t \;\; = \hat{\beta}_0 \;\; \sum_{i=1}^{n} \hat{\beta}_i x_{i,t} + \hat{\epsilon}_t, \qquad t = 1, 2, \ldots, n \tag{3.55}$$

and the estimate $\hat{\epsilon}_t$ for the error term ϵ_t is determined as the difference between the observed and the predicted dependent variable, that is, $\hat{\epsilon}_t = y_t - \hat{y}_t$.

The model (3.54) can be written as

$$\mathbf{y} = \beta \mathbf{X} + \epsilon \tag{3.56}$$

where

$$\mathbf{y} = [y_1 y_2 \ldots y_n]', \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & . & . & . & x_{n1} \\ x_{21} & x_{22} & . & . & . & x_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{n1} & x_{2n} & . & . & . & x_{nn} \end{bmatrix}, \qquad \beta = [\beta_0 \beta_1 \ldots \beta_n]$$

$\epsilon = [\epsilon_1 \epsilon_2 \ldots \epsilon_n]$,where the assumption that $y_1, y_2, ..., y_n$ is a random sample is equivalent to the assumption that $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2 \mathbf{I}_n$. It is also assumed that the independent variables $x_{i,t}$'s are chosen in such a way that $\text{rank}(\mathbf{X}) \leq n$.

Least squares method is used to estimate the parameters β. The least squares estimator $\hat{\beta}$ is such that $\sum_{t=1}^{n} \hat{\epsilon_t}^2 = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$ is a minimum.

Thus $\hat{\epsilon}'\hat{\epsilon} = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)$ must be minimised.

Theorem 3.13. *Let $\mathbf{y} = \beta\mathbf{X} + \epsilon$ where $\mathbf{X} : n \times p$ is of full rank, $\beta : p \times 1$ is a vector of unknown parameters, and $\epsilon : p \times 1$ is a random vector with mean zero and $cov(\epsilon'\epsilon) = var(\epsilon) = \sigma^2 \mathbf{I}_n$. The least squares estimator for β is denoted by $\hat{\beta}$ is given by*

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} \tag{3.57}$$

(Chen et al., 2010).

To evaluate the performance of the considered models, the measures of average error are applied and these are mean absolute percentage error (MAPE) and root mean square error (RMSE).

Scale-location plot

The scale-location plot is very similar to residuals vs fitted plot, but it simplifies analysis of the homoskedasticity assumption. It takes the square root of the absolute value of standardised residuals instead of plotting the residuals themselves. Homoskedasticity means constant variance in linear regression. That is, for the regression model given by the equation (3.56) homoskedasticity means that for each component ϵ_i of ϵ $\text{Var}(\epsilon_i) = \sigma^2$. The scale location plot has fitted values on the x-axis, and the square root of standardised residuals on the y-axis. Scale-location plot is used as a diagnostic tool for the multiple linear regression model for the NDVI.

3.8 Assessing the Model Performance

In this study, MAPE and RMSE are applied to evaluate the model performance. MAPE and RMSE are based on statistical summaries of $\epsilon_t \, (t = 1, 2, \ldots, n)$. The average model error is written as:

$$\bar{\epsilon}_t = \left[\frac{\sum_{t=1}^{n} v_t |\epsilon_t|^\lambda}{\sum_{t=1}^{n} v_t} \right]^{\frac{1}{\lambda}} \tag{3.58}$$

where $\lambda \geq 1$ and v_t is a scaling assigned to each $|\epsilon_t|^\lambda$ according to its hypothesised error (Willmott & Matsuura, 2005). RMSE is calculated by letting $\lambda = 2$ and $v_t = 1$ and is given by:

$$RMSE = \left[n^{-1} \sum_{t=1}^{n} |\epsilon_t|^2 \right]^{\frac{1}{2}} . \tag{3.59}$$

In order to obtain MAPE, we set $\lambda = 1$ and $v_t = 1$ and is given by:

$$MAPE = \left[n^{-1} \sum_{t=1}^{n} |100\epsilon_t| \right] . \tag{3.60}$$

The RMSE and MAPE values range from 0 to infinity, and smaller values indicate a better model.

3.9 Summary

In this chapter, the discussion of the methodology employed for this study was done. A method used to deal with missing data in this study was discussed.

Discussions of spatial variability and trend analysis methods employed are presented. These methods are used to find an area with similar rainfall pattern, for further investigation using other methods. The methods of time series decomposition, EMD and EEMD, were covered in details. EEMD is a noise assisted data adaptive method, which was adopted in previous studies to deal with the challenge of 'mode mixing'. The method used to find the optimal amplitude of noise to be added was discussed. EEMD is used to decompose the original time series into IMFs and residual. The derived IMFs is modelled in this study using GEVD. The methods used perform model diagnostics, selection and performance were mentioned. These methods include QQ-plot, PP-plot and goodness-of-fit tests. EEMD, synchronisation and multiple linear regression are used to investigate the climate variability impact on sugarcane and grapevines.

In conclusion, a hybrid method of EEMD and GEVD is adopted in this thesis for the first time and it is expected to give improved results. Previous studies revealed that hybrid methods normally perform much better than the primary methods.

In the next chapter, a discussion of the results according to methodology outlined in this chapter is carried out.

CHAPTER 4

SPATIAL VARIABILITY AND TIME SERIES DECOMPOSITION

4.1 Introduction

In chapter 3, the methodology that was used in this thesis was presented. In this chapter, the climate variables were decomposed into a number of time series called intrinsic mode functions (IMFs), using both Empirical Mode Decomposition (EMD) and Ensemble empirical mode decomposition (EEMD) methods. A method to determine the optimal amplitude of noise to be added is presented. The IMFs are modelled using generalised extreme value distribution (GEVD). Model diagnosis shows that the IMFs GEVD models are more robust than the original time series.

In section 4.2, rainfall spatial analyses for the selected region under study is done; thereafter, decomposition of the rainfall time series is done using EMD and EEMD in sections 4.3 and 4.4 respectively. The modelling of the decomposed data (IMFs) is done in section 4.6. Most of the results presented in this chapter have been published in the scientific journals listed on research output on page xx and published copies of the articles are attached in Appendix C.

4.1.1 Study Area and Data

South Africa is classified as a predominantly semi-arid country. The country has extreme climate, with desert and semi-desert in the northwestern region and very wet conditions in the eastern coastal area (du Plessis, 2017). The province of Western Cape is located at the south western tip of South Africa; it is bordered in north by the Northern Cape, in the east by Eastern Cape and in the south and west by the Indian and Atlantic Oceans. The south western and west part of the province comprises of the Mediterranean region, which typically receives winter rainfall (May to August). Due to its Mediterranean climate, various agricultural produce is farmed in this area. At least 50% of South Africa's agricultural exports is produced by the Western Cape province, and it supplies about 20% of the national agricultural production (du Plessis, 2017). The main produce is grapes, which is largely used for wine. The selected region for the study has been facing a lot of water challenges recently. There has been growing general interest in the winter rainfall, mainly due to the then threat of "day zero" in 2018 (Maxmen, 2018; Wolski, 2018).

The average monthly rainfall and temperature data were obtained from the South African Weather Service (SAWS) for the period from 1980 to 2018, for an area between 18-21ºE and 33-35ºS, which contains 15 weather stations as shown in Fig. 4.1. The number of missing data, mean, standard deviation and the highest rainfall for the stations are presented in Table 4.1. Paarl station has the highest number of missing values and the highest recorded maximum average monthly rainfall of 425.2mm. The monthly average rainfall for all the stations is between 29.6mm and 60.6mm.

In order to deal with the missing data in some of the stations, multiple imputation by chained equations (MICE) is used, which is described in section 3.2. The imputation method employed is predictive mean, which imputes

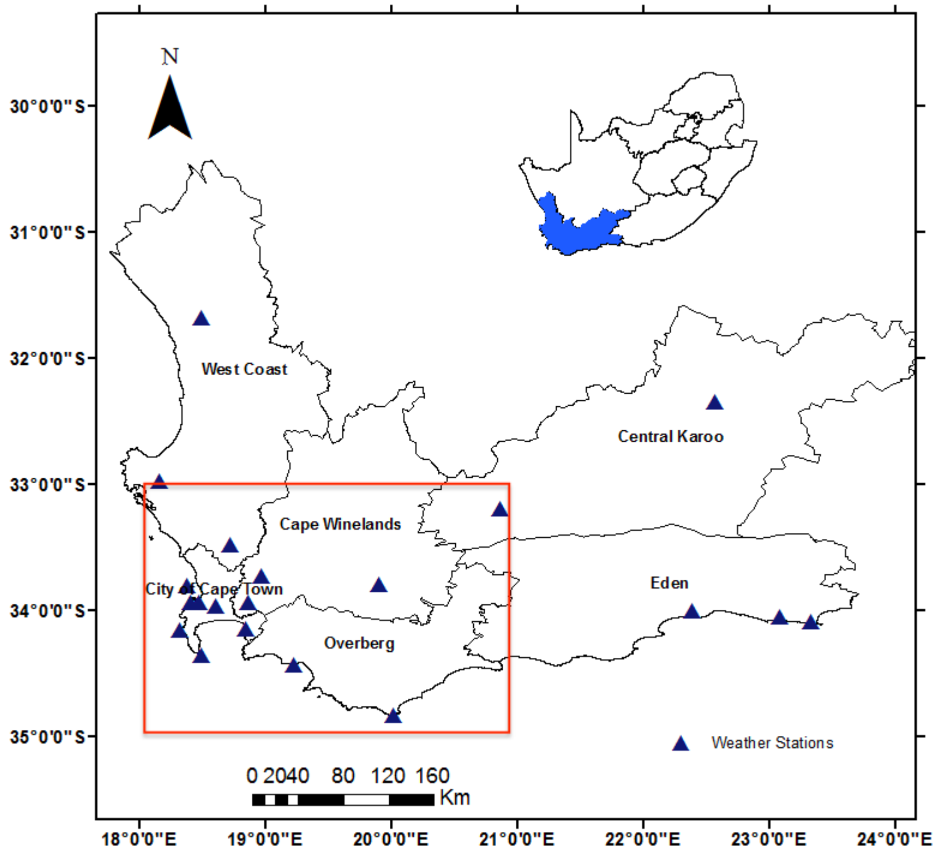


Figure 4.1: Study Area (square) in Western Cape South Africa with 15 weather stations.

missing values by means of the nearest-neighbour donor, with distance based on the expected values of the missing variables conditional on the observed covariates. In the data set for the stations, there are 33 missing variables and these were imputed. Paarl station has the highest number of missing values.

Rainfall variability in South Africa has been shown to be influenced by El Niño-Southern Oscillation (ENSO) and becomes nonlinear, with extreme weather conditions (Kane, 2009). Refer to section 1.2 (page 7) for the definitions of El Niño and La Niña. The other driver that has been shown to have an impact on weather patterns is the Quasi-biennial oscillation (QBO) (Begue et al., 2010). QBO is a regular variation of the winds that blow high above the equator. Strong winds in the stratosphere travel in a belt around the

Table 4.1: Descriptive statistics for the stations.

Station Name	**No. of Missing Values**	$\bar{x}_t$ **(mm)**	σ	**Highest rainfall (mm)**
Cape Agulhas	0	37.8	32.0	192.6
Cape Point	0	32.6	28.1	198.6
Cape Town Slangkop	0	41.6	36.7	189.0
Cape Town WO	0	43.8	40.8	229.4
Hermanus	0	40.6	33.8	233.6
Lainsburg	0	10.6	14.5	80.6
Langebaanweg AWS	0	22.3	22.4	133.1
Malmesbury	1	31.6	34.5	166.0
Molteno Reservoir	5	60.6	52.2	260.0
Paarl	17	67.5	68.5	425.2
Robbeneiland	1	29.6	31.3	180.8
Robertson	3	27.9	27.1	193.1
S.A. Astronomical Observatory	4	50.0	49.5	288.1
Stellenbosch	0	62.5	53.8	276.1
Strand	2	49.1	45.0	235.2

$\bar{x}_t$= mean and σ = standard deviation

planet, and these winds completely change direction in approximately 14 months (Kane, 2009). The data for Niño 3.4 and QBO is publicly available on Climate Explorer website (http://climexp.knmi.nl) (Van Oldenborgh & Burgers, 2005; Brönnimann et al., 2007). Using EEMD and synchronisation, the impact of ENSO and QBO on rainfall is investigated in later sections.

4.2 Rainfall Spatial Variability

South Africa generally experiences different types of rainfall, from winter rainfall to late summer rainfall, as has been discussed earlier in section 1.2. Water and disaster management may therefore vary according to the location. Studies have revealed that areas in the central and northeastern parts of South Africa are becoming drier and the eastern region around Drakensberg, which also includes Lesotho, has recorded an increase in rainy days (Jury, 2013; MacKellar et al., 2014). Additionally, the rainfall patterns have been changing with rainy days either becoming less, or more, in some instances. Groisman et al. (2005) showed that there has been an increase in rainfall in eastern, while MacKellar et al. (2014) revealed that there was no spatial coherence in the region. Therefore, rainfall trend and spatial analyses is very important to provide the spatial distribution of rainfall over the study area for the study period. The weather stations are located in specific areas, and spatial variability analyses reveal the spatial distribution for the un-sampled areas. This analyses may assist the Department of Agriculture and farmers to plan, knowing the expected amount of rainfall for their area. In this study, spatial and trend analysis was done, so that an area with similar rainfall pattern can be selected for the study, mainly due to the spatial differences in rainfall patterns in Western Cape (Jury, 2013). In this section, an analyses of rainfall variability in Western Cape for the period 1980 to 2018 is presented. This is done by using Mann-Kendall test (MK), coefficient of variation (CV) and kriging.

4.2.1 Rainfall Trend and Spatial Variability

Mann Kendall (MK) test results for the stations in the selected region, which mostly receives winter rainfall are shown in Table 4.2 below. Negative z-statistic shows that there is a monotonic downward trend in the rainfall, while a positive z-statistic shows that there is a monotonic upward trend. There is generally monotonic downward trend in all the stations for May, with significant trend for Paarl Station ($p = 0.041$) and Robertson ($p = 0.021$). However, in June, there is generally monotonic increasing trend with significant trend for Hermanus and Strand stations. There were no substantial change in the trend for July and August. For the overall season (MJJA column), Cape Town WO has a significant monotonic downward trend ($p = 0.041$), while Strand station has a significant monotonic upward trend ($p = 0.037$). The results from MK Test were generally not uniform for the region. In order to find out if the variability is consistent for the region for the months under study, spatial analyses was done.

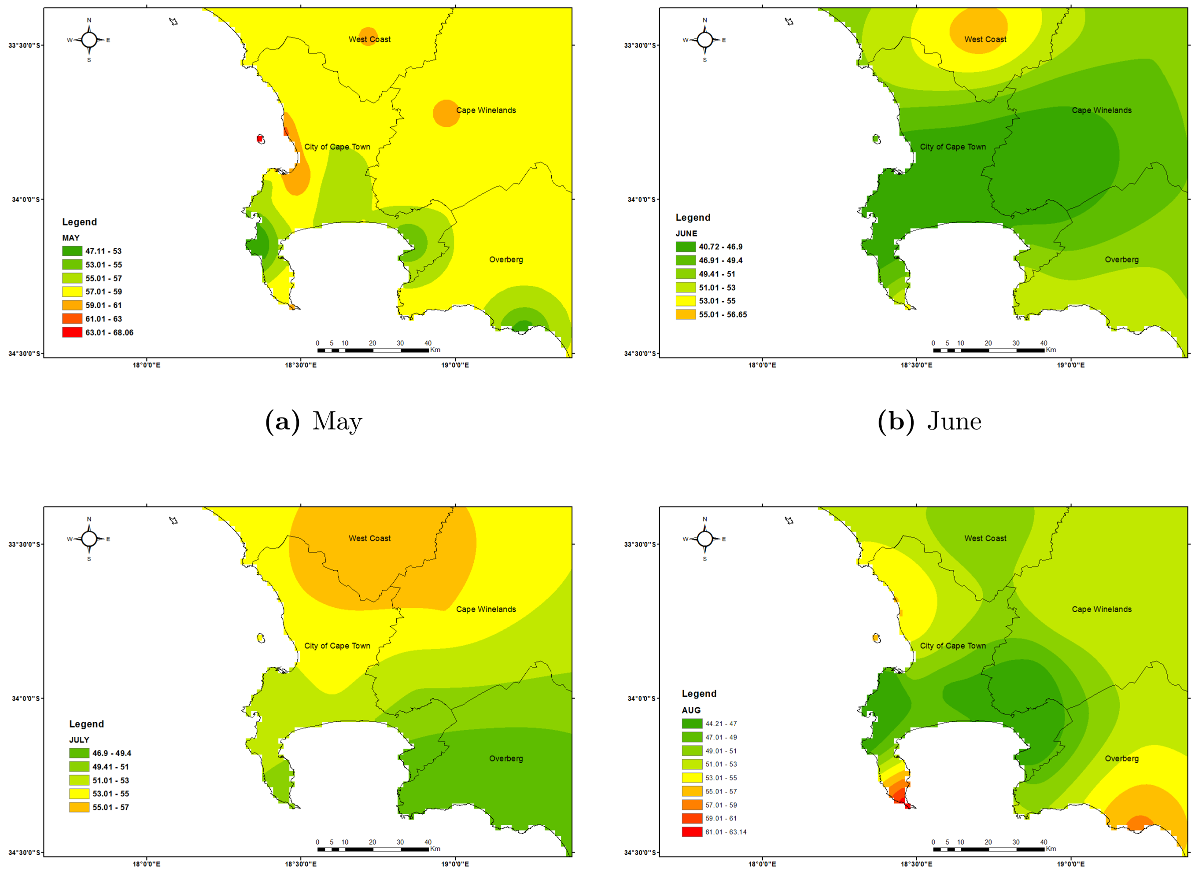
Table 4.2: z-statistic values of seasonal rainfall using MK for Stations in the study region in Western Cape Province.

Station Name	May	June	July	August	MJJA
Cape Agulhas	-1.221	0.565	0.512	1.322	-0.671
Cape Point	-0.100	0.352	1.018	1.634	0.327
Cape Town Slangkop	-0.451	0.056	0.394	0.480	0
Cape Town WO	-1.938	-0.397	-0.387	-0.350	-1.985*
Hermanus	-1.902	2.264*	1.268	0.604	1.479
Lainsburg	-1.223	0.655	0.544	0.444	1.232
Langebaanweg AWS	0.897	0.434	0.769	1.533	0.125
Malmesbury	-0.545	0.967	0.074	0.670	0.174
Molteno Reservoir	-1.265	0.099	-0.421	-0.272	-0.868
Paarl	1.881*	-0.600	-0.664	-0.049	1.508
Robbeneiland	-1.542	0.537	-0.140	0.724	-0.397
Robertson	1.981*	-0.500	-0.604	-0.149	1.408
S.A. Astronomical Observatory	-1.106	-0.377	-0.653	-0.050	-1.257
Stellenbosch	-1.220	1.132	-0.754	0.565	-0.352
Strand	-0.620	2.454*	1.410	0.338	2.087*

$\bar{x}_t$= mean and σ = standard deviation

CV was calculated for different stations, and kriging was used to find the spatial patterns of the rainfall. The spatial variability of the Western Cape map is presented in Fig. 4.2. June is the least variable with CV of between 40% and 59% and May has CV of up to 68%. The average CV for all the months is approximately 50%. May and August are more variable than June and July, which are the wettest months. In May, most part of the region selected has CV above 57%, while the CV for the greater part of the region is less than 50%.

Rainfall seasonal variability analyses were carried out, and the spatial patterns presented in a map are shown in the appendix in Fig. B1. Monthly rainfall data was used to generate annual rainfall, for the summer, covering the period December, January and February (DJF), autumn covering the period March, April and May (MAM), winter covering the period June, July and



(a) May **(b)** June

(c) July **(d)** August

Figure 4.2: Monthly winter rainfall variability (CV) for the region in Western Cape for the months a) May b) June c) July and d) August .

August (JJA) and spring covering the period September, October and November (SON). Generally, there was a variability of between 19% and 127%. Annual Rainfall is less variable in City of Cape Town, greater part of Overberg and part of West Coast and Cape Winelands. The wettest months JJA, compared to the northern areas, are also less variable for the same areas. A variability of between 25% and 33% was observed, while the northern part of West Coast is more variable, with a variability of up to 126% for DJF period. The Fig. B1b shows that MAM is more variable with most of the province having a CV of between 40% and 50%, while for JJA, most of the province has CV of between 30% and 40%. Annual rainfall, when compared with the seasons, is the least variable, with CV of between 19% and 36%.

The spatial variability analyses has shown that there is no uniform rainfall patterns, even for the region that receives rainfall during the same season. Therefore, in this study, an area with similar rainfall pattern of 11 weather stations is selected. The area consists of most of Cape Town Metropolitan Municipality and small areas of west Coast and Cape Winelands districts. In this study, the area is referred to as Cape Town Metropolitan area (CT Metro area). The results from Mann-Kendall test revealed that there is generally monotonic decreasing rainfall trend for May and August, and an increasing trend was observed for June and July. Furthermore, winter rainfall was shown to be more variable in the months of May and August, and less variable for June and July in the selected region. In order to find the source of the variability, the rainfall data is decomposed into multiple time scale time series, representing different periods for the selected region. In the next sections, the decomposition methods EMD and EEMD, are applied to the rainfall data. The issues that have arisen from EEMD method was first dealt with before applying it to the selected region under study.

4.3 Decomposition Using EMD

The monthly average rainfall of the 10 stations located in CT Metro area is averaged and the resultant time series plot is shown in Fig. 4.3. This average is used in this section and the sections that follow to carry out the study. The rainfall time series is standardised and decomposed using EMD into a number of IMFs and the results are presented in this section. The Cauchy-type convergence is used as stopping criteria, with standard deviation equal to 0.2. In order to determine this, a significance test is carried out. In this section, the rainfall data is decomposed using EMD to compare the results with EEMD.

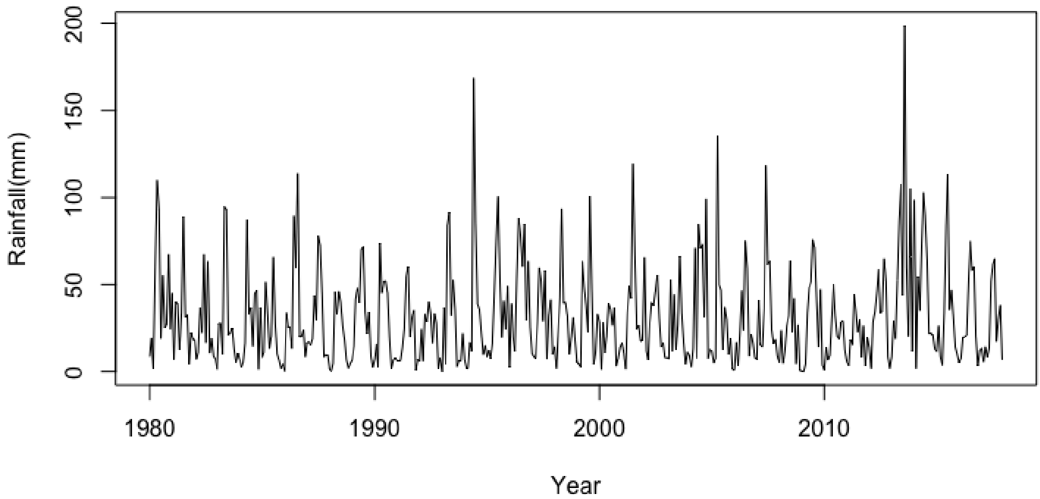


Figure 4.3: Time series plot of the monthly rainfall average of the 10 weather stations found between 18.2-19.2ºE and 33.5-34.5ºS.

4.3.1 EMD Decomposition Results and Discussion

The average monthly rainfall for the selected region (CT Metro area) was decomposed into 7 IMFs and residual as shown in Fig. 4.4. Time series plot of standardised rainfall and superimposed residual plot is presented in Fig. 4.5. The residual plot shows the general trend of the rainfall. The mean period of the IMFs is calculated by dividing the size of the data and the number of local maximum in any given IMF (that is 456/number of local maximum). Using this method, IMF1 has a period of about 3 months, IMF2 6.2 months, IMF3 12.3 months, IMF4 26 months, IMF5 52 months, IMF6 114 months and IMF7 228 months. IMF1 captures the noise found in the rainfall data, IMF2 inter-annual oscillation, IMF3 annual (seasonal) oscillation, IMF4 2-years oscillation, IMF5 4.5-year oscillation, IMF6 7-year oscillation and IMF7 16.5 year oscillation. The plot of the residual shows the general trend of the rainfall.

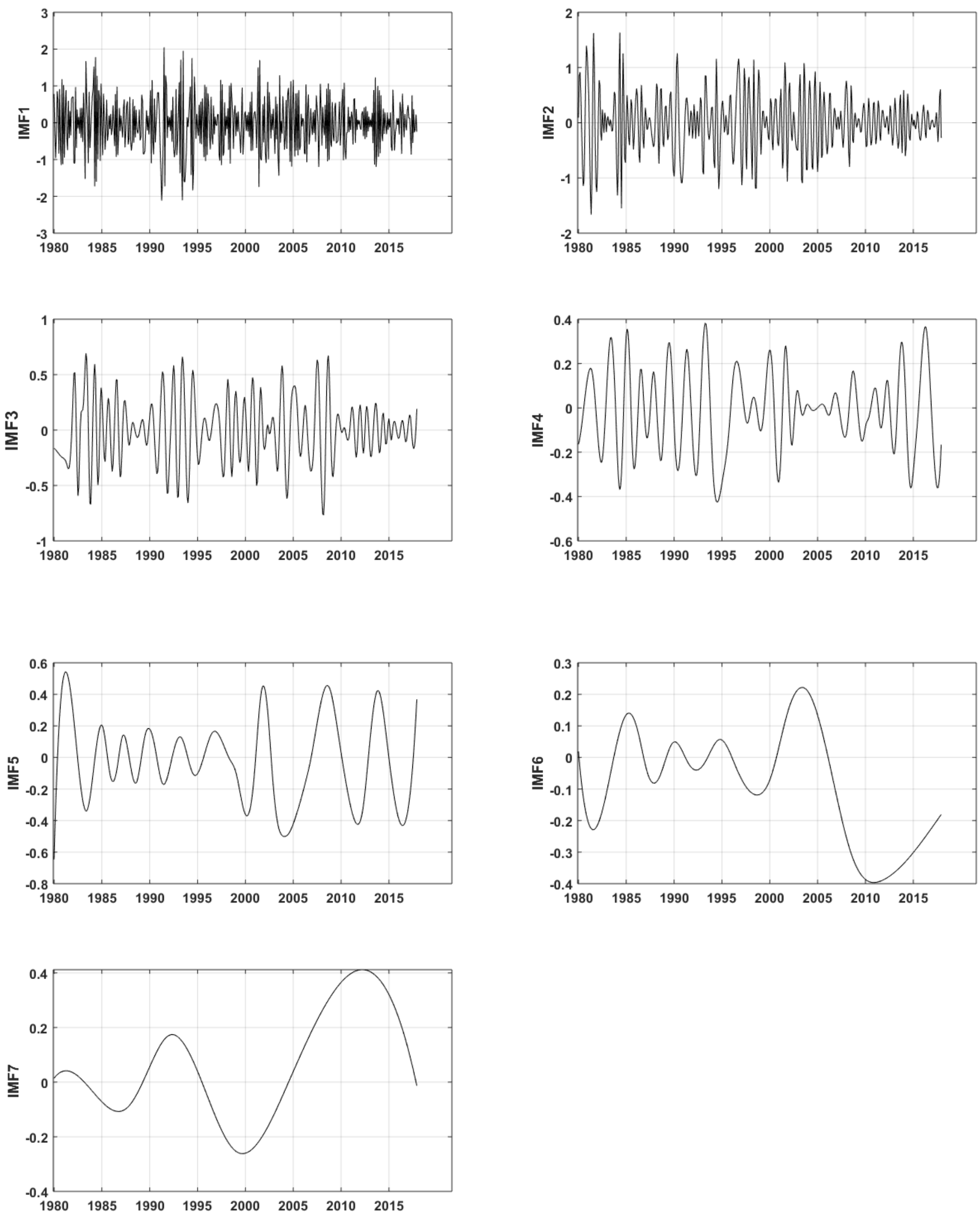


Figure 4.4: Decomposition of rainfall into IMF1 to 7. The x-axis represents the time in years and y-axis represents the frequency. The graphs are labelled on the y-axis from IMF1 (first graph on the left) to IMF7 (last graph on the bottom left).

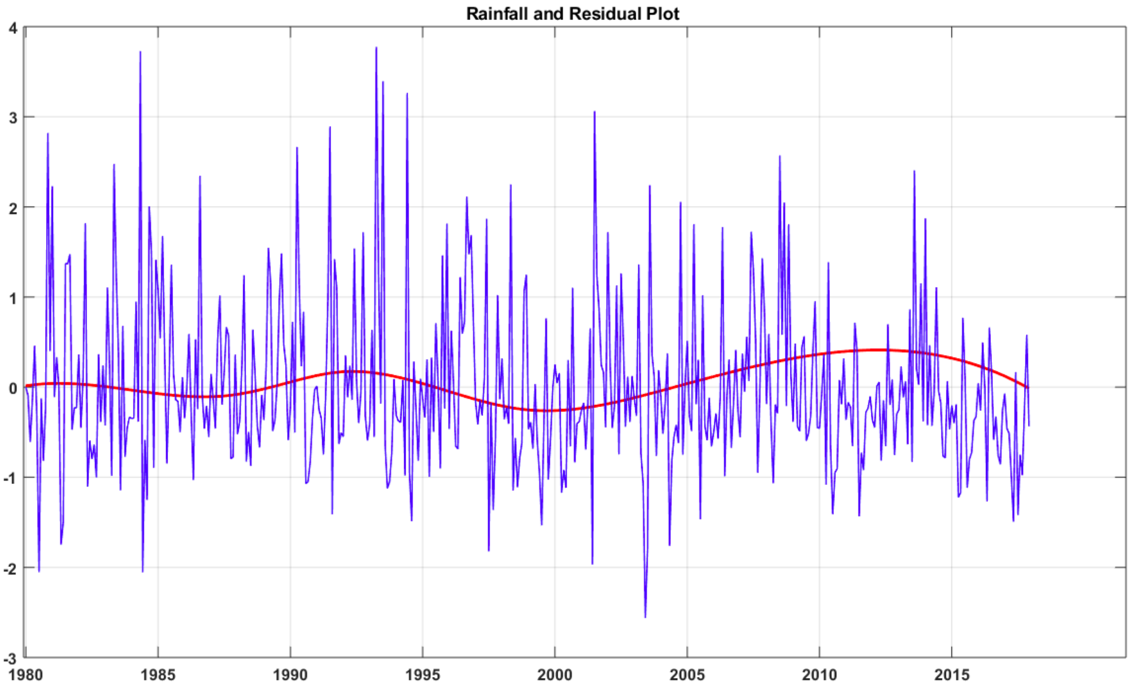


Figure 4.5: Time series plot of standardised rainfall and residual (red line) decomposed using EMD. The residual shows the rainfall trend.

The presence of noise in the IMFs is tested using a statistical significant test presented in section 3.4.3, and the results are displayed in Fig. 4.6. The test is carried out at 5% and 1% level of significance, with z-score $k = 1.96$ or 2.326 respectively. IMF1 is known to contain noise, therefore, it is used to gauge level of noise in other IMFs. The results presented in the Fig. 4.6 has IMF2 to 7. If an IMF (represented by asterisk on the graph) is located above the significant lines, it means they contain noise. The test shows that IMF3 and IMF6 are significantly different from white noise at 5% level of significance. Only IMF2, IMF3, IMF5 and IMF6 are significantly different from white noise at 1% level of significance.

EMD has "mode mixing" challenge, a case in which the signal of one IMF is found in the other, thereby giving a distorted signal. Additionally, EMD sometimes fails to identify intrinsic oscillations in the signal (Wu & Huang, 2009). For example, in the rainfall decomposition above on IMF5, the intrinsic

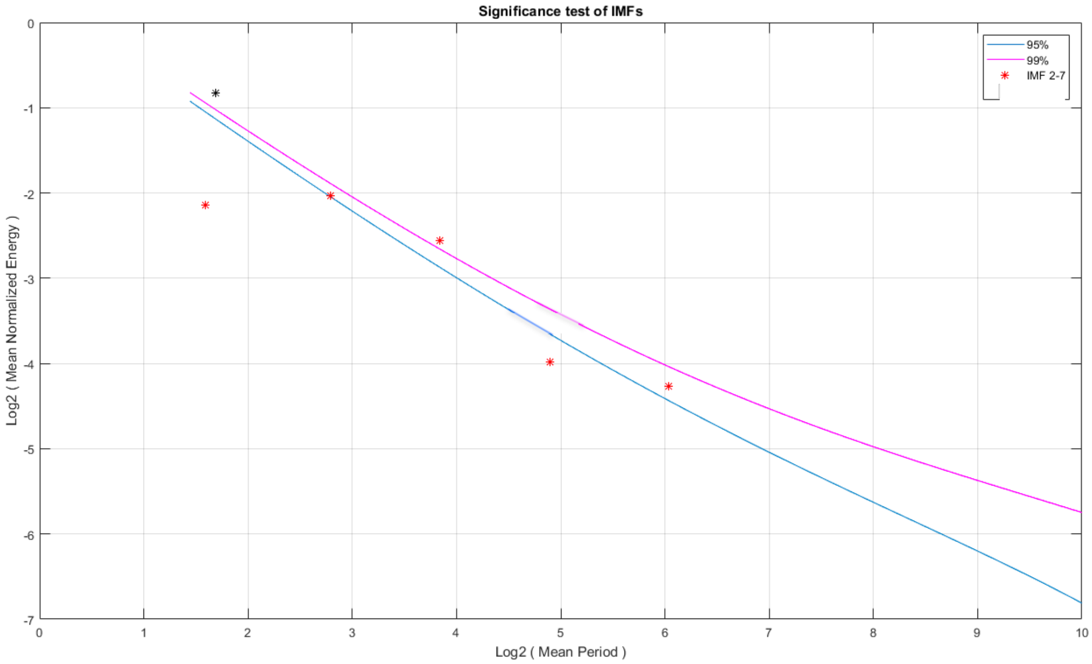


Figure 4.6: Significant test of the rainfall IMFs decomposed using EMD. The lines are 95th and 99th percentiles. The asterisks correspond to the pairs of the averaged mean energy density and the averaged mean period of IMF 2 to 7 (from left to right).

oscillation between 2005 and 2010 is not clearly represented. In order to cater for these weaknesses, data assisted method, Ensemble Empirical Mode Decomposition (EEMD), which consists of adding noise to the data before carrying out the EMD algorithm was developed by Wu & Huang (2009). In the next section, the same data is decomposed using EEMD.

