The Investigation of Trends, Analysis and Modelling of Daily Rainfall in Australia

Bright Emmanuel Owusu

A Thesis Submitted in Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Research Methodology

Prince of Songkla University

2018

Thesis Title	The Investigation of Trends, Analysis and Modelling of Daily Rainfall inAustralia
Author	Mr. Bright Emmanuel Owusu
Major Program	Research Methodology

Major Advisor:

…….....
(Emeritus Prof. Dr. Don McNeil)

Co- advisor:

……..
(Asst. Prof. Dr. Nittaya McNeil)

Examining Committee:

……..…… Chairperson
(Dr.Somporn Chuai-Aree)

…….....
(Emeritus Prof. Dr.Don McNeil)

……...
(Asst. Prof. Dr. Nittaya McNeil)

……....
(Asst. Prof. Dr. Suree Chooprateep)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfilment of the requirements for the Doctor of Philosophy Degree in Research Methodology.

……..……

(Assoc. Prof. Dr. Damrongsak Faroongsarng)

Dean of Graduate School

This is to declare that the work here submitted is the result of the candidate investigations. Due acknowledgements have been made of any assistance received.

……..Signature

(Emeritus Prof. Dr. Don McNeil)

Major Advisor

…….….Signature

(Mr. Bright Emmanuel Owusu)

Candidate

I at this moment declare that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

…….…….……Signature

(Mr. Bright Emmanuel Owusu)

Candidate

Thesis Title	The Investigation of Trends, Analysis and Modelling of Daily Rainfall in Australia
Author	Mr. Bright Emmanuel Owusu
Major Program	Research Methodology
Academic Year	2017

ABSTRACT

Daily accumulated rainfall data obtained from observational stations in Australia over a period of 64 years were explored using statistical methods to determine the patterns. The thesis composed three studies. In the first study, daily rainfall during 1950-2013 for 92 observational stations was obtained from Australian Bureau of Meteorology. Rainfall variability over successive 5-day periods from 1950-2013 was examined. The first model uses factor analysis to classify the 92 stations of factors, which represent distinct geographical regions. Gamma generalised linear models were then fitted to describe the patterns in the 5-day non-zero rainfall amount in each factor region with period and year as the predictors.

The factor analysis revealed eight factors in the data, which represent eight geographical regions. These regions comprise north, northeast, central east, northwest, central south, north southeast, southwest and south-southeast. The rainfall from each region exhibits different overall mean with various seasonal variations. The second study also uses daily accumulated rainfall from the 92 observational stations acquired from the Australian Bureau of Meteorology. Multiple linear regression and Gamma GLM models fitted to the 5-day rainfall observations were compared. A simple linear regression model was also fitted to explore the trends in the annual rainfall. The 5-day

rainfall observations in successive periods were serially correlated, and this was minimised by using the AR(1) technique in fitting the models. The Gamma GLM and the multiple regression models fitted the data quite well in all the regions, particularly in delineating the periodic seasonal patterns. The patterns of the models were well marked mainly in the tropical regions (lower latitudes) where the internal atmospheric variability is small relative to the forced change. However, the multiple regression models did better than the Gamma generalised linear model in the north, northeast and the southwest regions. These models could be used to infill data in areas where rainfall records are inadequate.

The linear regression model revealed substantial decreasing annual rainfall trends in the southwest and the north southeast regions. In contrast, significant increasing annual rainfall trends were found in the north and the northwest regions. The final study used daily accumulated rainfall from 105 observational stations acquired from the Australian Bureau of Meteorology. The 64 years data also span between 1950 and 2013. The stations were randomly selected to cover the whole of Australia.

Logistic regression was then applied to develop a model to predict 5-day rainfall probability of occurrence or non-occurrence in each station, but nine stations were carefully chosen as a case study. The nine stations were carefully selected to capture the patterns observed in the data from all the stations and to capture most of the six climate classification by the Australian Bureau of Meteorology.

The fitted logistic regression models predicted the occurrence and non-occurrence rainfall events quite well with good accuracies. The predictors significantly affected the models in all stations. Analysis of the levels of each of the predictors showed that

the parameters of the 5-day period factors were more influential in all the models, particularly during the rainy season in most stations relative to that of the annual factors.

Acknowledgements

I would like to express my sincere appreciation to my advisor, Professor Dr. Don McNeil, my co-advisor Asst. Prof. Dr. Nittaya McNeil, and Asst. Prof. Dr. Chamnein Choonpradub for their advice and support throughout the study. I am most grateful to my thesis committee members whose valuable comments, suggestions and advice helped in transforming the thesis to the required format.

My great thanks also go to lecturers and staff of the Department of Mathematics and Computer Science Prince of Songkla University Pattani Campus whose work directly or indirectly contributed to the completion of this thesis.

I would like to thank the Graduate school Prince of Songkla University and the Presbyterian University College Ghana for their financial support and providing an enabling environment for the successful completion of this thesis.

It is also my great thanks to my family and colleagues who have thrown their weight behind me throughout the study.

Bright Emmanuel Owusu

Contents

List of Tables

List of figures

CHAPTER 1

Introduction

1.1 Background and Rational

Rainfall is an important meteorological parameter and part of the hydrological cycle. It is one of the necessary quantities in meteorology and hydrology due to its high variability. It has essential effects on people's daily life, heavy rainfall possibly will bring tremendous losses in an economy, cause fatalities, and lack of it could render a place arid. Thus, rainfall is one of the top priorities in operational forecast and scientific research and is because of convective development under a favourable environment.

Australia's continent is bounded by the Indian and Pacific oceans, classified as the driest inhabited continent in the world. It is approximately 3,860 km long from its most northerly point to its southerly point and about 4,000 km wide, from east to west. The Bureau of Meteorology has classified the Australian climate into six regimes, which include equatorial, tropical, subtropical, desert, grassland, and temperate groups. The climates vary from tropical in the north to arid in the middle to temperate in the south and rainfall is highly erratic in both space and time and is limited usually to the coast. The largest part of the country is dry, mainly the middle and the west. About 80% of the country gets an annual rainfall of at most 600 mm, but the tropical region of the far north experiences an annual rainfall of over 4,000 mm (Sivakumar et al., 2014).

Precipitation refers to the various forms of water that flow from the atmosphere. It can be in the form of rain, snow, sleet, hail. If a portion of air rises, it expands in the lower pressure, cools, and therefore condenses moisture, producing cloud and, eventually, rainfall or snowfall. Rainfall processes are likely to be a complicated issue due to several factors involved. One of them is urbanisation in which surface modification of the natural surface is spatially replaced with human-made or artificial impervious materials, changing or disturbing surface energy balance and the atmosphere over the area (Oke, 1998).

Studies such as (Oke, 1998; Shepherd, 2005) revealed that increasing population resulting in urbanisation had modified land surface processes through the formation of 'urban heat island'. The modification of the land surface influences their local to global scale climate, as variations in land use and land cover constantly modify energy and moisture fields as well as circulation patterns.Moreover, a significant amount of anthropogenic heat is radiated from relatively large fuel usage. Air pollutants in the area can also play a role in increasing the number of cloud condensation nuclei, with a complex series of feedbacks to cloud formation and rainfall (Shem and Shepherd, 2009).

Hand and Shepherd (2009) revealed that increasing surface temperature enhances clouds formation and rainfall anomalies. Surface warming has a direct effect on rainfall. Increasing heating leads to greater evaporation resulting in an efficient drying of surfaces and increasing the duration and the intensity of drought. The water holding ability of air increases by about 7% per 1°C warming, which leads to increased water vapour in the atmosphere, which affects storm development. Also, storms supplied with increased moisture, yield more intense rainfall events (Trenberth, 2011).

Analysis of temperature change in Australia by Wanishsakpong and McNeil (2016) revealed similar increasing temperature patterns in many areas. This area includes the central, eastern, southern and southeastern parts of Australia.

Some countries including Australia have for the past few decades experienced extreme climatic events such as hurricanes, floods, drought and other disasters due to climate change, resulting in human casualties and massive damages to properties and infrastructure. A major concern of climate change regarding changes in rainfall will likely be the variability in amount, occurrence and the magnitude of these events. Moreover, variations in rainfall cause a lot of damage and loss of life in most countries every year throughout the globe.

A study by Trenberth et al.(2003) revealed that, for the past few decades, the global hydrological cycle has been experiencing major fluctuations, which comprise rainfall amount, frequency and duration.It is assumed that global warming will lead to a more dynamic atmosphere, which potentially leads to more frequent high-intensity rainfall events in many regions of the globe even in regions where rainfall is reducing.It is, therefore, necessary to investigate daily rainfall variability in Australia. Even though climatological characteristics are highly stochastic, most studies (Coe and Stern, 1982; Stern and Coe, (1984) used statistical methods to develop statistical models to predict future rainfall occurrences.

Statistical analysis of the temporal and spatial variability, trends, and a model fitting analysis of rainfall events are vital for water resource management, hydrologic and ecologic modelling, recharge assessment, agricultural and irrigation scheduling. The effects of increasing population and urbanisation on rainfall patterns are of scientific

concern because there are clear ties to contemporary research and prediction problems in meteorology, climatology, hydrology, and geography systems.This study investigates the spatial-temporal rainfall trends and applies appropriate statistical methods to model rainfall amounts and the probability of occurrences in Australia during 1950-2013.

1.2 Research Objective

The primary objective of this research is to classify rainfall variability, describe the spatial-temporal rainfall trends and apply an appropriate statistical method to model rainfall probability of occurrences and amounts.

1.3 Literature Review

Many scholars (Bärring, 1987, 1988; Wickramagamage, 2010; Um et al., 2011) have studied rainfall regions classification. Most of these studies described the spatial-temporal variability based on classification into distinct geographical regions by using principal component analysis (PCA) and factor analysis. These methods reasonably delineated rainfall regions in all their study areas. However, the latter gave easily interpretable results as compared to PCA. Darand and Daneshvar (2014) used PCA and hierarchical clustering analysis to regionalise rainfall regimes in Iran. Also, PCA and cluster analysis were used by Chambers(2003) to group rainfall into regions in South of Australia and observed that both methods gave similar clusters.

Some authors have modelled daily rainfall amount with various mathematical models. These include exponential (e.g. Todorovic and Woolhiser, 1975), lognormal (e.g. Cho et al., 2004) and Gamma distributions (e.g. Ison et al., 1971; Buishand, 1977; Stern and Coe, 1984; Husak et al., 2007).

Haan et al.(1976) tried to describe rainfall amounts by fitting Markov chain with many states each representing a range of amounts. The problems associated with these models are a large number of parameters to be estimated. Hession and Moore (2011) used spatial regression to evaluate the role of prognostic variables while integrating spatial autocorrelation in parameter estimation and hypothesis testing in rainfall data from Kenya. This methodology can yield a better understanding of associations among precipitation and multiple predictive variables with enhanced statistical rigour. The method showed that topographic variables such as elevation and slope strongly influence rainfall during the long and short rains, which are vital for Kenyan agriculture.

Rainfall data are known to be skewed, so a model is needed to take account of this skewness. One form of analysis, which has received much attention in the modelling of the daily rainfall amount, is the gamma GLM (generalised linear model). This model has been found by many authors (e.g. Coe and Stern 1982; Stern and Coe, 1984; Kenabatho et al., 2012) to fit rainfall observations extremely well. The GLM models are capable of reproducing rainfall features when predefined climatic zones are model independently and for infilling dataset in cases of missing observations.

Most authors have employed Markov chain models to analyse the occurrence of daily rainfall (Gabriel and Neumann, 1962; Haan et al., 1976; Coe and Stern, 1982; Stern and Coe, 1984). The used of this method in rainfall analysis became accepted after Gabriel, and Neumann (1962) used it to analyse the daily occurrence of rainfall at Tel Aviv, Israel, fitting a two-state, first-order Markov chain. Results from these studies revealed that Markov chain model was capable of simulating daily rainfall observations of any length centred on the estimated transitional probabilities and

frequency distributions of rainfall amounts and the simulated data was related to that of the historical data. However, the problem associated with this method is a large number of parameters that need to be estimated.

Currently, one of the widely used techniques used in the modelling of daily rainfall occurrence is a logistic regression, where the probability of a wet day at a specified site is a function of some predictors representing previous day's rainfall. Some studies (Coe and Stern, 1982; Stern and Coe, 1984; Chandler and Wheater, 2002; Kenabatho et al., 2012) used logistic regression to predict rainfall occurrences in their respective areas and concluded that the models did quite well. Prasad et al. (2010) used logistic regression to model monthly rainfall occurrences in India. The model performed well in capturing the extreme rainfall years and appear to achieve better than the direct model forecasts of total precipitation in respect of such years.

Analysis of nine years warm season mean daily cumulative rainfall by Hand and Shepherd (2009) used both satellite precipitation analysis and rain gauge stations to study the spatial variability in warm-season rainfall events around Oklahoma City (OKC). They applied correlation and linear regression on the urban rainfall ratio. They revealed that the North-Northeastern regions of the metropolitan OKC area remained statistically wetter than other regions. A study conducted in Mexico by Jauregui and Romales (1996) examined the urban and non-urban effect on convective rainfall trends. They fitted a regression model on the urban rainfall ratios and revealed that wet season (May-October) rainfall for Tacubaya an urban site showed a significant increasing trend while rainfall at the non urban site remained unchanged during 1941-1985. Rana et al. (2012) investigated the long term trend in monsoon rainfall by using Man-Kendall rank statistics and linear regression and the possible

effect of global climatic phenomena at Mumbai and Delhi in India with PCA and singular value decomposition. Thestudyshowedaninsignificantdecreasein southwest monsoon rainfall while the increase in winter and pre-monsoon seasonoverDelhiforthewholeperiod (1951-2004).A significant negative change for long-term rainfall was detected for different seasons and the whole year at Mumbai during the same season.Trends in southwest monsoon rainfall for both Delhi and Mumbai were shown to be related to the variability of climatic indices.

Boccolari and Malmusi (2013) assessed temperature and precipitation trends by using the linear regression. The existence of trends was then detected with both parametric t-test and the non-parametric Mann-Kendall test at 5% level to compare the results at Modena, during 1861-2010. They showed that rainfall is decreasing at 6.3 mm per decade, while their analysis for the last thirty years showed a significant increment of 74.8 mm per decade. A similar study on the analysis of seasonal and annual rainfall trends by Wickramagamage (2015) used linear regression to examine the trends in 5-day average rainfall in Sri Lanka. Currently, most studies on trend analyses of rainfall employ both linear regression and the Mann-Kendall statistics to compare results from the parametric and non-parametric test.

There have been some studies on the different aspects of Australian rainfall. For instance, investigations such as (Drosdowsky, 1993; Chambers, 2003) applied PCA and cluster analysis to describe rainfall variability and revealed that, rainfall was highly variable throughout the country. Trend analysis by Lavender and Abbs (2013) used linear regression technique and non-parametric Mann-Kendall test and observed significant changes with increases in the northwest and decreases in the east from 1970 to 2009. However, these increases may be partially associated with tropical

cyclones and other low-pressure systems. Studies such as (Wardle and Smith, 2004) suggested that the rising rainfall over northwest Australia is due to enhanced aerosols bring about as a result of human activity, especially from Asia. However, there is still lack of understanding of the atmospheric processes operating in this region. A study by England et al. (2006) examined the interannual rainfall extremes over southwest Western Australia (SWWA) about changes in tropical and subtropical sea surface temperature in the Indian Ocean. The SWWA has also experienced a reduction in rainfall since the mid-1960s, especially throughout austral winter (e.g. Allan and Haylock, 1993; Ansell et al., 2000; Li et al., 2003). This trend seems to be related to a combination of factors, including increased greenhouse gas concentrations, natural climate variability (Smith et al., 2000; Li et al., 2005) and land use change (Pitman et al., 2004).

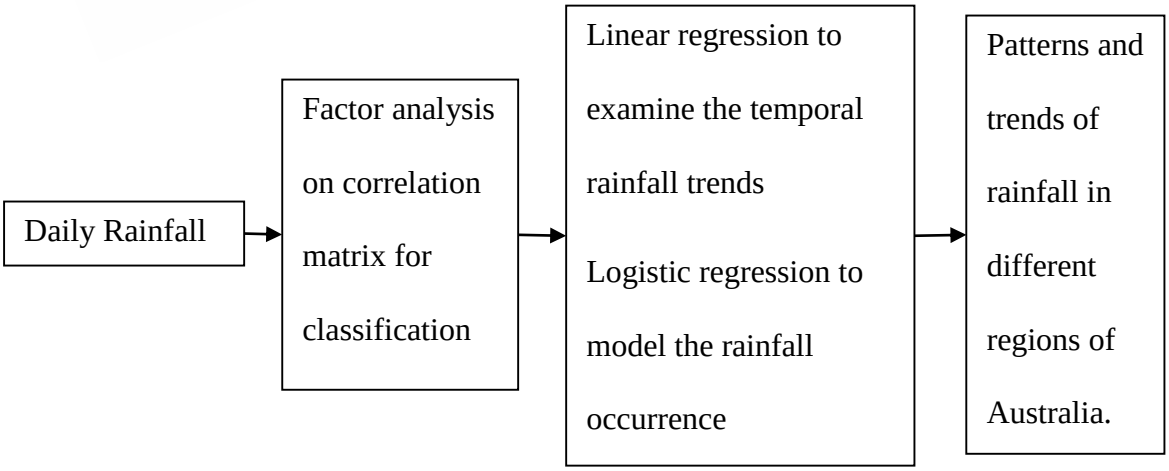
Studies (Srikanthan and Pegram 2009; Beecham et al., 2014) used multi-site two parts model to describe the daily variability at some sites. They described the need to model its occurrences and amounts with different models and observed an acceptable model fits in all the sites. Hasan and Dunn (2010) used a Poisson-gamma model to describe rainfall for agricultural planning. They used monthly observations from 220 stations using 6 of the 220 stations as a case study. In their study, important events such as the probability of prolonged dry spells and heavy erosive rainfall may be lost.