4.4 Decomposition Using EEMD

EEMD is a noise assisted method that consists of adding noise before decomposing the data using EMD. A different amplitude of noise is used for a certain number of ensembles. However, the amplitude of noise to be added to cater for the mode mixing in EMD, is a challenge faced when EEMD is employed. In this section, a method to determine the optimal noise to be added is presented. Using the results from this, the CT Metro area rainfall data is decomposed and the physical meaning of the IMFs is discussed.

4.4.1 Proposed Method to Determine the Amplitude of the Noise to be Added

A proportion of between 0.1 and 0.4 of standard deviation (SD) was proposed by Wu & Huang (2009), as an amplitude of noise that can be added before carrying out EMD. However, there is no clear method to determine the exact amplitude of noise to be added. To determine the size of the amplitude a threshold established by Ayenu-Prah & Attoh-Okine (2010), and variance contribution rate are used.

EEMD process is carried out for different amplitudes of between 0.1 and 0.4. Using the results from these different amplitudes, correlations are then calculated between the rainfall anomalies and respective IMFs. A threshold value using equation (3.26) was also calculated. The correlations and threshold, λ, are shown in Table 4.3. The first row contains correlations for EMD, second row (EEMD($\sigma_{0.1}$)) is EEMD with amplitude of 0.1 added, third row (EEMD($\sigma_{0.2}$)) is EEMD with amplitude of 0.2, and so on.

According to Ayenu-Prah & Attoh-Okine (2010), all the IMFs with correlations less than the threshold have insubstantial contribution to the overall variability of the rainfall pattern. IMFs 1, 2 and 3 were found to be important for the IMFs decomposed using EMD, since their correlations were above the threshold of λ= 0.2532. Now when using EEMD with amplitude of noise $\sigma_{0.1}$ added, IMFs 1 to 6 were found to be important, since they were all above the threshold value of 0.164071. For $\sigma_{0.2}$, $\sigma_{0.3}$ and $\sigma_{0.4}$, only IMFs 1 to 5 were found to be important, since they were all above the threshold value. The variable contribution rates for EEMD($\sigma_{0.2}$) (Table 4.4) showed that IMFs 1 to 6 had contribution rates that were sizeable, with IMF1 having a VCR of 55%, IMF2 26.57%, IMF3 12.32%, IMF4 8.03%, IMF5 3.55% and IMF6 had VCR of 0.86%. IMF6 is dimmed relevant only when $\sigma_{0.1}$ is used. Additionally, the periods captured by IMF 1 to 6 were found to sizeable contribution to the variability of rainfall in South Africa (Kane, 2009; Dieppois et al., 2016). Using this argument, we can determine the optimum amplitude of noise to be added as $\sigma_{0.1}$. The details of the contribution rates are shown in Table 4.4.

Table 4.3: Correlations between normalised rainfall data and IMFs

	IMFS								
	1	2	3	4	5	6	7	Res	λ
EMD	0.4957	0.3073	0.2080	0.0794	0.1628	0.0311	0.0671	0.0848	0.2532
EEMD($\sigma_{0.1}$)	0.7682	0.5738	0.3932	0.2393	0.2312	0.1761	0.0822	0.1296	0.1641
EEMD($\sigma_{0.2}$)	0.7726	0.5804	0.3965	0.2632	0.2418	0.1606	0.0744	0.1308	0.1634
EEMD($\sigma_{0.3}$)	0.7701	0.5757	0.4006	0.2494	0.2515	0.1591	0.0851	0.1302	0.1638
EEMD($\sigma_{0.4}$)	0.7684	0.5748	0.3949	0.2562	0.2497	0.1627	0.0885	0.1322	0.1640

Table 4.4: Periods and Variable Contribution Rate (%)

IMFs and Residue	**IMF1**	**IMF2**	**IMF3**	**IMF4**	**IMF5**	**IMF6**	**IMF7**	**Res**
Period (months)	2.98	6.16	12.32	25.33	50.67	114	228	
EMD	50.05	25.29	7.46	3.08	6.29	2.86	3.59	1.38
EEMD($\sigma_{0.1}$)	55.04	26.57	8.03	3.03	3.55	0.86	0.49	2.43
EEMD($\sigma_{0.2}$)	56.83	24.22	7.86	3.21	4.84	0.95	0.93	1.17
EEMD($\sigma_{0.3}$)	57.75	23.73	7.90	2.93	4.44	0.86	0.69	1.70
EEMD($\sigma_{0.4}$)	59.19	22.80	7.94	2.94	4.62	0.55	0.55	1.41

A significant test of the IMFs is carried out with only IMF2 not significant at 5% and 1% level of significance as shown in Fig. 4.7. There is an improvement in the importance of IMFs when compared to the IMFs found using EMD method (Fig. 4.6). In the following section, a discussion of results found using noise amplitude of 1% of standard deviation is done.

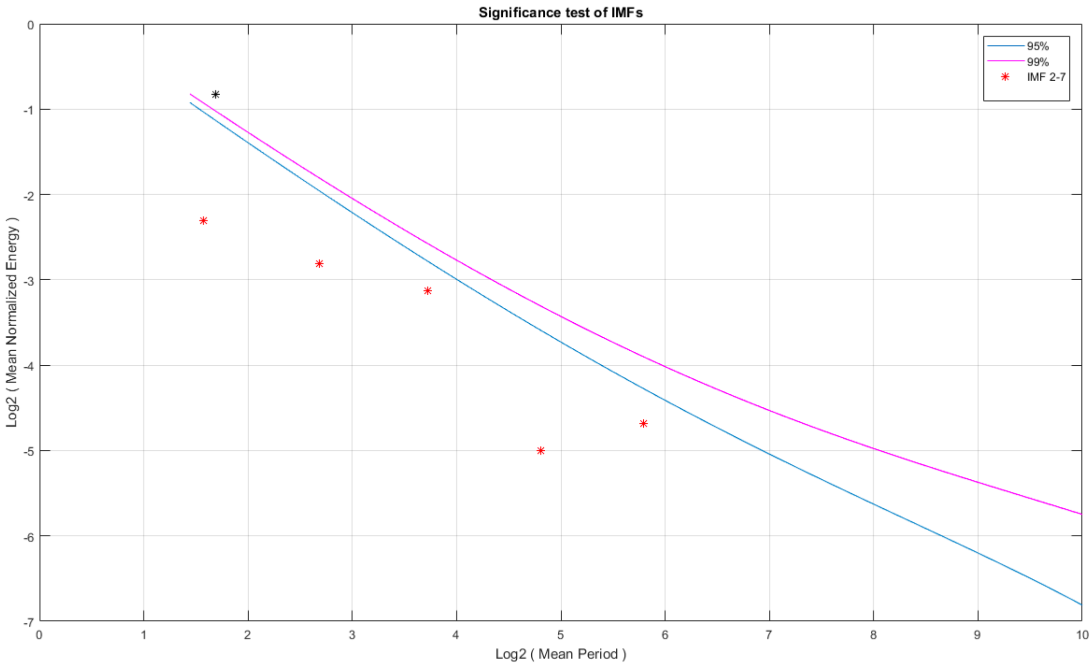


Figure 4.7: Significant test of the rainfall's IMFs decomposed using EEMD. The lines are 95th and 99th percentiles. The asterisks correspond to the pairs of the averaged mean energy density and the averaged mean period of IMF 2 to 7 (from left to right).

4.5 Rainfall Oscillations

The standardised monthly rainfall data was decomposed using EEMD with an amplitude of noise of 0.1 of standard deviation and an ensemble size of 100. It was noted that the ensemble size which is above 100 did not affect the general results considerably, except that it has increased computational time. The data was decomposed until when there was at most one maximum and one minimum in the residual. Time series plot of residual and standardised rainfall are presented in Fig. 4.8. The residual plot shows the general trend of the rainfall. The results from the trend are in agreement with results found using MK test, which showed that there was generally no considerable change in the rainfall. Previous studies have also made the same observation (Kane, 2009; MacKellar et al., 2014). However, it was observed that there has been an increase in the rainfall variability. Therefore, the decomposition that is done in this section is used to investigate the source of variability of the derived IMFs and give the interpretation of their physical meaning.

From the standardised rainfall data, 7 IMFs were found and are shown in Fig. 4.9. IMF1 contains the noise, IMF2 has a period of about 6 months, IMF3 has a period of about 12 months, which corresponds to an annual (seasonal) oscillation, IMF4 has a period of about 26 months which approximated a 2 year oscillation, IMF5 has an oscillation of about 54 months (4.5 years) and IMF 6 captures a quasi-decadal oscillation (7 year period). Previous studies have used wavelet analysis and identified similar oscillations in the South African rainfall that were also found in this study (Kane, 2009; Dieppois et al., 2016). In these previous studies, winter rainfall was found to be having substantial 2-3 year period and 3-4 year period, which contributed to the rainfall variability. Compared to wavelet analysis, EEMD is adaptive, intuitive and does not use basis functions. Additionally, the impact of different climate drivers, at different time scales can be shown.

The probability density function for each of the IMFs found is approximately normally distributed. The IMFs and residual are added together to reconstruct the data. The reconstructed data approximates the original data with Root Mean Square Error of order 10^{-9}. This clearly shows that EEMD is lossless decomposition, with minimal data being lost in the decomposition, and it manages to capture most of the oscillations in the data.

The orthogonality of the IMFs is calculated using equation (3.13) and (3.14). It is noted that the maximum value of orthogonality between the IMFs is found

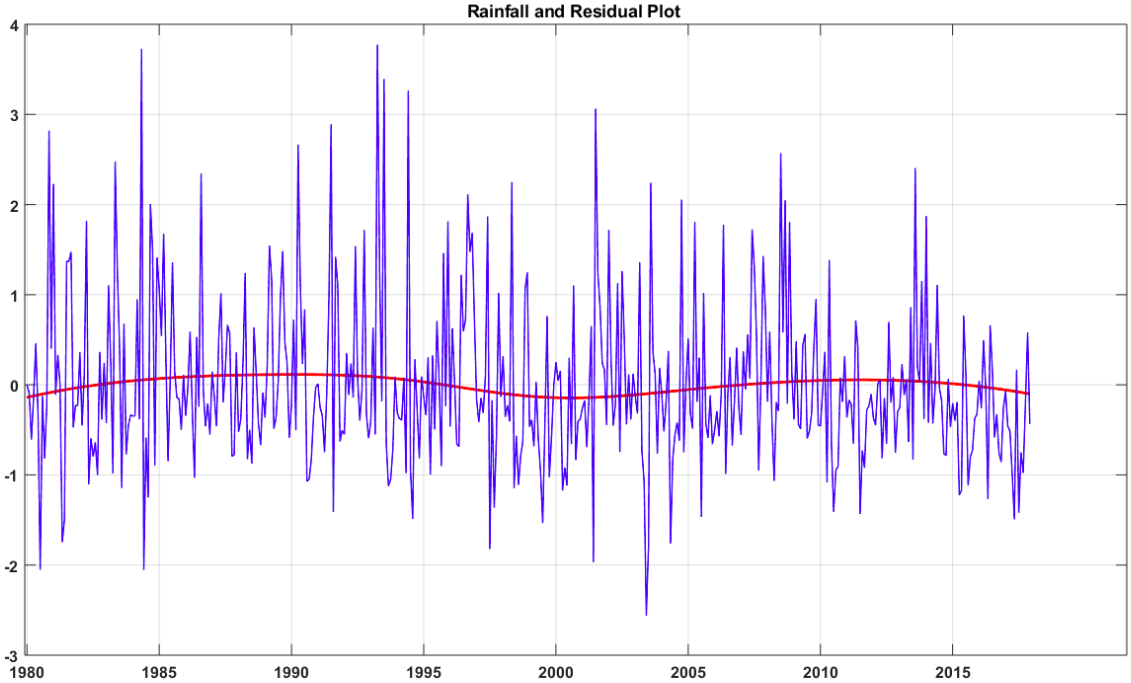


Figure 4.8: Time series plot of standardised monthly rainfall and residual (red line) decomposed using EEMD. The residual shows the rainfall trend.

to be approximately equal to 0.001, and it is way below the acceptable value of less than 0.1. The index of orthogonality for the IMFs is 0.594×10^{-4} for rainfall, which confirms that there is less amount of information leakage.

The derived IMFs need to be interpreted and their physical meaning deduced. Therefore, synchronisation, which finds the correlations between pair of oscillating systems is used. Synchronisation of coupled oscillating systems means appearance of certain relations between their phases and frequencies, whilst the amplitudes can remain non-correlated (Rosenblum et al., 2001). Here we use this concept in order to reveal the interaction between rainfall and other climatic drivers. Time series that are phase synchronised or locked have a prominent peak in the histogram plot, at a certain phase difference; and unrelated time series will have a uniform or diffuse distribution (Cazelles & Stone, 2003).

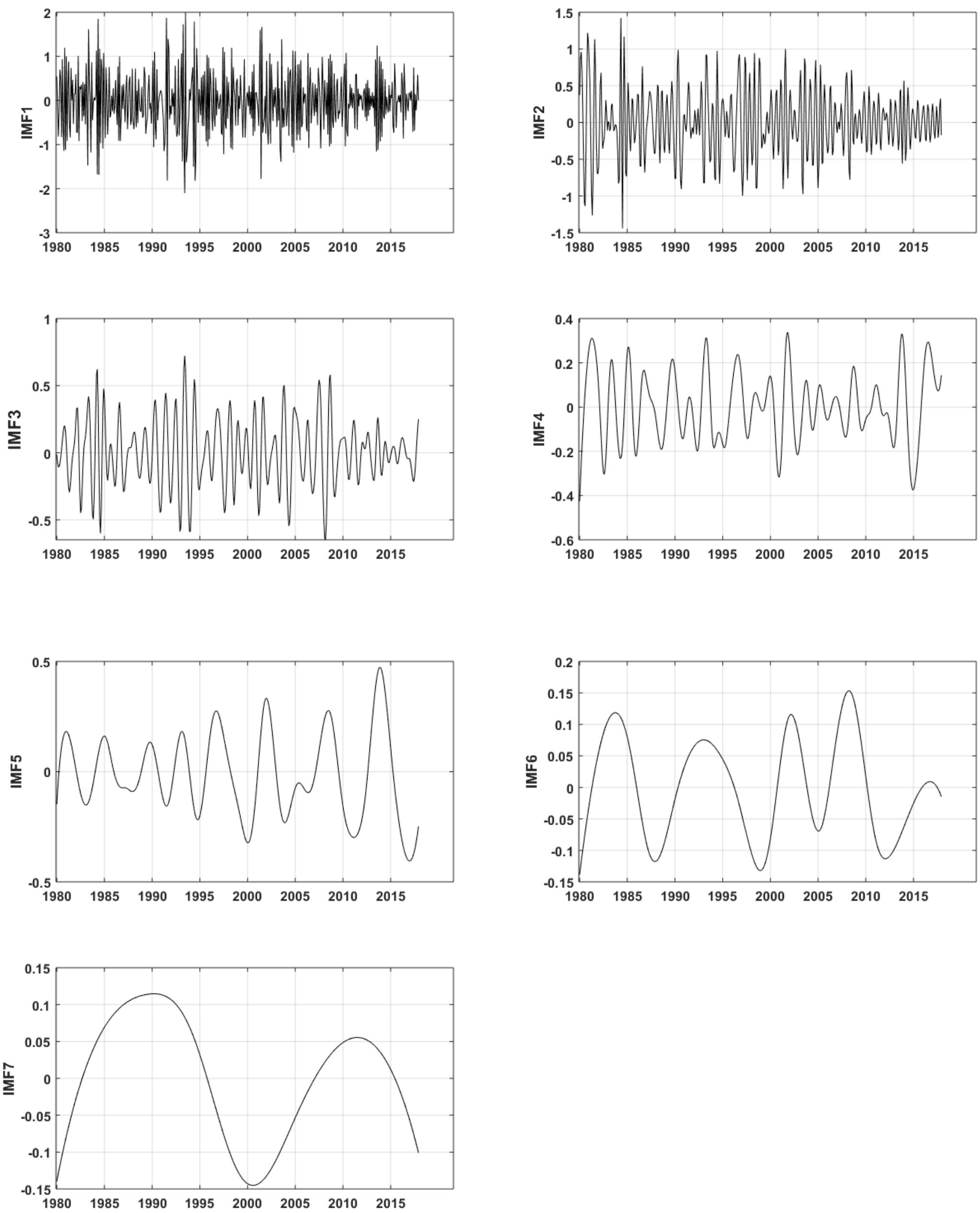


Figure 4.9: Decomposition of rainfall into IMF1 to 7. The x-axis represents the time in years and y-axis represents the frequency. The graphs are labelled on the y-axis from IMF1 (first graph on the left) to IMF7 (last graph on the bottom left).

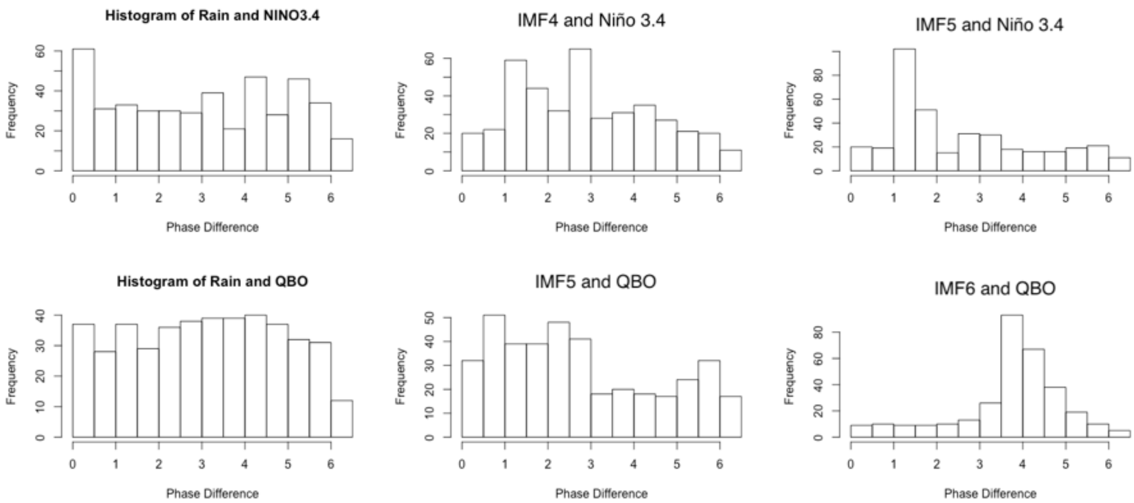


Figure 4.10: Histogram plot of synchronisation of rainfall, QBO and Niño 3.4 index IMFs. The x-axis represents the phase difference in radians and y-axis represents the frequency.

4.5.1 Influence of ENSO and QBO on Rainfall Variability

Synchronisation of the IMFs show that there is weak coupling between the original rainfall time series and Niño 3.4 or QBO, as shown in Fig. 4.10. The plots show a clear peak for rainfall's IMF5 with Niño 3.4 IMF5 (top right) which shows that there is phase locking. The last graph on the bottom right is a plot of rainfall's IMF6 and QBO IMF6, which shows that there is phase locking, since there is a clear peak. These results for Niño 3.4 shows that there may be an influence of ENSO on the rainfall pattern. This is in agreement with the results that were found by Philippon et al. (2012), which found that there was a sizeable association between ENSO and winter rainfall. They showed that there is a strong correlation of the May-June-July season with ENSO than any other period of the year. The QBO signal identified is in agreement with a study by Kane (2009),which also identified the presence of QBO and its considerable contribution to the variability of the winter rainfall for the same region. In this study, using EEMD, the correlation is further done at a monthly, rather than seasonal level.

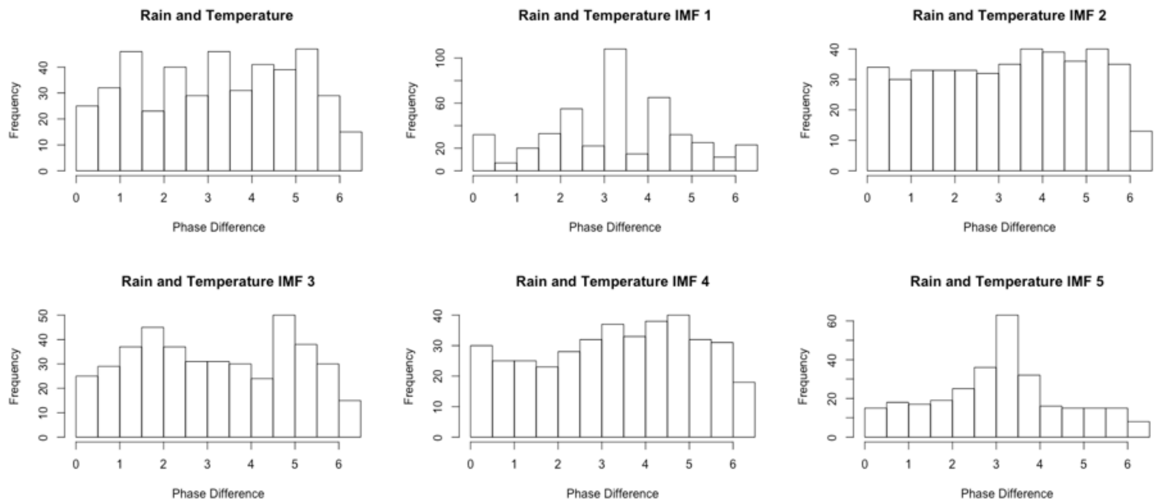


Figure 4.11: Histogram plot of the synchronisation of rainfall IMFs and temperature IMFs. The x-axis represents the phase difference in radians and y-axis represents the frequency

4.5.2 Influence of Temperature on Rainfall Variability

The temperature for the selected region was also decomposed into 7 IMFs and a residual. There is weak coupling between the original rainfall time series and temperature, as shown in Fig. 4.11 below. However, phase locking is observed for IMF1 and IMF5, since there is a clear peak for both of them. IMF1 captures the noise found in the signal and IMF5 is a 4.5-year oscillation. The phase-locking in IMF1 suggests that the temperature variability may have an impact on the rainfall patterns. The synchronisation on IMF5 shows that the temperature changes may have a long term impact on the rainfall variability. This is consistent with other studies that have used different climatic models to find the impact of increasing temperature in Southern Africa (Maúre et al., 2018; Nangombe et al., 2018; Nikulin et al., 2018). These models predict that there may be an increase in extreme rainfall patterns. In our study, we have managed to show the direct impact of increasing temperature using historical data.

4.6 Modelling Rainfall Using EEMD and GEVD

The block maxima approach taking months as independent and identically distributed (i.i.d.) is used in this section. A generalised extreme value distribution (GEVD) is fitted to the rainfall data and its derived IMFs, using maximum likelihood estimator. A number of studies used the original time series to find the GEVD model; however, in this study, the decomposed data is used. A model for the original data is presented and compared with the decomposed data model.

An assumption of independent and identically distributed (i.i.d.) (that is, randomness) is made when modelling data using statistical distributions. Testing of randomness is done using Bartels and the Ljung-Box tests. The null hypothesis for the tests is: the monthly mean rainfall is i.i.d. The results presented in Table 4.5 demonstrate that the data is i.i.d. and there is no serial correlation.

Table 4.5: Results of tests for i.i.d. of mean monthly rainfall data for the period 1980-2018

Test	***p*-value**
Bartels	0.1666
Ljung-Box	0.717

4.6.1 Modelling Rainfall Data Using GEVD

The block maxima approach in which months are taken as independent and identically distributed blocks was used. The normalised rainfall data is fitted a

GEVD model using maximum likelihood estimator. The maximum likelihood estimates of the GEVD with their corresponding standard errors in brackets and negative log-likelihood (NLL) value are presented in Table 4.6. The results show that the rainfall data can be modelled using a Weibull class of distribution, since $\hat{\xi}$ = -0.07085871 $<$ 0. Combining estimates and their corresponding standard errors, the 95% confidence intervals for μ, σ and ξ are [-0.4659007; -0.30904870], [0.7324351; 0.83925657] and [-0.1182223; -0.02349515] respectively.

Table 4.6: Maximum likelihood estimates (standard errors) of rainfall

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	NLL Value
-0.38747470	0.78584583	-0.07085871	587.6751
(0.04001400)	(0.02725088)	(0.02416553)	

Fig. 4.12 presents diagnostic plots of the GEVD model fitted to the monthly mean rainfall, using maximum likelihood estimator. The diagnostic plots show that each set of plotted points are close to the line of best fit, validating the use of the GEVD, although there are a number of outliers at the beginning and end of the line. Return level plot shows that there are outliers on the upper end of the tail, which implies the model may predict future return level with slightly higher error. The probability density plot shows that there is a minor difference in the theoretical and observed distributions.

4.6.2 Fitting GEVD Model on the IMFs

IMFs 5, 6 and residual are less orthogonal to each other and there is low correlation with the original time series as shown in Table 4.3. Therefore, according to Huang et al. (1998) and Ayenu-Prah & Attoh-Okine (2010), these can be added together to form one IMF. A new IMF (${c^*}_6$) is formed where

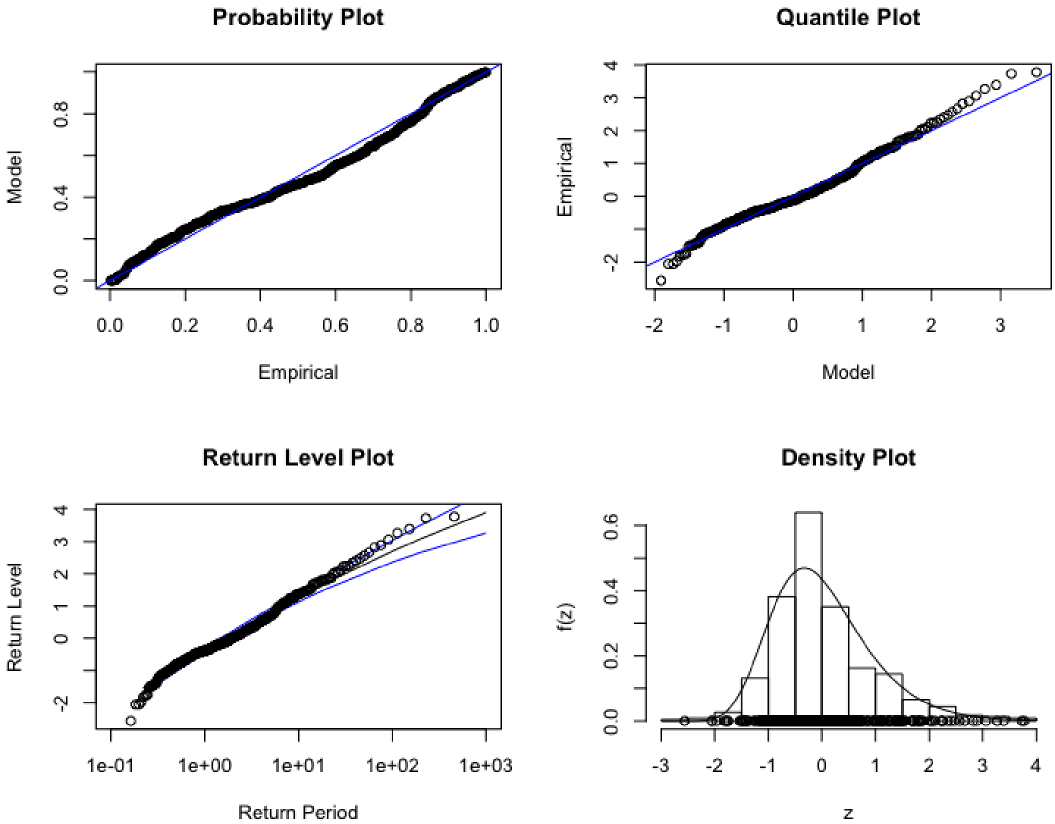


Figure 4.12: Diagnostic plots illustrating the fit of the mean monthly rainfall data for CT Metro area to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (bottom right panel)

Table 4.7: Parameter estimates, negative log-likelihood, AIC and BIC for the models

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	**NLL**	**AIC**	**BIC**
x	-0.387475	0.785845	-0.070859	587.6751	1181.35	1193.718
c_1	-0.28679	0.67913	-0.243483	477.3972	960.7943	973.1618
c_2	-0.16381	0.43964	-0.242588	276.4121	558.8243	571.1918
c_3	-0.090221	0.245886	-0.259559	9.7763	25.5526	37.9201
c_4	-0.047518	0.158145	-0.320800	-202.406	-398.812	-386.445
c_5	-0.085818	0.181877	-0.223071	-114.4572	-222.9144	-210.5469
c_6^*	0.008230	0.179920	-0.495781	-187.4498	-368.8995	-356.532

x=rainfall time series and c_i =i^{th} IMF.

${c^*}_6 = c_6 + c_7 +$ Res. The block maxima approach was used to fit GEVD on rainfall's IMFs using maximum likelihood estimator.

The diagnostic plots for c_1 to c_4 (Fig. 4.13 to 4.16) are presented in this chapter while the rest (c_5 and c_6^*) are presented in Appendices. The diagnosis of the model, using the QQ-plot and PP-plot, shows that there is an improvement in the fitting of the model. There is lesser divergence from the reference line for most of the IMFs. This is shown in Fig. 4.13 to 4.16 with (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (bottom right panel). There was an improvement in the models than in the original rainfall data, as shown by AIC and BIC for the decomposed data, and presented in Table 4.7. The AIC and BIC for the original rainfall time series is 1181.35 and 1193.718 respectively. IMF1 (c_1) has AIC = 960.794 and BIC = 973.1618 which is less compared to the rainfall values. According to the Table 4.7 the best GEVD model fit is for c_3 which has NLL = 9.77631, AIC = 25.55262 and BIC = 37.9201.

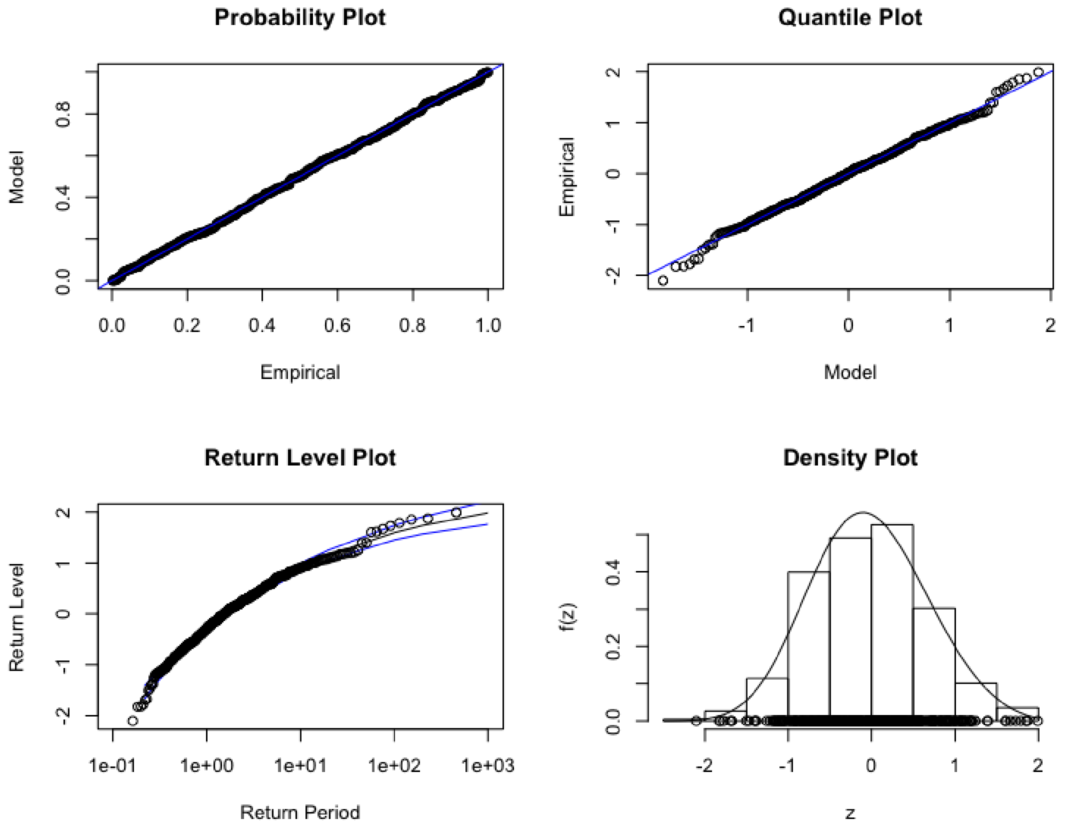


Figure 4.13: Diagnostic plots illustrating the fit of the rainfall IMF1 data to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

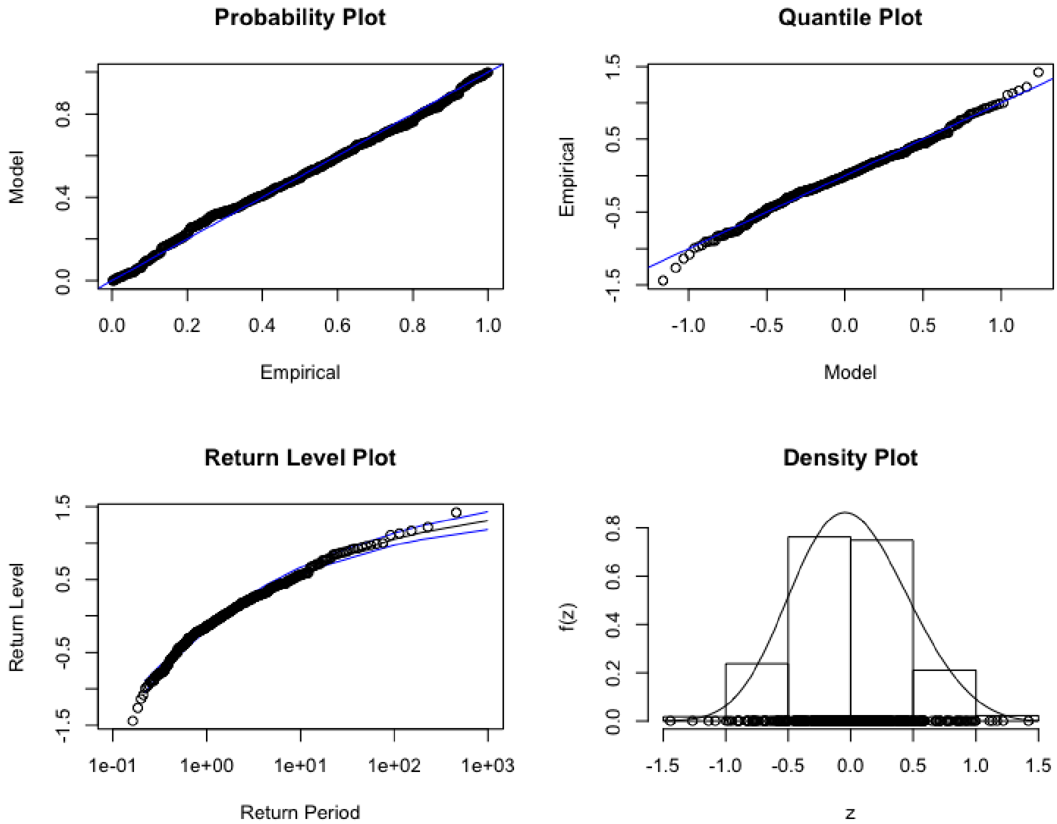


Figure 4.14: Diagnostic plots illustrating the fit of the rainfall IMF2 data to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

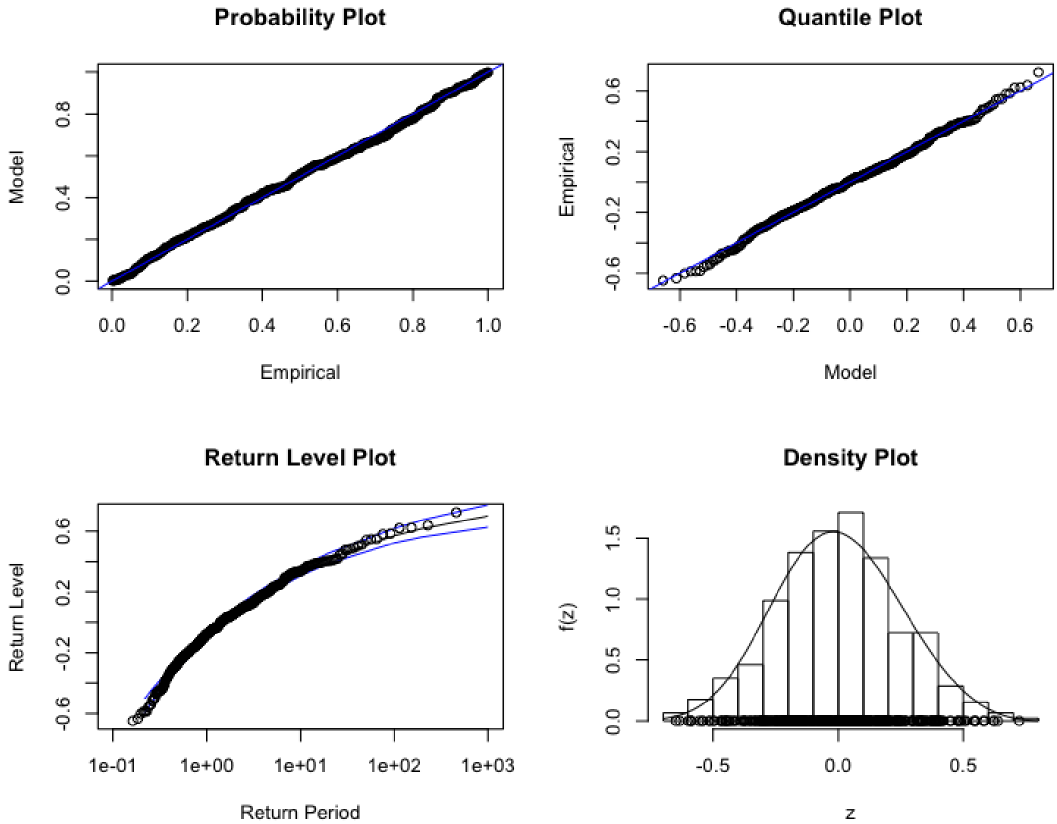


Figure 4.15: Diagnostic plots illustrating the fit of the rainfall IMF3 data to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

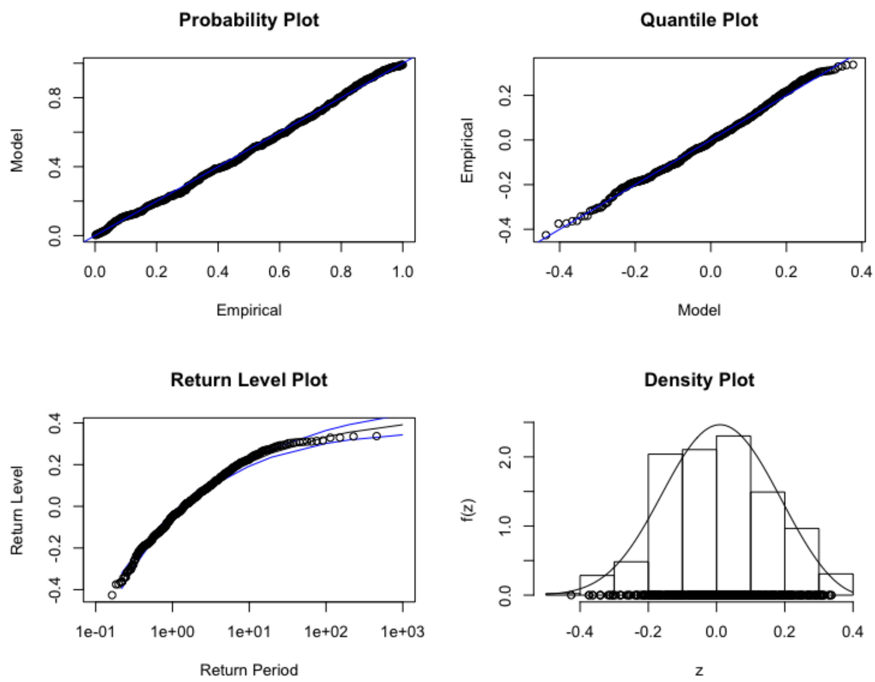


Figure 4.16: Diagnostic plots illustrating the fit of the rainfall IMF4 data to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (bottom right panel)

4.6.3 Goodness-of-fit Test for the Models

An assessment of goodness-of-fit (GoF) test of the models was carried out and the results are presented in Table 4.8. The Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests were used to test for the GoF of the fitted distributions; and hence, make a comparison with the original rainfall time series. From the results presented in the Table (4.8) the decomposed data have better fit compared to the original data except for c_2. Therefore, based on these results, the decomposed rainfall models provides better fit. This approach of using K-S and A-D test to test for GoF for the models is similar to what was done by Sukla et al. (2014) and Maposa (2016).

Table 4.8: Goodness-of-fit Test for the models

	K-S Test	**A-D Test**
x	0.06257	0.05630
c_1	0.99630	0.75506
c_2	0.04890	0.04005
c_3	0.68314	0.06817
c_4	0.743889	0.07948
c_5	0.326961	0.04106
c_6^*	0.07737	0.00561

x=rainfall time series and c_i =i^{th} IMF.

4.6.4 Return Level Estimates

The expected return levels/periods are estimated using the GEVD model for rainfall data and the IMFs. In Table 4.9 the return and corresponding probable high quantiles of monthly rainfall in millimetres (mm), at selected intervals of T = 2, 5, 10, 20 and 100 years are presented. It can be observed that the return level from the IMFs model are higher compared to the original rainfall time series. The average rainfall for the selected region for the winter rainfall period is about 90mm (du Plessis, 2017). Using the parameter estimates from the IMFs model, the return level of 164.7mm is associated with mean return period of 20 years, 155.1mm is associated with return level of 10 years and 181.3 return level of 100 years. This was in agreement to the findings by du Plessis (2017), who found the return period of 20 years for wetter months. The highest monthly rainfall recorded in June 1994 of 168.6mm has a return level of at least 20 years. Given the diagnostics and GoF test done, the IMFs GEVD model should provide a better estimate of the return level.

The estimated extreme monthly rainfall is useful for planning and general preparedness for floods and droughts. The knowledge of the extreme rainfall pattern may assist farmers, government and other stakeholders, as an early warning signal of floods and droughts in the coming seasons, in order to prepare for the type of crops to be planted, water harvesting and disaster management.

Table 4.9: Return level estimate from GEVD model

	$T = 2$	$T = 5$	$T = 10$	$T = 20$	$T = 100$
x	82.4987	98.3686	109.6122	120.2821	143.0778
$\sum_{i=1}^{6} c_i$	124.8351	143.6812	155.1385	164.6966	181.3186

x = rainfall time series and $\sum_{i=1}^{6} c_i$ summation of all IMFs Model

4.7 Summary

In this chapter, spatial variability analysis of the study area in Western Cape was carried out. The analyses of spatial variability, monthly and seasonal variability for the period 1980 to 2018 showed that there was a monotonic decrease in rainfall in most of the stations, except for stations located further north in the province. There was also a variability of up to 127% in the annual and seasonal rainfall in some of the stations. These rainfall trend analyses may help to predict future rainfall patterns for the purpose of planning and better water resources management. In order to investigate the source of the variability at different periods, decomposition methods EMD and EEMD were employed. However, before the application, a method to determine the optimal amplitude of noise to be added was presented. The rainfall data was decomposed into a number of IMFs, which represents the rainfall pattern at different periods. The physical meaning of the IMFs was shown by synchronisation with ENSO and QBO, which are climate oscillations known to influence the rainfall pattern in South Africa.

The IMFs, decomposed using EEMD, were modelled using GEVD. Model diagnosis and goodness-of-fit shows that the IMFs GEVD models have a better model than the original time series. The return periods for the model was presented. The results showed that 181.3mm is associated with a return level of 100 years. The return level of 20 years, which was 164.7mm is in agreement with previous studies that used different methods. Therefore, it can be concluded that the hybrid method of EEMD and GEVD were more robust compared to the primary method of GEVD only. This can be adopted to improve predictability of extreme weather conditions.

In the next chapter, using EEMD, synchronisation and regression, the climate variability impact on agriculture is investigated. This is done using case studies of grapevines in Western Cape and sugarcane in KwaZulu-Natal.

CHAPTER 5

IMPACT OF CLIMATE VARIABILITY ON AGRICULTURE

5.1 Introduction

The previous chapter 4, EMD and EEMD were applied to the study of rainfall and temperature trends or variability, using a selected area in Western Cape. The optimal amplitude of noise to be added in the EEMD was therefore proposed. This was adopted for the rest of the study and it is used in this chapter. The variability of rainfall and temperature was shown at multi-scale level. The climate variability that was discussed in the previous chapter will be investigated to know its impact on agriculture. In order to investigate the climate variability on agriculture, normalised difference vegetative index (NDVI) was used, which is a graphical indicator that can be used to monitor crop progression and expected yield. NDVI can be used as an indicator of the environmental and atmospheric stress on the crops or plants.

The NDVI is decomposed into IMFs, and synchronisation is used to assess the impact of rainfall and temperature variability on it. A multiple linear regression is used to investigate the influence of precipitation, temperature, vapour pressure, cloud cover and soil moisture (independent variables) on NDVI (dependent variable). The variables: precipitation, temperature, vapour pressure, cloud cover and soil moisture, were found to have considerable influence on plant growth (Zhao & Li, 2015; Miguez et al., 2018).

A case study of vineyards located in Western Cape is used to investigate the climate variability on agriculture. Global warming is known to affect the grapevines at different stages of growth. Changes in average temperatures have a direct impact on the the early development stages (phenology) and ripening process (Hayman et al., 2009). A combination of a temperature increase and reduction of soil water, or moisture, may precipitate the ripening process (Hayman et al., 2009; Webb et al., 2012). A different crop (sugarcane) and different area (KwaZulu-Natal) were used to find out if temperature and rainfall variability had similar influence on NDVI. Sugarcane was chosen since it has almost the same life-span as grapes and the area located in KwaZulu-Natal receives austral summer rainfall. The study area and data used is explained in section 5.2, multi-scale variability results and discussion are done in section 5.3 while the discussion of the environmental factors impact on NDVI is detailed in section 5.4. Most of the results presented in this chapter have been published in the scientific journals listed on research output on page xx and published copies of the articles are attached in Appendix C.

5.2 Study Areas and Data

Farming in South Africa is done according to the climate of the area. However, as discussed earlier, climate variability is felt in most parts of South Africa. Temperature increases over the decades is being experienced in both winter and summer rainfall regions (discussed on page 4 of this thesis). Rainfall has high

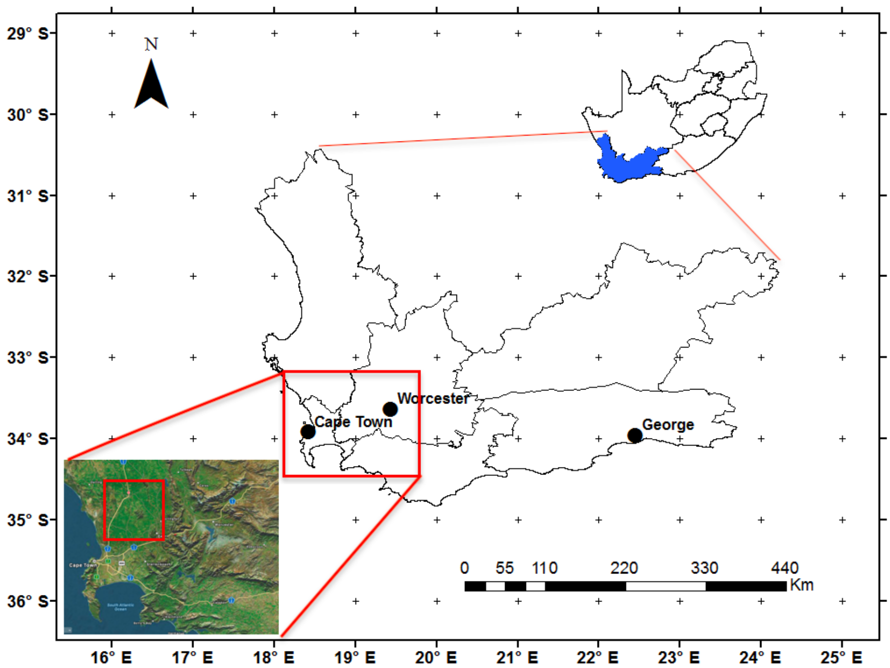


Figure 5.1: The location of the grapevines farms and study area in Western Cape province.
Inserted satellite image source: Google Maps

spatial variability with some regions becoming drier while others are becoming wetter. Climate variability is known to have an impact on agricultural production; therefore, in this study, two separate regions were selected to find out if climate variability has a similar impact. The two regions which we considered are Western Cape and KwaZulu-Natal in South Africa.

The first region is located in Western Cape, between -33ºS and -34ºS and 18.5ºE and 19.5ºE. This area has a number of vineyards (Fig. 5.1). The green vegetation selected in the satellite image mainly consists of the vineyards. The first grapevine in South Africa was planted in 1659 by the Dutch, and today, the industry has grown to more than 10 000 farmers (Van Schalkwyk, 2018). Western Cape has Mediterranean climate, which is suitable for grape farming. Grape production contributes about half of the Gross Domestic Product (GDP) for the province and provides about 8% of employment (Araujo et al., 2014).

Fig. 5.2 shows the second location, which is sugarcane production region

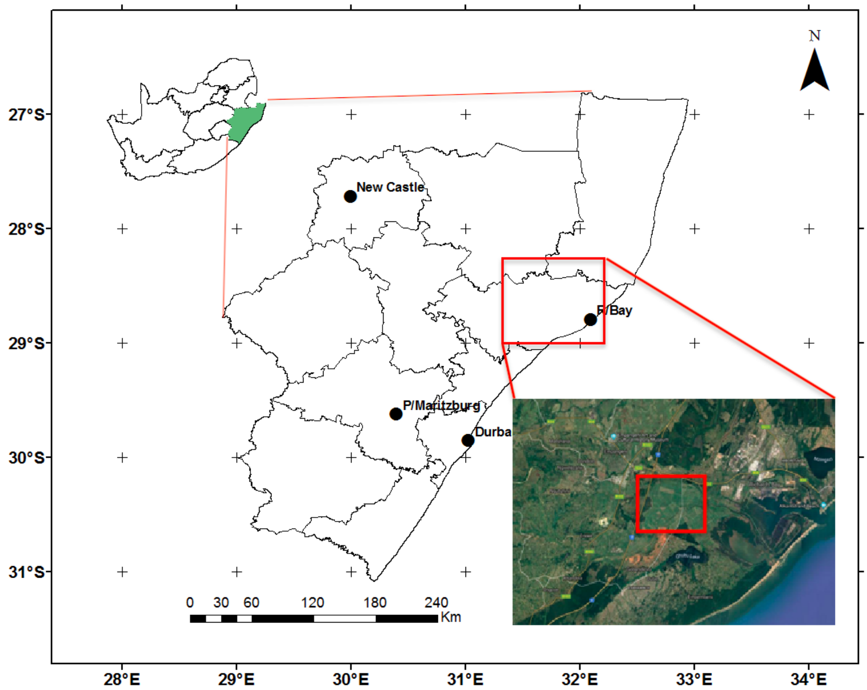
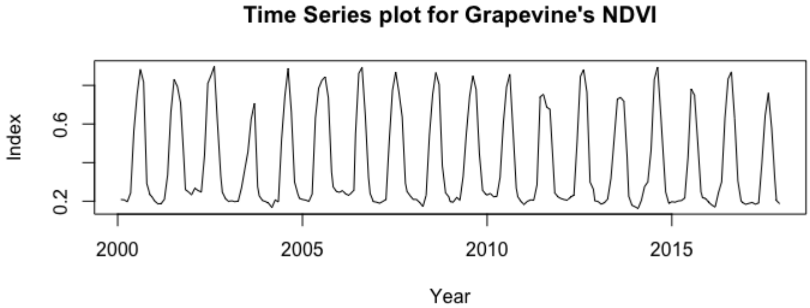


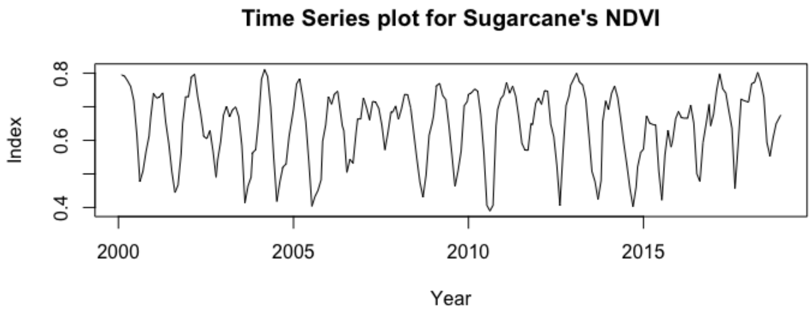
Figure 5.2: The location of sugarcane farms and study area in Felixton, KwaZulu-Natal Province.
Inserted satellite image source: Google Maps

located between -29ºS to -28.2ºS and 31.2ºE and 32.2ºE in Felixton, KwaZulu-Natal. The green vegetation selected in the satellite image mainly consist of the sugarcane plantations. The Felixton area contributes about 10% of sugarcane yield in South Africa and has one of the largest mills in the country. In 2017/18 season, about 1,7 million tons were crushed at the mill (SASA, 2019). Time series plots of the NDVI for both grapevines and sugarcane are presented in Fig. 5.3.

The NDVI is a simple graphical indicator that can be used to analyse remote sensing measurements. It measures the greenness of a plant and takes values between -1 and 1. In this study, the NDVI data is sourced from the NASA's Terra satellite using the instrument MOD13A3 vegetation index extracted from Appears Application (Didan, 2015). MOD13A3 version 6 provides Vegetation Index (VI) values at a per pixel basis of 500 meter spatial resolution and is taken every 16 days. The data set was averaged to find the monthly data. The other data variables used are: monthly observations of



(a) Grapevine's NDVI



(b) Sugarcane' NDVI

Figure 5.3: Time series plot for grapevine and sugarcane's NDVI for the from 2000 to 2018.

rainfall, temperature, NDVI, vapour pressure, cloud cover and soil moisture data, for the period 2000 to 2018 were used. The monthly average rainfall and temperature data was collected from South African Weather Service. The vapour pressure, cloud cover and soil moisture data are all freely available online on the Royal Netherlands Meteorological Institute (KNMI) website (https://climexp.knmi.nl) (Huang et al., 2015).

5.3 Climate Variability Impact on the Crop's NDVI

An investigation of the climate variability impact on NDVI is presented at different time scales using EEMD. Four IMFs were derived and these were synchronised to temperature and precipitation, and to identify the relationship between the time series. The plot of the IMFs from the decomposed grapevines' NDVI are shown in Fig. B5 in the appendix.

Time series that are phase synchronised or locked exhibit a modal distribution with a prominent peak at a given phase difference, whereas unrelated time series are characterised by a uniform or diffuse distribution. The synchronisation of the grapevines' NDVI and rainfall shows that there is phase locking for IMF2 and IMF3, since there is a prominent peak on one of the phase difference as shown in Fig. 5.4. IMF2 has a period of about 12 months, which represents the annual period, and IMF3 has a period of about 26 months (approximately 2 years). However, there is weak coupling for all other IMFs. This suggests that seasonal rainfall variability may have a direct impact on the grapevines' NDVI. The synchronisation of the grapevine's NDVI and temperature shows that there is weak coupling for all the IMFs since there is uniform distribution for all the histograms in Fig. 5.5. These results show that the increasing temperature that was observed for the period may not have an impact on the grapevine's NDVI. Previous studies revealed that increasing temperature is causing the quickening of the ripening stage and rainfall variability is having an impact on the phenology stage (Hayman et al., 2009; Kumar & Sharma, 2014; Araujo et al., 2014)

Given the results from the grapevines, a comparison study was carried out using sugarcane. An investigation was carried out to find out if this was the case with another plant in a slightly different climate. Sugarcane has almost similar lifespan as grapes. It is mostly grown in summer rainfall regions in KwaZulu-Natal. The sugarcane's NDVI was decomposed into 4 IMFs and the plots are shown in Fig. B6 in the appendix. The IMFs were synchronised to rainfall IMFs. The peaks for sugarcane are clearer than for grapevines. A histogram plot of synchronisation for sugarcane's NDVI shows a clear peak for IMF2 and IMF3, which provides evidence of phase locking of the time series (Fig. 5.6). However, there is weak coupling for all other IMFs. This shows that seasonal precipitation variability may have a direct impact on the NDVI. This agrees with studies by Kumar & Sharma (2014), who showed the direct impact of seasonal rainfall variability on sugarcane's yield in India. Fig. 5.7 reveals that temperature has an impact on the seasonal variability of NDVI, since a peak is also identified on the IMF2.

Increase in temperature has a positive effect on sugarcane, unlike on the grapes, where it quickens the ripening stage; thereby reducing the quality of the yield. The lesser impact of rainfall variability on grapevines may be associated with the farmer's intervention of using irrigation. Irrigation limits the impact of rainfall variability (Van Schalkwyk, 2018) .

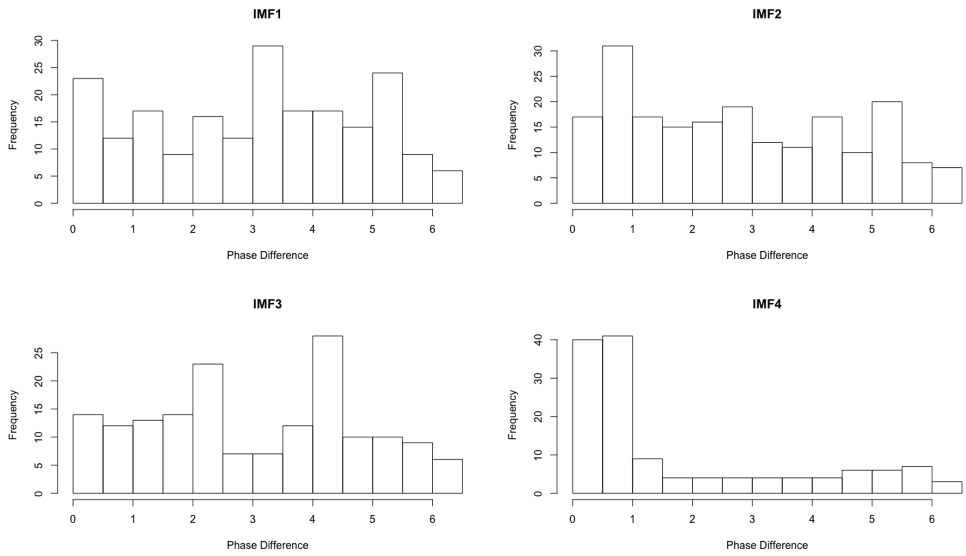


Figure 5.4: Histogram plot of synchronisation of grape NDVI and rainfall, with phase locking observed in IMF2 and IMF3 .

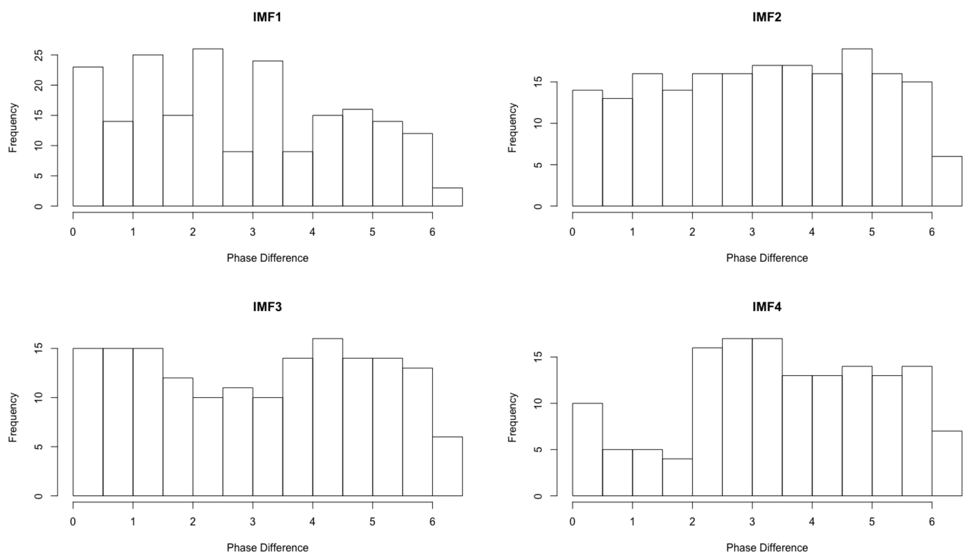


Figure 5.5: Histogram plot of synchronisation of grapevines's NDVI and temperature, showing no phase locking in any of the IMFs.

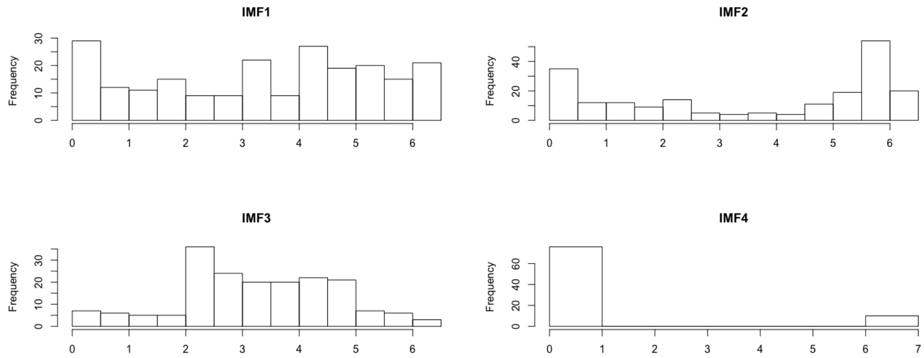


Figure 5.6: Histogram plot of synchronisation of sugarcane's NDVI and precipitation, with phase locking observed in IMF2 and IMF3.

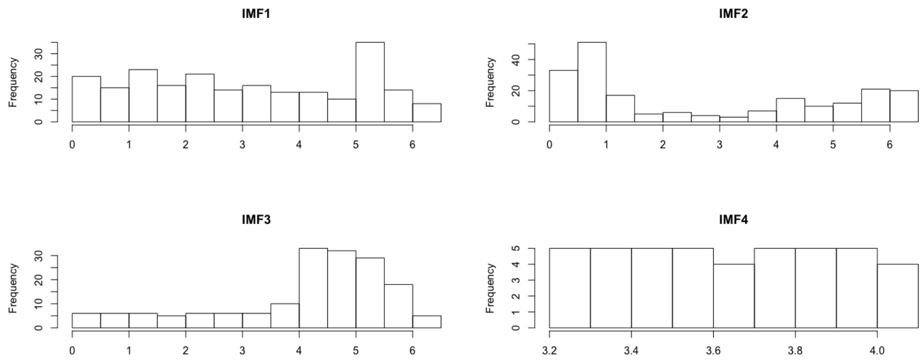


Figure 5.7: Histogram plot of synchronisation of sugarcane's NDVI and temperature, with phase locking observed in IMF2 and IMF3 ..

5.4 Impact of Environmental Factors

Multi-linear regression was used to investigate the environmental impact on plant's growth. The dependent variable is NDVI and the independent variables are rainfall, temperature, vapour pressure, percentage cloud cover and soil moisture. The results from grapevine's regression model revealed that there is no linear relationship between the grapevine's NDVI and the independent variables, with $r^2 = -0.007063$, F-statistic $= 0.6998$ and $p = 0.6242$. All the independent variables were found not to be significant except the constant. These results confirm earlier results found in section 5.3, which showed that rainfall and temperature variability have little or no influence over the grapevine's NDVI.

The results for sugarcane's regression modelling showed contrasting results with the grapevines. The model was significant, with F-statistic $= 60.74$ and $p < 0.001$, which implies that at least one of the predictor variables has an influence on the NDVI. The coefficient of determination for the model is $r^2 = 0.5788$ and adjusted $r^2 = 0.5693$. The regression model found is expressed as

$$\begin{aligned} y = 0.1602753 + 0.0084779x_1 + 0.0105461x_2 + 0.0228464x_3 \\ -0.0061422x_4 + 1.4131146x_5 \end{aligned} \tag{5.1}$$

where x_1 is the rainfall, x_2 is temperature, x_3 is vapour pressure, x_4 is percentage cloud cover and x_5 is soil moisture. The coefficients for vapour pressure, cloud cover and soil moisture were all found to be significant. Table 5.1 shows the diagnosis of the model's coefficients. The t-statistic for the coefficients for vapour pressure ($p < 0.001$), cloud cover ($p < 0.001$) and soil moisture (p $<$ 0.001) were significant. In order to evaluate the model, a series of diagnostic graphs are

Table 5.1: Coefficients for the sugarcane NDVI model

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**
Constant	0.1602753	0.0718909	2.229	0.0268
x_1	0.0084779	0.0085034	1.997	0.0319
x_2	0.0105461	0.0120964	1.872	0.0384
x_3	0.0228464	0.0037479	6.096	4.80e-09
x_4	-0.0061422	0.0006764	-9.081	$<$ 2e-16
x_5	1.4131146	0.2233190	6.328	1.37e-09

plotted and shown in Fig 5.8.

Residuals are values found from the difference between observed and expected values from the model. The points on the residuals vs fitted plot are spread out and this is an indication of a good model. The Normal Q-Q plot shows that the residuals are normally distributed, which is one of the assumptions that has to be met for a good model. It appears that the residuals do not violate the assumption of normality. The scale-location plot is showing that the residuals are more spread out, which is a requirement for a good model.

The residuals vs leverage plot helps to identify influential data points in the model. Points that are outside the dashed red Cook's distance line are influential on the model. The plot does not show many influential cases. The model provides further evidence of the impact of climate variability as suggested by Zhao & Li (2015) which highlighted the direct influence of the environmental factors modelled here.

It is important to check the in-sample forecasting power of the model. Fig. 5.9 shows the plot of the observed NDVI and predicted NDVI from the selected model. The model seems to show little variability between predicted and actual NDVI. The MAPE value of 0.09391 and RMSE of 0.067998 were found for this model, and which are reasonable.

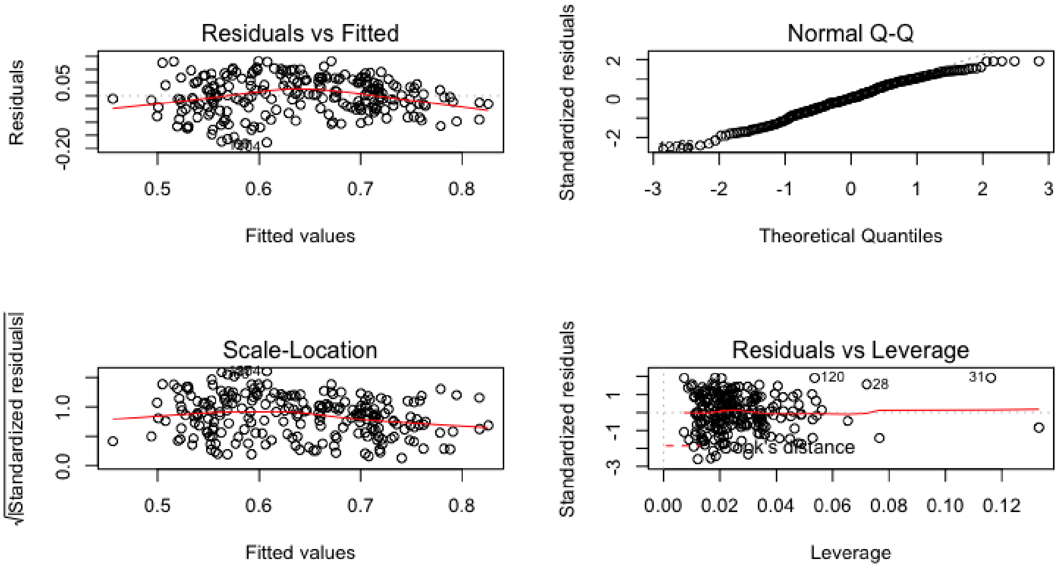


Figure 5.8: Residuals plots to evaluate the model. The plots are Residual vs Fitted (top left), QQ plot (top right), Scale-Location (bottom right) and Residuals vs Leverage (bottom right).