Most of the studies in Australia that used statistical models analysed rainfall using aggregated observations (10 days totals, monthly totals). In the modelling of rainfall characteristics, Coe and Stern (1982) emphasised that daily rainfall must be used to know important events such as, the probability of long dry spells, heavy rainfall which is crucial to agronomist, hydrologist, meteorologist in the planning of their

various operations. In assessing the possibility of climate change concerning rainfall, long-term seasonal patterns and annual trends must be examined. Knowledge of the trends in the spatial-temporal rainfall variability and their physical explanations are essential in climate change assessment (Kenabatho et al., 2012), and determining erosivity (Bärring, 1987). Even though some models have been used to describe rainfall in Australia, these models are not available for all areas, and most models used monthly data and have not given in-depth analysis of the daily patterns.

1.4 Conceptual Framework

Figure 1.1 is the conceptual diagram, and it shows the research processes of this study. Initially, factor analysis was used to reduce spatial correlation and group rainfall into distinct geographical regions. In determining the rainfall patterns and trends, a simple linear regression model was employed while Gamma GLM was then used to describe the rainfall amount and logistic regression was used to model its occurrences.



1.5 Organisation of the Thesis

This section highlights on how the thesis is organised. This write-up is made up of six chapters arranged as follows:

Chapter 1 introduces the background and rationale of rainfall, the study objectives, the literature review and the conceptual framework. In chapter 2, the description of the data source, management and the methods used for the study are provided while chapter 3 presents the preliminary analysis of the data. Classification and the modelling of daily rainfall amount in Australia during 1950-2013 are presented in chapter 4. In chapter 5, the probability of rainfall occurrences in Australia is provided. Chapter 6 gives the conclusion and discussions of the research findings together with recommendations for further studies.

CHAPTER 2

Methodology

This section presents the acquisition, sources and management of the data together with the various statistical methods that were employed in the analyses of the data. These methods include factor analysis, simple linear regression, gamma GLM and logistic regression. In this study, the R Project for Statistical Computing (2015) was used for all data analysis, graphical presentations and computations.

2.1 Data Sources and Management

In this study, single source long-term daily aggregated rainfall observations in Australia from 1950 to 2013wereacquired from the Australian Bureau of Meteorology at (http://www.bom.gov.au/climate/data). The first and second parts of the study use data from 92 observational stations (Figure 2.1) while the final part of the study employs data from 105 observational stations (Figure 2.2). These stations were selected to give a sample covering the whole area as evenly as possible, and they have continuous daily rainfall records that extend over a period of 64 years. However, there are no stations in the western desert, and most stations did not collect much data before 1950.

Data recorded on leap years were omitted to maintain same observations for each year, and therefore each station was made up of 23,360 observations for the 64 years records. For each station, the 5-day average rainfall was computed and used in this study as the response variable. This choice in the statistical analysis has the following advantages: the proportion of missing data is substantially reduced,and daily patterns

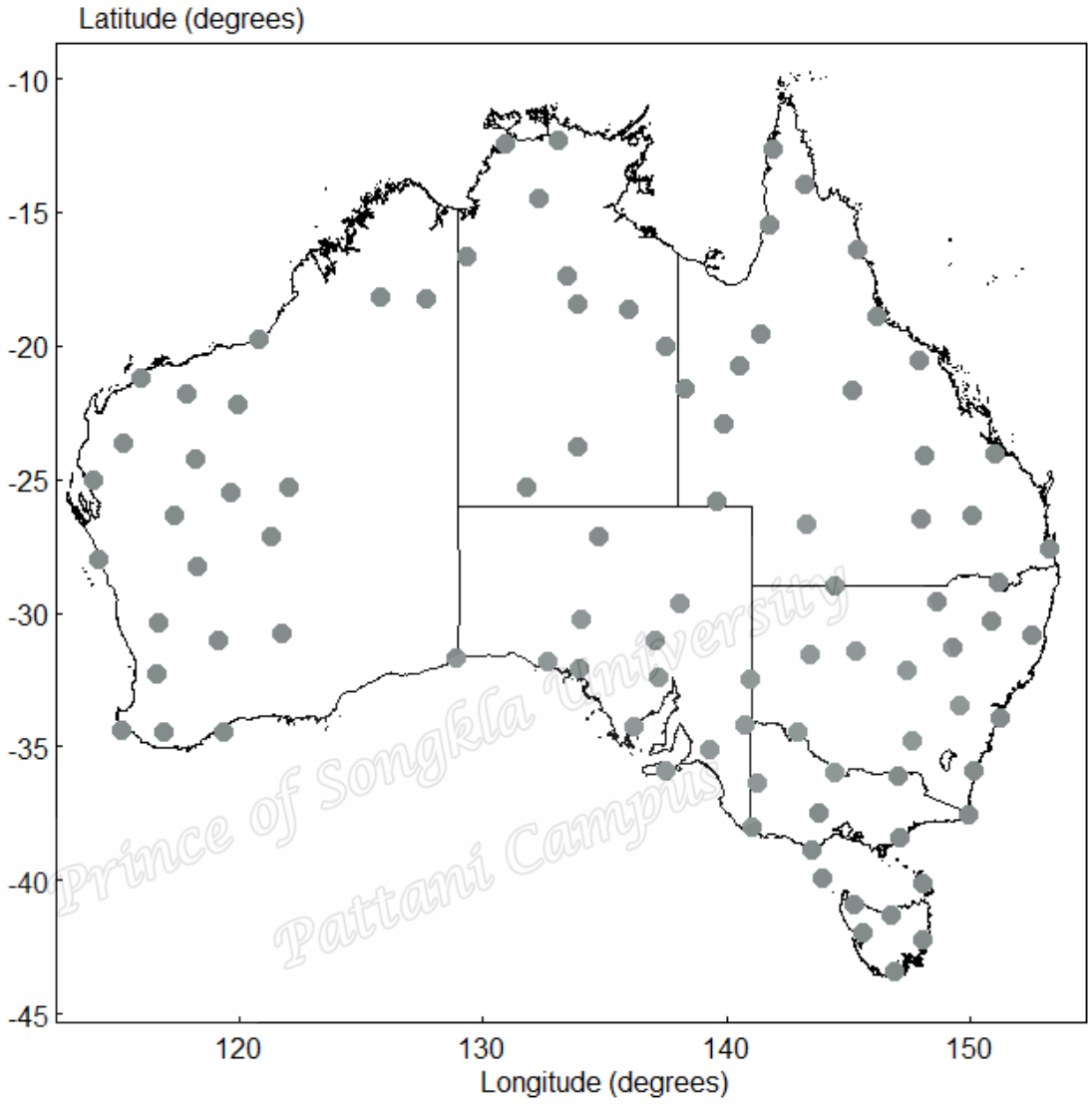


Figure 2.1 Map of the study area showing the locations of the 92 meteorological stations

are well represented graphically. Moreover, the correlation between data in consecutive periods is also substantially reduced. After omitting data for February 29 in leap years, each year contains precisely 73 data values, and the number of parameters in the model is reduced by a factor of 5, facilitating data management, computations time and providing substantially smoother model fit.

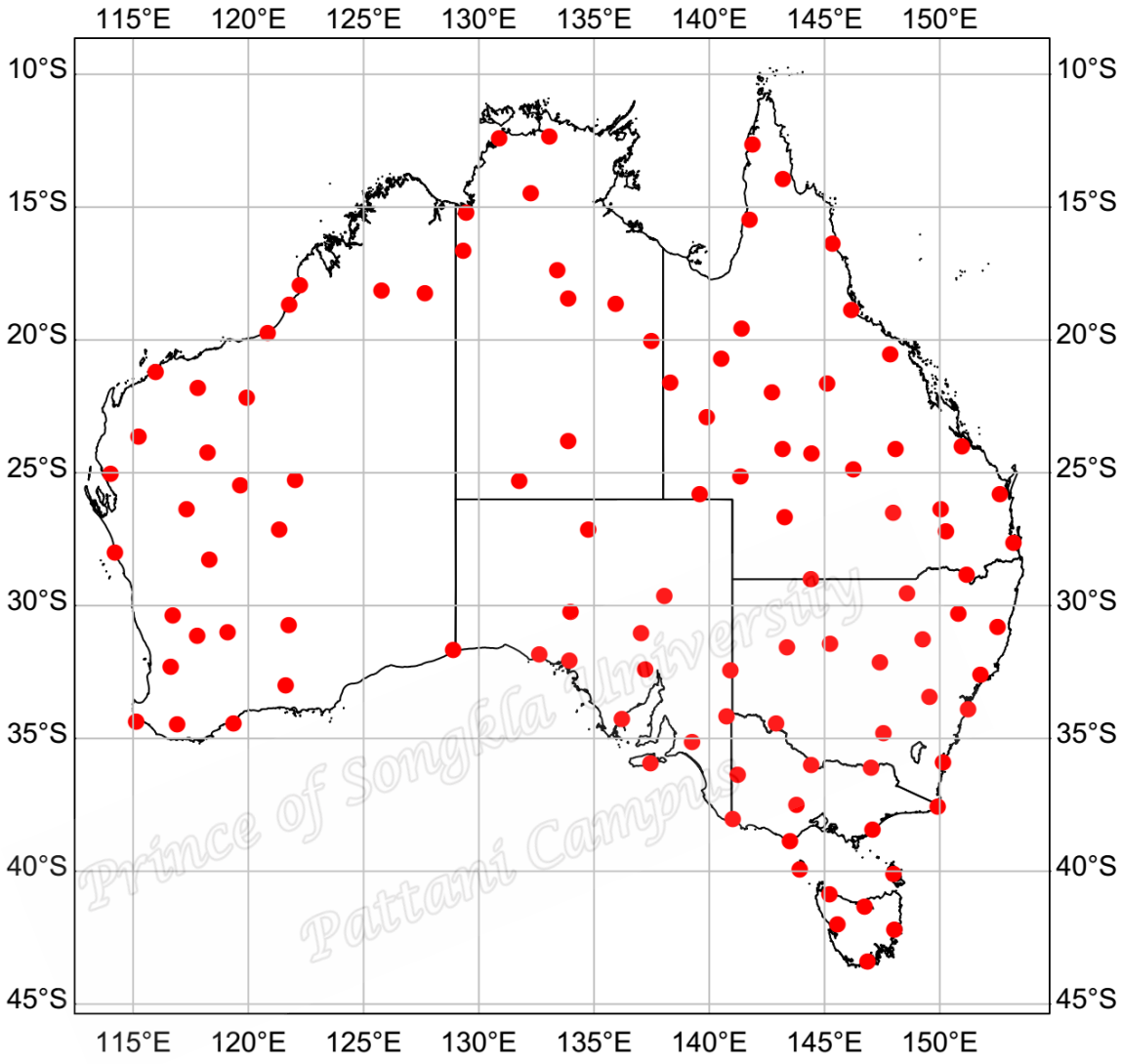


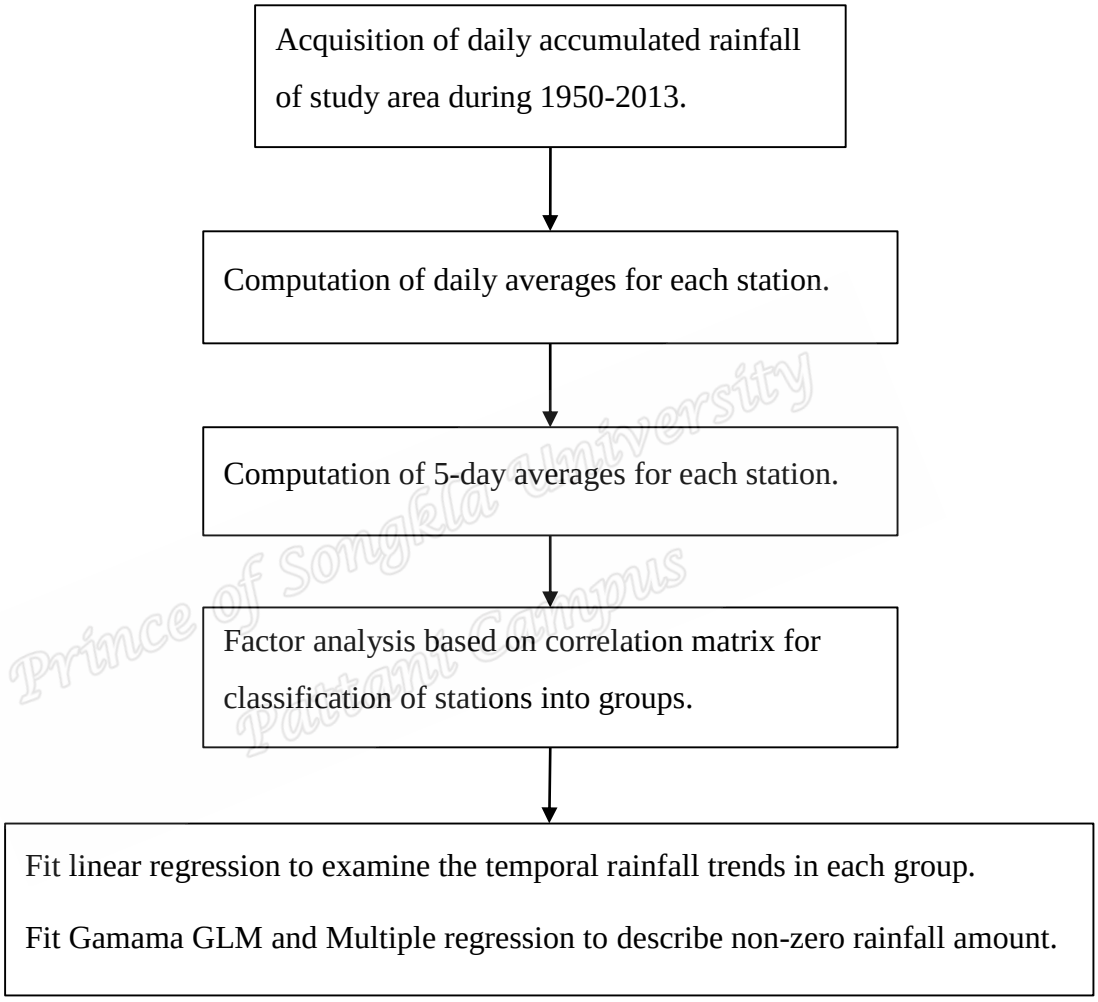
Figure 2.2 Map of the study area showing the locations of the 105 meteorological stations

2.1.1 Variables

The outcome variable of concern is the 5-day average rainfall while the predictor variables used in this study are the 5-day periods and the annual average rainfall factors. The first predictor comprises of 73 categories while the latter consists of 64.

2.2 Study diagrams

The processes involved in the analysis of the rainfall data in Australia during 1950-2013 is shown in Figure 2.3 and Figure 2.4.



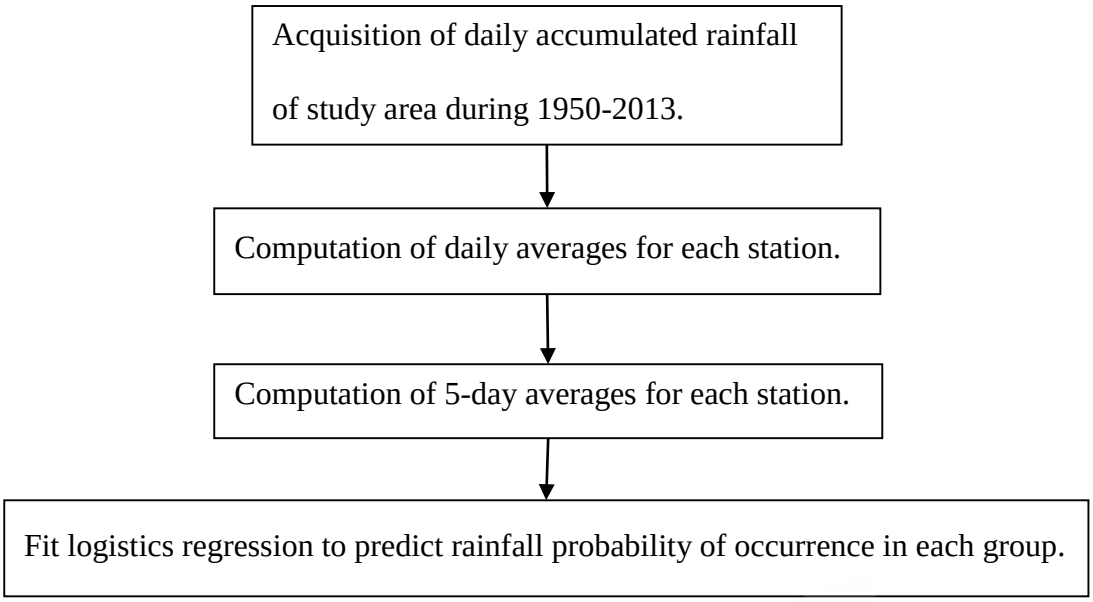


Figure 2.4 The diagram showing the steps involved in the data analysis in the second study.