Table 5.2: Observed NDVI and Predicted for out-of-sample

Date	Observed NDVI	Predicted NDVI
Jan 2019	0.76605	0.61956
Feb 2019	0.80265	0.74567
Mar 2019	0.8227	0.71530
Apr 2019	0.80615	0.67706
May 2019	0.80925	0.60008
June 2019	0.74315	0.57279

The out-of-sample forecast for 6 months, up to June, 2019 are shown in Table 5.2. The MAPE value of 0.17343 and RMSE of 0.14473 were found for this model and are reasonable. However, the model seems to under-forecast the NDVI. Therefore, this model maybe be used with combination with other methods to forecast the expected sugarcane's NDVI for Felixton area, and maybe be used as an early warning tool to inform the sugarcane farmers, Department of Agriculture and other stake-holders about the expected yield for the season.

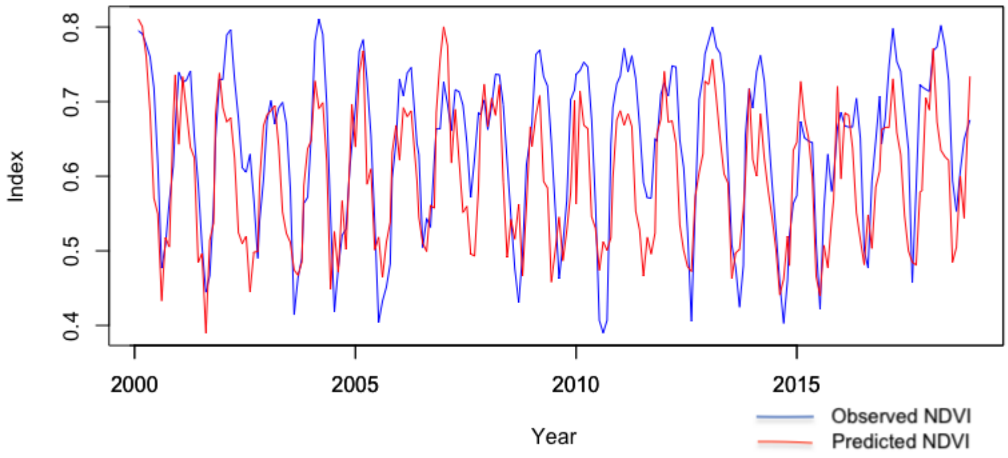


Figure 5.9: Observed NDVI versus predicted NDVI for sugarcane

5.5 Summary

The NDVI was used to show the impact of rainfall and temperature variability at different time scales on agriculture, using grapevines and sugarcane as case studies. This was obtained by decomposing the NDVI into different time scales and the resulting decomposed time series synchronised with precipitation and temperature. The variability of rainfall has more impact on the NDVI than temperature. Rainfall variability has a greater impact on the phenology stage of the grapevine. For the period under study, although there has been increase in temperature, it had no noticeable impact on the grapevines' NDVI. This was not the case with sugarcane, where it was shown that the variability of temperature has an impact on the NDVI at seasonal and 26 months (approximately 2 years) period. The results were also confirmed using regression, and it was shown that there was no linear relationship between grapevines' NDVI and rainfall, temperature and other factors. This was contrary to sugarcane where a model which encompasses precipitation, temperature, vapour pressure, cloud cover and soil moisture was proposed. The model had a small error for both in-sample and out-of-sample.

It can therefore, be concluded that the use of decomposition assisted in identifying intrinsic oscillations and their impact on the plants' NDVI. The regression model may be used for short and medium term planning, to predict the future NDVI and expected yields.

CHAPTER 6

SUMMARY AND RECOMMENDATIONS

Historical data was used to analyse rainfall and temperature characteristics and the impact of climate variables in selected regions in South Africa that receives winter and summer rainfall. The spatial variability analysis of the rainfall variability for selected regions was shown. This can be used for planning purposes as there are many un-sampled areas without weather stations in South Africa. The climate data was decomposed into different time scales and the analysis done. This was the first study that used decomposed data to model extreme weather conditions. Furthermore, the impact of climate variability on agriculture was shown. The NDVI was used to show the impact of variability during the growth phase of crops. A model that considered environmental factors, such as sunlight, water vapour and soil moisture, for sugarcane in Felixton, Kwazulu-Natal was proposed for the first time (section 5.4). In this chapter, the summary and recommendations made from this study are provided.

Variability and trend analysis of winter rainfall for selected region in Western Cape was carried out. MK Test revealed that there were few stations with significant change in the seasonal rainfall trend for the period 1980 to 2018 (section 4.2.1). Significant monotonic increasing rainfall trend was identified for the month of June while May had monotonic decreasing rainfall trend. The month of June was found to be less variable in the winter period, compared to other months. The spatial analysis shows that there is no uniform trend for both summer and winter rainfall.

EMD has had a number of applications in climate modelling, and other areas with oscillatory data. However, there are issues with mode mixing. Therefore, a data assisted EEMD was employed in this study. It was also observed that EEMD has issues on the amplitude of the noise to be added. In this study, a method to determine an optimal noise that can be added to rainfall data was then proposed. Using EEMD, with optimal noise, the effectiveness of EEMD to decompose data into multiple time series that are stationary was demonstrated. The rainfall and temperature data were decomposed into IMFs and residual data, which summed up to the original data. EEMD was effective in isolating the data into different time scales. Therefore, the variability of the rainfall pattern was identified, in the end, evidence of the effect of ENSO and QBO was provided (section 4.5.1). Phase synchronisation was used to find the relationship of the IMFs for the different time series under study.

The decomposed IMFs were modelled using generalised extreme value distribution, and the results showed that the model was better than the original time series model. It can be fairly concluded that models coupled with pre-processing or decomposing methods provides better fit.

The NDVI is a good measure of progressive plant growth. Using the sugarcane plant, the climate variability impact on its growth was shown. This was done by decomposing the NDVI into different time scales and the resulting decomposed time series synchronised with precipitation and temperature (section 5.3). The variability is easily identified in the seasonality of sugarcane. A model which encompasses precipitation, temperature, vapour pressure, cloud cover and soil moisture was proposed.

6.1 Recommendations

Based on the quantitative results found, the following recommendations are made towards the impact of climate variability:

- South Africa has large spatial differences in the rainfall patterns and variability; and therefore, any study that is to be carried out to find the climate variability impact must consider these differences.

- EEMD method is best suitable for identifying underlying signals in oscillating time series at different time scales. This method can be employed in nonlinear and non-stationary time series to identify hidden intrinsic signals in a time series.

- Climate variability has a direct impact on South Africa weather, as observed in the rainfall and temperature patterns. Therefore, more effective and relevant strategies must be in place to mitigate its impact.

- Hybrid methods, such as EEMD and GEVD, perform better than original standard methods. A combination of data pre-processing methods and other modelling methods improves the performance of the models. It is, therefore, recommended that hybrid methods should be considered, when creating models.

- Remote sensing data, such as normalised difference vegetative index (NDVI) can be used to measure the climate variability impact on crops. The NDVI data is freely available for any area in the world; and therefore, this may be used to carry out statistical analysis that will help farmers to adapt to climate change.

- Climate variability has an impact on both irrigated and non-irrigated crops, but it has far much more adverse effect on the non-irrigated crops. South Africa is classified as semi-arid; therefore, total reliance on irrigated agriculture may not be feasible. Therefore, there is need to invest more money on agricultural research by Department of Agriculture, agricultural research institutes and other stakeholders, so that adaptive crops may be created. It is also recommended that there should be focus on crops that require less water and can tolerate heat

6.2 Limitations of the Thesis

This thesis adds value to climate modelling using historical data. A concise, but detailed overview of EMD, EEMD and GEVD was provided as a quick and handy overview of the main theories and models for academics and researchers. One of the theoretical limitations of the study is that some topics and details, which some other researchers may deem important might have been left out. However, the results that were found to be relevant and most important to the study of South African climate variability were included.

The limitations of the empirical results is that, due to time and cost constraints, the study mainly focused on two chosen regions in South Africa. It would be of interest to do a study for the complete country and other regions that have slightly different weather. The advantage of this study is that it can be replicated for different regions in South Africa.

6.3 Future Perspectives

In view of the limitation mentioned above, it is suggested that since the statistical analysis was based on a small area of South Africa, further studies need to be carried out, using the same model in other areas of South Africa or other countries in Africa. Additionally, it will be of interest for future studies to carry out a related study, but using a longer period, to find the pattern of the rainfall over decades. The study focused on two agricultural products (sugarcane and grapevines), it will be of interest to extend the study to other agricultural crops.

REFERENCES

Aggarwal, P. K., Kalra, N., Chander, S., Bhatia, A., Bala, A., Rani, S., Chander, S., Pathak, H., & Kalra, N. (2006). Infocrop: A dynamic simulation model for the assessment of crop yields, losses due to pests, and environmental impact of agro-ecosystems in tropical environments. *Agricultural Systems*, 89, 1–25.

Ahmad, I., Tang, D., Wang, T., Wang, M., & Wagan, B. (2015). Precipitation trends over time using mann-kendall and spearman's rho tests in swat river basin, pakistan. *Advances in Meteorology*, 2015, 1–15.

Alexandrov, T. (2008). A method of trend extraction using singular spectrum analysis. *arXiv preprint arXiv:0804.3367*.

Alizadeh, F., Roushangar, K., & Adamowski, J. (2019). Investigating monthly precipitation variability using a multiscale approach based on ensemble empirical mode decomposition. *Paddy and Water Environment*.

Araujo, J. A., Abiodun, B. J., & Crespo, O. (2014). Impacts of drought on grape yields in western cape, south africa. *Theoretical and Applied Climatology*, 123(1-2), 117–130.

Ayenu-Prah, A. & Attoh-Okine, N. (2010). A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition. *Advances in Adaptive Data Analysis*, 2(01), 1–24.

Balocchi, R., Menicucci, D., Santarcangelo, E., Sebastiani, L., Gemignani, A., Gellarducci, B., & Varanini, M. (2004). Deriving the respiratory sinus arrhythmia from the heartbeat time series using empirical mode decomposition. *Chaos, Solitons and Fractals*, 20(1), 171–177.

Begue, N., Bencherif, H., Sivakumar, V., Kirgis, G., Leclair de Bellevue, J., et al. (2010). Temperature variability and trend estimates at tropopause and UT-LS over a subtropical site: Reunion (20.8 _ S, 55.5 _ E).

Berning, T. L. (2010). *Improved estimation procedures for a positive extreme value index*. PhD thesis, Stellenbosch: University of Stellenbosch.

Bhagwandin, L. (2017). *Multivariate Extreme Value Theory with an application to climate data in the Western Cape Province*. PhD thesis, University of Cape Town.

Blignaut, J., Ueckermann, L., & Aronson, J. (2009). Agriculture production's sensitivity to changes in climate in South Africa. *South African Journal of Science*, 105, 61 – 68.

Bonaventura, L. & Castruccio, S. (2005). Random notes on kriging: an introduction to geostatistical interpolation for environmental applications. *Milan: Politecnico de Milano*.

Bosman, P. A. & Thierens, D. (2000). Negative log-likelihood and statistical hypothesis testing as the basis of model selection in ideas.

Boudrissa, N., Cheraitia, H., & Halimi, L. (2017). Modelling maximum daily yearly rainfall in northern Algeria using generalized extreme value distributions from 1936 to 2009. *Meteorological Applications*, 24(1), 114–119.

Brönnimann, S., Annis, J. L., Vogler, C., & Jones, P. D. (2007). Reconstructing the quasi-biennial oscillation back to the early 1900s. *Geophysical Research Letters*, 34(22).

Cazelles, B. & Stone, L. (2003). Detection of imperfect population synchrony in an uncertain world. *Journal of Animal Ecology*, 72(6), 953–968.

Chang, C. Y., Huang, N. E., & Shen, Z. (1997). A Statistically Significant Periodicity in the Homestake Solar Neutrino Data. *Chinese Journal of Physics*, 35(6S), 818–831.

Chen, X., Tsao, M., & Zhou, J. (2010). Robust second-order least-squares estimator for regression models. *Statistical Papers*, 53(2), 371–386.

Chiew, F. H. S., Peel, M. C., Amirthanathan, G. E., & Pegram, G. G. S. (2005). Identification of oscillations in historical global streamflow data using empirical mode decomposition. *IAHS PUBLICATION*, (296), 53–64.

Chikobvu, D. & Chifurira, R. (2015). Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, 111(9-10), 01–08.

Claeskens, G. & Hjort, N. L. (2008). *Model selection and model averaging*. Technical report, Cambridge University Press.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.

Coughlin, K. & Tung, K. K. (2005). *Empirical mode decomposition and climate variability*. World Scientific Publishing.

de Carvalho, J. R. P., Monteiro, J. E. B. A., Nakai, A. M., & Assad, E. D. (2017). Model for multiple imputation to estimate daily rainfall data and filling of faults. *Revista Brasileira de Meteorologia*, 32(4), 575–583.

Deressa, T., Hassan, R., & Poonyth, D. (2005). Measuring the impact of climate change on South African agriculture: the case of sugarcane growing regions. *Agrekon*, 44(4), 524–542.

Didan, K. (2015). Mod13a1 modis/terra vegetation indices 16-day l3 global 500m sin grid v006.

Dieppois, B., Pohl, B., Rouault, M., New, M., Lawler, D., & Keenlyside, N. (2016). Interannual to interdecadal variability of winter and summer southern african rainfall, and their teleconnections. *Journal of Geophysical Research: Atmospheres*, 121(11), 6215–6239. 2015JD024576.

Dombry, C. (2015). Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli*, 21(1), 420–436.

du Plessis, J. A. (2017). An investigation into the evidence of seasonal rainfall pattern shifts in the western cape, south africa. *Journal of the South African Institution of Civil Engineering*, 59(4), 47–55.

Farrell, P. J. & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: extending the spiegelhalter test. *Journal of Statistical Computation and Simulation*, 76(9), 803–816.

Fauchereau, N., Pohl, B., Reason, C., Rouault, M., & Richard, Y. (2009). Recurrent daily olr patterns in the southern africa/southwest indian ocean region, implications for south african rainfall and teleconnections. *Climate Dynamics*, 32(4), 575–591.

Gbetibouo, G. & Hassan, R. (2005). Measuring the economic impact of climate change on major South African field crops: a Ricardian approach. *Global and Planetary Change*, 47(International Young Scientists' Global Change

Conference 2003), 143 – 152.

Gil-Alana, L. A., Yaya, O. S., & Fagbamigbe, A. F. (2018). Time series analysis of quarterly rainfall and temperature (1900–2012) in sub-saharan african countries. *Theoretical and Applied Climatology*, 137(1-2), 61–76.

Groisman, P. Y., Knight, R. W., Easterling, D. R., Karl, T. R., Hegerl, G. C., & Razuvaev, V. N. (2005). Trends in intense precipitation in the climate record. *Journal of Climate*, 18(9), 1326–1350.

Guo, B., Chen, Z., Guo, J., Liu, F., Chen, C., & Liu, K. (2016). Analysis of the Nonlinear Trends and Non-Stationary Oscillations of Regional Precipitation in Xinjiang, Northwestern China, Using Ensemble Empirical Mode Decomposition. *International Journal Of Environmental Research And Public Health*, 13(3).

Harrison, D. E. & Larkin, N. K. (1997). Darwin sea level pressure, 1876 to 1996: Evidence for climate change? *Geophysical Research Letters*, 24(14), 1779–1782.

Hartmann, D., Tank, A. K., Rusticucci, M., Alexander, L. V., Bronnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., & Zhai, P. M. (2013). Observations: Atmosphere and Surface. In T. [Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. . Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*: Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Hawinkel, P., Swinnen, E., Lhermitte, S., Verbist, B., Van Orshoven, J., & Muys, B. (2015). A time series processing tool to extract climate-driven interannual vegetation dynamics using ensemble empirical mode decomposition (EEMD). *Remote Sensing of Environment*, 169, 375–389.

Hawinkel, P., Thiery, W., Lhermitte, S., Swinnen, E., Verbist, B., Van Orshoven, J., & Muys, B. (2016). Vegetation response to precipitation variability in East Africa controlled by biogeographical factors. *Journal of Geophysical Research: Biogeosciences*, 121(9), 2422–2444.

Hayman, P., Leske, P., & Nidumolu, U. (2009). Climate change and viticulture. informing the decision making at a regional level. version 1.0, october 2009. south australian wine industry association and south australian research and development institute gwrdc project saw 06/01.

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., Smith, T. M., Thorne, P. W., Woodruff, S. D., & Zhang, H.-M. (2015). Extended reconstructed sea surface temperature version 4 (ERSST.v4). nino 3.4. *NOAA National Centers for Environmental Information*.

Huang, N. E., Shen, Z., & Long, S. R. (1999). A new review of non-linear water waves: the Hilbert spectrum. *Annual Review of Fluid Mechanics*, 31, 417–457.

Huang, N. E., Z.Shen, R.Long, S., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, V., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,. *Proceeding of the Royal Society of London, Series A*.

Huang, S., Chang, J., Huang, Q., & Chen, Y. (2014). Monthly streamflow prediction using modified EMD-based support vector machine. *Journal of Hydrology*, 511, 764–775.

Hughes, W. S. & Balling Jr, R. C. (1996). Urban influences on South African temperature trends. *International Journal of Climatology: A Journal of the*

Royal Meteorological Society, 16(8), 935–940.

Hurrell, J. (2015). CLIMATE AND CLIMATE CHANGE | climate variability. In *Encyclopedia of Atmospheric Sciences* (pp. 47–60). Elsevier.

IPCC (2014). IPCC, 2014: Annex II: Glossary. In K. Mach, S. Planton, & C. von Stechow (Eds.), *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 117–130). Geneva, Switzerland: IPCC IPCC.

Jones, M., Singels, A., et al. (2014). A preliminary assessment of mid-century climate change impacts on sugarcane production in south africa. In *Proceedings of the Annual Congress-South African Sugar Technologists' Association*, number 87 (pp. 290–297).: South African Sugar Technologists' Association.

Junsheng, C., Dejie, Y., & Yu, Y. (2006). Research on the intrinsic mode function (IMF) criterion in EMD method. *Mechanical Systems and Signal Processing*, 20(4), 817–824.

Jury, M. R. (2013). Climate trends in southern Africa. *South African Journal of Science*, 109(1-2), 1–11.

Kane, R. (2009). Periodicities, enso effects and trends of some south african rainfall series: an update. *South African Journal of Science*, 105(5-6), 199–207.

Kanga, Y., Khanb, S., & Maa, X. (2009). Climate change impacts on crop yield, crop water productivity and food security A review. *Progress in Natural Science*, 19, 1665–1674.

Kendall, M. (1980). Multivariate Analysis.

Kim, M. S., Kang, D. H., Kim, K. H., Kim, P. S., Jin, K. H., Seo, D. I., & Kwon, B. H. (2011). Rainfall distribution using kriging interpolation in typhoon rusa. In *Proceedings of the 11th WSEAS international conference on Applied computer science* (pp. 115–117).: World Scientific and Engineering Academy and Society (WSEAS).

Kim, T., Shin, J.-Y., Kim, S., & Heo, J.-H. (2018). Identification of relationships between climate indices and long-term precipitation in south korea using ensemble empirical mode decomposition. *Journal of Hydrology*, 557, 726–739.

Kruger, A. C. (2006). Observed trends in daily precipitation indices in South Africa: 1910?2004. *International Journal of Climatology*, 26(15), 2275–2285.

Kruger, A. C. & Nxumalo, M. (2017). Historical rainfall trends in South Africa: 1921–2015. *Water SA*, 43(2), 285–297.

Kruger, A. C. & Shongwe, S. (2004). Temperature trends in south africa: 1960-2003. *International Journal of Climatology*, 24(15), 1929–1945.

Kumar, A. & Sharma, P. (2014). Climate change and sugarcane productivity in india: An econometric analysis. *Journal of Social and Development Sciences*, 5(2), 111.

Le Roux, N. et al. (2009). Seasonal maize yield simulations for south africa using a multi-model ensemble system. Master's thesis, University of Pretoria.

Lebel, T., Bastin, G., Obled, C., & Creutin, J. (1987). On the accuracy of areal rainfall estimation: a case study. *Water Resources Research*, 23(11), 2123–2134.

Lehodey, P., Bertignac, M., Hampton, J., Lewis, A., & Picaut, J. (1997). El niño southern oscillation and tuna in the western pacific. *Nature*, 389(6652), 715.

Leon-Garcia, A. (2017). Probability, statistics, and random processes for electrical engineering.

Lin, L. & Hongbing, J. (2009). Signal feature extraction based on an improved EMD method. *Measurement*, 42(5), 796–803.

Linderhed, A. (2002). 2d empirical mode decompositions in the spirit of image compression. *Proceedings of SPIE*, (1), 1.

Linnenluecke, M. K., Nucifora, N., & Thompson, N. (2018). Implications of climate change for the sugarcane industry. *Wiley Interdisciplinary Reviews: Climate Change*, 9(1), e498.

Lisboa, I. P., Cherubin, M. R., Lima, R. P., Cerri, C. C., Satiro, L. S., Wienhold, B. J., Schmer, M. R., Jin, V. L., & Cerri, C. E. (2018). Sugarcane straw removal effects on plant growth and stalk yield. *Industrial crops and products*, 111, 794–806.

Liu, B., Feng, Y., Yin, Z., & Fan, X. (2019a). Radar signal emitter recognition based on combined ensemble empirical mode decomposition and the generalized s-transform. *Mathematical Problems in Engineering*, 2019, 1–15.

Liu, H., Zhan, Q., Yang, C., & Wang, J. (2019b). The multi-timescale temporal patterns and dynamics of land surface temperature using ensemble empirical mode decomposition. *Science of The Total Environment*, 652, 243–255.

Liu, Y., Wu, J., Liu, Y., Hu, B. X., Hao, Y., Huo, X., Fan, Y., Yeh, T. J., & Wang, Z.-L. (2015). Analyzing effects of climate change on streamflow in a glacier mountain catchment using an ARMA model. *Quaternary International*, 358, 137–145.

Longobardi, A. & Villani, P. (2010). Trend analysis of annual and seasonal rainfall time series in the Mediterranean area. *International journal of Climatology*, 30(10), 1538–1546.

Lu, Y. & Saniie, J. (2016). A comparative study of singular spectrum analysis and empirical mode decomposition for ultrasonic nde. In *2016 IEEE International Ultrasonics Symposium (IUS)* (pp. 1–4).: IEEE.

MacKellar, N., New, M., & Jack, C. (2014). Observed and modelled trends in rainfall and temperature for South Africa: 1960-2010. *South African Journal of Science*, 110(7-8), 1–13.

Mak, M. (1995). Orthogonal wavelet analysis: Interannual variability in the sea surface temperature. *Bulletin of the American Meteorological Society*, 76(11), 2179–2186.

Malherbe, J., Dieppois, B., Maluleke, P., Van Staden, M., & Pillay, D. (2016). South African droughts and decadal variability. *Natural Hazards*, 80(1), 657–681.

Malvić, T. & Balić, D. (2009). Linearity and lagrange linear multiplicator in the equations of ordinary kriging. *Nafta: exploration, production, processing, petrochemistry*, 60(1), 31–43.

Maposa, D. (2016). *Statistics of extremes with applications to extreme flood heights in the Lower Limpopo River Basin of Mozambique*. PhD thesis.

Maposa, D., Cochran, J. J., Lesaoana, M., & Sigauke, C. (2014). Estimating high quantiles of extreme flood heights in the lower limpopo river basin of mozambique using model based bayesian approach. *Natural Hazards and Earth System Sciences Discussions*, 2(8), 5401–5425.

Marques, C., Ferreira, J., Rocha, A., Castanheira, J., Melo-Gonçalves, P., Vaz, N., & Dias, J. (2006). Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth, Parts A/B/C*, 31(18), 1172–1179.

Masereka, E. M., Ochieng, G. M., & Snyman, J. (2018). Statistical analysis of annual maximum daily rainfall for nelspruit and its environs. *Jàmbá: Journal of Disaster Risk Studies*, 10(1).

Maúre, G., Pinto, I., Ndebele-Murisa, M., Muthige, M., Lennard, C., Nikulin, G., Dosio, A., & Meque, A. (2018). The southern african climate under 1.5 c and 2 c of global warming as simulated by cordex regional climate models. *Environmental Research Letters*, 13(6), 065002.

Maxmen, A. (2018). As cape town water crisis deepens, scientists prepare for 'day zero'. *Nature*, 554(7690), 13–14.

Miguez, F., Archontoulis, S., & Dokoohaki, H. (2018). Nonlinear regression models and applications. *Applied statistics in agricultural, biological, and environmental sciences*, (appliedstatistics), 401–448.

Molla, K. I., Rahman, S., Sumi, A., & Bank, P. (2005). Emperical Mode Decomposition Analysis of Climate Changes with Special Reference to rainfall data. *Discrete Dynamics in Nature and Society*, 2006, 1–17.

Molla, K. I., Sumi, A., & Rahman, S. (2007). Analysis of Temperature Change under Global Warming Impact using Empirical Mode Decomposition. *International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering*, 1.

Mossel, E. & Vigoda, E. (2006). Limitations of markov chain monte carlo algorithms for bayesian inference of phylogeny. *The Annals of Applied Probability*, 16(4), 2215–2234.

Namitha, M. R. & Ravikumar, V. (2018). Analysis of extreme rainfall events and calculation of return levels using generalised extreme value distribution. *International Journal of Pure & Applied Bioscience*, 6(6), 1309–1316.

Nangombe, S., Zhou, T., Zhang, W., Wu, B., Hu, S., Zou, L., & Li, D. (2018). Record-breaking climate extremes in africa under stabilized 1.5 c and 2 c global warming scenarios. *Nature Climate Change*, 8(5), 375.

Nema, M. K., Khare, D., Adamowski, J., & Chandniha, S. K. (2018). Spatio-temporal analysis of rainfall trends in chhattisgarh state, central india over the last 115 years. *Journal of Water and Land Development*, 36(1), 117–128.

Nemukula, M. M. & Sigauke, C. (2015). Modelling average minimum daily temperature using extreme value theory with a time varying threshold. In *Annual Proceedings of the South African Statistical Association Conference*, volume 2015 (pp. 57–64).: South African Statistical Association (SASA).

New, D. A. (2013). *A probabilistic framework for the empirical mode decomposition and applications to precipitation data*. PhD thesis, San Diego State University.

Niang, I., Ruppel, O., Abdrabo, M., Essel, A., Lennard, C., Padgham, J., & Urquhart, P. (2014). Africa. Climate Change 2014:. In *Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1199–1265).: Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Nicholls, N., Gruza, G. V., Jouzel, J., Karl, T. R., Ogallo, L. R., & Parker, D. E. (1996). Observed climate variability and change. In J. T. Houghton, L. G. M. Filho, B. A. Callander, N. Harris, A. Kattenberg, & K. Maskell

(Eds.), *In Climate Change1995: The Science of Climate Change* (pp. 133?192).: Cambridge University Press: Cambridge.

Nikulin, G., Lennard, C., Dosio, A., Kjellström, E., Chen, Y., Hänsler, A., Kupiainen, M., Laprise, R., Mariotti, L., Maule, C. F., et al. (2018). The effects of 1.5 and 2 degrees of global warming on africa in the cordex ensemble. *Environmental Research Letters*, 13(6), 065003.

Norazizi, N. A. A. & Deni, S. M. (2019). Comparison of artificial neural network (ANN) and other imputation methods in estimating missing rainfall data at kuantan station. In *Communications in Computer and Information Science* (pp. 298–306). Springer Singapore.

O Ibeje, A., Osuagwu, J., & R Onosakponome, O. (2018). A markov model for prediction of annual rainfall. *Journal of Engineering and Applied Science*, 3, 1–5.

Ouyang, Q., Lu, W., Xin, X., Zhang, Y., Cheng, W., & Yu, T. (2016). Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. *Water Resources Management*, 30(7), 2311–2325.

Peel, M., Srikanthan, R., McMahon, T., & Karoly, D. (2011). Ensemble empirical mode decomposition of monthly climatic indices relevant to australian hydroclimatology. In *MODSIM2011, 19th International Congress on Modelling and Simulation, December* (pp. 3615–3621).

Peralta, N., Assefa, Y., Du, J., Barden, C., & Ciampitti, I. (2016). Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. *Remote Sensing*, 8(10), 848.

Philippon, N., Rouault, M., Richard, Y., & Favre, A. (2012). The influence of enso on winter rainfall in south africa. *International Journal of Climatology*, 32(15), 2333–2347.

Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). Synchronization?a unified approach to nonlinear science.

Qiwei, X., Bo, X., Jianping, L., Weixuan, X., & Hua, H. (2007). EMD algorithm based on bandwidth and the applicationon one economic data analysis. In *Signal Processing Conference, 2007 15th European* (pp. 2419–2423).: IEEE.

Rilling, G., Flandrin, P., Goncalves, P., et al. (2003). On empirical mode decomposition and its algorithms. In *IEEE-EURASIP workshop on nonlinear signal and image processing*, volume 3 (pp. 8–11).: IEEER.

Robinson, T. & Metternicht, G. (2003). A comparison of inverse distance weighting and ordinary kriging for characterising within-paddock spatial variability of soil properties in western australia. *Cartography*, 32(1), 11–24.

Rosenblum, M., Pikovsky, A., Kurths, J., Schäfer, C., & Tass, P. A. (2001). Phase synchronization: from theory to data analysis. In *Handbook of biological physics*, volume 4 (pp. 279–321). Elsevier.

Sanches, G. M., Duft, D. G., Kölln, O. T., Luciano, A. C. d. S., De Castro, S. G. Q., Okuno, F. M., & Franco, H. C. J. (2018). The potential for rgb images obtained using unmanned aerial vehicle to assess and predict yield in sugarcane fields. *International journal of remote sensing*, 39(15-16), 5402–5414.

SASA (2019). Facts and figures. www.sasa.org.za.

Schulte, J. A. (2016). Wavelet analysis for non-stationary, nonlinear time series. *Nonlinear Processes in Geophysics*, 23(4), 257.

Schulze, R. E. (1997). *South African Atlas of Agrohydrology and Climatology: Contribution Towards a Final Report to the Water Research Commission on*

Project 492: Modelling Impacts of the Agricultural Environment on Water Resources: TT82-96. Water Research Commission (WRC).

Sinclair, S. & Pegram, G. G. S. (2005). Empirical Mode Decomposition in 2-D space and time: a tool for space-time rainfall analysis and nowcasting. *Hydrology & Earth System Sciences Discussions*, 2(1), 289–318.

Srikanthan, R., Peel, M., McMahon, T., & Karoly, D. (2011). Ensemble empirical mode decomposition of australian monthly rainfall and temperature data. In *MODSIM2011, 19th International Congress on Modelling and Simulation, December* (pp. 3643–3649).

Sukla, M. K., Mangaraj, A., & Sahoo, L. (2014). An Investigation on the Stochastic Modeling of Daily Rainfall Amount in the Mahanadi Delta Region, India. *Research Journal of Mathematical and Statistical Sciences*, 2(9), 1–8.

Sun, C. & Ma, Y. (2015). Effects of non-linear temperature and precipitation trends on loess plateau droughts. *Quaternary International*, 372, 175–179.

Tabrizi, A., Garibaldi, L., Fasana, A., & Marchesiello, S. (2014). *Influence of Stopping Criterion for Sifting Process of Empirical Mode Decomposition (EMD) on Roller Bearing Fault Diagnosis*, (pp. 389–398). Springer Berlin Heidelberg: Berlin, Heidelberg.

Tang, B., Dong, S., & Song, T. (2012). Method for eliminating mode mixing of empirical mode decomposition based on the revised blind source separation. *Signal Processing*, 92(1), 248 – 258.

Tarnavsky, E., Garrigues, S., & Brown, M. E. (2008). Multiscale geostatistical analysis of avhrr, spot-vgt, and modis global ndvi products. *Remote Sensing of Environment*, 112(2), 535–549.

Tettey, M., Oduro, F. T., Adedia, D., & Abaye, D. A. (2017). Markov chain analysis of the rainfall patterns of five geographical locations in the south eastern coast of ghana. *Earth Perspectives*, 4(1).

Torrence, C. & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1), 61–78.

Trenberth, K. E. & Hoar, T. J. (1996). The 1990?1995 El Niño?Sourthern Oscillation event: Longest on record. *Geophysical Research Letters*, 23 (1), 57?60.

UNFCC (1992). Framework Convention on Climate Change. *Article 1*, (pp.$\tilde{7}$).

Unnikrishnan, P. & Jothiprakash, V. (2018). Daily rainfall forecasting for one year in a single run using singular spectrum analysis. *Journal of Hydrology*, 561, 609 – 621.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 1–67.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in r. jstat softw 45 (3): 1–67.

Van Oldenborgh, G. J. & Burgers, G. (2005). Searching for decadal variations in enso precipitation teleconnections. *Geophysical Research Letters*, 32(15).

Van Schalkwyk, D. (2018). The impact of climate change on the bud and flowering dates of grapevine cultivars at nietvoorbij in stellenbosch. *Wineland[online]: https://www. wineland. co. za/the-impact-of-climate-change-on-the-bud-and-flowering-dates-ofgrapevine-cultivars-at-nietvoorbij-in-stellenbosch/Date of access*, 25.

Varikoden, H., Al-shukaili, H. S. A., Babu, C. A., & Samah, A. A. (2016). Rainfall over Oman and its teleconnection with El Niño Southern Oscillation. *Arabian Journal of Geosciences*, 9(8), 520.

Wang, Z.-Y., Qiu, J., & Li, F.-F. (2018). Hybrid models combining EMD/EEMD and ARIMA for long-term streamflow forecasting. *Water*, 10(7), 853.