Some of the data collected from some observational stations were accumulated to some indicated periods. Initially, the daily averages were computed based on the period of accumulation indicated. The 5-day average of the data was then computed for all the stations. In the first study (Figure 2.3), factor analysis was applied to the correlation matrix of the 5-day average rainfall of each station to put them into groups. The temporal rainfall trends in each of the groups revealed by the factor analysis were determined by the simple linear regression model. Further analyses involve fitting Gamma GLM and multiple linear regression models to describe the non-zero rainfall amount.In the second study (Figure 2.4), logistic regression models were applied to predict the non-occurrence and occurrence 5-day rainfall probability in all the stations.

2.3 Statistical methods

2.3.1 Serial Correlation

Serial correlation refers to meteorological persistence or the tendency for the weather in consecutive times to be similar. For continuous variables (e.g., rainfall), persistence typically is characterised regarding serial correlation or temporal autocorrelation. The prefix “auto” in autocorrelation denotes the correlation of a variable with itself, so that temporal autocorrelation indicates the correlation of a variable with its future and past values. Sometimes such correlations are referred to as lagged correlations. Usually, autocorrelations are computed as Pearson product-moment correlation coefficients, despite the fact that there is no motivation behind why different types of slacked connection cannot also be processed. Representing the sample mean of the first n - 1 values with the subscript “-” and that of the last n -1 values with the subscript “+,” the lag-1 autocorrelation is

$$r_1 = \frac{\sum_{i=1}^{n-1}\left[\left(x_i - \bar{x}_-\right)\left(x_i - \bar{x}_+\right)\right]}{\left[\sum_{i=1}^{n-1}\left(x_i - \bar{x}_-\right)^2 \sum_{i=2}^{n}\left(x_i - \bar{x}_+\right)\right]^{1/2}} \qquad (2.1)$$

Equation 2.1 can be generalised to lag-k autocorrelation coefficient using

$$r_k = \frac{\sum_{i=1}^{n-k}\left[\left(x_i - \bar{x}_-\right)\left(x_{i+k} - \bar{x}_+\right)\right]}{\left[\sum_{i=1}^{n-k}\left(x_i - \bar{x}_-\right)^2 \sum_{i=2}^{n}\left(x_{i+k} - \bar{x}_+\right)\right]^{1/2}} \qquad (2.2)$$

where $\bar{x}$ is the mean of $N-k$.

2.3.2 The Correlogram

In the analysis of time series data, a correlogram is an image of correlation statistics. It is a plot of the sample autocorrelations versus the time lags. Correlogram is essential in the interpretation of a set of autocorrelation coefficients. For an entirely random time series, 5% of the values of r_k would be expected to be within $\pm 2/\sqrt{N}$ as shown in Figure 2.5.

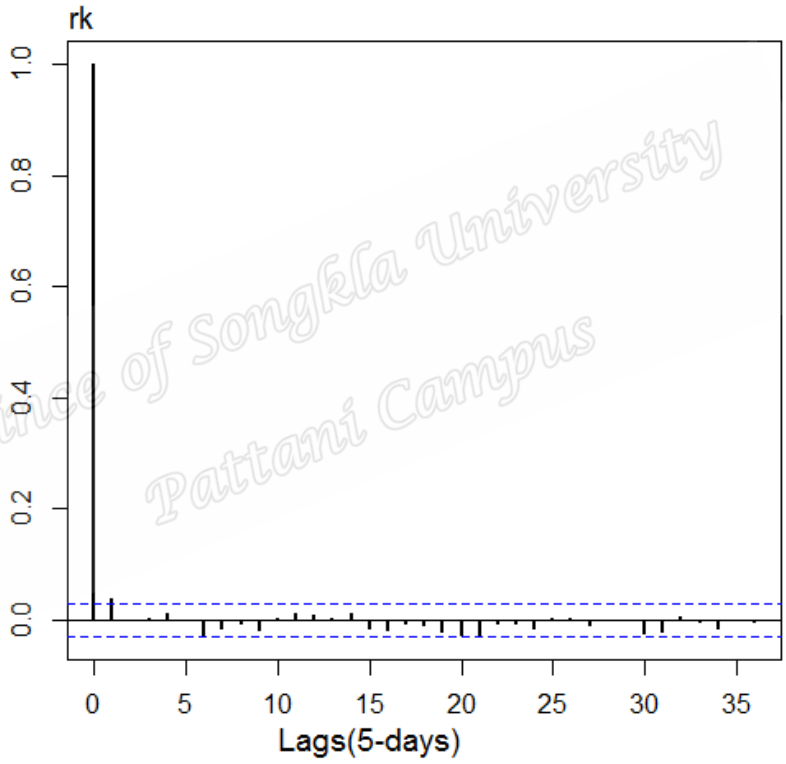


Figure 2.5 The correlogram, the horizontal dotted lines depict the $\pm 2/\sqrt{N}$.

2.3.3 Factor analysis

Factor analysis technique involves putting variables into groups (factors) that maximise correlations between variables within each group and minimise correlations between variables in different groups. It is widely used as a method for analysing multiple outcomes and is ideally suited for meteorological data collected from stations

spread over a large region such as a continent or nation (Bukantis, 2002, Unkel et al., 2010). The factor model with m factors $(f_1, f_2, f_3, ..., f_m)$ can be written as:

$$y_t = \mu + \sum_{j=1}^{m} \lambda_j f_j \tag{2.3}$$

where $_t$ is the rainfall for time t, μ is the average rainfall amount of each station, λ_j is the factor loadings on the j^{th} factor, and f_j is the j^{th} common factor. The factor model also provides uniqueness values to the variables where variables with high uniqueness values cannot be assigned to any factor (Rencher, 2002). The factor analysis technique was applied to the correlation matrix of the 5-day rainfall observations.

The model parameters were estimated by the maximum likelihood factor extraction method. In order to obtain simple structure pattern of loadings that are easy to interpret with few large and many small coefficients, the factors were rotated. Two main types of rotations are used in factor analysis; orthogonal when the new axes are also at right angles to each other and oblique when the new axes are not forced to be orthogonal to each other, but the latter (specifically the Promax oblique method) was used in this study. The Promax rotation has the benefit of being fast and conceptually simple and attempts to fit an objective matrix which has a simple structure (Abdi, 2003). The estimated loadings are used to identify each factor. Stations were assigned to factors if it has at least 0.333 loadings and it is the largest among the loadings of all the other factors (Cheung et al., 2015).

2.3.4 Linear regression analysis

Linear regression analysis is a technique that allows for finding the best relationship between a dependent variable and several independent variables. The general form of this method can be represented as:

$$Y = \alpha + \beta x + \varepsilon \qquad (2.4)$$

where the dependent variable (Y) is the rainfall time series while x is the time, α is a constant, β is the slope and ε the error term which is assumed to be normally distributed and has constant variance. The trends in the rainfall time series were explored with equation (2.4).

2.3.6 Generalised Linear Models

In linear modelling, data are expected to be normally spread.In perspective of the non-normality of the vast majority of meteorological data (e.g. rainfall), the GLM models and other continuous distributions are considered. In GLM modelling, consider a vector

$y = [y_1, ..., y_n]$ of independent random variables with distribution in the natural exponential family and with a mean vector $\mu = [\mu_1, ..., \mu_n]$ A vector of parameters $\beta = [\beta_1, ..., \beta_p]$ and relate to the mean (μ) by $g(\mu) = X\beta$ where X is a $p \times 1$ known matrix of covariates, and g is the link function.Aside from the Gaussian, GLM can deal with a bigger class of distributions for the response variable Y. Some of the distributions include the binomial, Poisson and Gamma (Mccullagh and Nelder 1989). The relationship of the response variable Y.to the linear predictor through the link function $g(\mu)$.

2.3.5 Multiple regression

Multiple linear regression (MR) is the broader situation of simple linear regression. Regarding the latter, there is still single predictand**,** Y, but there is more than one predictor (*x*) in multiple linear regression.

$$Y = \alpha + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i \qquad (2.5)$$

The MR and Gamma GLM models were then applied to the non-zero 5-day rainfall means in all the factor regions. These procedures used the mean period rainfall from each region as the response variable, the year and period factors as the independent variables in fitting the models. A detailed account of these techniques are given underneath:

If μ_{kst} is the mean period rainfall observation for every region k (k,…, 8) at period *s*and year *t*, then the MR is defined as

$$h(\mu_{kst}) = \mu_{0k} + \sum_{i=1}^{64} \eta_{ik} x_{ti} \sum_{j=1}^{73} \beta_{jk} w_{sj} + h(\mu_{t(s-1)k}) + \varepsilon_{kst} \qquad (2.6)$$

where*h(.)* signifies the dependent variable fourth root transformation which is defined as $h(\mu_{tsk}) = \sqrt[4]{\mu_{tsk}}$, ε_{tsk} are random error terms which are normally distributed having mean 0 and variance σ^2. x_{ti} are predictors for relating years *t* and year *i*, w_{sj} are predictors for relating periods *s* and period *j*. Thus, x_{ti} becomes 1 if *t* =*i* and 0 elsewhere and w_{sj} becomes 1 if $s = j$ and 0 elsewhere. The year 1950 and period 1 are the starting year and period respectively. For region k, μ_{0k} denotes the mean period rainfall, which is the starting year and period, $\eta_{ik}$is the year *i* effect and β_{jk} the period

j effect.Fitted values $h(\mu_{kst})$ from the model (2.6) have to be transformed back to give the fitted values of μ_{kst}. Thus the backwards transformed can be written as μ_{kst}. Let $\hat{\mu}$ be the fitted value for μ and $h(\mu)$ the above-defined transformation. Then $\hat{\mu}$ can be estimated from the fitted value h($\hat{\mu}$) as $\hat{\mu} = (h(\hat{\mu}))^4$.

2.3.7 Logistic regression

The probability of 5-day rainfall occurrence and non-occurrence were predicted by using logistic regression. Logistic regression is a statistical method that can be used to describe the association of numerous independent variables to a dichotomous dependent variable. It is one of the generalised linear models with a binomial distribution for the response variable. The usual link function is a logit function, which in this study relates the logarithm of the odds ratio of the expected value of rainfall probability to linear predictors. It describes the occurrence probability of wet and dry days. To model the pattern of wet and dry days for each station using logistic regression, let p_i denote the probability of rain for the *i*th day in the data set, conditional on the variables x_i, then the model is given by:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \qquad (2.7)$$

where α is a constant, β_i $(i = 1,2,3,...,k)$ are the slope coefficients of the model and x_i $(i = 1,2,3,...,k)$ are the predictors and k is the number of predictors. In fitting the models, the response variable was dichotomous with zero (0) representing non-occurrence and one (1) representing the occurrence of 5-day rainfall. The first 50% of

the rainfall observations (2336) were used as the training subset, and the remaining 50% of the observations (2336) was used as the validation subset.

2.3.8 Evaluation of the fitted Models

The logistic regression models were assessed by using the confusion matrices, Receiver Operating Characteristic (ROC) diagram and the Akaike Information Criteria (AIC). The ROC diagram is a discrimination-based graphical forecast verification display used to assess the goodness of fit of a model predicting a dichotomous outcome. It comprises a plot of true positive and false positive classification rates (1-specificity) for various probable thresholds of the model. Numerous indices of precision have been suggested to summarise ROC curves. Specifically, the area under the ROC curve (AUC) index, is a standout amongst the most regularly utilised. A detailed account of the underlying principles of the ROC curve is given by Metz (1978), Hosmer and Lemeshow (2000) and Fawcett (2006).

The maiden application of ROC diagram in meteorology was by Mason (1982), and it is derived from signal detection theory in electrical engineering. A perfect forecast would have AUC value of 1, depicting forecast outcome as 1 (5-day rainfall) if $P_{ij} \geq c$ and 0 otherwise. A study by Vilar del Hoyo et al. 2011 revealed that an AUC value of 0.5 depicts no discrimination; a value among 0.5 and 0.69 shows poor discrimination; a value of 0.7-0.79 shows sound discrimination; a value of 0.8-0.9 designates excellent discrimination, and a value of 0.9 or higher specifies exceptional discrimination. The ROC curve composed of plots of the fraction of positive outcomes correctly forecast by the model alongside the false positive rate (proportion of all outcomes incorrectly predicted) as *c* varies. The choice of *c* is made to draw a

compromise between the number of the predicted and observed to make sure that equal weights are allocated to false positive and false negative prediction errors. ROCdiagrams have gained much attention in recent years to assess the probability forecasts for binary predictands. It is useful to emphasise that they do not provide an entire representation of the joint distribution of predictions and observations. In comparing the goodness of fit concerning the AIC values, models with smaller values fit better than those with higher AIC values.

2.3.9 Gamma generalised linear model

The probability density function of the gamma function is written as

$$f(x \mid \lambda, \gamma) = \frac{1}{\Gamma(\lambda)} \gamma^{\lambda} x^{\lambda-1} e^{-\gamma x},\ x, \lambda, \gamma > 0 \qquad (2.8)$$

where λ and γ are the shape and the scale parameters respectively, and x is the 5-daily rainfall amount.The shape parameter of the gamma function is assumed to be constant throughout the data set for each region. If μ_i is the mean rainfall amount for any period *i* connected with predictors ξ_i then the 5-day mean rainfall amount is estimated by equation (2.9) where τ_0 is constant and τ 's the slopes.

$$\ln(\mu_i) = \tau_0 + \tau_1\xi_1 + \tau_2\xi_2 + \ldots + \tau_k\xi_k \qquad (2.9)$$

The predictors used in this model were also the 5-day rainfall period (seasonal effect) and the annual daily rainfall (year effect). The fitting of the gamma GLM is similar to the fitting of multiple linear regression models where other parametric transformation or polynomial terms can be incorporated in the set of explanatory variables to take account of the multi-modal and the non-linear responses. Fitting the gamma GLM

involves finding an appropriate estimate for the model parameters. The models will be fitted using the treatment contrast which makes the first coefficient for each factor a reference; such that each coefficient represents a comparison of that coefficient with first factor (Venables and Ripley, 2002).

The fitted models will be assessed by the deviance residual (Q-Q) plots. The deviance residuals were plotted against the theoretical quantiles to determine if the GLMs are appropriate for the fitted models in all the factor regions from 1950-2013. The sum contrasts (Venables and Ripley, 2002; Tongkumchum and McNeil, 2009) was applied to obtain 95% confidence intervals (CI) of the means of the fitted models to compare it with the overall rainfall mean. It gives criteria to assess the distinction between the CI of the fitted models with the overall mean.

CHAPTER 3

Preliminary Analysis of Data

This chapter reports on the primary analysis of daily accumulated rainfall data during 1950-2013 in Australia.

3.1 The study area and data

In this study, long term daily accumulated rainfall data for 105 observational weather stations spread over Australia from 1950 to 2013wereused for the study (Figure 3.1). These 105 stations were selected to give a sample covering the whole area as evenly as possible, and they have continued daily rainfall records that extends over a period of 64 years. Data recorded on leap years were excluded to maintain same observations for each year, and therefore each station is made up of 23,360 observations for the 64 years records.

3.2 Summary of the daily rainfall data

The statistical summaries of the 5-day rainfall for all the stations are presented in Table 3.1. The 5-day mean ranges from 0.45 to 9.41 mm. The highest value occurred in Lake Margaret Dam while the least occurred in Woomera. Over 98% of the stations have less than 10% missing data. Figure 3.2 and 3.3 gives a detailed account of seasonal and annual patterns of 5-day mean rainfall for nine selected stations together with the probability of occurrence. The nine stations were carefully selected to capture the significant climate classification concerning the Köppen climate classification (Köppen, 1918). These stations comprised several climatic regions of Australia with various rainfall distributions. Darwin and Weipa are composed of

tropical savanna climate with different wet and dry seasons with an annual rainfall of about 1,110.2 mm and over 2000 mm respectively. Kowanyama also experiences tropical savanna climate with the wet season running from December to April and is characterised by frequent torrential downpours and high humidity.

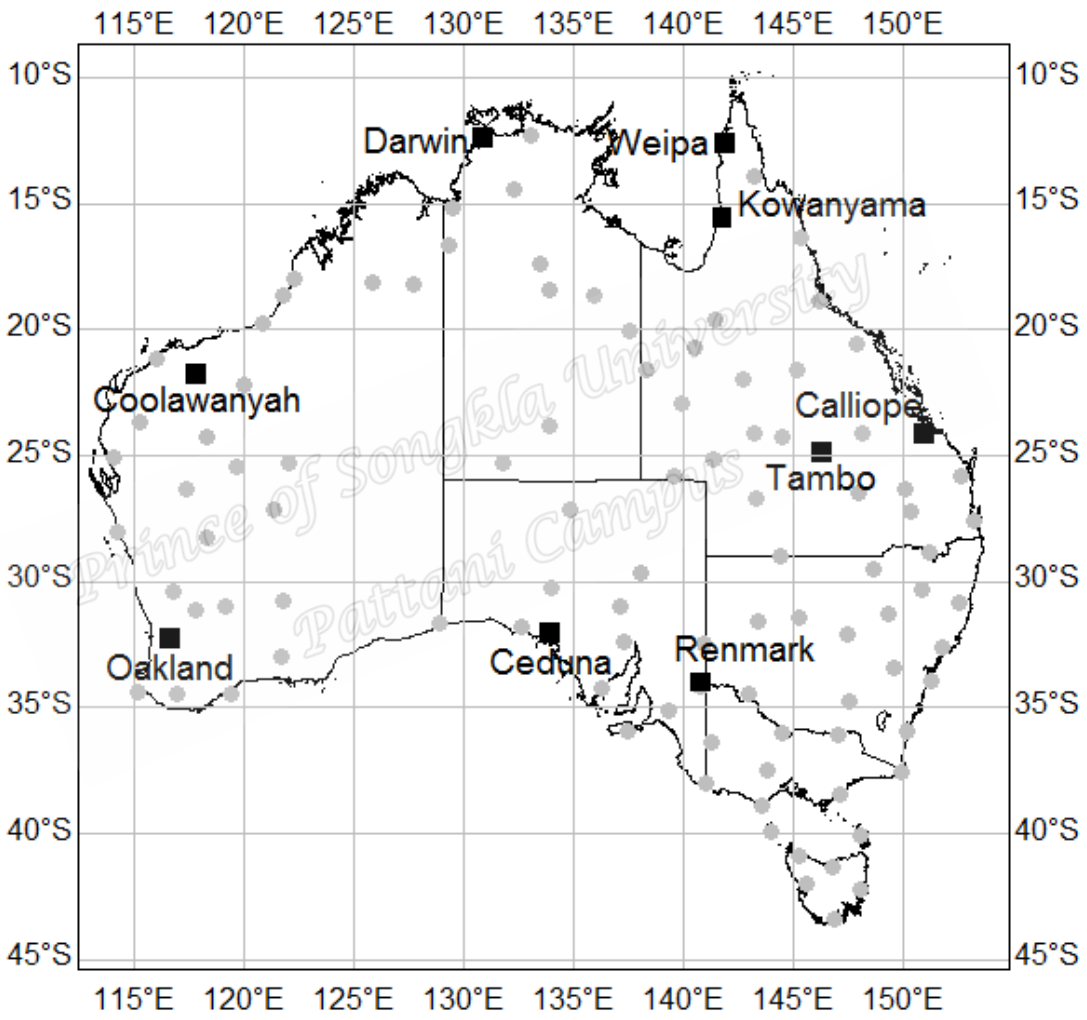


Figure 3.1 Location of the stations used for the study; with the nine selected named and designated using black squares while the grey dots signify the remaining stations

Table 3.1 Summary statistics of the 5-day rainfall for the dataset (1950-2013), for the all the stations

Station Name	Station ID	Lat. (° S)	Lon.(° E)	Elev.(m)	maximum	Mean	Median	% Missing
Hall Creek Airport	2012	18.23	127.66	422	94.80	1.58	0.00	1.8
Fossil Downs	3027	18.14	125.78	120	73.40	1.59	0.00	0.3
Bonney Downs	4006	22.18	119.94	480	69.04	0.91	0.00	1.8
Mandora	4019	19.74	120.84	8	66.56	1.07	0.00	11.5
Coolawanyah	5001	21.8	117.81	360	80.6	1.07	0.00	0.3
Mardie	5008	21.19	115.98	11	69.04	0.86	0.00	0.3
Callagiddy Station	6008	25.05	114.03	25	55.90	0.59	0.00	0.5
Lyndon	6029	23.64	115.24	210	71.20	0.60	0.00	1.3
Mileura	7049	26.37	117.33	400	42.84	0.61	0.00	1.1
Mount Vernon	7059	24.23	118.24	350	81.72	0.68	0.00	2.9
Neds Creek	7103	25.48	119.65	600	52.68	0.65	0.00	0.8
Challa	7197	28.28	118.31	400	36.64	0.60	0.00	0.0
Balline	8004	28.01	114.22	100	53.64	1.15	0.00	0.2
Elena	8230	30.38	116.72	290	32.24	0.79	0.00	0.1
Bangalup	9506	34.47	116.92	180	25.92	1.92	0.88	0.1
Cape Leeuwin	9518	34.37	115.14	13	36.04	2.62	1.12	0.0
Peppermint Grove	9594	34.44	119.36	60	43.02	1.88	0.76	0.5
Oaklando	10620	32.30	116.64	240	37.70	1.37	0.00	0.8
Eucla	11003	31.68	128.9	93	23.88	0.82	0.16	3.6
Bullfinch	12011	30.99	119.11	362	27.80	0.86	0.00	0.1
Bulong	12013	30.75	121.75	380	47.32	0.76	0.00	0.2
Wonganoo	12108	27.14	121.34	500	39.36	0.64	0.00	0.1
Glen- Ayle	13002	25.27	122.04	480	48.00	0.71	0.00	0.6
Darwin Airport	14015	12.42	130.89	30	162.68	4.81	0.20	0.0
Oenpelli	14042	12.33	133.06	7	123.38	4.08	0.00	1.8
Waterloo	14815	16.63	129.33	140	162.68	1.98	0.00	2.5
Katherine Council	14902	14.46	132.26	107	88.78	2.84	0.00	1.5
Avon Downs	15005	20.03	137.49	205	70.80	1.02	0.00	6.2
Helen Springs	15015	18.43	133.88	288	148.00	1.41	0.00	3.4
Brunette Downs	15085	18.64	135.94	218	53.20	1.24	0.00	0.2
Newcastle Waters	15086	17.38	133.41	209	73.70	1.55	0.00	3.4
Curtin Springs	15511	25.31	131.76	490	33.44	0.64	0.00	7.3
Alice Springs Airport	15590	23.80	133.89	546	52.40	0.78	0.00	0.0
Woomera (Arcoona)	16000	31.02	137.05	125	37.80	0.45	0.00	0.1
Port Augusta (C.W)	16005	32.40	137.23	179	32.00	0.66	0.00	0.4
Tarcoola (Mulgathing)	16031	30.24	133.99	198	28.70	0.50	0.00	0.7
Todmorden	16047	27.14	134.76	200	44.92	0.50	0.00	1.3
Marree Comparison	17031	29.65	138.06	50	33.40	0.47	0.00	0.2
Penong (Penalumba)	18002	31.83	132.63	44	15.56	0.81	0.00	1.0
Ceduna Amo	18012	32.13	133.7	15	21.02	0.80	0.16	0.3
Tumby Bay	18091	34.26	136.21	45	30.48	0.86	0.24	0.1
Murrays Lagoon	22806	35.93	137.29	29	20.34	1.46	0.60	1.6
Renmark Irrigation	24003	34.17	140.75	20	21.92	0.73	0.08	0.0
Murray Bridge Comp	24521	35.12	139.26	33	23.04	0.97	0.30	0.0
Coen Post Office	27005	13.95	143.2	199	94.00	3.37	0.03	1.6
Weipa Eastern Avenue	27042	12.63	141.88	20	116.76	5.25	0.16	2.2
Cloncurry Mcillwraith	29008	20.71	140.52	196	90.48	1.42	0.00	7.0
Kowanyama Airport	29038	15.48	141.75	10	133.00	3.57	0.00	0.3
Whyanbeel Valley	31062	16.39	145.35	6	252.44	8.01	2.12	0.7
Bambaroo	32001	18.88	146.17	13	173.30	5.18	0.76	0.2
Burdekin Shire Council	33001	19.58	147.41	11	77.26	2.83	0.00	0.3
Collinsville Post Office	33013	20.55	147.85	196	75.96	1.98	0.00	0.1
Giligulgul	35029	26.36	150.05	318	52.02	1.76	0.00	0.0
Springsure Comet	35065	24.12	148.09	345	64.32	1.99	0.00	0.2
Springsure Comet	35065	24.12	148.09	345	64.32	1.99	0.00	0.2
Tiree	36046	21.63	145.12	282	50.44	1.36	0.00	0.2

Continuation of **Table 3.1**

Station Name	Station ID	Lat. (° S)	Lon.(° E)	Elev.(m)	Maximum	Mean	Median	% Missing
Urandangi	37043	21.61	138.31	174	46.60	0.90	0.00	0.1
Boulia Airport	38003	22.91	139.9	162	48.46	0.74	0.00	0.1
Roseberth Station	38020	25.79	139.59	47	32.14	0.45	0.00	0.7
Calliope Station	39020	24.02	150.97	58	104.94	2.31	0.00	0.4
Mount Cotton West	40141	27.62	153.2	148	155.72	3.64	0.84	1.0
Texas Post Office	41100	28.85	151.17	295	43.68	1.91	0.24	0.0
Mitchell Post Office	43020	26.49	147.98	337	52.00	1.54	0.00	0.1
Hungerford	44181	29.00	144.41	139	40.58	0.92	0.00	1.6
Eromanga	45006	26.67	143.27	156	54.20	0.87	0.00	2.5
Wilcannia (Reid Street)	46043	31.56	143.37	75	33.72	0.77	0.00	0.1
Collarenebri	48031	29.54	148.58	145	94.52	1.49	0.00	0.2
Cobar (Tambua)	48115	31.42	145.25	240	33.74	1.03	0.00	0.0
Euston (Beningtoon)	49023	34.45	142.91	64	32.00	0.90	0.00	0.0
Tottenham (Burdenda)	50011	32.13	147.41	198	38.20	1.39	0.00	0.1
Barraba (Neranghi)	54023	30.29	150.81	563	44.00	2.00	0.32	0.1
Bellbrook (East Street)	59000	30.81	152.51	95	80.56	3.26	0.80	1.0
Bathurst Agricultural	63005	33.43	149.56	713	34.18	1.86	0.52	0.1
Coonabarabran	64008	31.27	149.27	505	61.30	2.23	0.30	0.1
Randwick	66052	33.91	151.24	75	84.00	3.44	1.02	0.0
Moruyua Heads Pilot	69018	35.91	150.15	17	86.46	2.807	0.64	0.0
Hume Reeservoir	72023	36.10	147.03	184	29.24	1.967	0.66	0.0
Old Junee (Millbank)	73025	34.79	147.56	270	28.72	1.479	0.40	0.1
Kaniva	78078	36.37	141.24	142	18.60	1.23	0.44	0.1
Patho West	80044	35.99	144.43	90	29.08	1.06	0.20	0.1
Gabo Island Lighthouse	84016	37.57	149.92	15	61.24	2.58	1.04	0.1
Giffard	85033	38.42	147.09	15	41.40	1.61	0.72	0.2
Ballarat Aerodrome	89002	37.51	143.79	435	50.10	1.88	0.96	0.0
Cape Otway Lighthouse	90015	38.86	143.51	82	24.40	2.57	1.64	0.1
Nelson	90059	38.03	141.02	20	36.36	2.25	1.32	0.0
Frankford	91033	41.32	146.74	240	4.44	2.96	1.40	0.0
South Forest	91034	40.88	145.22	60	26.40	2.94	1.92	0.2
Swansea (Kelvedon)	92018	42.20	148.04	6	44.20	1.62	0.56	0.6
Hastings Chalet	94027	43.41	146.87	35	50.22	3.77	2.32	0.2
Lake Margaret Dam	97006	41.99	145.57	665	56.36	9.41	7.80	2.3
Harwell	98003	39.92	143.93	74	26.28	2.87	1.88	0.6
Flinders Island Airport	99005	40.09	148	9	27.52	2.04	1.06	1.1
Trayning	10126	31.12	117.79	290	27.48	0.89	0.08	0.1
Ayrshire Downs	37001	21.97	142.72	188	71.12	1.12	0.00	0.3
Bauple	40013	25.82	152.62	71	79.92	3.04	0.68	0.2
Belah Park	42002	27.21	150.28	292	38.72	1.62	0.00	0.4
Bidyadanga	3030	18.68	121.78	11	106.62	1.5	0.00	2.1
Broome Airport	3003	17.95	122.24	7	132.92	1.7	0.00	0.1
Clarence Town	61010	32.59	17.95	24	77.48	2.95	0.88	0.0
Currawilla Station	38007	25.14	141.35	93	45.96	0.66	0.00	0.3
Isisford Post Office	36026	24.26	144.44	203	42.06	1.29	0.00	0.1
Legune	14803	15.21	129.45	26	109.02	3.14	0.00	11.6
Mutooroo	20017	32.45	140.92	210	48.76	0.64	0.00	0.3
Noonbah	37098	24.11	143.19	180	51.80	0.97	0.00	0.0
Salmon Gums Resident	12071	32.99	121.62	249	26.08	0.98	0.26	0.9
Tambo Post Office	35069	24.88	146.26	395	49.88	1.49	0.00	0.0

During the dry season, almost no rainfalls and days are warm to hot but humidity is low and the nights can get quite cold. Tambo experiences a hot semi-arid climate while Renmark and Ceduna have cold semi-arid climates and these areas receive little rainfall throughout the year. For instance, Ceduna has an annual rainfall of about 300

mm with hot, dry summers and cold, slightly wetter winters and July is the wettest month while Oakland experiences temperate oceanic climate.

In comparing the 5-day mean rainfall station wise revealed substantial spatial variability in the study area. Among the wettest stations, Darwin Weipa and Kowanyama show similar patterns with wet months from December to March (Figure 3.2). In Oakland and Ceduna, the wettest months were observed from May to August. However, Renmark experiences uniform rainfall throughout the year. Analysis of the occurrence probability patterns of the mean rainfall amount for the stations revealed three main groupings. Darwin Weipa and Kowanyama can be put in a group. They show similar rainfall probability of occurrence with mean lying between the probability of 0.44-0.56 and higher probability values are observed between December and March. However, the rainfall occurrence probability is mostly below 0.2% throughout the remaining periods of the year. The probability of rainfall occurrence in Coolawanyah, Tambo and Calliope can also be put into a group. The mean ranges from 0.46-0.76. The occurrence probability is quite stable relative to the other stations.

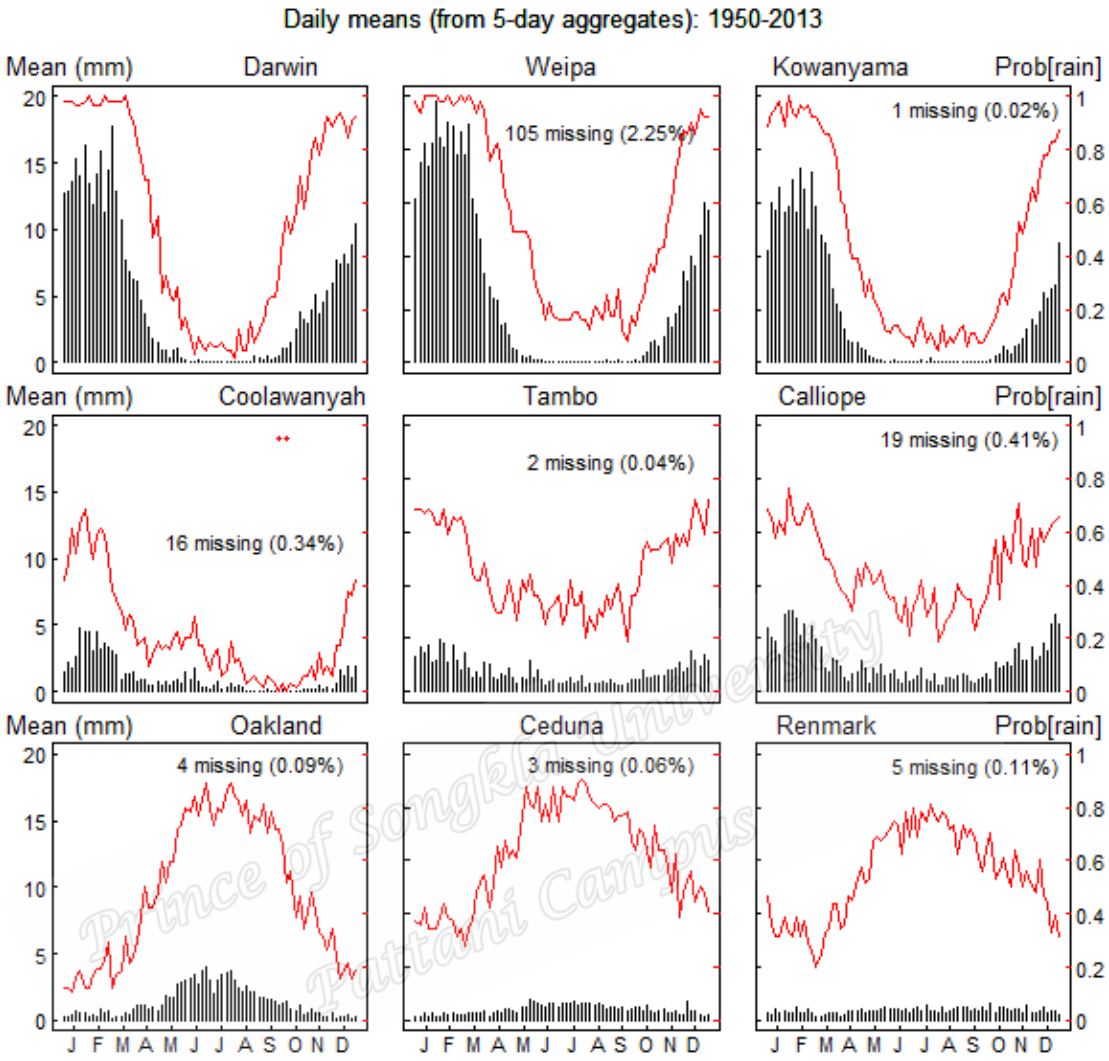


Figure 3.2 The probability of occurrence 5-daily rainfall means

Oakland, Ceduna and Renmark can also be put in a group. They have similar probability patterns with mean ranging from 0.48 to 0.62. The average occurrence probability is mostly above 0.2% throughout the year with higher probabilities values occurring between May and August.Relative to the 5-day means, station-wise