Webb, L., Whetton, P., Bhend, J., Darbyshire, R., Briggs, P., & Barlow, E. (2012). Earlier wine-grape ripening driven by climatic warming and drying and management practices. *Nature Climate Change*, 2(4), 259.

WHO (2002). World Health Organisation. In *Global Health Observatory (GHO)*.

WHO (2015). World Health Organisation. In *Global Health Observatory (GHO)*.

Wilks, D. S. (2011). Principal component (eof) analysis. In *International Geophysics*, volume 100 (pp. 519–562). Elsevier.

Willmott, C. J. & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1), 79–82.

Wolski, P. (2018). How severe is cape town's "day zero" drought? *Significance*, 15(2), 24–27.

Wu, Z. & Huang, H. E. (2004). A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. Roy. Soc. London*, 460A 1597? 1611, 1597? 1611.

Wu, Z. & Huang, N. E. (2005). Statistical significance test of intrinsic mode functions. *Hilbert–Huang Transform and Its Applications*, 5, 107–127.

Wu, Z. & Huang, N. E. (2009). Ensemble Empirical Mode Decomposition: A noise-assisted data analysis method. *Advanced Adaptive Data Analysis*, 1, 1–41.

Wu, Z. & Huang, N. E. (2014). Statistical significance test of intrinsic mode functions. In *Hilbert–Huang Transform and Its Applications* (pp. 149–169).

Yang, Q., Ma, Z., & Xu, B. (2016). Modulation of monthly precipitation patterns over east china by the pacific decadal oscillation. *Climatic Change*, 144(3), 405–417.

Yang, X., Xie, X., Liu, D. L., Ji, F., & Wang, L. (2015). Spatial interpolation of daily rainfall data for local climate impact assessment over greater sydney region. *Advances in Meteorology*, 2015, 1–12.

Yanming, Z., Jun, W., & Xinhua, W. (2012). Study on the change trend of precipitation and temperature in kunming city based on mann-kendall analysis. In *Future Computer, Communication, Control and Automation* (pp. 505–513). Springer.

Yiou, P., Baert, E., & Loutre, M. F. (1996). Spectral analysis of climate data. *Surveys in Geophysics*, 17(6), 619–663.

Yusuf, A. U. (2014). Markov chain model and its application to annual rainfall distribution for crop production. *American Journal of Theoretical and Applied Statistics*, 3(2), 39.

Zhang, R. R., Ma, S., & Hartzell, S. (2003). Signatures of the seismic source in EMD-based characterization of the 1994 Northbridge, California, earthquake recording. *Bulletin of the Seismological Society of America*, 93(1), 501–518.

Zhang, X., Peng, Y., Zhang, C., & Wang, B. (2015). Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? some experiment evidences. *Journal of Hydrology*, 530, 137–152.

Zhang, Z. & Moore, J. C. (2015). Empirical orthogonal functions. In *Mathematical and Physical Fundamentals of Climate Change* (pp. 161–197). Elsevier.

Zhao, D. & Li, Y.-R. (2015). Climate change and sugarcane production: Potential impact and mitigation strategies. *International Journal of Agronomy*, 2015, 1–10.

Zhao, X.-h. & Chen, X. (2015). Auto Regressive and Ensemble Empirical Mode Decomposition Hybrid Model for Annual Runoff Forecasting. *Water Resources Management*, 29(8), 2913–2926.

Zhu, S., Zhou, J., Ye, L., & Meng, C. (2016). Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of yangtze river, china. *Environmental Earth Sciences*, 75(6).

Ziervogel, G., New, M., Archer van Garderen, E., Midgley, G., Taylor, A., Hamann, R., Stuart-Hill, S., Myers, J., & Warburton, M. (2014). Climate change impacts and adaptation in South Africa. 5(5), 605 – 620.

APPENDICES

This page is left blank on purpose.

Appendix A

EMD/EMD Matlab Code

```
%function allmode=eemd(Y,Nstd,NE)
%
% This is an EMD/EEMD program
%
% INPUT:
%       Y: Inputted data;1-d data only
%       Nstd: ratio of the standard deviation of the added noise and that of Y;
%       NE: Ensemble number for the EEMD
% OUTPUT:
%       A matrix of N*(m+1) matrix, where N is the length of the input
%       data Y, and m=fix(log2(N))-1. Column 1 is the original data, columns 2, 3, ...
%       m are the IMFs from high to low frequency, and column (m+1) is the
%       residual (over all trend).
%
%
% Additional function: extrema.m which is found below
% Reference Wu et al. (2009)

function allmode=eemd(Y,Nstd,NE)

%part1.read data, find out standard deviation ,divide all data by std
xsize=length(Y);
dd=1:1:xsize;
Ystd=std(Y);
Y=Y/Ystd;

%part2.evaluate TNM as total IMF number,ssign 0 to TNM2 matrix
TNM=fix(log2(xsize))-1;
TNM2=TNM+2;
for kk=1:1:TNM2,
    for ii=1:1:xsize,
        allmode(ii,kk)=0.0;
    end
end

%part3 Do EEMD  -----EEMD loop start
for iii=1:1:NE,   %EEMD loop -NE times EMD sum together

    %part4 --Add noise to original data,we have X1
    %for i=1:xsize,
     %   temp=randn(1,1)*Nstd;
      %  X1(i)=Y(i)+temp;
    %end
    X1=awgn(Y,Nstd,'measured');

    %part4 --assign original data in the first column
    for jj=1:1:xsize,
        mode(jj,1) = Y(jj);
    end

    %part5--give initial 0 to xorigin and xend
    xorigin = X1;
    xend = xorigin;

    %part6--start to find an IMF-----IMF loop start
    nmode = 1;
```

```
        while nmode <= TNM,
         xstart = xend; %last loop value assign to new iteration loop
                        %xstart -loop start data
         iter = 1;      %loop index initial value

         while iter<=10,
             [spmax, spmin, flag]=extrema(xstart);  %call function extrema
             %the usage of  spline ,please see part11.
             upper= spline(spmax(:,1),spmax(:,2),dd); %upper spline bound of this sift
             lower= spline(spmin(:,1),spmin(:,2),dd); %lower spline bound of this sift
             mean_ul = (upper + lower)/2;%spline mean of upper and lower
             xstart = xstart - mean_ul;%extract spline mean from Xstart
             iter = iter +1;
         end


         %part8--subtract IMF from data ,then let the residual xend to start to find next IMF
          xend = xend - xstart;

          nmode=nmode+1;

         %part9-- xstart is this time IMF
         for jj=1:1:xsize,
             mode(jj,nmode) = xstart(jj);
         end

    end
    %part6--start to find an IMF-----IMF loop end

    %part 10--after gotten  all(TNM) IMFs ,the residual xend is over all trend
    %                        put them in the last column
    for jj=1:1:xsize,
        mode(jj,nmode+1)=xend(jj);
    end
    %after part 10 ,original +TNM-IMF+overall trend  ---those are all in mode
     allmode=allmode+mode;

end
%part3 Do EEMD  -----EEMD loop end

%part10--devide EEMD summation by NE,std be multiply back to data
allmode=allmode/NE;
allmode=allmode*Ystd;

%
%

function [spmax, spmin, flag]= extrema(in_data)

flag=1;
dsize=length(in_data);

%part1.--find local max value and do end process

%start point

spmax(1,1) = 1;
spmax(1,2) = in_data(1);
```

```
%Loop --start find max by compare the values
%when [ (the jj th value > than the jj-1 th value ) AND (the jj th value > than the jj+1 th value

jj=2;
kk=2;
while jj<dsize,
    if ( in_data(jj-1)<=in_data(jj) & in_data(jj)>=in_data(jj+1) )
        spmax(kk,1) = jj;
        spmax(kk,2) = in_data (jj);
        kk = kk+1;
    end
    jj=jj+1;
end

%end point
spmax(kk,1)=dsize;
spmax(kk,2)=in_data(dsize);

%End point process-please see reference about spline end effect
%extend the slpoe of neighbor 2 max value ---as extend value
%original value of end point -----as original value
%compare extend and original value

if kk>=4
    slope1=(spmax(2,2)-spmax(3,2))/(spmax(2,1)-spmax(3,1));
    tmp1=slope1*(spmax(1,1)-spmax(2,1))+spmax(2,2);
    if tmp1>spmax(1,2)
        spmax(1,2)=tmp1;
    end

    slope2=(spmax(kk-1,2)-spmax(kk-2,2))/(spmax(kk-1,1)-spmax(kk-2,1));
    tmp2=slope2*(spmax(kk,1)-spmax(kk-1,1))+spmax(kk-1,2);
    if tmp2>spmax(kk,2)
        spmax(kk,2)=tmp2;
    end
else
    flag=-1;
end

msize=size(in_data);
dsize=max(msize);
xsize=dsize/3;
xsize2=2*xsize;


%part2.--find local min value and do end process

%local max
spmin(1,1) = 1;
spmin(1,2) = in_data(1);
jj=2;
kk=2;
while jj<dsize,
    if ( in_data(jj-1)>=in_data(jj) & in_data(jj)<=in_data(jj+1))
        spmin(kk,1) = jj;
        spmin(kk,2) = in_data (jj);
        kk = kk+1;
    end
    jj=jj+1;
end
```

```
%local min
spmin(kk,1)=dsize;
spmin(kk,2)=in_data(dsize);

if kk>=4
    slope1=(spmin(2,2)-spmin(3,2))/(spmin(2,1)-spmin(3,1));
    tmp1=slope1*(spmin(1,1)-spmin(2,1))+spmin(2,2);
    if tmp1<spmin(1,2)
        spmin(1,2)=tmp1;
    end

    slope2=(spmin(kk-1,2)-spmin(kk-2,2))/(spmin(kk-1,1)-spmin(kk-2,1));
    tmp2=slope2*(spmin(kk,1)-spmin(kk-1,1))+spmin(kk-1,2);
    if tmp2<spmin(kk,2)
        spmin(kk,2)=tmp2;
    end
else
    flag=-1;
end

flag=1;
```

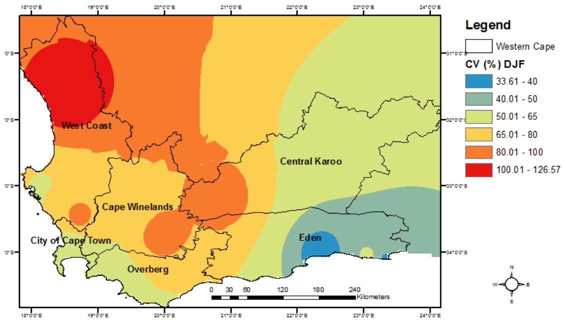
```
%##################################################################
%  function [spmax, spmin, flag]= extrema(in_data)
%
% This is a utility program for cubic spline envelope,
%   the code is to  find out max values and max positions
%                            min values and min positions
%    (then use matlab function spline to form the spline)
%
%   function [spmax, spmin, flag]= extrema(in_data)

% Reference Wu et al. (2009)

function [spmax, spmin, flag]= extrema(in_data)

flag=1;
dsize=length(in_data);


spmax(1,1) = 1;
spmax(1,2) = in_data(1);

%Loop --start find max by compare the values
%when [ (the jj th value > than the jj-1 th value ) AND (the jj th value > than the jj+1 th value
)
%the value jj is the position of the max
%the value in_data (jj) is the value of the max
%do the loop by index-jj
%after the max value is found,use index -kk to store in the matrix
%kk=1,the start point
%the last value of kk ,the end point






jj=2;
kk=2;

while jj<dsize,
    if ( in_data(jj-1)<=in_data(jj) & in_data(jj)>=in_data(jj+1) )
```

```
        spmax(kk,1) = jj;
        spmax(kk,2) = in_data (jj);
        kk = kk+1;
    end
    jj=jj+1;
end

%end point
spmax(kk,1)=dsize;
spmax(kk,2)=in_data(dsize);

%End point process-please see reference about spline end effect
%extend the slope of neighbour 2 max value ---as extend value
%original value of end point -----as original value
%compare extend and original value

if kk>=4
    slope1=(spmax(2,2)-spmax(3,2))/(spmax(2,1)-spmax(3,1));
    tmp1=slope1*(spmax(1,1)-spmax(2,1))+spmax(2,2);
    if tmp1>spmax(1,2)
        spmax(1,2)=tmp1;
    end

    slope2=(spmax(kk-1,2)-spmax(kk-2,2))/(spmax(kk-1,1)-spmax(kk-2,1));
    tmp2=slope2*(spmax(kk,1)-spmax(kk-1,1))+spmax(kk-1,2);
    if tmp2>spmax(kk,2)
        spmax(kk,2)=tmp2;
    end
else
    flag=-1;
end

%these 4 sentence seems useless.
msize=size(in_data);
dsize=max(msize);
xsize=dsize/3;
xsize2=2*xsize;

%local max
spmin(1,1) = 1;
spmin(1,2) = in_data(1);
jj=2;
kk=2;
while jj<dsize,
    if ( in_data(jj-1)>=in_data(jj) & in_data(jj)<=in_data(jj+1))
        spmin(kk,1) = jj;
        spmin(kk,2) = in_data (jj);
        kk = kk+1;
    end
    jj=jj+1;
end


%local min
spmin(kk,1)=dsize;
spmin(kk,2)=in_data(dsize);
```

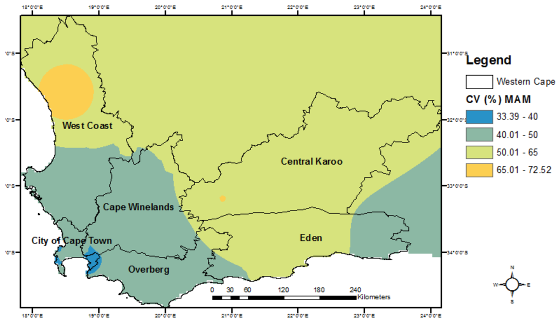
```
if kk>=4
    slope1=(spmin(2,2)-spmin(3,2))/(spmin(2,1)-spmin(3,1));
```

```
    tmp1=slope1*(spmin(1,1)-spmin(2,1))+spmin(2,2);
    if tmp1<spmin(1,2)
        spmin(1,2)=tmp1;
    end

    slope2=(spmin(kk-1,2)-spmin(kk-2,2))/(spmin(kk-1,1)-spmin(kk-2,1));
    tmp2=slope2*(spmin(kk,1)-spmin(kk-1,1))+spmin(kk-1,2);
    if tmp2<spmin(kk,2)
        spmin(kk,2)=tmp2;
    end
else
    flag=-1;
end

flag=1;
```

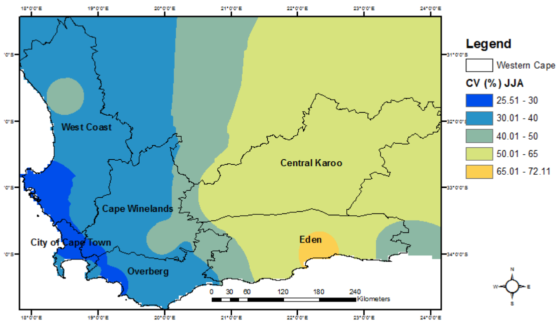
Appendix B



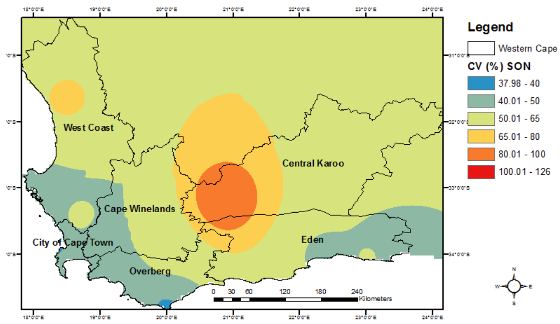
(a) DJF season variability



(b) MAM season variability



(c) JJA season variability



(d) SON season variability

Figure B1: Seasonal rainfall variability in Western Cape a) DJF b) MAM c) JJA d) SON.

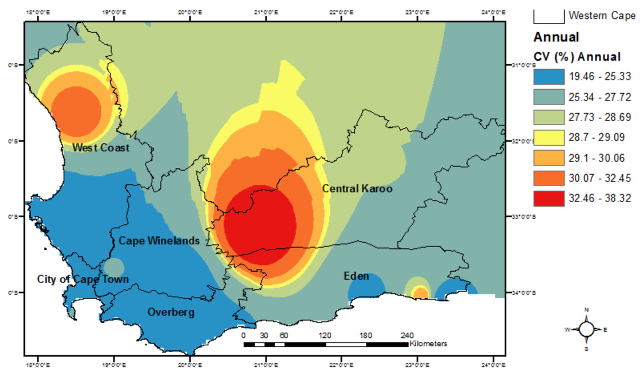


Figure B2: Annual rainfall variability in Western Cape

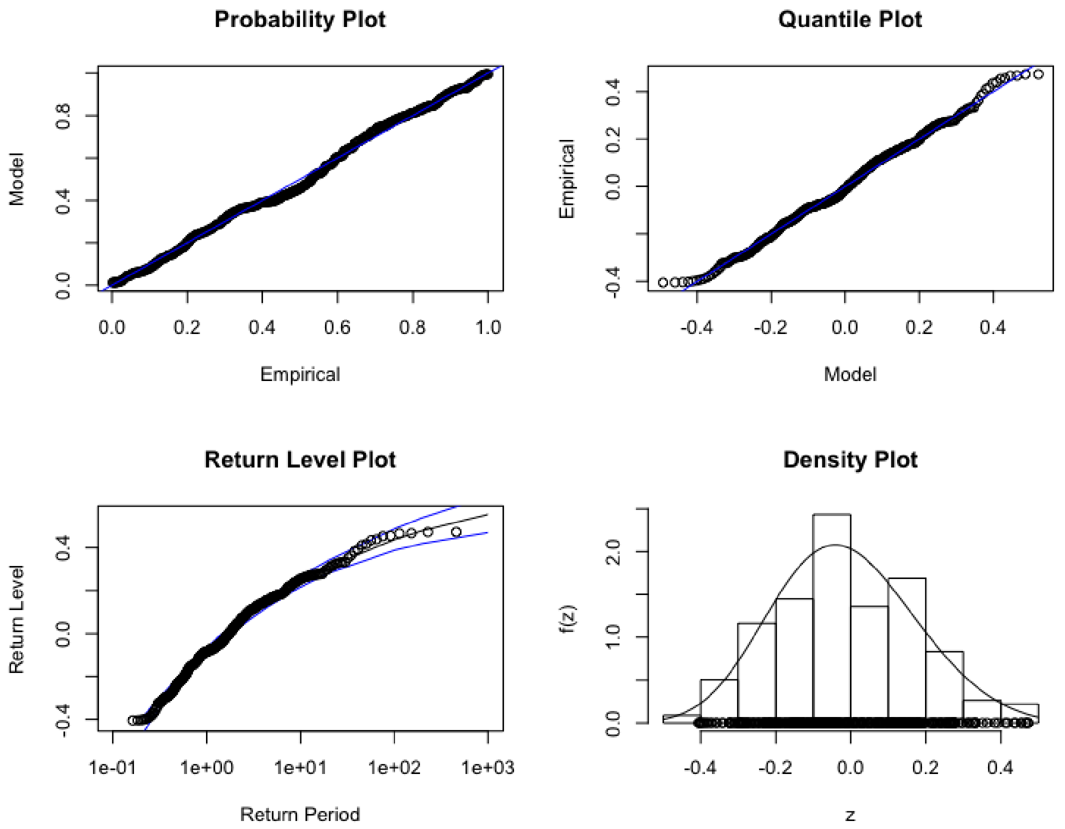


Figure B3: Diagnostic plots illustrating the fit of the rainfall IMF5 data to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (bottom right panel)

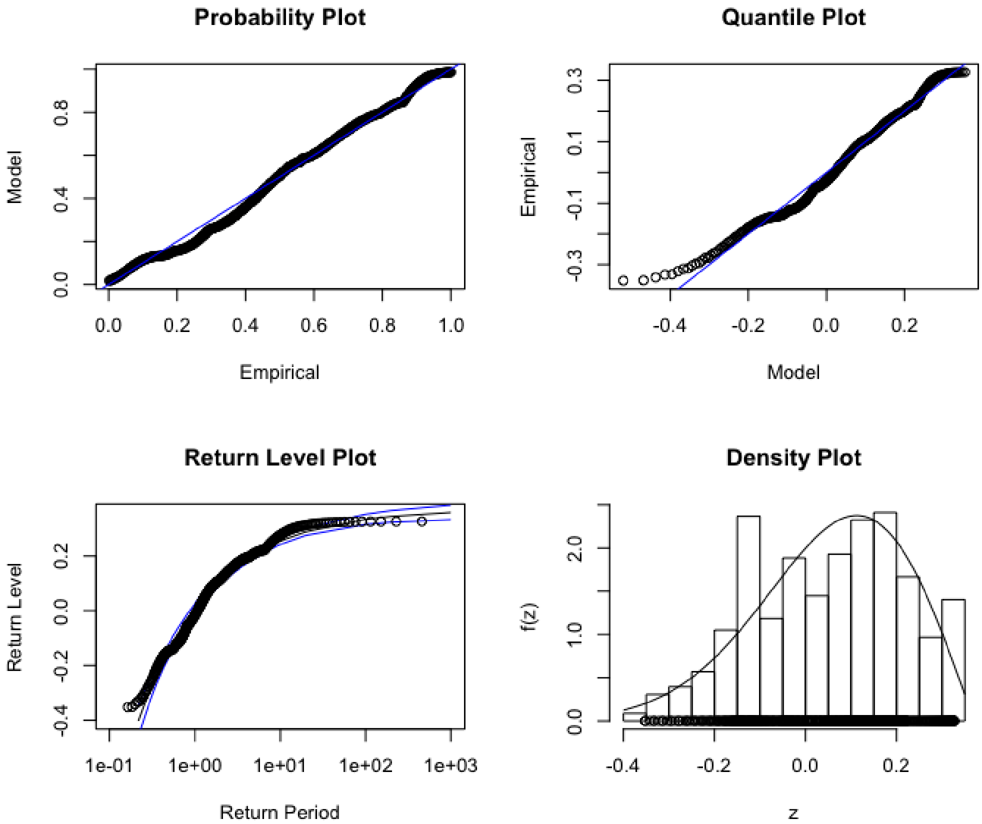


Figure B4: Diagnostic plots illustrating the fit of the summation IMF6, IMF7 and residual data to the GEVD, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (bottom right panel)

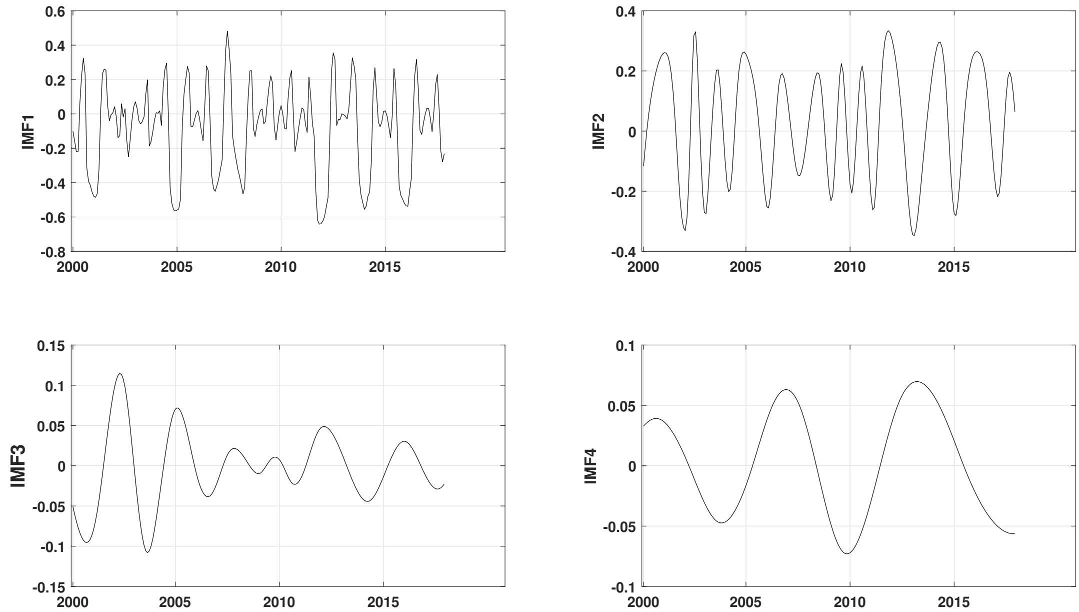


Figure B5: Decomposition of grapevines' NDVI into IMF1 to 4.

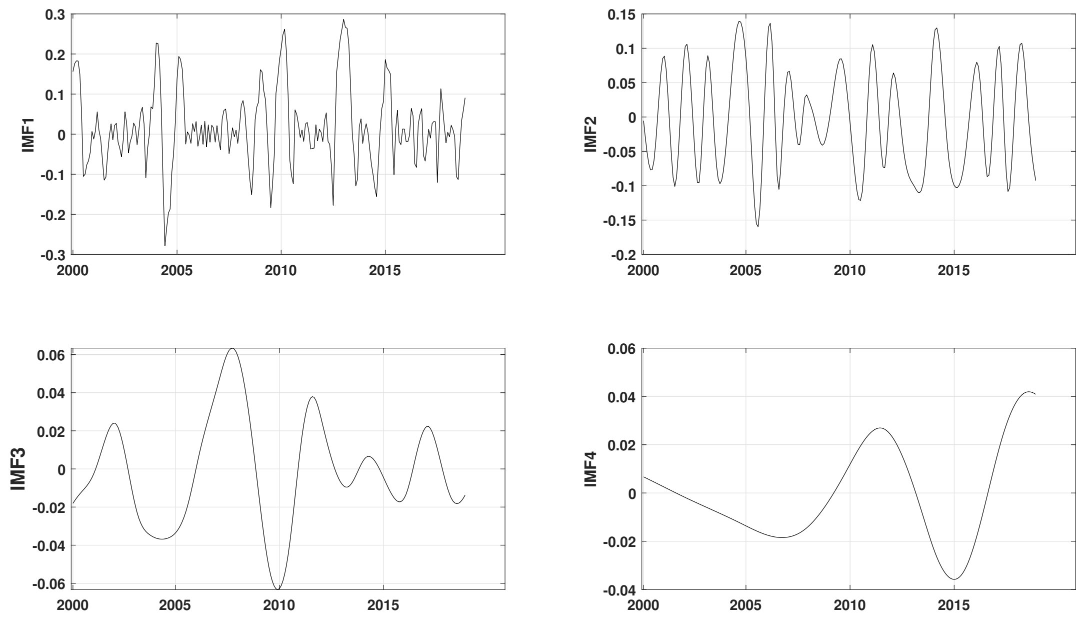


Figure B6: Decomposition of sugarcane's NDVI into IMF1 to 4.

Appendix C

The peer reviewed paper published as articles and conference proceedings listed on page xx are attached in the following pages.

Zvarevashe, W, et al. 2019. Analysis of Rainfall and Temperature Data Using Ensemble Empirical Mode Decomposition. *Data Science Journal,* 18: 46, pp. 1–9. DOI: https://doi.org/10.5334/dsj-2019-046

RESEARCH PAPER

Analysis of Rainfall and Temperature Data Using Ensemble Empirical Mode Decomposition

Willard Zvarevashe[1], Syamala Krishnannair[1] and Venkataraman Sivakumar[2]

[1] Department of Mathematical Sciences, University of Zululand, KwaDlangezwa, ZA

[2] Discipline of Physics, School of Chemistry and Physics, University of Kwazulu-Natal, Durban, ZA

Corresponding author: Willard Zvarevashe (wzvarevashe@gmail.com)

Climatic variables such as rainfall and temperature have nonlinear and non-stationary characteristics such that analysing them using linear methods inconclusive results are found. Ensemble empirical mode decomposition (EEMD) is a data-adaptive method that is best suitable for data with nonlinear and non-stationary characteristics. The average monthly rainfall and temperature data for a selected region in South Africa are decomposed into intrinsic mode functions (IMFs) at different time scales using EEMD. The IMFs exhibit an inter-annual to inter-decadal variability. The influence of climatic oscillations such as El-Niño Southern Oscillation (ENSO) and quasi-biennial oscillation (QBO) is identified. The influence of temperature variability on rainfall is also shown at different time scales. Based on the results obtained, the EEMD method is found to be suitable to identify different oscillations in the rainfall and temperature data.

Keywords: Rainfall; Temperature; EEMD; Climate Change; Data analysis

1. Introduction

The annual variation of rainfall in Southern Africa, directly and indirectly, affect human livelihoods and ecosystem through droughts, temperature changes, water supply problems and reduced agricultural production. Most of Southern Africa experience Austral summer rainfall (October to March) and north-eastern to south-western regions experience austral winter rainfall (May to August) (Dieppois *et al.*, 2016). The summer and winter rainfall variability has been shown to be influenced by El Niño-Southern Oscillation (ENSO) and becomes nonlinear, with extreme weather conditions (Phillippon *et al.*, 2012; Fauchereau *et al.*, 2009; Kane, 2009). The temperature in South Africa has increased by over one and half times more than the globally observed temperature increases (Kruger & Shongwe, 2004; DEA, 2011; Mackellar *et al.*, 2014).

El Niño describes the warming of sea surface temperature that occurs periodically, typically concentrated in the central-east equatorial Pacific (Lehodey *et al.*, 1997). La Niña is the term adopted which describes the opposite side of the fluctuations. The Niño 3.4 index is one of the key atmospheric indices used to gauge the strength of El Niño and La Niña. The other driver that has been shown to have an impact on weather patterns is the Quasi-Biennial Oscillation (QBO) (Begue *et al.*, 2010). QBO is a regular variation of the winds that blow high above the equator. Strong winds in the stratosphere travel in a belt around the planet, and in about 14 months these winds completely change direction (Kane, 2009). In the study by Kane (2009), it was shown that the warming of climate and sea surface temperatures has an impact on the rainfall patterns. Due to the influences of these climatic drivers and others, the weather patterns become nonlinear and non-stationary such that linear models are sometimes inconclusive (Huang *et al.*, 1999; Molla *et al.*, 2006). Fourier based methods assume that the data is linear and the data must be strictly periodic which is not the case with climate data (Schulte, 2016).

Empirical mode decomposition (EMD), was introduced by Huang *et al.* (1998), which does not make assumptions about linearity and stationarity of the time series and it is best suitable to analyse climate data. The time series is decomposed into different time scales called intrinsic mode functions (IMFs), which can reveal intrinsic changes in the climate system (Huang *et al.*, 1998; Wu *et al.*, 2009). EMD has a challenge of mixing the signals from one IMF to another, therefore, to cater for this, a noise assisted method named Ensemble Empirical Mode Decomposition (EEMD) was introduced (Wu *et al.*, 2009).

EEMD has been applied widely in the hydrological and atmospheric studies. Chiew *et al.* (2005) applied EMD to annual streamflow to find significant oscillations in the data. Twenty unimpaired catchments from different parts of the world were used and 3, 6–7, 11–15 and 20–25 year oscillations in the stream flows were identified. In another study, the climate variability was observed in the case study of East and Central Africa, which was demonstrated by mapping the coupling between precipitation variables, inter-annual vegetation changes and the ENSO and Indian Ocean Dipole (IOD). EEMD was adopted because of its ability to breakdown normalised vegetation index (NDVI) series into multiple time scales components and its generic ability to be used with other time series analysis tools (Hawinkel *et al.*, 2015, 2016). The Pacific Decadal Oscillation (PDO) is known to have an influence over East China's annual and summer rainfall. Using EEMD the monthly rainfall influence of PDO was identified (Yang *et al.*, 2017). The intrinsic oscillations in the land surface temperature of Wuhan, China were revealed using EEMD by decomposing the data into annual, inter-annual, noise and trend (Liu *et al.*, 2019). In South Korea several climate drivers have been shown to have an impact on the precipitation, however many studies did not take into consideration the inherent cycles in the long term precipitation. Using EEMD, cross-correlation and multiple linear regression on the influence of ENSO, QBO, Arctic Oscillation, Atlantic Meridional Mode and others was shown at the monthly level (Kim *et al.*, 2018). Some notable applications of EEMD include finding the effects of temperature and precipitation trends in Plateau droughts (Sun & Ma, 2015) and identifying the variability of monthly precipitation in Iran (Alizadeh *et al.*, 2019).

In this study EEMD is used to decompose a 38-year rainfall and temperature data for a selected region in Western Cape, South Africa to reveal underlying physical signals. The influence of ENSO, QBO and temperature on the rainfall pattern is identified. The selected region chosen has been facing a lot of water challenges recently. There has been growing general interest in the winter rainfall mainly due to a threat of "day zero" in 2018 (Kruger *et al.*, 2017; Maxmen, 2018; Wolski, 2018). However, very few studies have used statistical analysis to investigate the rainfall variability during winter. This is the first study that uses EEMD and synchronisation to investigate the influence of ENSO and QBO on the rainfall pattern for the region. This paper is organized as follows: in section 2, the data set and methodology are described; in section 3, the analysis and discussion of the results is done and conclusion is done in section 4.

2. Data and Method

2.1 Data

The average monthly rainfall and temperature data were obtained from the South African Weather Service (SAWS) for the period 1980 to 2018 for an area between 18.2–19.2°E and 33.5–34.5°S, which contains 11 weather stations as shown in **Figure 1**. The area selected receives winter rainfall from April to August.

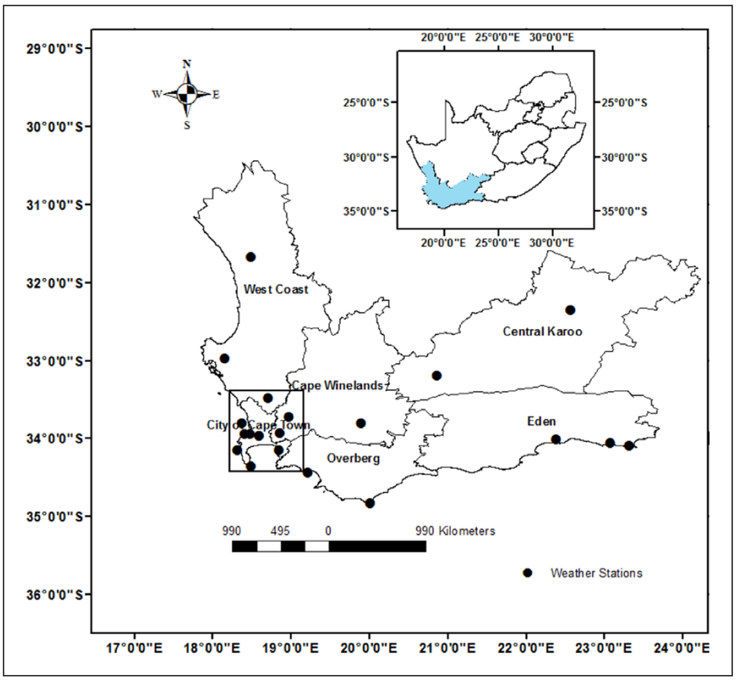


Figure 1: Study Area (square) in Western Cape South Africa with 10 weather stations.

The data for Niño 3.4 and QBO is publicly available on Climate Explorer website (http://climexp.knmi.nl) (Oldenborgh & Burgers, 2005).

Rainfall and temperature data from weather stations has challenge of having missing data. Therefore, to cater for the missing data multivariate imputation by chained equations (MICE) method is used to impute the missing data (Van Buuren & Groothuis-Oudshoorn, 2011). The MICE method was chosen for this study because it does not assume normal distribution of the data and it also assumes missing at random (MAR). The R package '*mice*' developed by Van Buuren & Groothuis-Oudshoorn (2011) is used for imputation.

2.2 Ensemble Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is an adaptive time-frequency data representation technique, which requires only that the data must consist of a simple intrinsic mode of oscillations (Huang *et al.*, 1998). It is most suitable for nonlinear and non-stationary data. The EMD methodology is based on a shifting process, which identifies local extrema (maxima and minima) and results in the formation of intrinsic mode functions (IMFs). In order to decompose a given time series x_t into IMFs the following algorithm is used:

(i) Construct the upper x_t^{max} and lower x_t^{min} envelopes connecting via cubic spline interpolation all the maxima and minima of x_t, respectively.
(ii) Compute $\Delta x_t = x_t - \frac{x_t^{\max} - x_t^{\min}}{2}$;
(iii) Repeat steps (i) and (ii) for Δx_t until the resulting signal possess the properties that the number of extrema is equal (or differ at most by one) to the number of zero crossings, and the mean value between the upper and lower envelope is equal to zero at any point. Denote the resulting signal by $h_t(1)$, which is the first IMF.
(iv) Take the difference $x_t(1) = x_t - h_t(1)$ and repeat steps from (i) to (iii) to obtain the second IMF $h_t(2)$ (Huang *et al.*, 1998; Guo *et al.*, 2016).

The first IMF contain the highest fluctuations and this is subtracted from the original data and subsequent IMFs are then derived from the subtracted data. The IMFs and residual data approximate the original data when they are summed together (Huang *et al.*, 1998). A time series is decomposed into IMFs, $h_t(i)$ ($i = 1,2, \ldots n$) and residual r_t so that the original data is approximated by the sum of IMFs and residual.

$$x_t \approx \sum_{i=1}^{n} h_t(i) + r_t. \tag{2.1}$$

EMD has a challenge of mode mixing, where signals from one IMF is found in another. Therefore, a noise assisted method, Ensemble Empirical Mode Decomposition (EEMD), which consist of adding white noise before carrying out EMD algorithm was introduced (Wu & Huang, 2009). For a given time series x_t an ensemble of white noise of size m, ε_j ($j = 1,2\ldots,m$), is introduced to each data point, x_i, such that the i^{th} "artificial" value becomes

$$y_i = \varepsilon_j + x_i. \tag{2.2}$$

An average of the IMFs found from the data with noise becomes the final IMF, that is

$$c_t(i) = \frac{1}{m}\sum_{j=1}^{m} d_t(j) \tag{2.3}$$

where $d_t(j)$ is the IMF of the time series with added noise, y_t and $c_t(i)$ is the final i^{th} IMF for the original time series x_t. The average of residuals from y_t gives the final residual that is,

$$r_t = \frac{1}{m}\sum_{j=1}^{m} r_t(j) \tag{2.4}$$

where r_t^j are residuals from the time series with added noise.

The IMFs must be mutually orthogonal to each other. Higher orthogonality corresponds to less amount of information leakage. The index of orthogonality (IO) is used to calculate the orthogonality which is given by

$$IO = \frac{1}{n}\sum_{t}\frac{1}{\left(x_t\right)^2}\left(\sum_{i=1}^{n+1}\sum_{j=1}^{n+1} h_t(i)\cdot\ h_t(j)\right) \tag{2.5}$$

where i and j represents the i^{th} and j^{th} IMFs and n is the size of the IMF (Molla *et al.,* 2006).

2.3 Synchronisation

Synchronisation of coupled oscillating systems means appearance of certain relations between their phases and frequencies (Rosenblum *et al.*, 2001). Here we use this concept in order to reveal the interaction between rainfall and other climatic drivers. R package '*synchrony*' is used which measures phase synchrony between quasiperiodic times series (Cazelles & Stone 2003). Time series that are phase synchronised or locked exhibit a modal distribution with a prominent peak at a given phase difference, whereas unrelated times series are characterized by a uniform or diffuse distribution.

3. Results and Discussion

The average monthly rainfall for the region located between 18.2–19.2°E and 33.5–34.5°S was standardised for easy computation and comparison. The multivariate imputation by chained equations is used to impute the missing rainfall and temperature data. The data was decomposed until when there is at most one maximum and one minimum in the residual. From the standardised rainfall data, 7 IMFs were found and are shown in **Figure 2**. IMF 3 has a period of about 12 months which corresponds to an annual (seasonal)

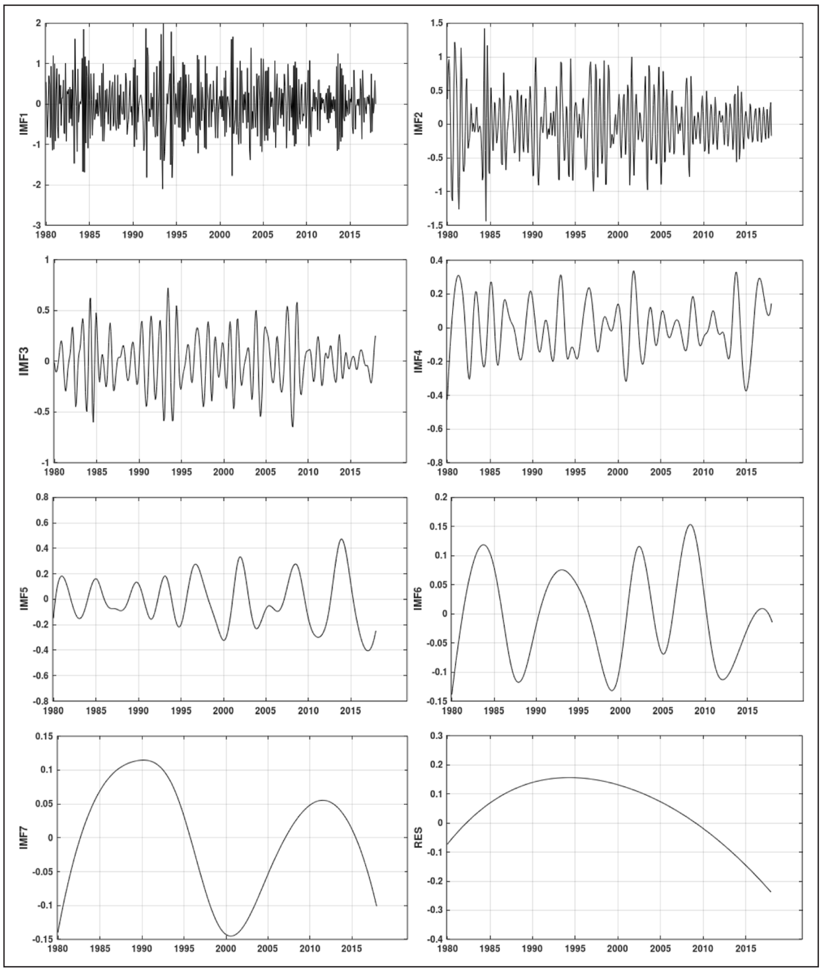


Figure 2: Decomposition of rainfall into IMF1 to 7 and residual using EEMD. The *x*-axis represents the time in years and *y*-axis represents the frequency. The graphs are labelled on the y-axis from IMF1 (first graph on the left) to RES (last graph on the bottom right) which is a graph of the residual. IMF1 captures the noise found in the rainfall data, IMF2 inter-annual oscillation, IMF3 annual oscillation, IMF4 2-years oscillation, IMF5 4.5-year oscillation, IMF6 7-year oscillation and IMF7 16.5 year oscillation. The plot of the residual shows the general trend of the rainfall.

oscillation, IMF 4 has a period of about 26 months which approximated a 2 year oscillation, IMF 5 has an oscillation of about 54 months (4.5 years) and IMF 6 captures a quasi-decadal oscillation (7 year period). Previous studies have used Wavelet Analysis, identified similar oscillations in the South African rainfall that were also found in this study (Kane, 2009; Dieppois *et al.*, 2016). In these previous studies, winter rainfall was found to be having significant 2–3 year period and 3–4 year period which contributed to the rainfall variability. As compared to Wavelet Analysis, EEMD is adaptive, intuitive and does not use basis functions. Additionally, the impact of different climate drivers at different time scales can be shown. The graph on the bottom right of **Figure 2** illustrates the residual plot, which shows the general trend of the rainfall. A study by Maúre *et al.* (2018) used several climatic models to predict the rainfall trend over Southern Africa under global warming and the study pointed to a decreasing daily rainfall for Western Cape, which is in agreement with the obtained residual plot.

The probability density function for each IMF is approximately normally distributed. The IMFs and residual are added together to reconstruct the data. The reconstructed data approximates the original data with Root Mean Square Error of order 10–14. This clearly shows that EMD is lossless decomposition with minimal data being lost in the decomposition and managing to capture most of the oscillations in the data.

It is noted that the maximum value of orthogonality between the IMFs is found to be approximately equal to 0.001 and it is way below the acceptable value of less than 0.1. The index of orthogonality for the IMFs is 0.594 × 10–4 for rainfall, 0.154 × 10–6 for Niño 3.4 and 0.235 × 10–5 for QBO. It confirms that there is less amount of information leakage.

The cross-correlation between rainfall and QBO and Niño 3.4 shows that there is no correlation as shown in the auto-correlation function (ACF) plot in **Figure 3**. However, when the time series were decomposed correlation is identified for IMF 3 for both Niño 3.4 and QBO as shown in **Figure 4**. There is a correlation of rainfall's IMF3 and Niño 3.4 index at lag –4, –5, –6, 2 and 3. These results confirm the influence of ENSO on the seasonal rainfall and also the quasi-biennial oscillation which is consistent with results found by Philippon *et al.* (2012) and Kane (2009). Additionally, the general pattern of the rainfall at different time scales is identified up to quasi-decadal oscillation.

Cross-correlation can be used only when the time series is stationary, the Augmented Dicky Fuller Test (ADF) and Phillips-Perron Unit Root Test shows that IMF 4 to 7 are not stationary therefore cross-correlation cannot be used. The synchronisation does not require that the time series to be stationary hence it was used to identify any relationship for those IMFs. These IMFs found are further synchronised with Niño 3.4 and QBO and the results are shown in **Figure 5** below. The results show that there is weak coupling between the original rainfall time series and Niño 3.4 or QBO. However, there is phase-locking identified for IMF 5 for Niño 3.4, since there is a clear peak. These results for Niño 3.4 shows that there may be an influence of ENSO on the rainfall pattern. This is in agreement with the results that were found by Philippon *et al.* (2012), which found that there was a significant association between ENSO and winter rainfall. They showed that there is a strong correlation of the May-June-July season with ENSO than any other period of the year. In this study using EEMD, the correlation is further done at a monthly level than seasonal level. The clear peak

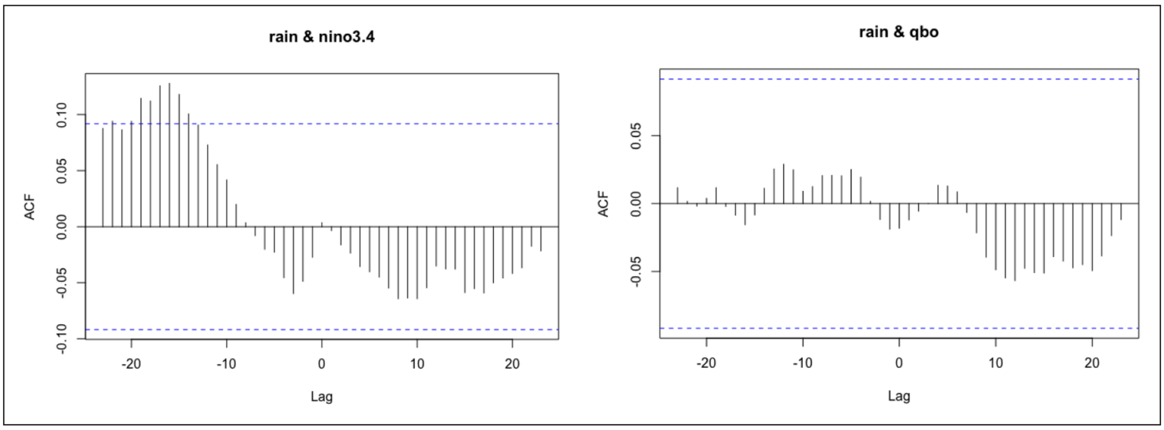


Figure 3: Cross-correlation for rainfall with Niño 3.4 and QBO. The *x*-axis represents the time gap in months and *y*-axis represents the correlations. The spikes that are above or below the dotted blue line indicate significant correlation. In the left graph shows that there is correlation of rainfall and Niño 3.4 index at lag –14 to –19. The right graph shows that there is no correlation of rainfall and QBO. The raw data is not showing any association of ENSO and QBO with rainfall data.

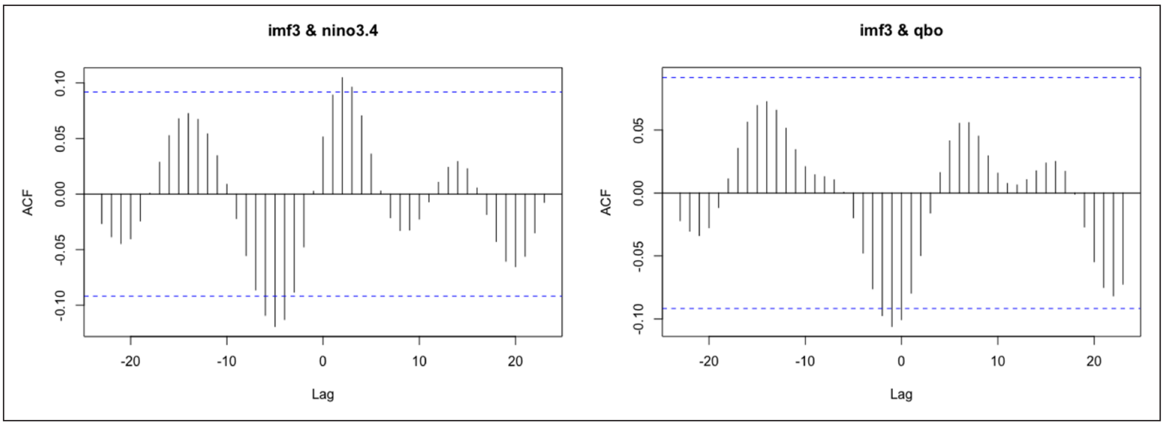


Figure 4: Cross-correlation for IMF3 with Niño 3.4 and QBO. The *x*-axis represents the time gap in months and *y*-axis represents the correlations. The spikes that are above or below the blue line indicate significant correlation. In the left graph shows that there is correlation of rainfall's IMF3 and Niño 3.4 index at lag –4, –5, –6, 2 and 3. The right graph shows that there is correlation of rainfall's IMF3 and QBO at lag –1 and –2.

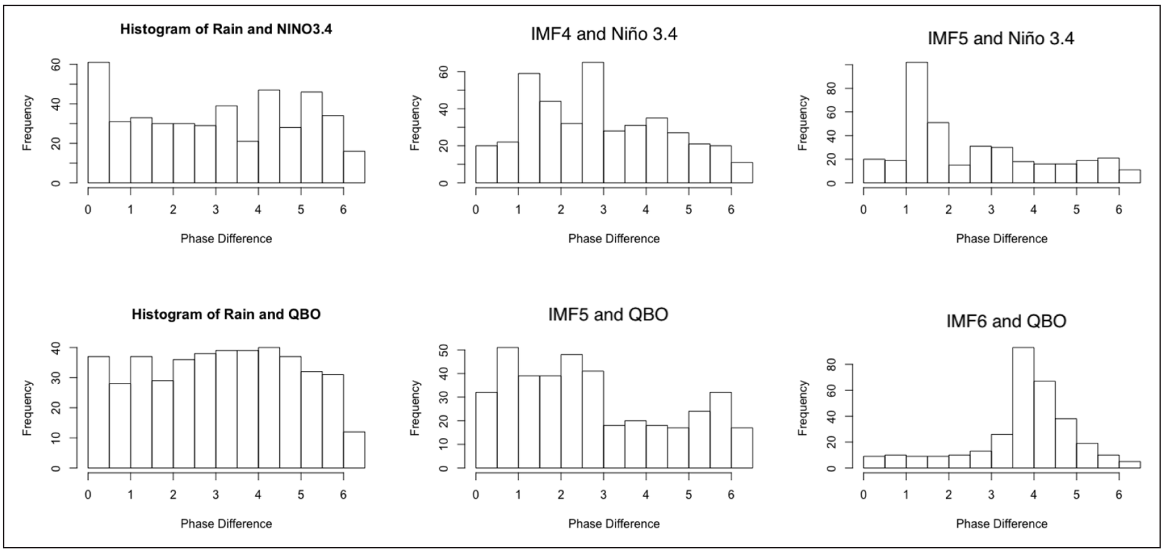


Figure 5: Histogram plot of synchronisation of rainfall, QBO and Niño 3.4 index IMFs. The *x*-axis represents the phase difference in radians and *y*-axis represents the frequency. The plots shows a clear peak for rainfall IMF5 with Niño 3.4 IMF5 (top right) which shows that there is phase locking. The last graph on the bottom right is a plot of rainfall's IMF6 and QBO IMF6 which shows that there is phase locking since there is a clear peak.

on the histogram of rainfall's IMF6 and QBO (**Figure 5**) shows that there is phase locking. This QBO signal identified is in agreement with a study by Kane (2009) which also identified the presence of QBO and it was shown that it contributed significantly to the variability of the winter rainfall for the same region.

Temperature for the selected region was also decomposed into 7 IMFs and a residual. There is weak coupling between the original rainfall time series and temperature as shown in **Figure 6** below. However, phase locking is observed for IMF 1 and IMF5, since there is a clear peak for both of them. IMF1 captures the noise found in the signal and IMF5 is a 4.5-year oscillation. The phase-locking in IMF1 suggest that the temperature variability may have an impact on the rainfall patterns. The synchronisation on IMF5 shows that the temperature changes may have a long term impact on the rainfall variability. This is consistent with other studies that have used different climatic models to find the impact of increasing temperature (Maúre *et al.*, 2018; Nangombe *et al.*, 2018; Nikulin *et al.*, 2018). These models predict that there will be an increase in extreme rainfall patterns. In our study, we have managed to show the direct impact of increasing temperature using historical data.

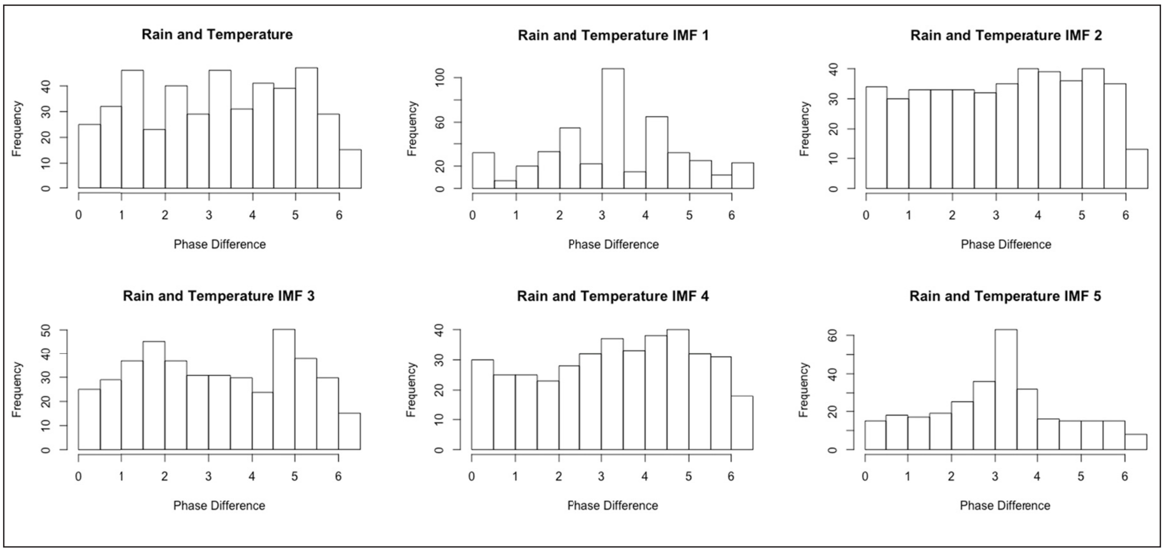


Figure 6: Histogram plot of the synchronisation of rainfall IMFs and temperature IMFs. The *x*-axis represents the phase difference in radians and *y*-axis represents the frequency. The plots shows a clear peak for Rainfall and Temperature for IMF1 (second graph on top), and IMF5 (bottom right). This shows that there is phase locking for these IMFs. IMF1 captures the noise in the data and IMF5 captures the 4.5 year-oscillation in the data.

4. Conclusions

The effectiveness of EEMD to analyse nonlinear and non-stationary data was demonstrated. The rainfall and temperature data were decomposed into IMFs and residual data, which summed up to the original data. The decomposed IMFs found can be used with other methods such as regression and neural networks to predict the impact of climate drivers in the future. EEMD was effective in isolating the data into different timescales and therefore the variability of the rainfall pattern was identified, in the end, evidence of the effect of ENSO and QBO was provided. Cross-correlation and phase synchronisation was used to find the relationship of the IMFs from the different time series under study. It will be of interest for future studies to carry out a study for a longer period to find the pattern of the rainfall over decades.

Additional Files

The additional files for this article can be found as follows:

- **Niño 3.4 Index.** A measurement of the strength of El-Niño and La-Niña. DOI: https://doi.org/10.5334/dsj-2019-046.s1
- **QBO Data.** Measurement of the variation of the wind above the equator. DOI: https://doi.org/10.5334/dsj-2019-046.s2
- **Rainfall Data.** Original rainfall data for the study area and the decomposed rainfall into IMFs. DOI: https://doi.org/10.5334/dsj-2019-046.s3
- **Temperature Data.** Original temperature data for the study area and the decomposed temperature into IMFs. DOI: https://doi.org/10.5334/dsj-2019-046.s4

Acknowledegement

The author would like to acknowledge South African Weather Service for the data that was used in this study and National Research Fund (Grant number 112979) for the funding of the work.

Author Contributions

Willard Zvarevashe is the main author of the work. Venkataraman Sivakumar edited and contributed on the atmospheric part of the study. Syamala Krishnannair provided final approval of the version to be published.

Competing Interests

The authors have no competing interests to declare.

References

Alizadeh, F, Roushangar, K and **Adamowski, J.** 2019. Investigating monthly precipitation variability using a multiscale approach based on ensemble empirical mode decomposition. *Paddy and Water Environment*, 1–19. DOI: https://doi.org/10.1007/s10333-019-00754-x

Cazelles, B and **Stone, L.** 2003. Detection of imperfect population synchrony in an uncertain world. *Journal of Animal Ecology*, 72(6), 953–968. DOI: https://doi.org/10.1046/j.1365-2656.2003.00763.x

Chiew, FHS, Peel, MC, Amirthanathan, GE and **Pegram, GGS.** 2005. Identification of oscillations in historical global streamflow data using empirical mode decomposition. *IAHS PUBLICATION*, 296: 53–64.

DEA 2011. South Africa's Second National Communication under the United Nations Framework Convention on Climate Change.

Dieppois, B, Pohl, B, Rouault, M, New, M, Lawler, D and **Keenlyside, N.** 2016. Interannual to interdecadal variability of winter and summer southern African rainfall, and their teleconnections. *Journal of Geophysical Research: Atmospheres*, 121(11): 6215–6239. DOI: https://doi.org/10.1002/2015JD024576

Fauchereau, N, Pohl, B, Reason, CJC, Rouault, M and **Richard, Y.** 2009. Recurrent daily OLR patterns in the Southern Africa/Southwest Indian Ocean region, implications for South African rainfall and teleconnections. *Climate Dynamics*, 32(4): 575–591. DOI: https://doi.org/10.1007/s00382-008-0426-2

Hawinkel, P, Swinnen, E, Lhermitte, S, Verbist, B, Van Orshoven, J and **Muys, B.** 2015. A time series processing tool to extract climate-driven interannual vegetation dynamics using ensemble empirical mode decomposition (EEMD). *Remote Sensing of Environment*, 169: 375–389. DOI: https://doi.org/10.1016/j.rse.2015.08.024

Hawinkel, P, Thiery, W, Lhermitte, S, Swinnen, E, Verbist, B, Van Orshoven, J and **Muys, B.** 2016. Vegetation response to precipitation variability in East Africa controlled by biogeographical factors. *Journal of Geophysical Research: Biogeosciences*, 121(9): 2422–2444. DOI: https://doi.org/10.1002/2016JG003436

Huang, NE, Shen, Z and **Long, SR.** 1999. A new review of non-linear water waves: the Hilbert spectrum. *Annual Review of Fluid Mechanics*, 31: 417–457. DOI: https://doi.org/10.1146/annurev.fluid.31.1.417

Huang, NEZ, Shen, R, Long, S, Wu, MC, Shih, HH, Zheng, Q, Yen, NC, Tung, V and **Liu, HH.** 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceeding of the Royal Society of London, Series A.* DOI: https://doi.org/10.1098/rspa.1998.0193

Kane, R. 2009. Periodicities, ENSO effects and trends of some South African rainfall series: an update. *South African Journal of Science*, 105(5–6): 199–207. DOI: https://doi.org/10.4102/sajs.v105i5/6.90

Kim, T, Shin, JY, Kim, S and **Heo, JH.** 2018. Identification of relationships between climate indices and long-term precipitation in South Korea using ensemble empirical mode decomposition. *Journal of Hydrology*, 557: 726–739. DOI: https://doi.org/10.1016/j.jhydrol.2017.12.069

Kruger, AC and **Nxumalo, MP.** 2017. Historical rainfall trends in South Africa: 1921–2015. *Water SA*, 43(2): 285–297. DOI: https://doi.org/10.4314/wsa.v43i2.12

Kruger, AC and **Shongwe, S.** 2004. Temperature trends in South Africa: 1960–2003. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 24(15): 1929–1945. DOI: https://doi.org/10.1002/joc.1096

Lehodey, P, Bertignac, M, Hampton, J, Lewis, A and **Picaut, J.** 1997. El Niño southern oscillation and tuna in the western pacific. *Nature*, 389(6652): 715–718. DOI: https://doi.org/10.1038/39575

Liu, H, Zhan, Q, Yang, C and **Wang, J.** 2019. The multi-timescale temporal patterns and dynamics of land surface temperature using Ensemble Empirical Mode Decomposition. *Science of the Total Environment*, 652: 243–255. DOI: https://doi.org/10.1016/j.scitotenv.2018.10.252

MacKellar, N, New, M and **Jack, C.** 2014. Observed and modelled trends in rainfall and temperature for South Africa: 1960–2010. *South African Journal of Science,* 110(7–8): 1–13. DOI: https://doi.org/10.1590/sajs.2014/20130353

Maúre, G, Pinto, I, Ndebele-Murisa, M, Muthige, M, Lennard, C, Nikulin, G, Meque, A, et al. 2018. The southern African climate under 1.5 C and 2 C of global warming as simulated by CORDEX regional climate models. *Environmental Research Letters*, 13(6): 065002. DOI: https://doi.org/10.1088/1748-9326/aab190

Maxmen, A. 2018. As Cape Town water crisis deepens, scientists prepare for 'Day Zero'. *Nature*, 554(7690). DOI: https://doi.org/10.1038/d41586-018-01134-x

Molla, MKI, Rahman, MS, Sumi, A and **Banik, P.** 2006. Empirical mode decomposition analysis of climate changes with special reference to rainfall data. *Discrete dynamics in Nature and Society, 2006.* DOI: https://doi.org/10.1155/DDNS/2006/45348

Nangombe, S, Zhou, T, Zhang, W, Wu, B, Hu, S, Zou, L and **Li, D.** 2018. Record-breaking climate extremes in Africa under stabilized 1.5 C and 2 C global warming scenarios. *Nature Climate Change*, 8(5): 375. DOI: https://doi.org/10.1038/s41558-018-0145-6

Nikulin, G, Lennard, C, Dosio, A, Kjellström, E, Chen, Y, Hänsler, A, van Meijgaard, E, et al. 2018. The effects of 1.5 and 2 degrees of global warming on Africa in the CORDEX ensemble. *Environmental Research Letters*, 13(6): 065003. DOI: https://doi.org/10.1088/1748-9326/aab1b1

Philippon, N, Rouault, M, Richard, Y and **Favre, A.** 2012. The influence of ENSO on winter rainfall in South Africa. *International Journal of Climatology*, 32(15): 2333–2347. DOI: https://doi.org/10.1002/joc.3403

Rosenblum, M, Pikovsky, A, Kurths, J, Schäfer, C and **Tass, PA.** 2001. Phase synchronization: from theory to data analysis. In: *Handbook of biological physics*, 4: 279–321. North-Holland. DOI: https://doi.org/10.1016/S1383-8121(01)80012-9

Schulte, JA. 2016. Wavelet analysis for non-stationary, nonlinear time series. *Nonlinear Processes in Geophysics*, 23(4): 257. DOI: https://doi.org/10.5194/npg-23-257-2016

Sun, C and **Ma, Y.** 2015. Effects of non-linear temperature and precipitation trends on Loess Plateau droughts. *Quaternary International*, 372: 175–179. DOI: https://doi.org/10.1016/j.quaint.2015.01.051

Van Buuren, S and **Groothuis-Oudshoorn, K.** 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3): 1–67. DOI: https://doi.org/10.18637/jss.v045.i03

van Oldenborgh, GJ and **Burgers, G.** 2005. Searching for decadal variations in ENSO precipitation teleconnections. *Geophys. Res. Lett.*, 32(15): L15701. DOI: https://doi.org/10.1029/2005GL023110

Wolski, P. 2018. How severe is Cape Town's "Day Zero" drought? *Significance*, 15(2): 24–27. DOI: https://doi.org/10.1111/j.1740-9713.2018.01127.x

Wu, Z and **Huang, NE.** 2009. Ensemble Empirical Mode Decomposition: A noise-assisted data analysis method. *Advanced Adaptive Data Analysis*, 1: 1–41. DOI: https://doi.org/10.1142/S1793536909000047

Yang, Q, Ma, Z and **Xu, B.** 2017. Modulation of monthly precipitation patterns over East China by the Pacific Decadal Oscillation. *Climatic change*, 144(3): 405–417. DOI: https://doi.org/10.1007/s10584-016-1662-9

How to cite this article: Zvarevashe, W, Krishnannair, S and Sivakumar, V. 2019. Analysis of Rainfall and Temperature Data Using Ensemble Empirical Mode Decomposition. *Data Science Journal,* 18: 46, pp. 1–9. DOI: https://doi.org/10.5334/dsj-2019-046

Submitted: 12 January 2019 **Accepted:** 12 September 2019 **Published:** 25 September 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS

Use of Satellite-Based NDVI Time Series to Detect Sugarcane Response to Climate Variability

Willard Zvarevashe[1*], Syamala Krishnannair[1] and Venkataraman Sivakumar[2]

[1]Department of Mathematical Sciences, University of Zululand, Private Bag X1001, KwaDlangezwa, 3886, South Africa
[2]Department of Chemistry and Physics, University of Kwazulu-Natal, Private Bag X54001, Durban, South Africa

*Corresponding author: wzvarevashe@gmail.com, *Tel: +27 633961263*

The Normalised Difference Vegetative Index (NDVI) is positively correlated to the yield in many crops. In recent times, it has been used to gauge the expected yield for crops. In this study, we have investigated the climate variability of sugarcane in a selected region in Felixton, Kwazulu-Natal province. The NDVI time series is decomposed to identify the seasonal, trend and random series. In order to find the periods and the impact of precipitation and temperature on the NDVI, a data adaptive method, ensemble empirical mode decomposition and synchronisation is applied. The results show that the seasonal rainfall and temperature variability has an influence on the NDVI. However, the rainfall variability has more direct impact. A non-linear regression model to find the effect of variability of precipitation, temperature, vapour pressure, cloud cover and soil moisture on the sugarcane growth is proposed.

Keywords: Precipitation, Temperature, Climate Change, Statistical Analysis, Sugarcane, Felixton

Introduction

The agriculture sector has not been spared from precipitation variability and warming climate. Driven by the growing concern about climate change impact on agricultural sector, there has been a growing interest on the subject. From 1990s onwards, there are a number of studies that have been carried out which mainly focused on the impact of climate variables such as air temperature, precipitation and CO_2 on the sugarcane yield (Jones & Singels, 2014; Linnenluecke et al., 2015; Miguez et al., 2018). Many of the studies used experimental analyses, agro-ecological models and process-based dynamic crop growth models and a few used statistical analysis (Kurma & Sharma, 2014; Jones & Singels, 2014; Zhao & Li, 2015; Linnenluecke et al., 2015). Therefore, there is need to carry out more statistical based analysis. Furthermore, a number of the statistical methods use yield or production as a basis of analysis of the impact of climate variability and do not look at the growth phase of the sugarcane.