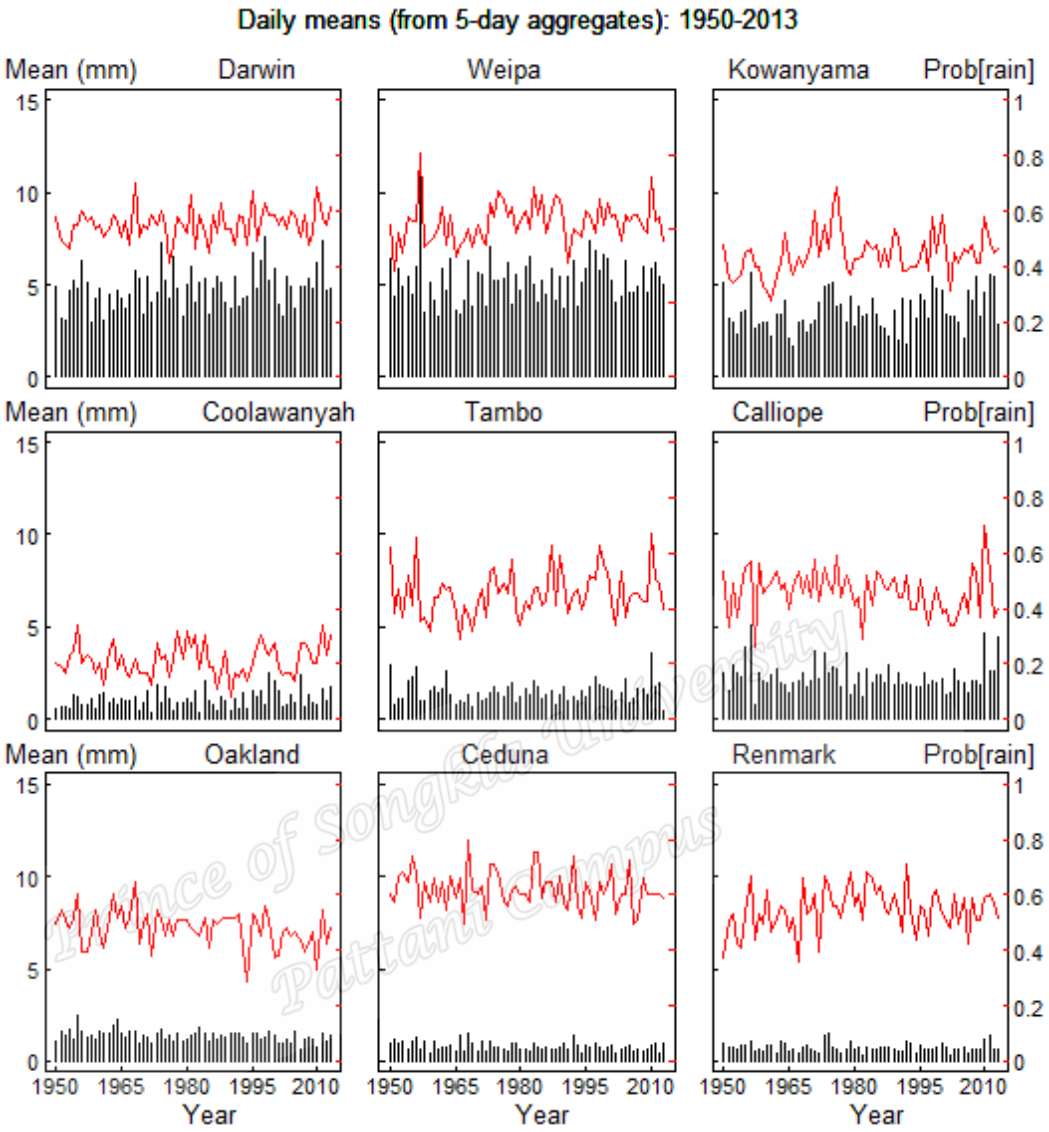


Figure 3.3 Annual rainfall and occurrence probability patterns

comparison of the annual mean rainfall did not show any considerable spatial variability in the study area throughout the 64 years. There were no apparent patterns of the mean annual rainfall and little variations in the occurrence probabilities (Figure 3.3). The annual means were higher in Weipa, Darwin and Kowanyama relative to the other stations. Low annual rainfall means can be observed in Oakland, Ceduna and Renmark.

CHAPTER 4

Classification and the modelling of daily rainfall Amount in Australia

This chapter involves the classification and modelling of the spatial and temporal variability of the 5-daily rainfall amount in Australia during 1950-2013 using factor analysis and linear regression modes. Gamma GLMs and multiple regression are also fitted to describe the non-zero amount. The result of this study also appears in the author's first and second manuscripts (Appendix I and II).

4.1 Factor analysis results

To begin with, factor analysis (equation 2.3) was applied to the correlation matrix of the 5-day average rainfall in all the meteorological stations. The Factor analysis model grouped Australia into eight geographical rainfall regions, and these regions are made up of groups of stations that load with just one factor. Table 4.1 reveals the loadings of the factors structured ina downward order within each factor and the uniqueness values without the mixed factors. The analysis was conducted on several factors, but the factor loadings showed that eight factors were adequate and accounted for 52% of the observed correlations in the data with large correlation values of at least 0.34.

Moreover, 15 stations out of the 92 correlated with factor 1 (F1), 13 with factor 2 (F2), 10 with factors 3, 4 and 6 (F3, F4, and F6). Also, correlated 7 with factor 5 (F5), 6 with factor 7 (F7), and 4 with factor 8 (F8) while 17 stations correlated to more than one factor in this study (mix-factors). Figure 4.1 depicts the map of Australia and the

Table 4.1 The factor loading of the 92 stations with the identified dominating factors in bold

Stations	Factor Loadings								Uniqueness
	F1	F2	F3	F4	F5	F6	F7	F8	
15085	**0.828**								0.381
15086	**0.810**								0.401
15015	**0.788**								0.464
37043	**0.744**			0.115	0.163	-0.132			0.435
14815	**0.731**	-0.145				0.101		0.109	0.405
15005	**0.728**				0.127	-0.106			0.510
2012	**0.703**				-0.119	0.205			0.400
14902	**0.690**	-0.145			-0.103		0.159	0.105	0.309
29008	**0.620**			0.119	0.116	-0.117	0.196		0.489
38003	**0.602**	0.111		0.192	0.156	-0.146			0.516
15590	**0.584**	0.316		0.106			-0.214	-0.107	0.512
3027	**0.562**	-0.114			-0.118	0.286			0.489
14015	**0.512**	-0.118			-0.219	0.107	0.205	0.104	0.349
38020	**0.347**	0.260				-0.114			0.746
14042	**0.499**	-0.112		-0.122	-0.178	0.107	0.325	0.122	0.280
18012	-0.204	**0.804**	0.148						0.353
16005		**0.804**							0.461
18002	-0.234	**0.775**				0.106			0.449
16031		**0.746**	-0.126			0.127			0.536
16000		**0.739**	-0.124						0.533
18091	-0.205	**0.712**	0.221						0.386
11003	-0.205	**0.655**				0.212			0.616
24003		**0.650**	0.176	-0.107				0.254	0.364
17031	0.198	**0.646**	-0.101						0.520
20017		**0.593**						0.187	0.566
16047	0.317	**0.518**					-0.161		0.604
46043		**0.434**		0.163				0.321	0.495
48115		**0.346**		0.249				0.259	0.578
98003			**0.886**						0.188
90015			**0.874**						0.210
91034	0.107		**0.860**				-0.102		0.293
91033	0.111		**0.772**						0.387
90059	-0.103	0.228	**0.755**						0.232
99005			**0.752**					0.207	0.382
94027		-0.184	**0.714**				-0.124	-0.162	0.557
97006		-0.174	**0.688**		-0.121		-0.126	-0.279	0.549
89002		0.197	**0.661**					0.290	0.263
92018			**0.444**					0.253	0.728
35029				**0.806**					0.414
41100				**0.803**				0.152	0.395
43020				**0.792**					0.377
54023				**0.742**			-0.173	0.297	0.373
48031				**0.721**				0.249	0.357
35065	F5	F6	0.115	**0.682**			0.298	-0.128	0.411
39020				**0.580**			0.344		0.534
40141	-0.145		-0.124	**0.559**			0.207		0.594
59000		-0.118	-0.254	**0.552**		0.109		0.199	0.520
44181	0.144	0.225		**0.380**		-0.112		0.168	0.569
10620		-0.169			**0.918**				0.282
8230		-0.126			**0.849**	0.137			0.333
8004					**0.809**	0.102			0.336
9518					**0.775**	-0.156			0.299
12011					**0.736**	0.265			0.420
9594		0.165			**0.554**				0.563
7059						**0.724**			0.511
7103		0.113				**0.718**			0.514

Continuation of **Table 4.1**

Stations	Factor loading								Uniqueness
	F1	F2	F3	F4	F5	F6	F7	F8	
12011					**0.736**	0.265			0.420
9594		0.165			**0.554**				0.563
7059						**0.724**			0.511
7103		0.113				**0.718**			0.514
12108		0.171				**0.674**			0.538
13002		0.183			-0.153	**0.637**			0.588
4006	0.175				-0.156	**0.628**			0.536
7049					0.267	**0.615**			0.526
5001	0.134				-0.116	**0.601**			0.572
5008						**0.594**			0.630
6029					0.129	**0.554**	0.107		0.651
4019	0.290				-0.101	**0.487**			0.615
33001		0.128		0.152			**0.741**		0.394
33013				0.284			**0.711**	-0.118	0.401
32001	0.163	0.125	-0.108				**0.698**		0.353
31062	0.181	0.129	-0.150	-0.201			**0.645**		0.471
27042	0.328			-0.132	-0.188		**0.582**		0.308
27005	0.261			-0.139			**0.477**		0.401
73025		0.146	0.218	0.124				**0.592**	0.352
69018			-0.193	0.268				**0.542**	0.565
84016			0.114	0.116				**0.530**	0.661
80044		0.333	0.329					**0.455**	0.348

eight regions as determined by the factor model. The small size points of Figure 4.1 shows stations that correlated with only one factor and the large size point's shows stations that correlated with more than one factor. These regions are F1: North (N), F2: Central South (CS), F3: South-Southeast (SSE) and F4: Central East (CE). The remaining regions include F5: Southwest (SW), F6: Northwest (NW), F7: Northeast (NE) and F8: North-Southeast (NSE).

4.2 Annual rainfall patterns

The annual 5-day mean rainfall amount (black bars) and its probability of occurrence (red curve) in all the eight regions are shown in Figure 4.2. There are no clear annual patterns revealed by the graph. Among all the regions, the NE received the highest average of 4.43 mm d^{-1} followed by SSE with an average of 3.22 mm d^{-1} while the CE received an average of 2.08 mm d^{-1}. NSE, N and SW followed in that order with averages 1.98, 1.76 and 1.51 mm d^{-1} respectively. The NW received 0.79 mm d^{-1}

while the least of 0.67 mm $d^{-1}$was received in the CS. Similar annual probability patterns in some of the regions were also revealed by Figure 4.2.

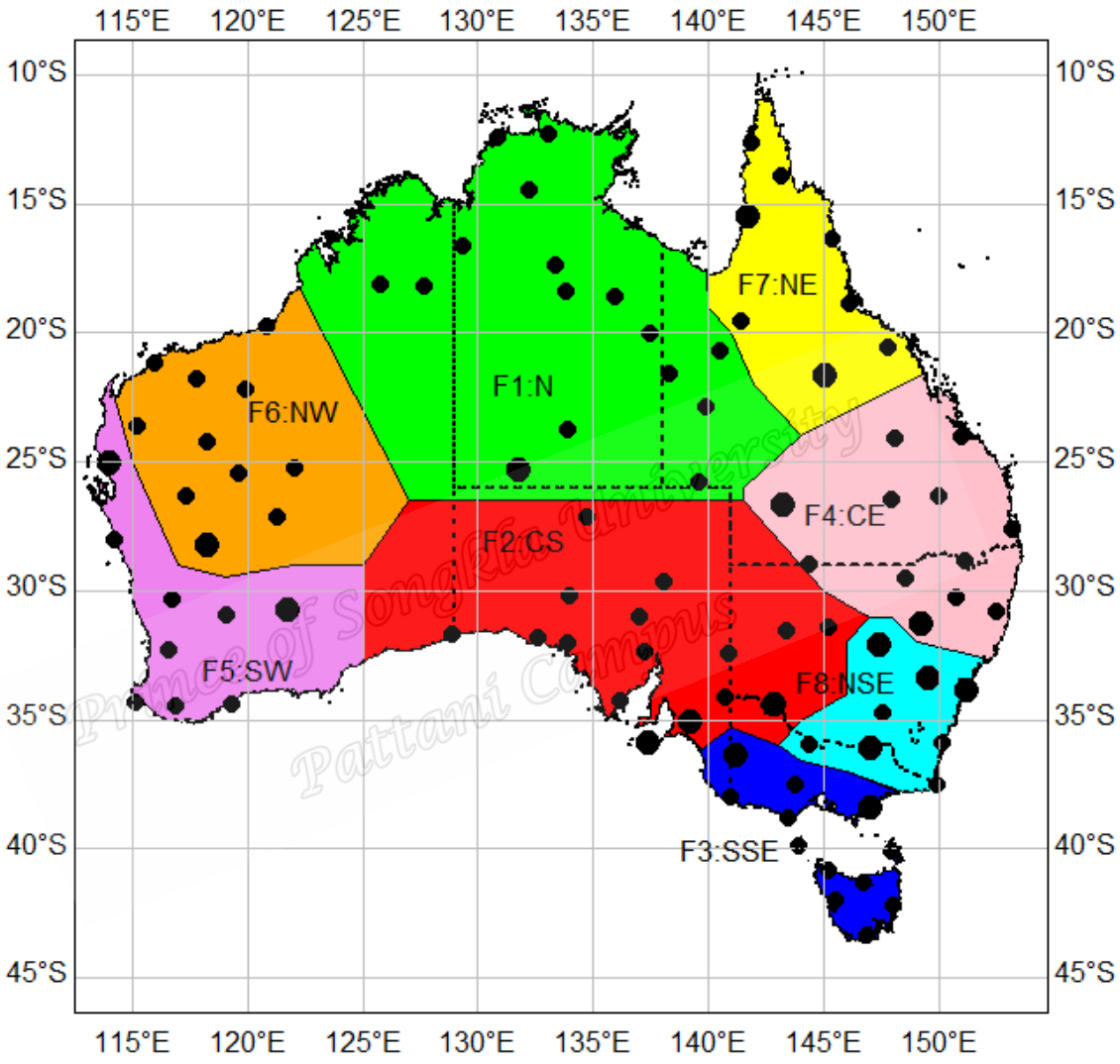


Figure 4.1 Map of rainfall factor regions of Australia during 1950-2013

The black vertical bars denote the mean annual rainfall amount while the red curve denotes the probability of its occurrence.These probability patterns show the likelihood of obtaining the corresponding 5-day rainfall in that particular year. The analysis of the probability curve revealed that SSE was anticipated to experienced

rainfall throughout the year, whereas SW, NSE, CS and CE showed varying probabilities between 0.8-1.0. However, N and NE had a probability between 0.59 and 1.0. The least annual rain probability of 0.59 was observed with high variations throughout the 64 years in NW.

4.3 Seasonal rainfall patterns

Figure 4.3. shows the observed seasonal rainfall amount (black bars) and the probability of occurrence (red curve) in Australia from 1950-2013. Each bar corresponds to 5-day average in each particular month. The study revealed four distinct seasonal rainfall patterns in Australia, but among these regions, similar patterns with different seasonal rainfall amount were seen at CS and NSE, N and NE, NW and CE, SW and SSE.

It is remarkable to note that, while these seasonal rainfall patterns had very low variations in the CS and NSE, the remaining regions experienced high seasonal variations. NE received the highest maximum seasonal rainfall of 14.50 mm $d^{-1}$and this occurred in February with an average of 4.43 mm d^{-1} followed by the SSE with 3.22 mm d^{-1}average.

Also, the CE received an average of 2.08 mm d^{-1} while NSE, N and SW followed with averages of 1.98, 1.76, and 1.51 mm d^{-1} respectively. Both the NW and CS recorded average values less than 1 mm d^{-1}. The probability of rainfall was approximately 1.0 throughout the season in SSE while it varied from 0.8-1.0 in CE, CS, NSE and SW. Seasonal rainfall probability of 1.0 was observed in the N and NE from December- March. However, this probability dropped to 0.3 in the N in August

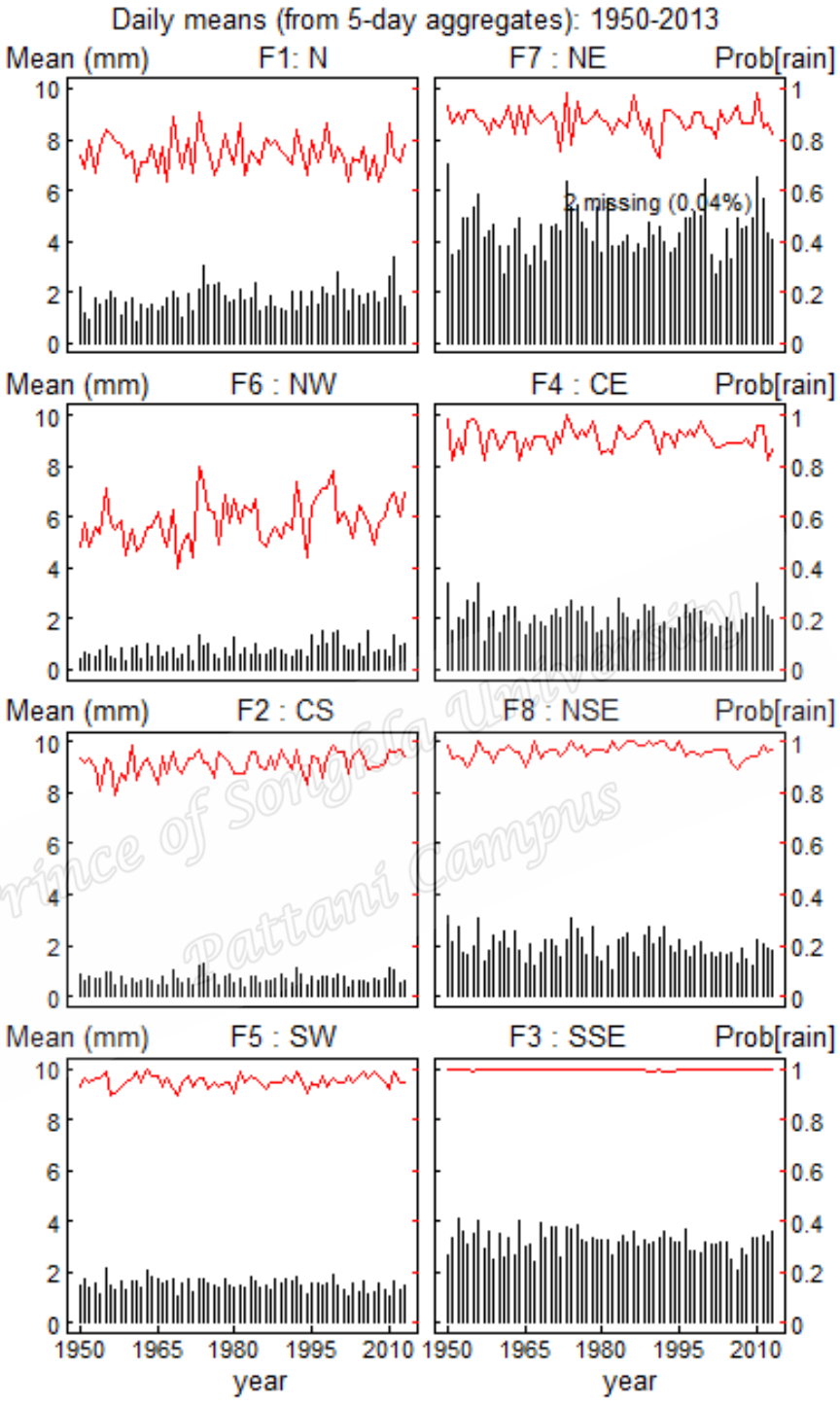


Figure 4.2 Annual rainfall amount and probability of occurrences in Australia during 1950-2013.

and 0.6 in NE during September. In NW, rainfall probability varied from 0.2-0.9. Considering all the eight-factor regions, N, CE, NW and NE receive a substantial

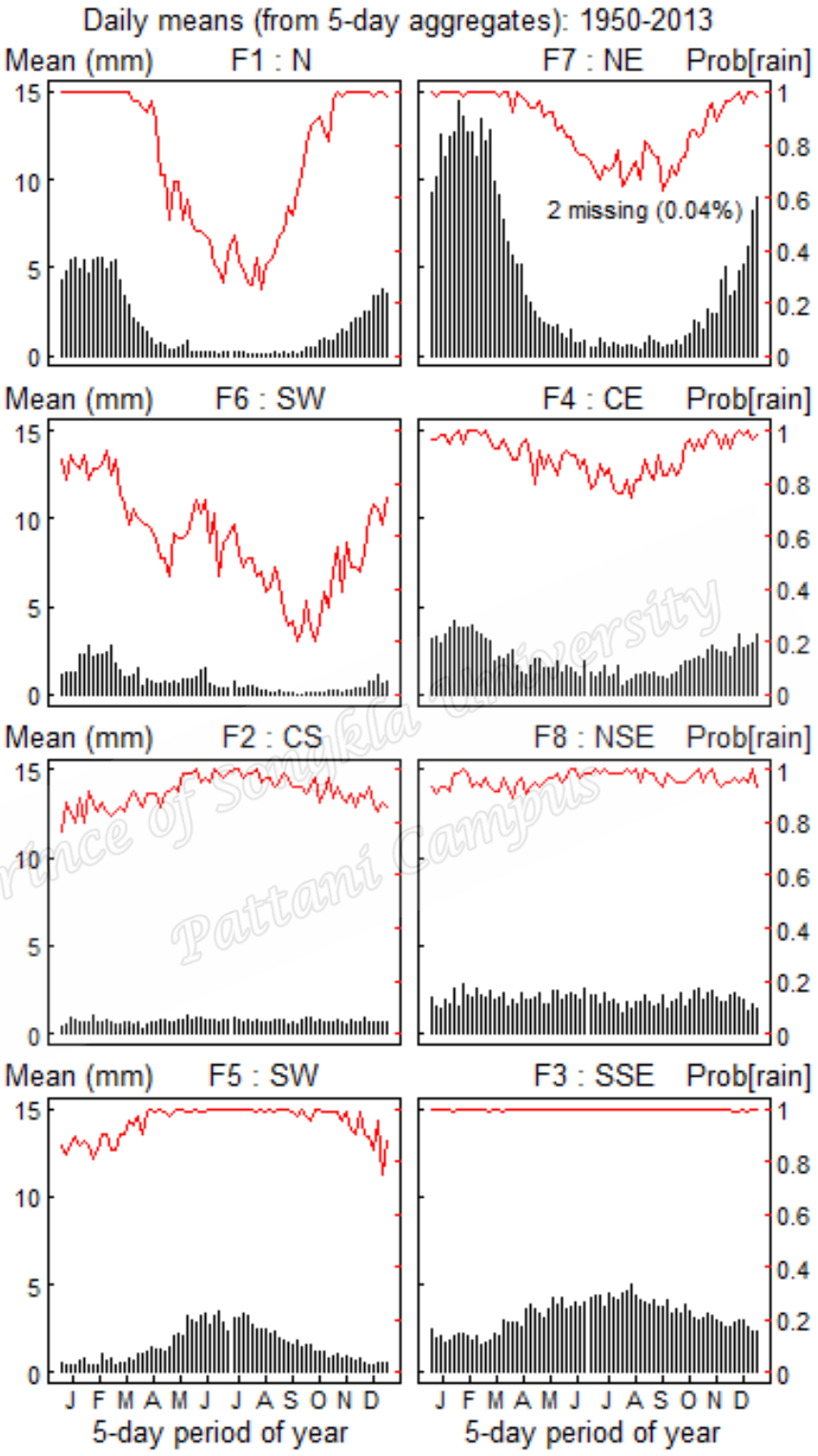


Figure 4.3 Seasonal rainfall amount and probability of occurrences in Australia during 1950-2013.

amount of rainfall from December-March (wettest months) while SSE and SW have their wettest months between August and September.On the other hand, uniform monthly rainfall distributions were observed in CS and NSE.

4.4 Annual rainfall trends in Australia

The analysis of trends in the 64 years annual rainfall time series by the linear regression model (equation 2.4) showed negative and positive slopes for the regions (Table 4.2). The CE, CS, NSE, SW and SSE regions were seen to be having negative slopes, which shows declining trends (Figure 4.4). Significant declining trends were apparent in the NSE and SW regions. Conversely, significant increasing annual rainfall trends were apparent in the N and NW regions of Australia while a non-significant rising trend at 95% confidence level was apparent in the NE region. Contrary to the N and the NW regions, the NE region is mostly liable to extreme weather conditions, for example, tropical cyclones that lead to lots of convective clouds. It explains more significant part of cloud cover in NE as related to that in the N and NW regions. It possibly will also explain part of the increasing rainfall trend in this region since cloud formation and rainfall are interrelated.

4.5 The analysis of the fitted gamma GLM models

The result of the fitted Gamma model is presented graphically in Figure 4.5. The model (equation 2.9) was made up of 136 parameters which composed of 73 seasonal and 64

Table 4.2 The results of the analysis of annual rainfall trends

Region	N	NE	CE	NW	CS	NSE	SW	SSE
slope	0.0076	0.0065	-0.0029	0.0001	-0.0001	-0.0075	-0.0035	-0.0050
p-value	0.0302	0.8757	0.4725	0.0026	0.7898	0.0235	0.0444	0.0964

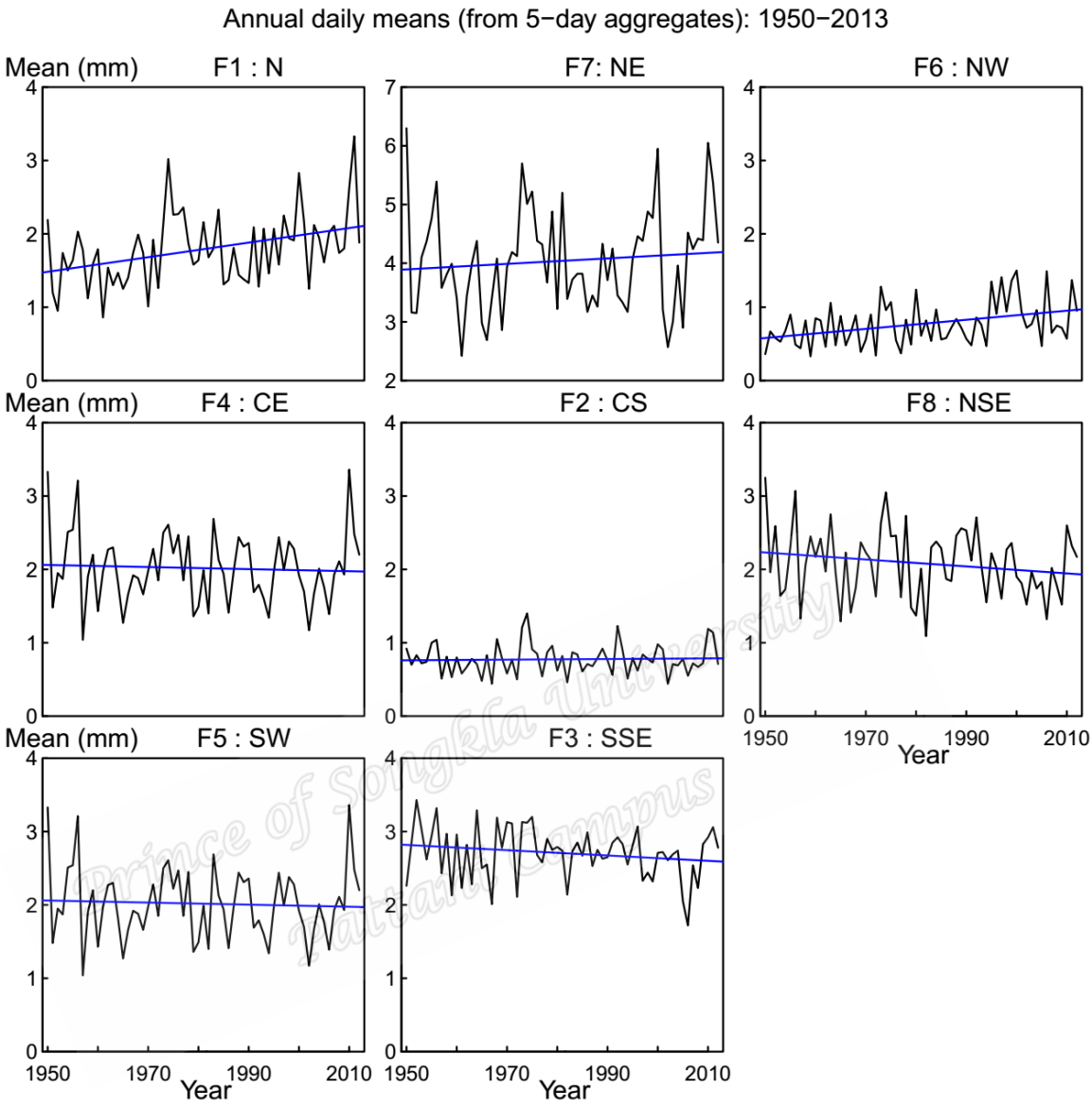


Figure 4.4 Trends in the annual rainfall in Australia

years effect. The predictors were seen to be significant in the models in the N, SSE, and CE but the models in SW, NW and NE were only significantly affected by the seasonal effect. In CS and NSE the model was seen to be significantly affected by the annual effect.

Figures 4.5 show the result of the fitted 5-day means rainfall amount. The top panel represents the models for N, NE, NW and the CE while the bottom panel represents

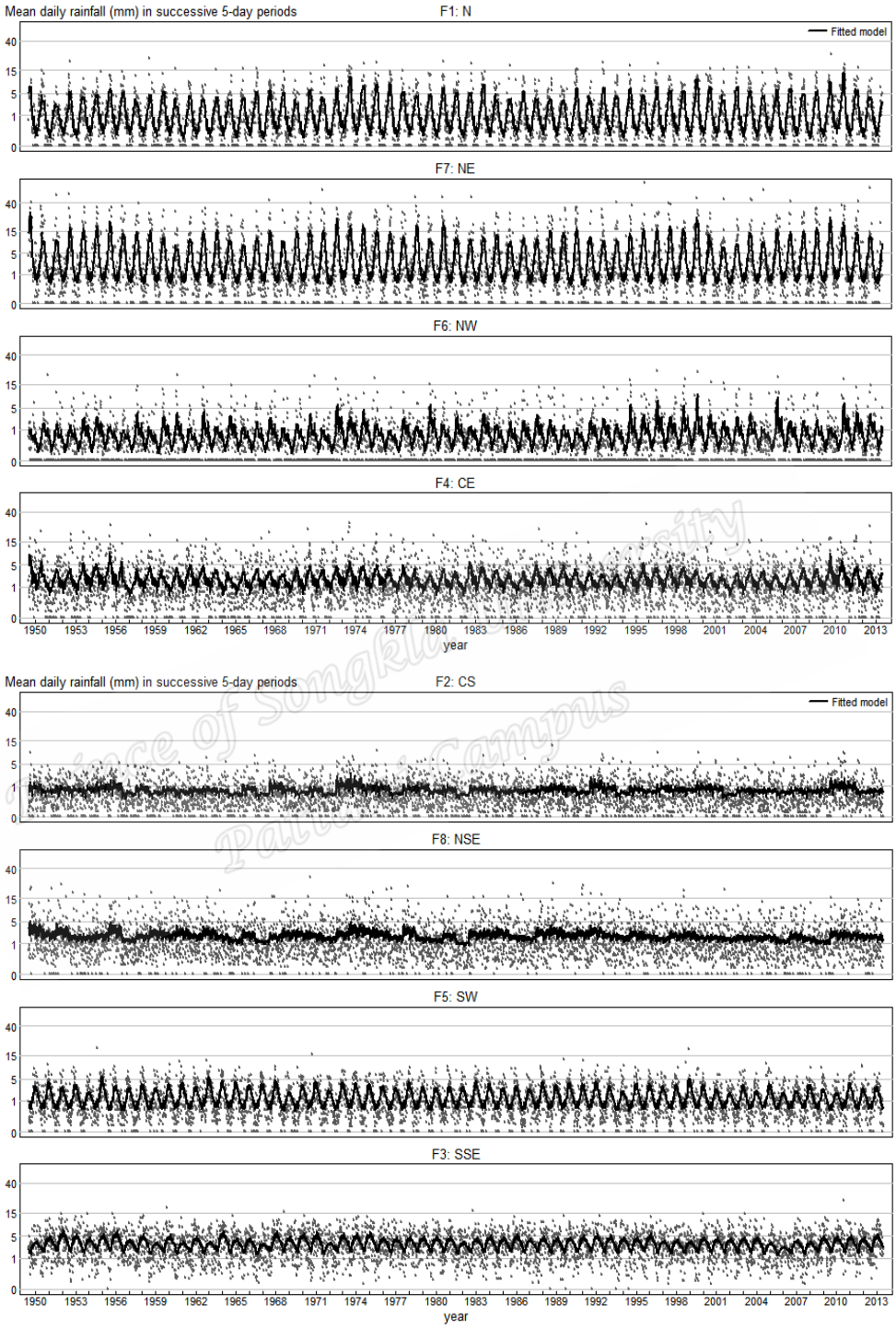


Figure 4.5 Fitted gamma model for the 5-daily rainfall means in Australia from 1950-2013

the models in the CS, NSE, SW and SSE. The grey points on both figures are the observed daily rainfall, and the black lines denote the fitted gamma model. The mean rainfall amount varies throughout the 64 years in all the eight regions with minimum variations observed in CS and NSE from 1950-2013. However, the models showed similar patterns in some of the regions, but the shape parameters and the means of the models were different. For instance, the mean rainfall patterns in N and NE were similar while that of NW was also similar to CE.

There was not much difference observed between the patterns in CS and NSE, and rainfall patterns in SSE and SW were similar. Besides, in N, NE, NW and CE, most of the annual rainfall was observed in the summer months. However, the CS and NSE experiences uniform rainfall throughout the season.