The use of vegetation index to monitor crop progression and yield has increased over the years since it has been found to be positively correlated to yield for crops such as maize, soybean and sugarcane (Tarnavsky et al., 2008; Peralta, 2016; Sanches et al., 2018). One of the most commonly used index is Normalised Difference Vegetative Index (NDVI). Using this index, the stress on the leaves can be diagnosed through spectral responses. The NDVI can show nitrogen status and chlorophyll at micro level. Therefore, any limiting factors in the environment can be identified through the physiological stress in the leaves because of the variation in the concentration of nutrients (Lisboa et al., 2018).

The warming climate which has resulted in the increase of CO_2 has been found to have a positive correlation on the sugarcane by several studies (Jones & Singels, 2014; Zhao & Li, 2015; Linnenluecke et al., 2015). However, the variability of precipitation has had a negative impact on the growth of sugarcane (Kumar & Shamar, 2014). Sunlight, vapour pressure and soil moisture have all been found to have a significant influence on the plant growth (Zhao & Li, 2015; Miguez et al., 2018).

In this study, using the NDVI, the impact of climate variability on sugarcane of a selected region in Felixton, northern KwaZulu-Natal is shown. A data adaptive method, Ensemble Empirical Mode Decomposition and synchronisation is used to identify the variability at different time scales. A multiple non-linear regression model is proposed to find and predict the influence of air temperature, precipitation, vapour pressure, soil moisture and sunlight on NDVI.

Data and Method of Analysis

Study Area

A sugarcane production region located between -29ºS to -28.2ºS and 31.2ºE and 32.2ºE in Felixton, KwaZulu-Natal province was selected for the study (see Fig. 1). The Felixton area contributes about 10% of sugarcane yield in South Africa and has one of the largest mills in the country. In 2017/18 season about 1,7 million tons were crushed at the mill (SASA, 2019).

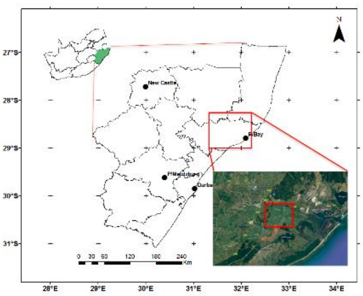


Figure 1: The location of study area in KwaZulu-Natal Province.

Data

In this study, monthly observations of precipitation, temperature, NDVI, vapour pressure, cloud cover and soil moisture data for the period from 2000 to 2018 were used. For precipitation the Global Precipitation Climatology Project (GPCC) is used (Schneider et al., 2018). The GPCC, vapour pressure, cloud cover and soil moisture data are all freely available online on the Royal Netherlands Meteorological Institute (KNMI) website (https://climexp.knmi.nl) (Huang et al., 2014). The Global Historical Climatology Network version 2 and the Climate Anomaly Monitoring System (GHCN + CAMS) were used for temperature. The NDVI data set MOD13A3 version 6 was extracted from Appears Application (Didan, 2015). MOD13A3 version 6 provides Vegetation Index (VI) values at a per pixel basis at 500 meter spatial resolution and is taken every 16 days. The data set was averaged to find the monthly data.

Ensemble Empirical Mode Decomposition
Ensemble Empirical Mode Decomposition (EMD) is a data adaptive time-frequency representation method, which does not have a lot of underlying assumptions and it only requires that the data must consist of simple intrinsic oscillations (Huang et al., 1998). The NDVI is decomposed into intrinsic mode functions (IMFs) and a residual. The IMFs represents the original data at different time scale and the residual shows the general trend of the NDVI.

In order to investigate the variability of the NDVI at different time scales, the index was decomposed into different time scales using ensemble empirical mode decomposition. Five IMFs were derived and these were synchronised to temperature and precipitation and to identify the relationship between the time series.

Synchronisation of coupled oscillating systems means appearance of certain relations between their phases and frequencies (Rosenblum et al., 2001). Here, we use this concept in order to reveal the interaction between NDVI and the climatic variables precipitation and temperature. R package '*synchrony*' is used which measures phase synchrony between quasi-periodic times series (Cazelles & Stone, 2003).

Multiple Non-Linear Regression
Least squares method is used to find the linear and non-linear relationship between NDVI as dependent variable and independent variables; precipitation, temperature, vapour pressure, cloud cover and soil moisture. The general model is given by,

$$y = \beta_0 + \sum_{i=0}^{n} \beta_i f_i(x, t) \qquad (1)$$

Where y is the dependent variable NDVI, is error term, is a coefficient and is the independent climatic variables (Miguez et al., 2018).

Results and Discussion
Statistical Characteristics of the NDVI
Using the R package '*Forecast*' a simple additive statistical model was derived from the NDVI time series which is given by:

$$y_t = m_t + s_t + e_t \qquad (2)$$

where t=time, y_t is the observed time series, m_t is the trend series, s_t is the seasonal time series and e_t is the error term. The results of the decomposition using R-package '*Forecast*' is plotted in Fig. 2 below. The trend plot shows a dip in the NDVI around 2004 and 2015. The published results from Department of Agriculture, Fisheries and Forestry (DFF) shows that the yield for 2015/2016 dropped to 14,861 thousand tons from 17,756 thousand tons in 2014/2015 season (DAFF, 2016). This was mainly attributed to the drought that was experienced in South Africa (SASA, 2019).

Multi-scale Variability
Time series that are phase synchronised or locked exhibit a modal distribution with a prominent peak at a given phase difference, whereas unrelated times series are characterized by a uniform or diffuse distribution.

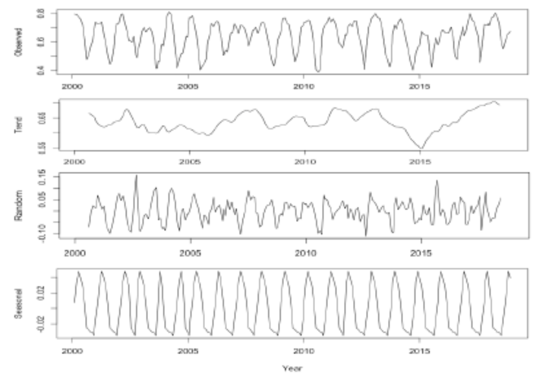


Figure 2. Plot of the decomposition of the NDVI time series into trend, random and seasonal series.

The results show that there is phase locking for IMF2 for precipitation since there is a clear peak (Fig. 3). IMF2 has a period of about 12 months which represents the annual period. However, there is weak coupling for all other IMFs. This shows that seasonal precipitation variability has a direct impact on the NDVI. This agrees with studies by Kumar & Shamar (2014) which showed the direct impact of rainfall variability. Fig. 4 reveals that temperature has an impact on the seasonal variability of NDVI, since a peak is also identified on the IMF2.

Impact of Environmental Variability
Multi non-linear regression was used to find the model of the environmental impact on the sugarcane growth. The percentage cloud cover was used to represent the sunlight. The model found is expressed as

$$y = 0.0826450 + 0.1790437x_1 - 0.0718855x_1^2 + 0.0038270x_2^2 + 0.0225784x_3 - 0.0063429x_4 + 1.4087915x_5$$

where x_1 is the precipitation, x_2 is temperature, x_3 is vapour pressure, x_4 is percentage cloud cover and x_5 is soil moisture. The coefficients for vapour pressure, cloud cover and soil moisture were all found to be significant. The Pearson Coefficient of correlation for the model is . In order to evaluate the model a series of histogram are plotted and shown in Fig. 5.

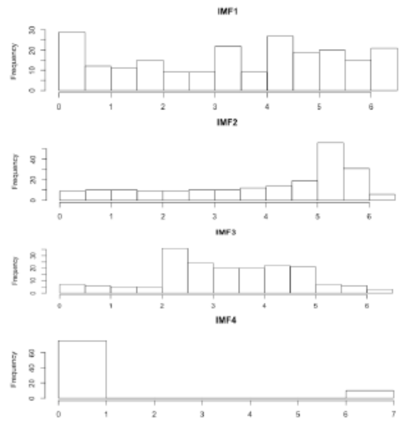


Figure 3. Histogram plot of synchronisation of NDVI and precipitation. The graphs show that there is clear peak for only IMF2 (second graph) which shows that there is phase locking.

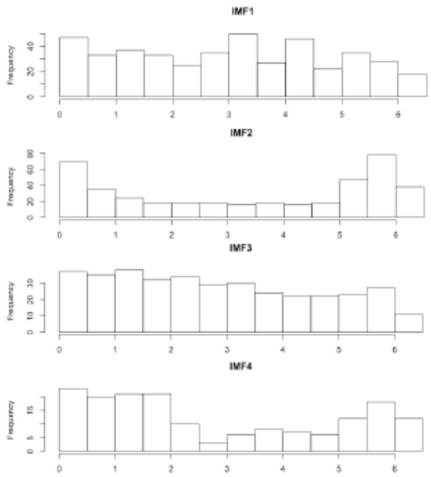


Figure 4. Histogram plot of synchronisation of NDVI and temperature. The graphs show a clear peak for only IMF2 (second graph) which shows that there is phase locking.

Residuals are values found from the difference between observed and expected values from the model. The points on the Residuals Vs Fitted plot are spread out and this shows that as a good model. The Normal Q-Q plot shows that the residuals are normally distributed, which is one of the assumptions that has to be met for a good model. The Scale-location plot is showing that the residuals are more spread out.

The Residuals vs Leverage plot helps to identify influential data points in the model. Points that are outside the dashed red Cook's distance line are influential on the model. The plot does not show any influential cases. Therefore, the model found to predict the plant growth using the cited external environmental factors.

The model provides further proof of the impact of climate variability as suggested by Zhao & Li (2015) which highlighted the direct influence of the environmental factors modelled here.

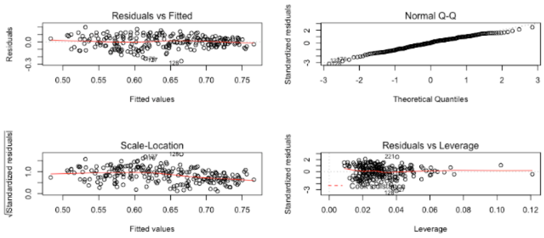


Figure 5. Residuals plots to evaluate the model. The plots are Residual vs Fitted (top left), QQ plot (top right), Scale-Location (bottom left) and Residuals vs Leverage (bottom right).

Conclusion

The NDVI was used to show the variability of the sugarcane at different time scales. This was done through decomposing the NDVI into different time scales and the resulted decomposed time series synchronised with precipitation and temperature. The variability of rainfall has more impact on the NDVI than temperature. A model which encompasses precipitation, temperature, vapour pressure, cloud cover and soil moisture was proposed.

Reference

Cazelles, B., & Stone, L. (2003). Detection of imperfect population synchrony in an uncertain world. *Journal of Animal Ecology*, 72(6), 953-968.

DAFF (Department of Agriculture, Forestry and Fisheries). (2016). A Profile Of The South African Sugar Market Value Chain. Pretoria. [Accessed online on 27 May 2019]. www.nda.agric.za.

Didan, K. (2015). *MOD13A1 MODIS/Terra Vegetation Indices 16-Day L3 Global 500m SIN Grid V006* [NDVI]. NASA EOSDIS Land Processes DAAC. doi: 10.5067/MODIS/MOD13A1.006

Huang, B., V.F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T.C. Peterson, T.M. Smith, P.W. Thorne, S.D. Woodruff, and H.-M. Zhang. (2014): Extended Reconstructed Sea Surface Temperature version 4. NINO 3.4. doi:10.1175/JCLI-D-14-00006.1.

Huang, N. E., Z.Shen, S. R.Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, V. Tung, and H. H. Liu. (1998). 'The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,'. *Proceeding of the Royal Society of London*, Series A.

Jones, M. R., & Singels, A. (2014). A preliminary assessment of mid-century climate change impacts on sugarcane production in South Africa. In *Proceedings of the Annual Congress-South African Sugar Technologists' Association* (No. 87, pp. 290-297). South African Sugar Technologists' Association.

Kumar, A., & Sharma, P. (2014). Climate change and sugarcane productivity in India: An econometric analysis. *Journal of Social and Development Sciences*, *5*(2), 111.

Linnenluecke, M. K., Nucifora, N., & Thompson, N. (2018). Implications of climate change for the sugarcane industry. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(1), e498.

Lisboa, I. P., Cherubin, M. R., Lima, R. P., Cerri, C. C., Satiro, L. S., Wienhold, B. J., ... & Cerri, C. E. (2018). Sugarcane straw removal effects on plant growth and stalk yield. *Industrial crops and products*, *111*, 794-806.

Miguez, F., Archontoulis, S., & Dokoohaki, H. (2018). Nonlinear regression models and applications. *Applied statistics in agricultural, biological, and environmental sciences*, (appliedstatistics), 401-448.

Peralta, N., Assefa, Y., Du, J., Barden, C., & Ciampitti, I. (2016). Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. *Remote Sensing*, *8*(10), 848.

Rosenblum, M., Pikovsky, A., Kurths, J., Schäfer, C., & Tass, P. A. (2001). Phase synchronization: from theory to data analysis. In *Handbook of biological physics* 4, 279-321. North-Holland.

Sanches, G. M., Duft, D. G., Kölln, O. T., Luciano, A. C. D. S., De Castro, S. G. Q., Okuno, F. M., & Franco, H. C. J. (2018). The potential for RGB images obtained using unmanned aerial vehicle to assess and predict yield in sugarcane fields. *International journal of remote sensing*, *39*(15-16), 5402-5414.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., (2018): GPCC Full Data Monthly Product Version 2018 at 0.25°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historical Data. DOI: 10.5676/DWD_GPCC/FD_M_V2018_025.

SASA (South African Sugar Association). (2019). *Facts and figures.* [Accessed online on 29 May 2019]. www.sasa.org.za.

Tarnavsky, E., Garrigues, S., & Brown, M. E. (2008). Multiscale geostatistical analysis of AVHRR, SPOT-VGT, and MODIS global NDVI products. *Remote Sensing of Environment*, *112*(2), 535-549.

Zhao, D., & Li, Y. R. (2015). Climate change and sugarcane production: potential impact and mitigation strategies. *International Journal of Agronomy*, *2015*.

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 51-5 (2018) 132–137

Analysis of Austral Summer and Winter Rainfall Variability in South Africa Using Ensemble Empirical Mode Decomposition

Willard Zvarevashe[1]*, Syamala Krishnannair[1] **, Venkataraman Sivakumar ***

[1]Department of Mathematical Sciences, University of Zululand, Private Bag X1001, KwaDlangezwa, 3886, South Africa
**(Tel: +27 633 961 263; email: wzvarevashe@gmail.com)*
*(**Email: KrishnannairS@unizulu.ac.za)*
****Department of Chemistry and Physics, University of KwaZulu-Natal, Private Bag X54001, Durban, South Africa (Email:venkataramans@ukzn.ac.za)*

Abstract: Rainfall exhibits nonlinear and non-stationary characteristic such that analysis done using linear methods are inconclusive. It is, therefore, necessary to understand the multi-scale variations of the rainfall. In this study, a comparison is done of two different regions which receive two distinct rainfall types. Northeastern to southwestern South Africa receives austral winter rainfall whilst the rest of the country receives austral summer rainfall. The multi-scale rainfall variability is done using data adaptive method, ensemble empirical mode decomposition. Changes of rainfall in both regions exhibited inter-annual and quasi-decadal oscillation. We found that there were distinctive spatial differences in the patterns.

Keywords: Climate Change, Time Series Modelling, Statistical Data Analysis, Ensemble Empirical Mode Decomposition, Rainfall

1. INTRODUCTION

It is reported that climate is warming, with reported temperature of 0.85 ^{0}C in the past 130 years (Hartmann et al., 2013). According to studies, temperature in South Africa has increased by approximately one and half times more than the global temperature increases. This has led to high fluctuations of rainfall in South Africa (Kruger, 2006; Midgley, 2010). This climate change will have a greater impact on the South African economy as a large proportion of the crops are rain-fed (Niang et al., 2014).

The north-eastern to southwestern regions in South Africa experiences austral winter rainfall and the rest of the country experiences austral summer rainfall (Dieppois et al., 2016). The austral summer starts in October to March and austral winter from May to August. Different regions records different wettest months between October and March (Schulze, 1997).

The inter-annual variability (2-8 years) has been shown to be influenced by El Niño-Southern Oscillation (ENSO) and causes fluctuations of rainfall from wet and dry seasons (Kane, 2009; Fauchereau et al., 2009). Therefore, this adds to the complexity of the climate system. It is generally known that most of the long-term variations in climatic factors such as rainfall and temperature exhibit nonlinear and non-stationary characteristic (Frazke et al., 2014; Lee et al., 2010). Linear models become limited in helping to understand the systems. It will therefore require the climatic system to be broken down into components which can be analysed separately. The conventional methods such as Wavelet Analysis and Fourier transform uses basis functions which may be able to explain some segments but fails on other segments of the non-stationary time series (Guo et al., 2016). There is still need to carry out more studies on an effective method which can reveal the intrinsic changes in the climate system.

Ensemble empirical decomposition (EEMD) is a data adaptive method that is best suited for nonlinear and non-stationary time series and can reveal some of the intrinsic changes in the climate system (Wu et al., 2009). The method decomposes the time series into different time scales called intrinsic mode functions (IMFs). EEMD does not use *a priori* determined functions. It can eliminate the mode mixing problem of the empirical mode decomposition (EMD).

The main objectives of this study are to (1) provide better understanding of the oscillations embedded in the rainfall time series in austral summer and winter regions and therefore help improve seamless predictions at different time scales (2) to explore the contributions of oscillations on different time scales to overall rainfall variability in the two regions. In section 2, the data set and methodology are described and results and discussion in done in section 3.

2. DATA AND METHODS

2.1 Data Description

Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.06.223

It was observed that there are a lot of fluctuations in the South African rainfall which is characterised by high inter-annual variability of up to 200% of the mean for both summer and winter rainfall (Kane, 2009; Malherbe et al., 2016). Therefore, to carry out a closer analysis of the rainfall fluctuations, two areas were selected, one which experiences winter rainfall (A) and the other, which experiences summer rainfall (B). The area selected which experiences winter rainfall is between 19-20ºE and 31-32ºS which is found in Northern Cape and Western Cape provinces. The second area selected is between 29-30ºE and 23-24ºS which is found in Free State Province and receives summer rainfall. Rainfall data for June, July and August (JJA) period for winter rainfall season was used and for summer rainfall season December, January and February (DJF) data was also used which are the wettest months for both winter and summer rainfall. The seasonality of rainfall in South Africa is shown in figure 1 below.

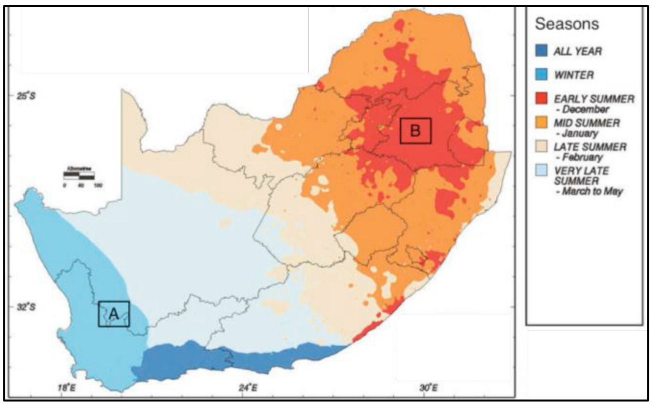


Figure 1. Rainfall Seasonality in South Africa (Source: Schulze, 1997)

In this study, the seasonal monthly observations of precipitation data for the period 1980 to 2016 was used. The Global Precipitation Climatology Project (GPCP) data was obtained from KNMI Climate explorer (https://climexp.knmi.nl). The GPCP data set that was established by the World Climate Research Programme combines precipitation data from Infrared and microwave satellite estimates measurements from rain gauge data from the Global Precipitation Climatology Centre's.

2.2 Ensemble Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is an adaptive time-frequency data analysis method, which requires only that the data must consist of simple intrinsic mode of oscillations (Huang et al., 1998). It is most suitable for nonlinear and non-stationary data. The EMD methodology is based on a sifting process, which identifies local extrema (maxima and minima) which results in the formation of intrinsic mode functions (IMFs). The first IMF will contain the highest fluctuations and this is subtracted from the original data and subsequent IMFs are then derived from the subtracted data. The IMFs and residual data approximate the original data when they are summed together (Huang et al., 1998). That is, for a given time series $x(t) = (x_1, x_2, \dots, x_N)^T \in \mathbb{R}$ of length N, which is decomposed into n IMFs and residual then

$$x(t) \approx \sum_{i=1}^{n} c_i(t) + r_n(t). \quad (1)$$

where $c_i(t)$ is the i^{th} IMF , n is the number of IMFs and $r_n(t)$ is the n^{th} residual.

However, the shortcoming of EMD is that of mode mixing. To overcome this a noise assisted method was introduced called Ensemble empirical mode decomposition (Wu et al., 2009). Noise is introduced to each data set in the time series $x(t)$. This noise is treated as random noise that can be encountered in the process. Therefore, the i^{th} "artificial" value becomes

$$y(t) = \varepsilon_i + x(t), for\ i = 1,2, \dots, m \quad (2)$$

where ε_i is the random noise and m is the ensemble size.

EMD process is carried out on the new time series $y(t)$. This process is repeated for different realization of noise until the pre-set ensemble size is reached. That is, for each value x_i different white noise is added. The new IMF is found by averaging the IMFs found from the different realization of noise (ensemble).

The accepted range of noise to be added is between 0.1-0.4 times the standard deviation of the original data (Wu et al., 2009). In this study, there was no noticeable changes when different ensembles were used from 100 going upwards. However, the proportion of the noise added had a small effect and this will be further studied in the future research. A noise amplitude of 0.2 of standard deviation was used and ensemble size of 100 for the EEMD.

The effects of each IMF fluctuations to overall variability of the rainfall is calculated using Variance contribution rate (VCR) which is given by

$$VCR_i = \frac{var(c_i(t))}{\sum_{i=1}^{n} var(c_i(t)) + var(r_n(t))} X100 \quad (3)$$

where $var(c_i(t))$ is variance of i^{th} IMF $c_i(t)$ and $var(r_n(t))$ is variance of residual (Guo et al., 2016).

3. RESULTS AND DISCUSSION

The average seasonal monthly rainfall was standardized for easy computation and comparison to find the rainfall anomalies. The precipitation time series plot for an area in Free State province which experiences summer rainfall is shown in figure 2 and another time series plot for an area in Northern Cape and Western Cape provinces which experiences winter rainfall is shown in figure 4. The 2 series are decomposed into IMFs and the graphs are shown in figure 3 and figure 5 respectively. When the IMFs were added together they managed to approximate the original data with an error of order 10^{-14}.

Each IMF represents fluctuations in different time scales from highest frequency to lowest frequency. The decomposition was done on 3 months' rain season data (DJF and JJA). A period of 3 months for summer and 3.1 months for winter was identified which represented the annual oscillation of the rainfall. This time scale (IMF1) contains the highest frequency and the last IMF called residual has the lowest frequency and normally represents the overall trend of the data. Generally, each IMF represents the inherent oscillations deduced from the original data (Huang et al., 1998).

Some of the oscillations not only reflect periodic variations of the climatic systems under external forcing but also non-linear feedback of the climatic system (Guo et al., 2016). An example is the IMF2 with a period of 5.7 (summer) and 6.4 months (winter) which is an oscillation of about 2 seasons and therefore resembles basic characteristics of quasi-biennial oscillation (QBO). IMF3 has a period of 13.5 months which corresponds with 4.5 seasons or years and this therefore exhibits a quasi-triennial oscillation (QTO).

Previous studies which pre-dominantly used Wavelet Analysis, detected similar cyclic periods in South African rainfall that were also found in this study. These were QBO, QTO, and 6-11 year oscillations (Kane, 2009; Dieppois et al., 2016). The QBO and QTO accounted for 30-50% of total variance which is consistent with what was found in this study. As compared to Wavelet Analysis EEMD is adaptive, intuitive and does not use basis functions. Additionally, this study went further to show the variance contribution to overall variability of the rainfall for summer and rainfall separately.

The effect of the oscillations at different time scales on the overall variability of the original signal can be shown by variance contribution rate. Table 1 shows the periods and their variance contribution rates to the overall variability of the rainfall anomalies in the winter and summer rainfall.

There are noticeable differences in the variance contribution rate for the IMFs for summer and winter rainfall. The variance contribution rate for IMF1 (annual oscillation) in the summer is 52.5% whilst in winter rainfall it is 62.12%, which is a difference of more than 10%. The QBO (IMF2) is contributing 20.6% in summer rainfall and 15.1% in winter rainfall. There are different periods for IMF4, with 5-6 year oscillation for summer rainfall and 7-8 year oscillation. However, the variance contribution rate is similar (4.3%) for both. IMF5 has an average period of 49.5 months which is similar to quasi-decadal oscillation and the contribution rates are slightly different.

The winter and summer rainfall for the selected areas have different variance contribution rate to the overall variability of the rainfall anomalies. The IMF1 contributed higher to the variability in the winter rainfall and there is also a noticeable difference for other IMFs. Generally, it was observed that there is a noticeable difference in the contribution rates for the IMFs except only IMF4.

4. CONCLUSSION

The JJA and DJF (winter and summer respectively) rainfall seasons' anomalies data was decomposed into different time scales using ensemble empirical mode decomposition. The first time scale (IMF1) managed to capture the annual oscillation and the second time scale (IMF2) identified the QBO in the rainfall anomalies. Various other time scales were also identified up to decadal oscillation. When these different time-scales (IMFs) were added together they managed to approximate the original data with very minimal error. The study showed that the variance contribution rates for the winter and summer rainfall seasons were different.

However, there is still need for a further research on the parameter selection on the noise to be added and its amplitude. Additionally, the climatic drivers of the oscillations identified needs to be studied.

Table 1: Variance Contribution rate for winter and summer rainfall

Rainfall		**IMF1**	**IMF2**	**IMF3**	**IMF4**	**IMF5**	**Res**
Summer	Period (months)*	3	5.7	13.5	21.1	49.3	
	Contribution Rate (%)	52.5	20.6	17.4	4.3	3	0.74
Winter	Period (months)	3.1	6.4	13.5	29.6	49.3	
	Contribution Rate (%)	62.1	15.1	13.3	4.3	1.2	3.5

*The decomposition was done on 3 months' rain season data, therefore 3 months equates to a season or year, 6 months equates to 2 seasons and so on.

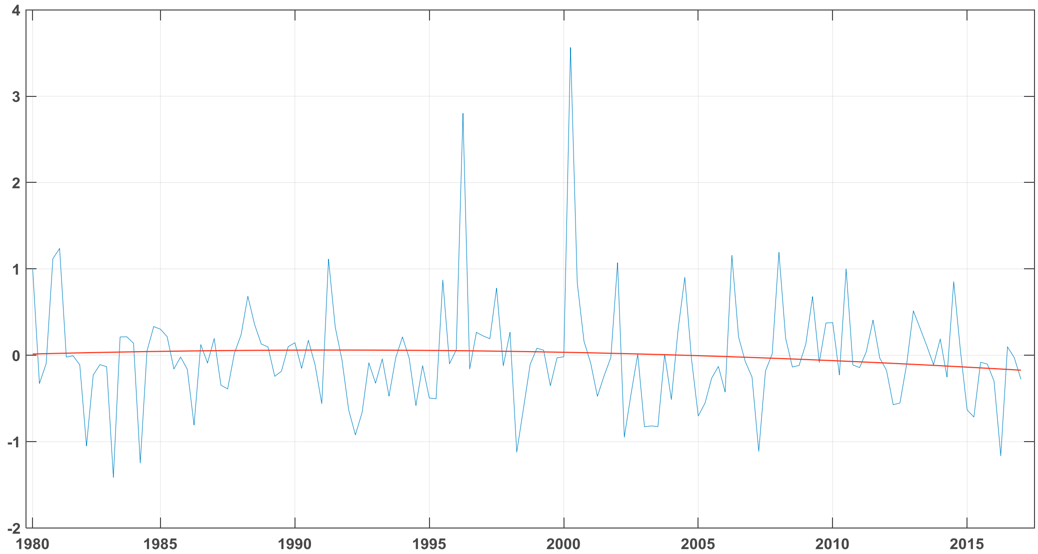


Figure 2. Seasonal GPCP v2.3 precipitation for the area between 29-30E and 23-24S which receives summer rainfall. A residual plot (red line) is done which is shows the general trend of the rainfall. This is showing slight reduction of rainfall over the years.

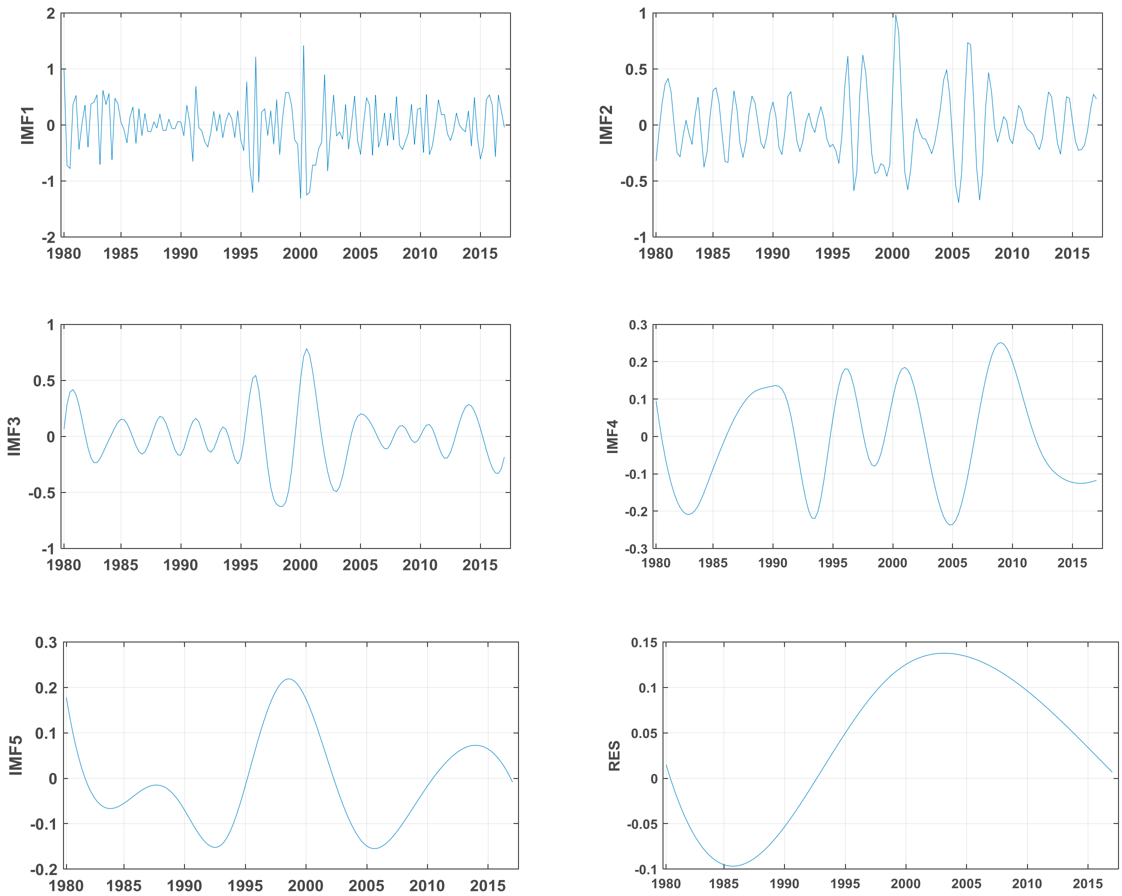


Figure 3. Decomposition of Seasonal GPCP v2.3 precipitation for the area between 29-30E and 23-24S (winter rainfall). IMFs into IMFs using EEMD.