4.6 The adequacy of the fitted gamma models

The deviance residuals were plotted against the theoretical quantiles (Q-Q plots) to determine if the gamma GLM were appropriate for the fitted models (Figure 4.6). The black lines denote the residual plots, and the grey line denotes the line of best fit for the correct gamma model. Most of the residuals plots had low variations concerning the factor regions apart from little departures at the upper and lower tails of some of the models.

4.7 The 95% confidence intervals for the adjusted means

These delineated how the fitted rainfall means (seasonal and annual) varied from the overall mean between 1950-2013 in Australia (Figures 4.7 and 4.8). Figure 4.7 displays the95% confidence intervals for the adjusted means for the models in N, NE, NW and CE while figure 4.8 shows that in the CS, NSE, SW, and SSE. In both

graphs, it can be seen that all the models were within the 95% CI signifying that the models were represented quite well in all the regions but high variations in most regions. In the SW and SSE regions, seasonal rainfall rises sharply from February to June and declines from July to December.

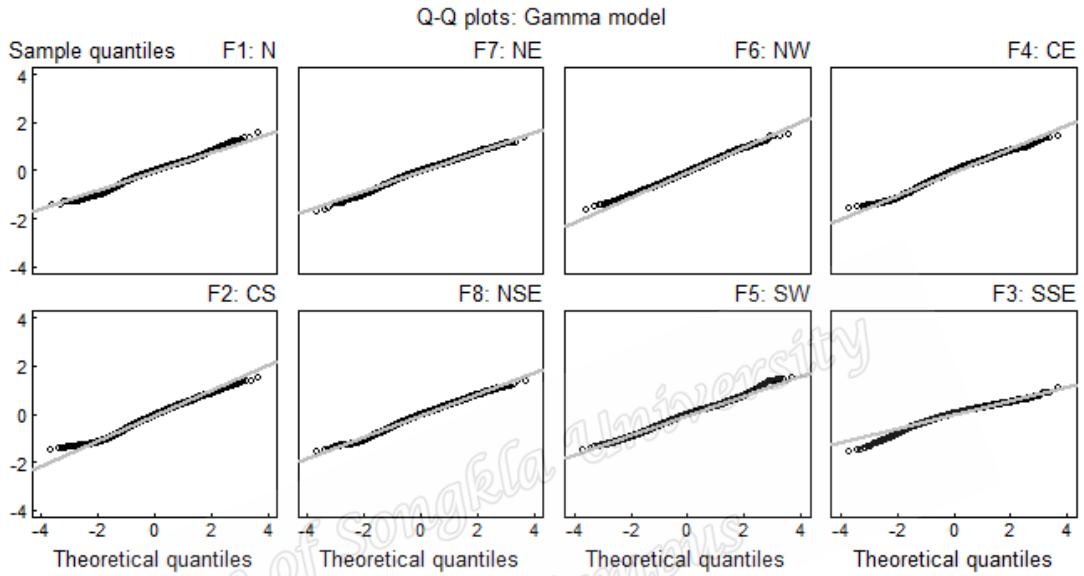


Figure 4.6 Plots of deviance residuals versus theoretical quantiles based on the gamma GLM for predicting the 5-day rainfall means in Australia during 1950-2013.

However, seasonal rainfall increases sharply from January to February where it attains the maximum, decreases to its minimum in August and increases gradually until December in the N and NE.The NW receives above average rainfall mostly between January and June then experiences below average until December while in the CE below average rainfall mostly occurs between April and October.However, no clear seasonal patterns were observed in the CS and the NSE regions. These seasonal patterns were significant in all the factor regions ($p< 0.001$) except for CS and NSE regions (first and second panels of Figure 4.8).

4.8 Comparing the Gamma GLM and the MR models in fitting models to the 5-day rainfall amount

The models for the Gamma GLM and the MR models were compared to ascertain which of the model fitted the data very well and in which area.In comparing the two methods, one major concern for the modelling of time series observations is the dependency between the response observations, which violate the assumption of independent errors (Figure 4.9). Figure 4.10 is the autocorrelation function (ACF) which is applied to assess the autocorrelation of mean period rainfall. The top panel (a) depicts the ACF in fitting the MR models, and the bottom panel (b) gives that of fitting the Gamma GLM models. Some of the sample lag values (the black vertical lines) are beyond the 95% confidence interval line (the horizontal dotted lines) which specifies significant dependencies among the response variables.

The dependences among the period rainfall in each region were minimised by adding the AR(1) term (Chatfield, 1996) to both the MR and Gamma GLM models (Figure 4.11). Most of the sample lag values are within the 95% confidence interval consequently signifying a drastic reduction of serial auto-correlation in the models.

The results of the pattern in fitting the MR and the Gamma GLM models to the 5-day rainfall amount for all the eight regions are shown in Figure 4.11. The grey dots represent the observed period's rainfall, while the black and red curves represent their estimated mean from models (2.5) and (2.6) respectively.Both models showed an obvious seasonal periodic pattern, and this is seen mainly in the N and NE regions where over 52% of the variations were explained by the models(coefficient of determination of 0.54 and 0.52 respectively). Moreover, the fluctuations and it

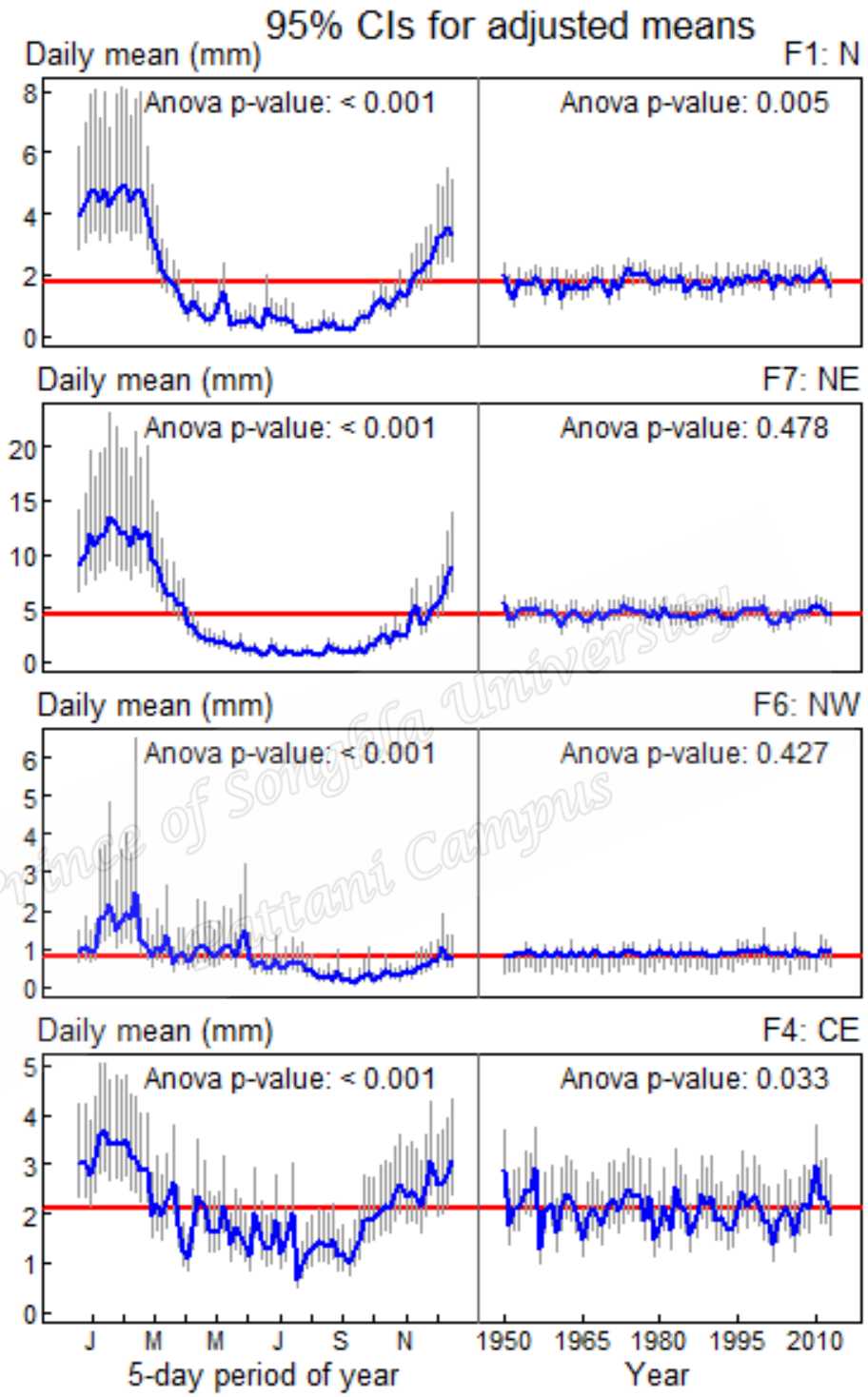


Figure 4.7 The 95% confidence intervals for the adjusted means for N, NE, NW and CE

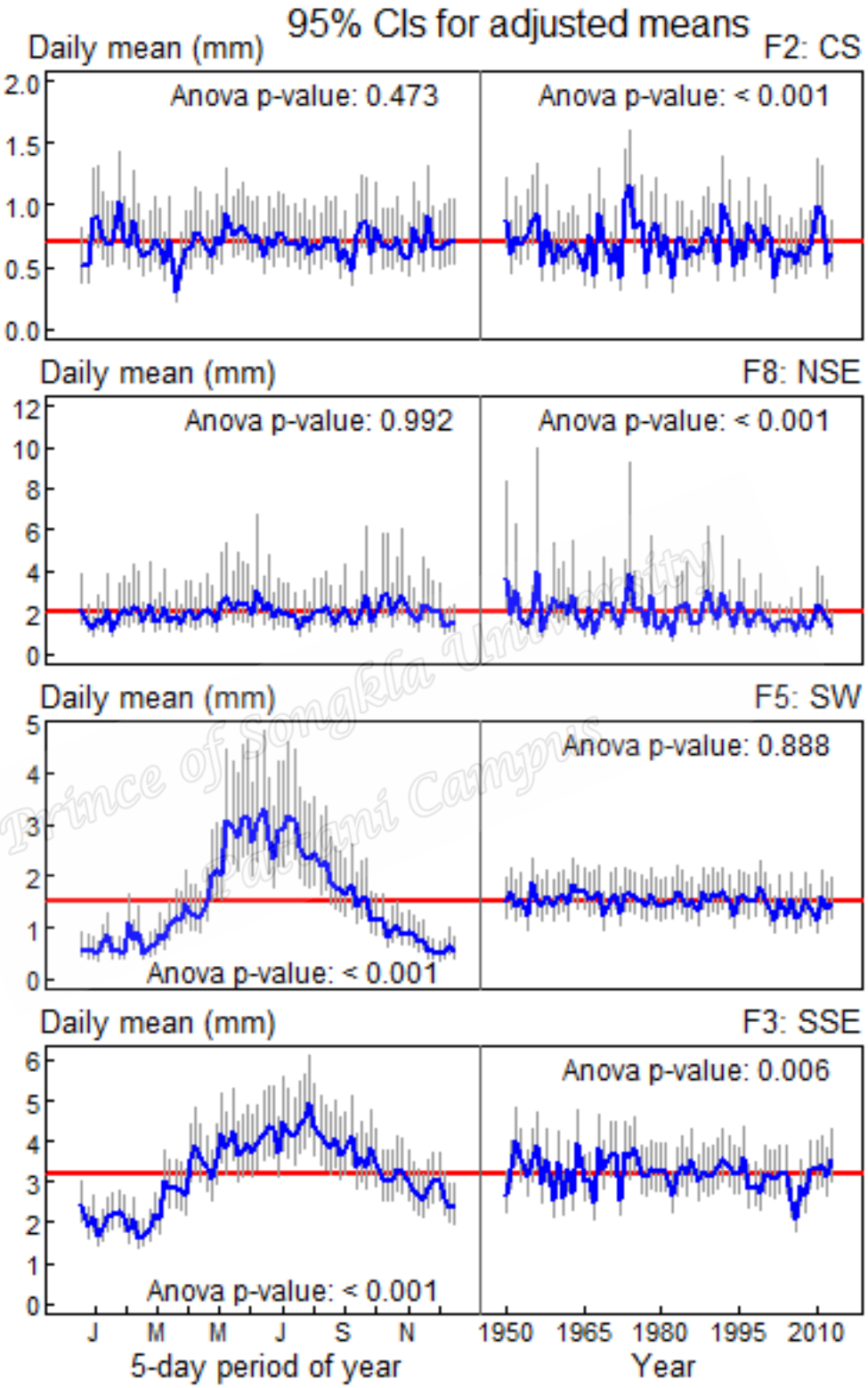
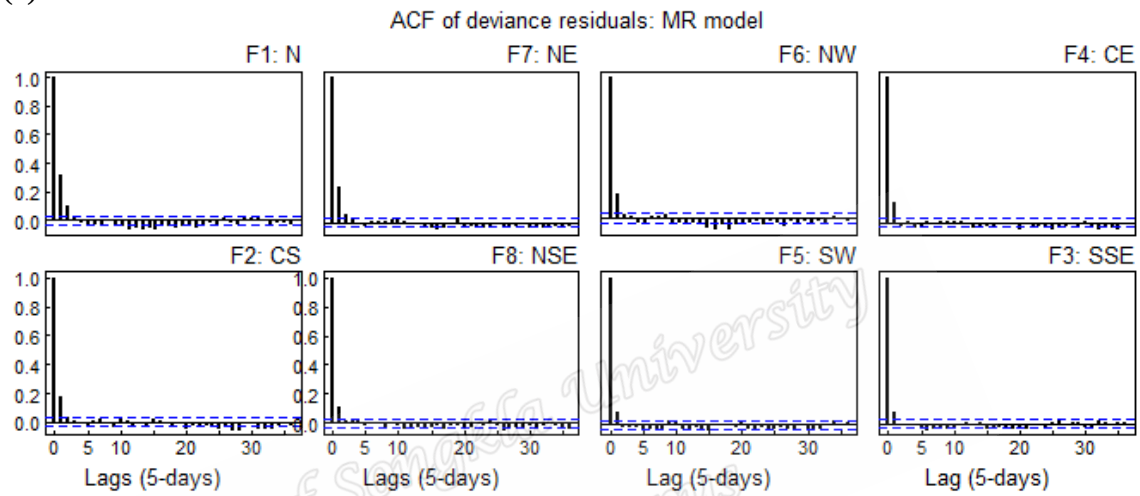


Figure 4.8 The 95% confidence intervals for the adjusted means for CS, NSE, SW and SSE.

timings in the data were well displayed by the models as well as the 64 years mean. However, the models underestimated the magnitude of the fluctuations.

Relatively, there is not much difference between the models from both methods in the remaining regions as shown in Figure 4.11

(a)



(b)

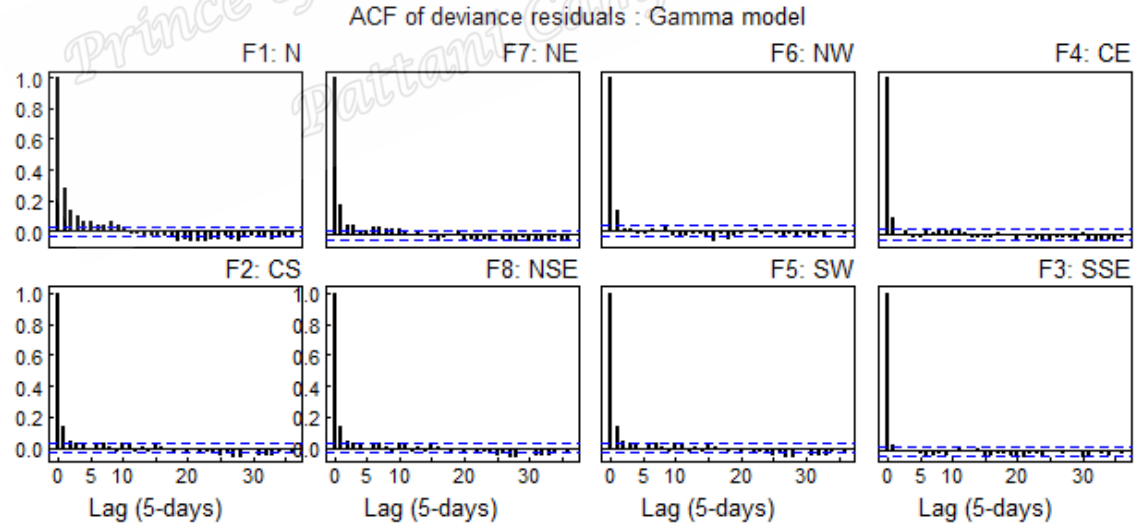


Figure 4.9 The autocorrelation function for the models before minimising serial correlation.

The eight-factor regions display various rainfall variability and trends for all the fitted models.The N and NE, which is in the tropical region of Australia, have similar period rainfall deviations so as in the NW and the CE. The pattern in CS and NSE was

also similar so as in the SW and SSE.The eight-factor regions show different rainfall variability and trends for all the fitted models.The N and NE, which is in the tropical region of Australia, have similar period rainfall variation so as in the NW and the CE. The pattern in CS and NSE was also similar so as in the SW and SSE.In general, the NE region relative to the other regions had the highest amount of period rainfall in the seasonal span, which can be due to a substantial number of temporal meteorological conditions such as fronts and low-pressure systems, which possibly can disturb rainfall in this area.