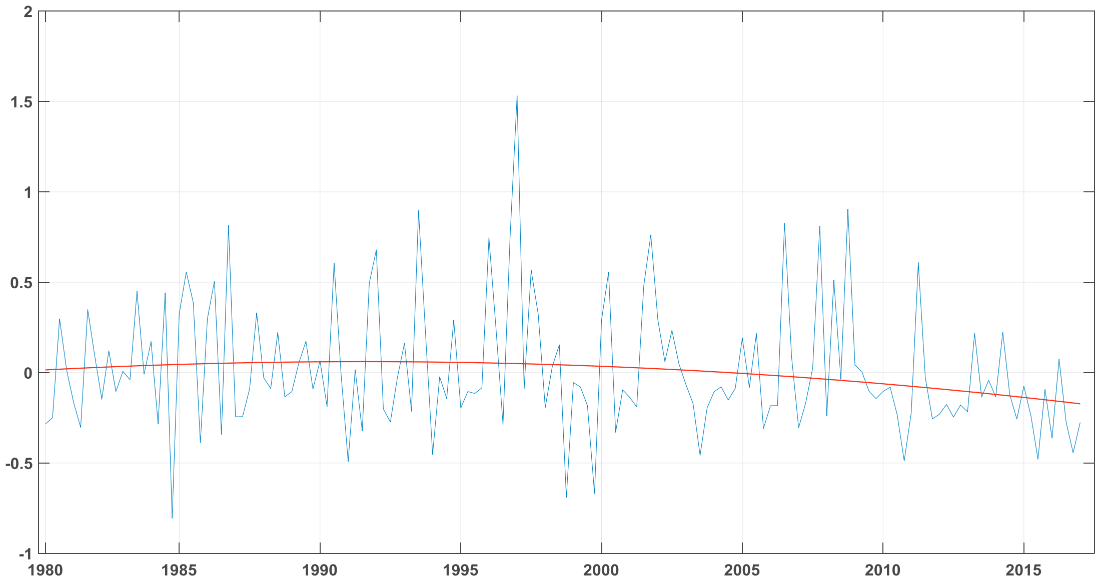


Figure 4. Seasonal GPCP v2.3 precipitations for the area between 19-20E and 31-32S which receives winter rainfall. A residual plot (red line) is done which is shows the general trend of the rainfall. This is showing slight reduction of rainfall over the years

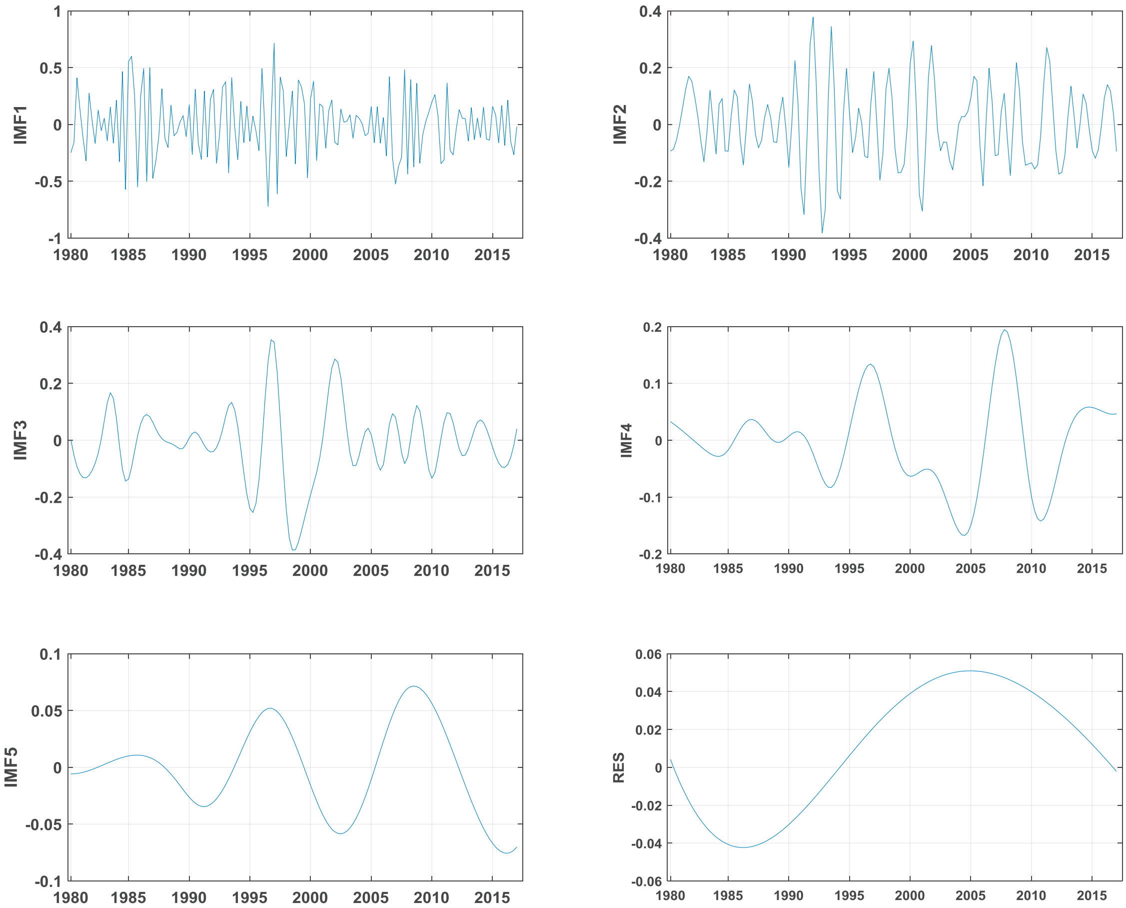


Figure 5. Decomposition of Seasonal GPCP v2.3 precipitations for the area between 19-20E and 31-32S IMFs into IMFs using EEMD for the area, which receives winter rainfall.

REFERENCES

Dieppois, B., Pohl, B., Rouault, M., New, M., Lawler, D., & Keenlyside, N. (2016). Interannual to interdecadal variability of winter and summer southern African rainfall, and their teleconnections. *Journal of Geophysical Research: Atmospheres*, *121*(11), 6215-6239.

Fauchereau, N., Pohl, B., Reason, C. J. C., Rouault, M., & Richard, Y. (2009). Recurrent daily OLR patterns in the Southern Africa/Southwest Indian Ocean region, implications for South African rainfall and teleconnections. *Climate Dynamics*, *32*(4), 575-591.

Franzke, C. (2012). Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of Climate*, *25*(12), 4172-4183.

Guo, B., Chen, Z., Guo, J., Liu, F., Chen, C., & Liu, K. (2016). Analysis of the Nonlinear Trends and Non-Stationary Oscillations of Regional Precipitation in Xinjiang, Northwestern China, Using Ensemble Empirical Mode Decomposition. *International Journal Of Environmental Research And Public Health*, 13(3).

Hartmann, D., Tank, A. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., & Zhai, P. M. (2013). Observations: Atmosphere and Surface. In T. [Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. . Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*: Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Huang, N. E., Z.Shen, R.Long, S., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, V., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proceeding of the Royal Society of London, Series A.*

Kane, R. (2009). Periodicities, ENSO effects and trends of some South African rainfall series: an update. *South African Journal of Science*, 105(5-6), 199–207.

Kruger, A. C. (2006). Observed trends in daily precipitation indices in South Africa: 1910–2004. *International Journal of Climatology*, 26(15), 2275–2285.

Lee, T., & Ouarda, T. B. M. J. (2011). Prediction of climate nonstationary oscillation processes with empirical mode decomposition. *Journal of Geophysical Research: Atmospheres*, *116*(D6).

Midgley, G., Van Wilgen, B., & Mantlana, B. (2010). South Africa's Second National Communication under the United Nations Framework Convention on Climate Change. *Department of Environmental Affairs Republic of South Africa*.

Niang, I., Ruppel, O., Abdrabo, M., Essel, A., Lennard, C., Padgham, J., & Urquhart, P. (2014). Africa. Climate Change 2014:. In V. R. Barros, C. Field, D. Dokke, M. Mastrandre, K. Mach, T. Bilir, M. Chatterjee, K. Ebi, Y. Estrada, R. Genova, B. Girma, E. K. A. Levy, S. MacCracken, P. Mastrandrea, & L. White (Eds.), *Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1199–1265).: Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Schulze, R. E. (1997). *South African Atlas of Agrohydrology and Climatology: Contribution Towards a Final Report to the Water Research Commission on Project 492: Modelling Impacts of the Agricultural Environment on Water Resources: TT82-96*. Water Research Commission (WRC).

Philippon, N., Rouault, M., Richard, Y., & Favre, A. (2012). The influence of ENSO on winter rainfall in South Africa. *International Journal of Climatology*, *32*(15), 2333-2347.

Wu, Z. & Huang, N. E. (2009). Ensemble Empirical Mode Decomposition: A noise-assisted data analysis method. *Advanced Adaptive Data Analysis*, 1, 1–41

Annual and Seasonal Spatial Analysis of Rainfall Variability in Western Cape, South Africa

Willard Zvarevashe[1*], Syamala Krishnannair[1] and Venkataraman Sivakumar[2]

[1]*Department of Mathematical Sciences, University of Zululand, Private Bag X1001, KwaDlangezwa, 3886, South Africa*

[2]*Department of Chemistry and Physics, University of Kwazulu-Natal, Private Bag X54001, Durban, South Africa*

Abstract

An understanding of the trend and the spatial shifts of the rainfall is very important for the development and management of water resources so that there is greater sustainability especially in the rainfed agriculture. Man-Kendell test was applied to 20 stations scattered in Western Cape. An analysis of spatial variability, monthly and seasonal variability for the period from 1980 to 2017 showed that there was a monotonic decrease in rainfall in most of the stations except for stations located further north in the province. There was also a variability of up to 127% in the annual and seasonal rainfall in some of the stations. These rainfall trend analyses may help to predict future rainfall patterns and thereby allowing planning and better water resources management.

Keywords: Rainfall, Trend Analysis, Spatial Analysis

1. *Introduction*

Change in the rainfall patterns have adverse effects on all the earth's spheres particularly the hydrosphere, biosphere and lithosphere. Therefore, study of the climate trends such as temperature and rainfall trends becomes very essential, which is the first step towards systematic study of climate change (Nema *et al.,* 2018). The whole of Southern Africa is prone to relatively high rainfall variability and there are numerous recorded floods and draughts occurrences (Usman and Reason, 2004). Most of agricultural production in Africa depends heavily on the rainfall (Kijazi and Reason, 2005). Western Cape in South Africa, has not been spared from this rainfall variability. Some of the regions in the province have been facing water challenges in the last decade because of general decrease of rainfall (Kruger and Nxumalo, 2017; Zvarevashe *et al.,* 2018). It becomes important to carry out an analysis of the rainfall trends for different rainfall periods and different regions to find out if the trend is similar throughout the province and for different seasons.

In view of the precarious position that Western Cape find itself in, this study seeks to investigate precipitation patterns in the province. Although there have been several studies that have investigated the variability of rainfall patterns in South Africa, very few concentrated on Western Cape. Furthermore, there are few spatial statistics based studies for South Africa in general. In this study, Mann-Kendell Test (MK) will be used, which is a non-parametric method that can be used to detect trends in a time series for any period (MacKellar *et al.,* 2014; Nema *et al.,* 2018). Additionally, coefficient of variation (CV%) is calculated and used to carry out spatial analysis using Inverse Distance Weighted (IDW) method.

2. *Data and Method*

2.1 *Study Area*

Monthly rainfall data was used to generate annual, summer covering the period December, January and February (DJF), autumn covering the period March, April and May (MAM), winter covering the period June, July and August (JJA) and spring covering the period September, October and November (SON). The data was obtained from South African Weather Service. It was collected from the weather stations in Western Cape for the period from 1980 to 2017. The study area and location of the weather stations are shown in figure 1.

2.2 *Mann-Kendell and Coefficient of Variation*

Trends in the rainfall data for the period from 1980 to 2017 were evaluated using Mann-Kendell trend Test (MK) and variability was calculated using Coefficient of Variation (CV%). MK has an advantage of not having any underlying assumptions about the distribution of the data. It can indicate the temporal patterns in yearly and season precipitation (Kendall, 1980). A correlation coefficient, τ, which denotes relative strength of the trend of the time series is computed. It takes on a value between -1 and 1. The probability of this trend occurring by chance is also estimated, from which a measure of statistical significance can be assigned.

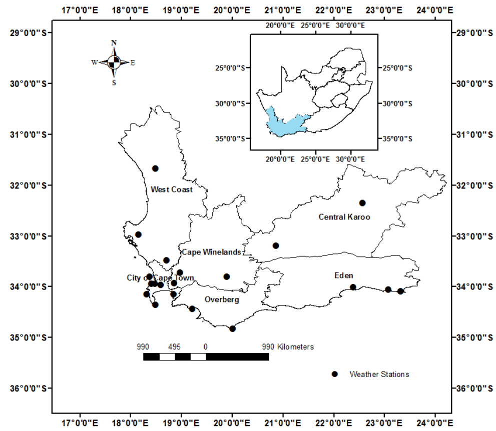


Figure 1: The location of Western Cape province in South Africa, its magisterial districts and the location of weather stations used in the study.

A 5% level of significance is used such that if the probability estimate was less than 0.05, the trend was deemed to be significant. The trend estimates were calculated for seasonal and totals for all the stations.

Given the time series $x_1, x_2, x_3, \dots, x_n$ then the MK statistic, *S*, is defined as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_j - x_i) \qquad (1)$$

Where x_j and x_i are the data values in years *j* and *i*, respectively, with $j > i$, *n* is the total number of years and sgn() is the sign function. Let $x_j - x_i = \theta$ then

$$\operatorname{sgn}(\theta) = \begin{cases} 1 & if\ \theta > 0 \\ 0 & if\ \theta = 0 \\ -1 & if\ \theta < 0 \end{cases} \qquad (2)$$

The statistics S is approximately normally distributed, with mean zero and variance given by:

$$\mathrm{Var}(S) = N(N-1)(2N+5)/18.$$

The standard normal variable *Z* is then formulated as

$$z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}\ if\ S > 0 \\ 0\ if\ \theta = 0 \\ \frac{S+1}{\sqrt{Var(S)}}\ if\ S < 0 \end{cases} \qquad (3)$$

The coefficient of variation (CV%) is then calculated. CV%, also known as "relative variability", equals the standard deviation divided by the mean. It measures how a value is varied around the mean (Label et al., 1987).

Using the CV% for each station the Inverse Distance Weighted method (IDW) is used to investigate spatially fleeting precipitation variability over the regions using ArcGIS (Singh *et al.,* 1987; Zuo *et al.,* 2016). IDW is a deterministic method for multivariate interpolation which uses exact points to calculate the nearest points using a weight. Therefore, using the rainfall data from the stations available an estimate of monthly rainfall for areas without stations can be done using IDW. (Moeletsi *et al.,* 2016).

3. Results and Discussion

3.1 Statistical Characteristics of Annual and Seasonal Rainfall

Table 1 shows the z-statistics of the seasonal and annual rainfall of selected stations in Western Cape province using MK. Negative z-statistic shows that there is a monotonic downward trend in the rainfall whereas a positive z-statistic shows that there is a monotonic upward trend. There is generally monotonic downward trend in most of the stations for different seasons. Paarl station has a significant downward trend for annual (*p* = 0.003), DJF (*p* =0.006), MAM (*p*=0.001) and SON (*p*=0.046). Stellenbosch station which is also located in the same district with Paarl station has almost similar trend with a significant downward trend in DJF (*p*=0.008) and MAM (*p*=0.004). It was also noted that significant upward trend only recorded in Strand (*p*=0.009) and Hermanus (*p*=0.007) stations for the period JJA and Cape Point (*p*=0.033 for the period SON. George WO and Knysna stations have significant monotonic downward trend for 2 consecutive seasons of DJF and MAM. This is expected as they are also located in the same district. Almost similar trend was observed at Plettenbergbaai station which is another station located in the same district. These results agreed with what was found by MacKellar et al. (2014) and Kruger et al. (2017).

The results from MK Test showed that there is generally monotonic decreasing rainfall trend. A further investigation of this decrease is then done to analyse the variability of the rainfall over different seasons.

Table 1: Z-statistic values of seasonal rainfall using MK for Stations in Western Cape Province

Station	**Annual**	**DJF**	**MAM**	**JJA**	**SON**
Beaufort-Wes	-1.0666	-0.372	-1.215	0.397	-0.223
Cape Agulhas	-0.955	-0.528	-2.489*	0.163	0.427
Cape Point	0.453	-0.653	-1.333	0.805	2.138*
Cape Town Slangkop	-0.113	-0.733	-0.733	0.508	0
Cape Town WO	-1.986*	-0.864	-1.191	-0.911	-0.864
George WO	-1.985	-1.939*	-1.985*	0.981	-0.490
Hermanus	0.27177	-0.513	-1.782*	2.689**	-0.393
Knysna	-1.238	-2.386*	-1.962*	-0.634	0.030
Laingsburg	-0.620	-1.213	-0.226	0.056	0.338
Langebaanweg Aws	-1.685	-0.805	-2.578**	-0.289	-0.226
Malmesbury	-0.521	-0.447	0.024	0.174	-0.620
Molteno Reservoir	-1.612	-1.662	-1.860	-0.620	-0.174
Paarl	-3**	-3.438**	-3.147**	-0.924	-1.995*
Plettenbergbaai	-1.3778	-1.985*	-0.537	0.210	-1.098
Robbeneiland	-0.537	0.093	-0.210	0.817	0.817
Robertson	0.42257	-0.950	-2.774**	1.321	0.898
S. A. Astronomical Observatory	-1.609	-1.911	-1.660*	-0.981	0.063
Stellenbosch	-1.785	-2.665**	-2.879**	0	-0.289
Strand	1.1279	-0.931	0.649	2.594**	-0.620
Vredendal	-1.195	-1.245	-1.597	-0.101	-1.685

*z-statistic is significant at 0.05 level and ** z-statistic is significant 0.01 level

3.2 Annual and Seasonal Analysis

CV% was calculated for the different stations and IDW was used to find the spatial patterns of the rainfall. The results are shown in figure 2 which shows the variability of rainfall for annual and seasonal rainfall. Generally, there is a variability of between 19% to 127%. Annual Rainfall is less variable in City of Cape Town, greater part of Overberg and part of West Coast and Cape Winelands. The wettest months JJA, are also less variable for the same areas as compared to the northern areas. A variability of between 25 to 33% is observed. Whilst the northern part of West Coast is more variable with a variability of up to 126% for DJF period. Figure 2 (c) shows that MAM is more variable with most of the province having a CV% of between 40% to 50% whilst for JJA most of the province has CV% of between 30% and 40%. Annual rainfall is least variable as compared to the seasons, with CV% of between 19% and 36%.

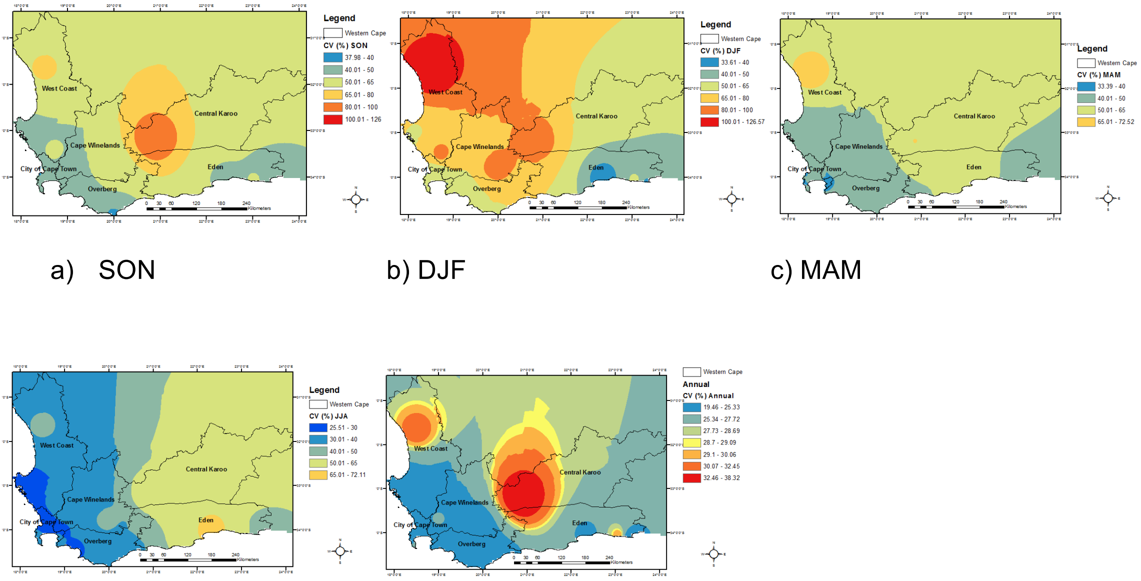
4. Conclusions

An analysis of the rainfall trend was carried out for annual, DJF, MAM, JJA and SON rainfall totals for the period 1980 to 2017 over Western Cape. Mann-Kendall test and inverse distance weighted method were used to explore these variability. The results showed that there is generally downward trend of rainfall in most part of Western Cape. Coefficient of variation for the whole province showed that there is variability of between 19% and 126%. The highest variability was recorded in the driest period of DJF and the lowest variability were recorded in the wettest months of JJA.

It was observed that although there is generally monotonic decreasing rainfall pattern (from table 1) over the study period the variability is mostly observed in the seasonality of the rainfall than in the annual rainfall.

Acknowledgement

The author would like to acknowledge South African Weather Service for the data that was used in this study and NRF for the funding of the work.



a) SON b) DJF c) MAM

d) JJA e) Annual

Figure 2: Annual and seasonal rainfall variability (CV%) in Western Cape a) annual b) DJF c) MAM d) JJA e) SON.

4. Reference

Kijazi A.L. and C.J.C. Reason. 2005, Relationships between intraseasonal rainfall variability of Coastal Tanzania and ENSO, *Theoretical and Applied Climatology*, 82: 153-176.

Kendall, M. (1980). Multivariate Analysis, (London, Charles Griffin).

Kruger, A. C., & Nxumalo, M. P. (2017). Historical rainfall trends in South Africa: 1921–2015. *Water SA*, *43*(2), 285-297.

Lebel, T., Bastin, G., Obled, C., & Creutin, J. D. (1987). On the accuracy of areal rainfall estimation: a case study. Water Resources Research, 23(11), 2123-2134.

MacKellar, N., New, M., & Jack, C. (2014). Observed and modelled trends in rainfall and temperature for South Africa: 1960-2010. South African Journal of Science, 110(7-8), 1-13.

Moeletsi, M. E., Shabalala, Z. P., De Nysschen, G., & Walker, S. (2016). Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over the Free State Province, South Africa. Water SA, 42(3), 466-474.

Nema, M. K., Khare, D., Adamowski, J., & Chandniha, S. K. (2018). Spatio-temporal analysis of rainfall trends in Chhattisgarh State, Central India over the last 115 years. *Journal of Water and Land Development*, *36*(1), 117-128.

Singh, V. P., & Chowdhury, P. K. (1986). Comparing Some Methods Of Estimating Mean Areal Rainfall 1. Jawra Journal of the American Water Resources Association, 22(2), 275-282.

Usman M.T. and C.J.C. Reason. 2004, Dry spell frequencies and their variability over southern Africa, *Climate Research*, 26: 199-211.

Zuo, R., Carranza, E. J. M., & Wang, J. (2016). Spatial analysis and visualization of exploration geochemical data. *Earth-Science Reviews*, 158, 9-18.

Zvarevashe, W., Krishnannair, S., & Sivakumar, V. (2018). Analysis of Austral Summer and Winter Rainfall Variability in South Africa Using Ensemble Empirical Mode Decomposition. IFAC-PapersOnLine, 51(5), 132-1.

Time Series Analysis of Rainfall Data at Cape Point Using Empirical Mode Decomposition

Willard Zvarevashe[1*], Syamala Krishnannair[1] and Venkataraman Sivakumar[2]

[1]*Department of Mathematical Sciences, University of Zululand, Private Bag X1001, KwaDlangezwa, 3886, South Africa*

[2]*Department of Chemistry and Physics, University of Kwazulu-Natal, Private Bag X54001, Durban, South Africa*

Abstract

Empirical Mode Decomposition (EMD) is a data adaptive method, which is best suited to analyse nonlinear and non-stationary data. The original data is decomposed into several intrinsic mode functions(IMFS) which represent different time scales.

A 31-years of monthly rainfall data is measured at Cape Point weather station is decomposed into 10 IMFs and residual. The residual shows a general pattern of the expected rainfall variations over the site. The inter-annual variability (2-8 years) is identified which is known to be primarily influenced by El-Niño Southern Oscillation (ENSO). The Southern Oscillation Index (SOI), which is one of the indices used to measure the strength of ENSO, is further used to find the correlation with rainfall.

We found a high correlation between IMFs from SOI and rainfall, which evidences the influence of ENSO on the rainfall pattern through the EMD method. The identified IMFs are orthogonal to each other and their sum approximates the original data, which shows that it is a lossless decomposition and there is less amount of information leakage. Based on the results obtained, the EMD method is found to be suitable to identify different oscillations in the rainfall data.

Keywords: Rainfall, EMD, ENSO, Time series analysis

1. Introduction

The annual variation of rainfall in Southern Africa directly and indirectly affect human livelihoods and ecosystem through draughts, temperature changes, water supply and agricultural production. Most of Southern Africa experience Austral summer (October to March) and northeastern to southwestern regions experience austral winter months (May to August) (Dieppois et al., 2016). The summer rainfall variability has been shown to be influenced by El Niño-Southern Oscillation (ENSO) and becomes nonlinear with extreme weather patterns of wet (La Niña) and dry (El Niño) weather patterns (Kane, 2009; Fauchereau et al., 2009). There is also influence of ENSO on the w inter rainfall in South Africa (Phillippon et al., 2012). The Southern Oscillation Index (SOI) is one of the several indices used to gauge the intensity of ENSO. SOI is a standardised index based on the fluctuations in sea level pressure between Tahiti and Darwin, Australia measured during El Niño and La Niña episodes (Lehodey et al., 1997). Due to the volatility of climatic conditions which result in nonlinearity and non - stationarity characteristic of the data linear models are sometimes inconclusive (Huang et al., 1999; Molla et al., 2005).

Empirical mode decomposition (EMD), was therefore introduced by Huang et al. (1998), which does not make assumptions about linearity and stationarity. The intrinsic mode functions (IMFs) derived from the data are usually easy to interpret and relevant to the underlying physical system. EMD is a nonlinear technique and adaptively represents non-stationary time series data in different time scales (Huang et al., 1998). In this study 31-year rainfall in Cape-Point is decomposed using EMD and underlying physical signals are identified. Specifically, the influence of ENSO on the rainfall pattern is identified. This paper is organized as follows. In section 2, the data set and methodology are described. In section 3, the analysis and discussion of the results is done and conclusion is done in section 4.

2. Data and Method

2.1 Data

The monthly rainfall data were obtained from the Cape Point weather station, Cape Town, South Africa. It was collected from the South African Weather Services. It covers the period from the 1980 to 2011. The Southern Oscillation Index (SOI) represents fluctuations in sea level pressure between Tahiti and Darwin, Australia. It is one of the key atmospheric indices used to gauge the strength of El Niño. The data for SOI is publicly available on National Center for Environmental Information (NCEI) (https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/).

2.2 Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is an adaptive time-frequency data analysis method, which requires only that the data must consist of simple intrinsic mode of oscillations (Huang et al., 1998). It is most suitable for nonlinear and non-stationary data. The EMD methodology is based on a sifting process, which identifies local extrema (maxima and minima) which results in the formation of intrinsic mode functions (IMFs). The first IMF will contain the highest fluctuations and this is subtracted from the original data and subsequent IMFs are then derived from the subtracted data. The IMFs and residual data approximate the original data when they are summed together (Huang et al., 1998). The rainfall time series is decomposed into local orthogonal modes. A given series x(t) is decomposed into IMFs, $c_i(t)$, and residual r(t) so that the original data is approximated by the sum of IMFs and residual.

$$x(t) \approx \sum_{i=1}^{n} c_i(t) + r(t). \qquad (2.1)$$

The IMFs and residual were added together to reconstruct the original data. Root Mean Square (RMSE) was used to evaluate the original data and reconstructed decomposed data.

The IMFs must be mutually orthogonal to each other. Higher orthogonality corresponds to less amount of information leakage. Index of orthogonality (IO) is used which is given by

$$IO = \frac{1}{T}\sum_t \frac{1}{x^2(t)}\left(\sum_{i=1}^{n+1}\sum_{i=1}^{n+1} c_i(t)\cdot c_j(t)\right) \qquad (2.2)$$

where i and j represents the i^{th} and j^{th} IMFs.

Orthogonality between two IMFs is given by

$$IO_{(i,j)} = \frac{1}{T}\sum_t \frac{c_i(t)\cdot c_j(t)}{c_i{}^2(t)+c_j{}^2(t)}\,. \qquad (2.3)$$

The orthogonality between two IMFs must be less than 0.100 for it to be acceptable (Molla et al., 2005).

3. Results and Discussion

The average monthly rainfall at Cape-Point was standardized for easy computation and comparison. The data was decomposed until when there is at most one maximum and one minimum in the residual. From the standardized rainfall data 10 IMFs were found and 8 for SOI. The decomposed rainfall and SOI showed a higher correlation in some of the IMFs as shown in figure 3.1. It is noted that IMF 4 (rainfall) against IMF 3 (SOI) is having positive correlation of 0.285 and -0.193 against IMF 4 (SOI). IMF 3 is annual oscillation whilst IMF 4 has average cycle of about 18.2 months. Figure 3.2 shows the general pattern and oscillation of IMF 4 for rainfall and SOI. Seven-year oscillation represented by IMF 6 and 7 (rainfall) also have significant correlations with IMF 6 and 7 from SOI.

The rainfall data for May, June and July period was decomposed into 7 IMFs and residual. On figure 3.2 Annual oscillation (IMF 1), quai-biannual occilllation (IMF 2) and ENSO like oscillation (IMF 6) are identified in the rainfall. The probability density function for each IMF is approximately normally distributed. The IMFs and residual are added together to reconstruct the data. The reconstructed data approximates the original data with Root

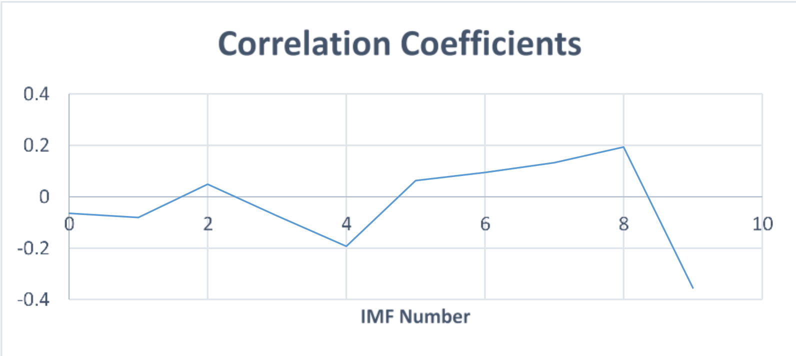


Figure 3.1 The correlation coefficients of the SOI and Cape Point Monthly Rainfall and their corresponding IMF-like components for a 12 month period. IMF 0 here represents original signal.

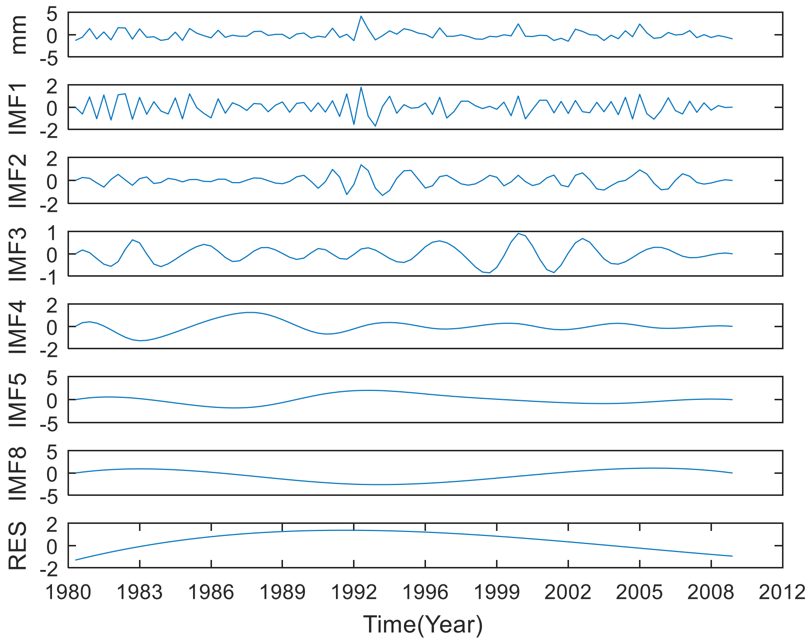


Figure 3.2 Rainfall IMFs for May, June and July at Cape Point station.

Mean Square Error of order 10^{-14}. This clearly shows that EMD is lossless decomposition with minimal data being lost in the decomposition and managing to capture most of the oscillations in the data.

The overall index of orthogonality (IO) for decomposed rainfall is $0.594\text{x}10^{-4}$ and SOI $0.154\text{x}10^{-4}$. It is noted that the maximum value of orthogonality between the IMFs is found to be ≈ 0.001 and it is way below the acceptable value of less than 0.1. It confirms that there is less amount of information leakage

4. Conclusions

The effectiveness of EMD to analyse nonlinear and

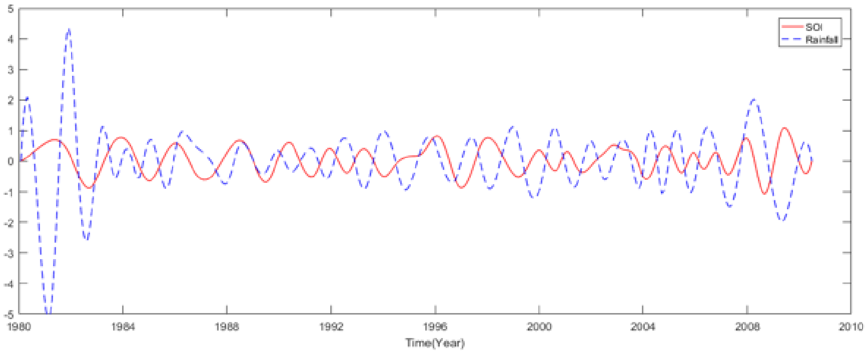


Figure 3.3 IMF 4 for SOI and rainfall

non-stationary data was demonstrated. The rainfall data was decomposed into IMFs and residual data, which summed up to the original data. EMD was effective in isolating the data into different timescales and therefore the variability of the rainfall pattern was identified, in the end evidence of the effect of ENSO was provided. This is consistent with what was found by Phillippon et al. (2012)

References

Dieppois, B., Pohl, B., Rouault, M., New, M., Lawler, D., & Keenlyside, N. (2016). Interannual to interdecadal variability of winter and summer southern African rainfall, and their teleconnections. *Journal of Geophysical Research: Atmospheres* , *121*(11), 6215-6239.

Fauchereau, N., Pohl, B., Reason, C. J. C., Rouault, M., & Richard, Y. (2009). Recurrent daily OLR patterns in the Southern Africa/Southwest Indian Ocean region, implications for South African rainfall and teleconnections. *Climate Dynamics*, *32*(4), 575-591.

Huang, N. E., Shen, Z., & Long, S. R. (1999). A new review of non-linear water waves: the Hilbert spectrum. *Annual Review of Fluid Mechanics*, 31, 417–457.

Huang, N. E., Z.Shen, R.Long, S., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, V., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,. *Proceeding of the Royal Society of London, Series A.*

Lehodey, P., Bertignac, M., Hampt on, J., Lewis, A., & Picaut, J. (1997). El niño southern oscillation and tuna in the western pacific. Nature, 389(6652), 715–718.

Kane, R. (2009). Periodicities, enso effects and trends of some south african rainfall series: an update. *South African Journal of Science*, 105(5-6), 199–207.

Molla, M. K. I., Rahman, M. S., Sumi, A., & Banik , P. (2006). Empirical mode decomposition analysis of climate changes with special reference to rainfall data. *Discrete dynamics in Nature and Society, 2006.*

Molla, K. I., Sumi, A., & Rahman, S. (2007). Analysis of Temperature Change under Global Warming Impact using Empirical Mode Decomposition. *International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering*, 1.

Philippon, N., Rouault, M., Richard, Y., & Favre, A. (2012). The influence of ENSO on winter rainfall in South Africa. *International Journal of Climatology* , *32*(15), 2333-2347.