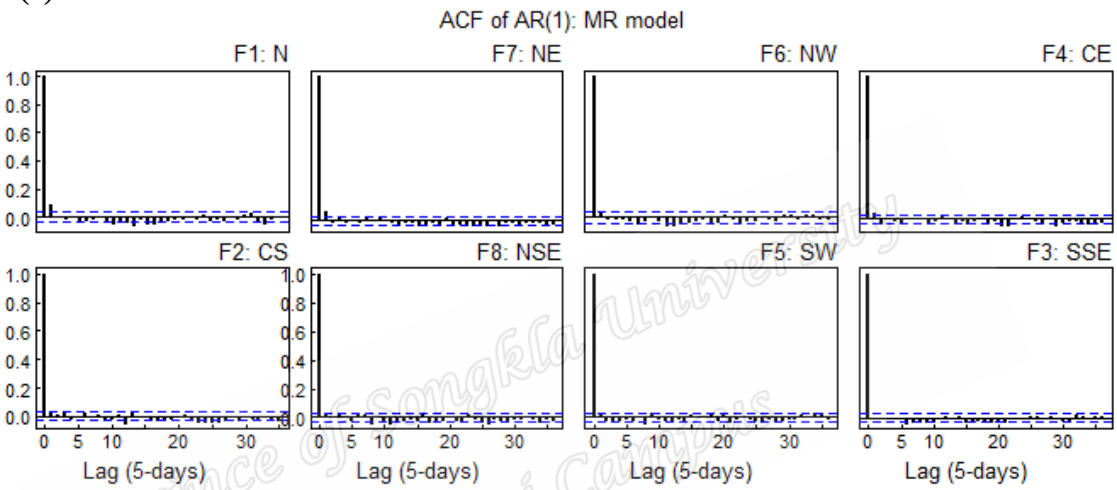
The fitted GLM and MR models were assessed by plotting residuals alongside the theoretical quantiles (Figure 4.12). The top graph shows that of the MR (response variable is the fourth root transformed) and the bottom shows the Gamma. Both models were filtered with AR(1) term.The figure demonstrates that the residuals plots had low variations concerning the expected line of best fit indicating that, both models fitted the data quite well in the factor regions except little departures at the upper and lower tails of some of the models. The results of the MR model seems to be superior to that of the Gamma GLM model in the N, NE and SW regions.

4.9 Modelling the 64 years mean seasonal and annual rainfall with multiple regression

Multiple regression, equation (2.5) was used to model the 64 years mean period rainfall (to show the monthly) and annual rainfall patterns for all the regions. This model was applied because it is one of the most straightforward methods used to describe monthly rainfall patterns (Hughes & Saunders, 2002; Oettli & Camberlin, 2005). It also fitted the period mean rainfall was superior to that of the Gamma GLM

models inmost of the regions. Figure 4.13 shows the results of the estimated 64 years periods rainfall variations for all the regions.Analysis of the graph indicated three main categories regarding the period rainfall estimates.The models in the N, NE, CE and NW regions can be put into a category (first group) having related patterns while the models in the CS and NSE can also be put in a group (second group).

(a)



(b)

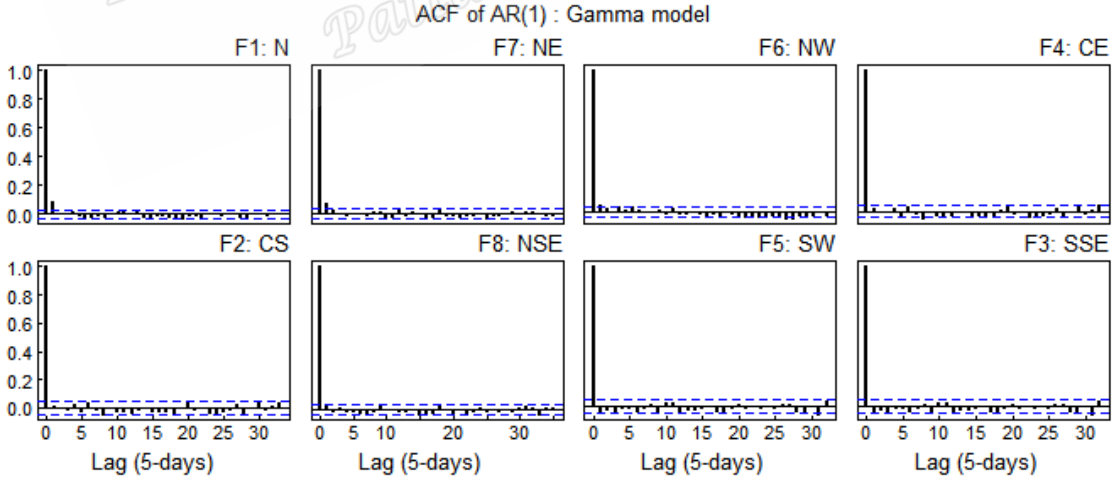


Figure 4.10 The autocorrelation function for the models after minimising serial correlation.

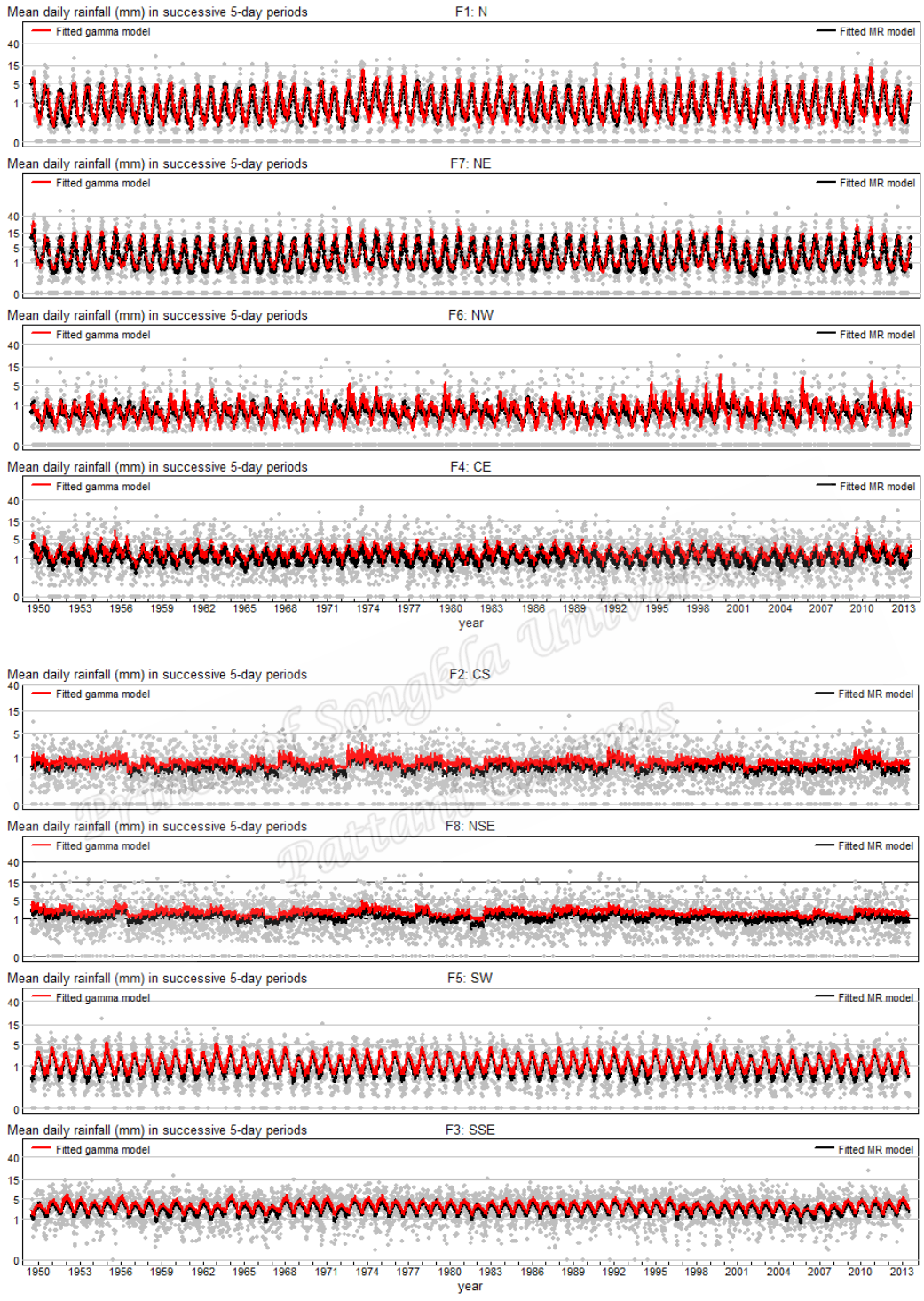


Figure 4.11 The plots of fitted 5-day rainfall and their estimated mean patterns in eight regions during 1950-2013

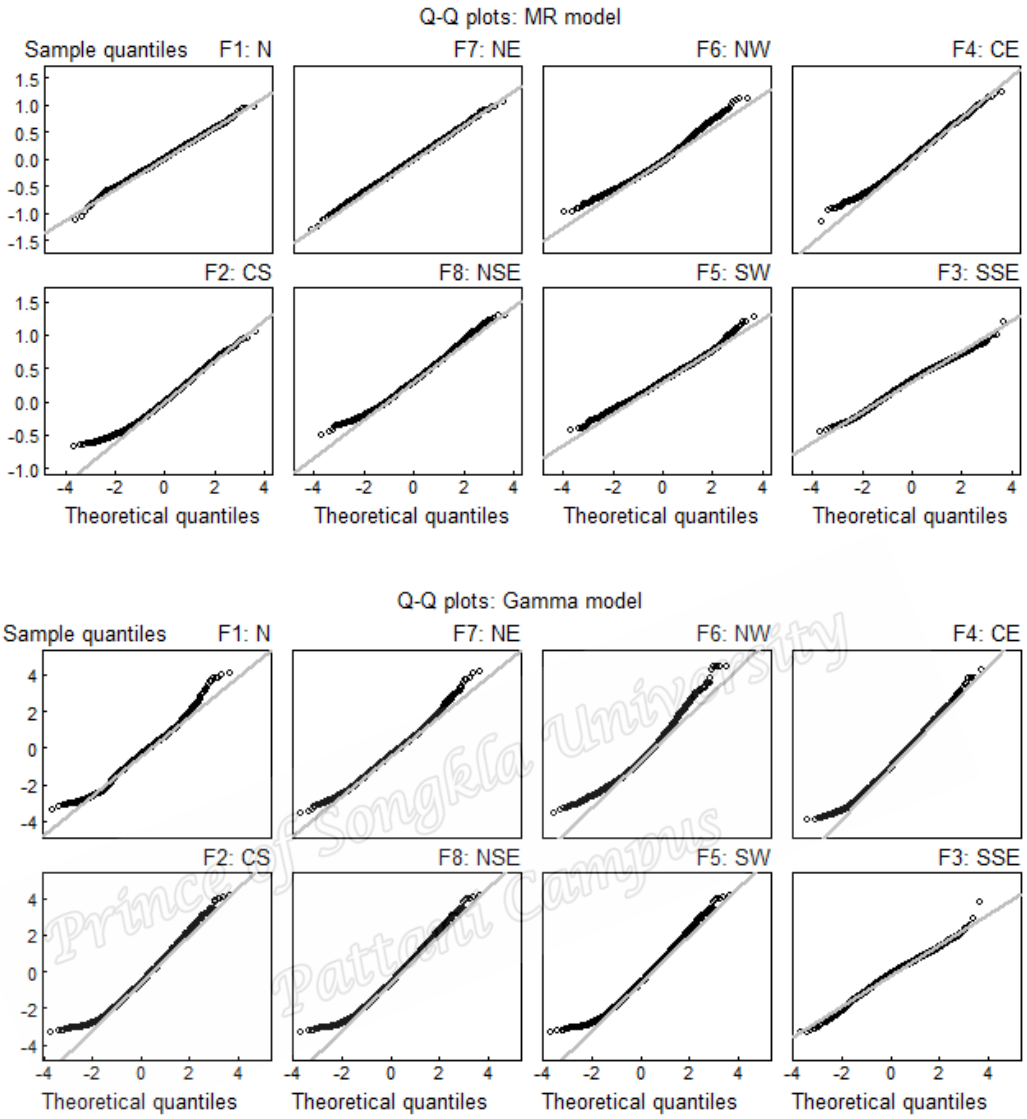


Figure 4.12 The residual quantile-quantile (Q-Q) plots for the models.

The SW and SSE also form a group (final group) with similar patterns. In the first group, the period rainfall rises sharply from January to March, reaches its maximum between this period then declines gradually until it minimum between August and September. The period rainfall observed in the second group is quite stable throughout the year (it varies from 0.6 to 2.2 mm d^{-1}). The most substantial projected period rainfall in these regions occurred between May and July, but the minimum was observed during March and May in the CS and occurred between August and October

in the NSE. On the other hand, the final group received their minimum period rainfall between January and February. The period rainfall increased gradually from February until it reaches the largest between June and August.

The projected period rainfall is different in each region: the NE, which encompasses equatorial and tropical regions, has the highest in Austral summer during the active part of the monsoon whereas the mid-latitude region of SSE has the most during Austral winter.

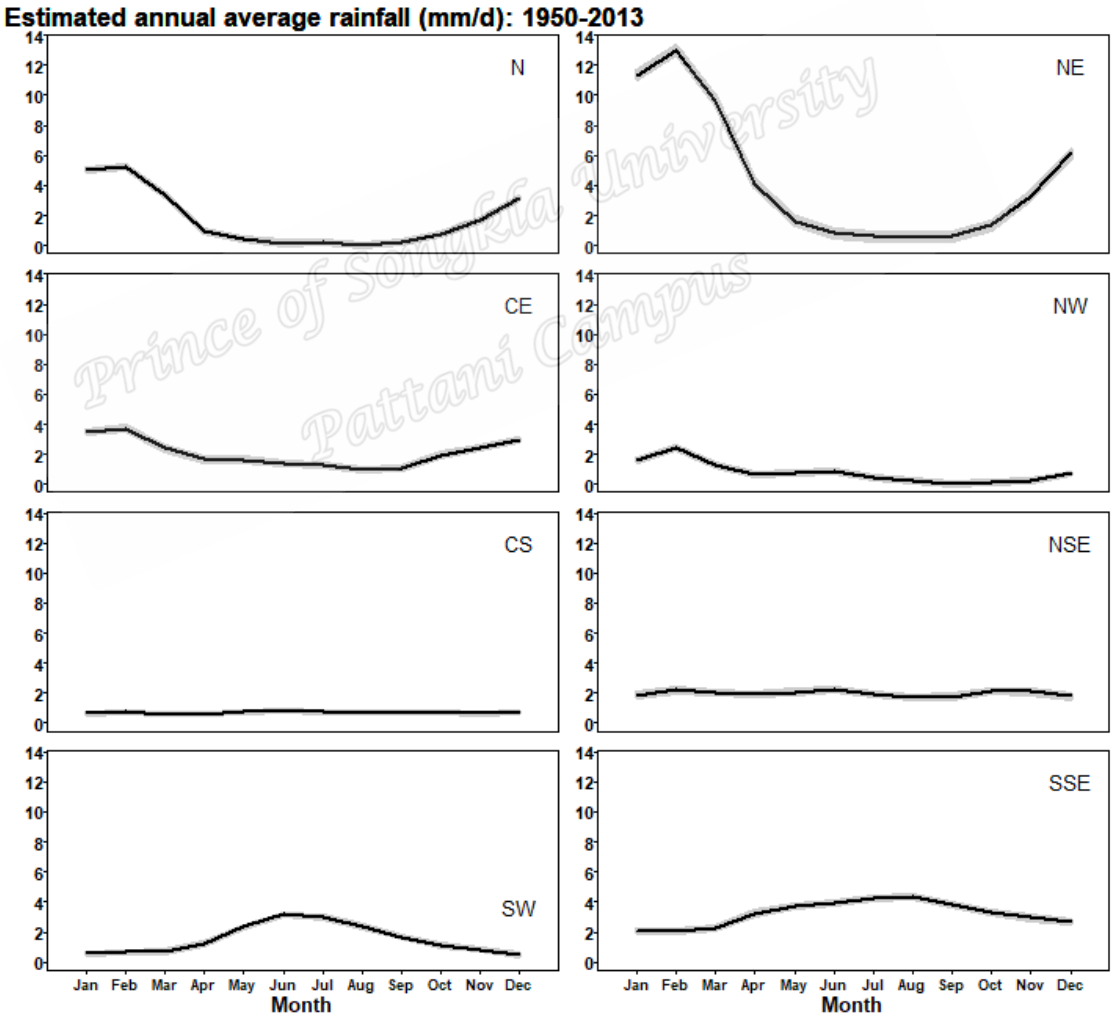


Figure 4.13 The estimated monthly rainfall (line) and the 95% confidence interval (shaded) in each month of a year in the factor regions.

Figure 4.14 also displays the results of the estimated inter-annual rainfall patterns. In the eight regions, the patterns varied between 0.3 and 7.1 mm d^{-1}, reached the lowest value between 1950 and 1975 in the NW, but the highest value took place during 1950 and 1960 in the NE. Generally, the highest annual rainfall values in most of the regions also happened among 1950 and 1970 but for the N, CS and NW regions where it happened between 1970 and 2012.The estimated annual rainfall amount was low particularly in CS and the NW relative to the remaining regions.

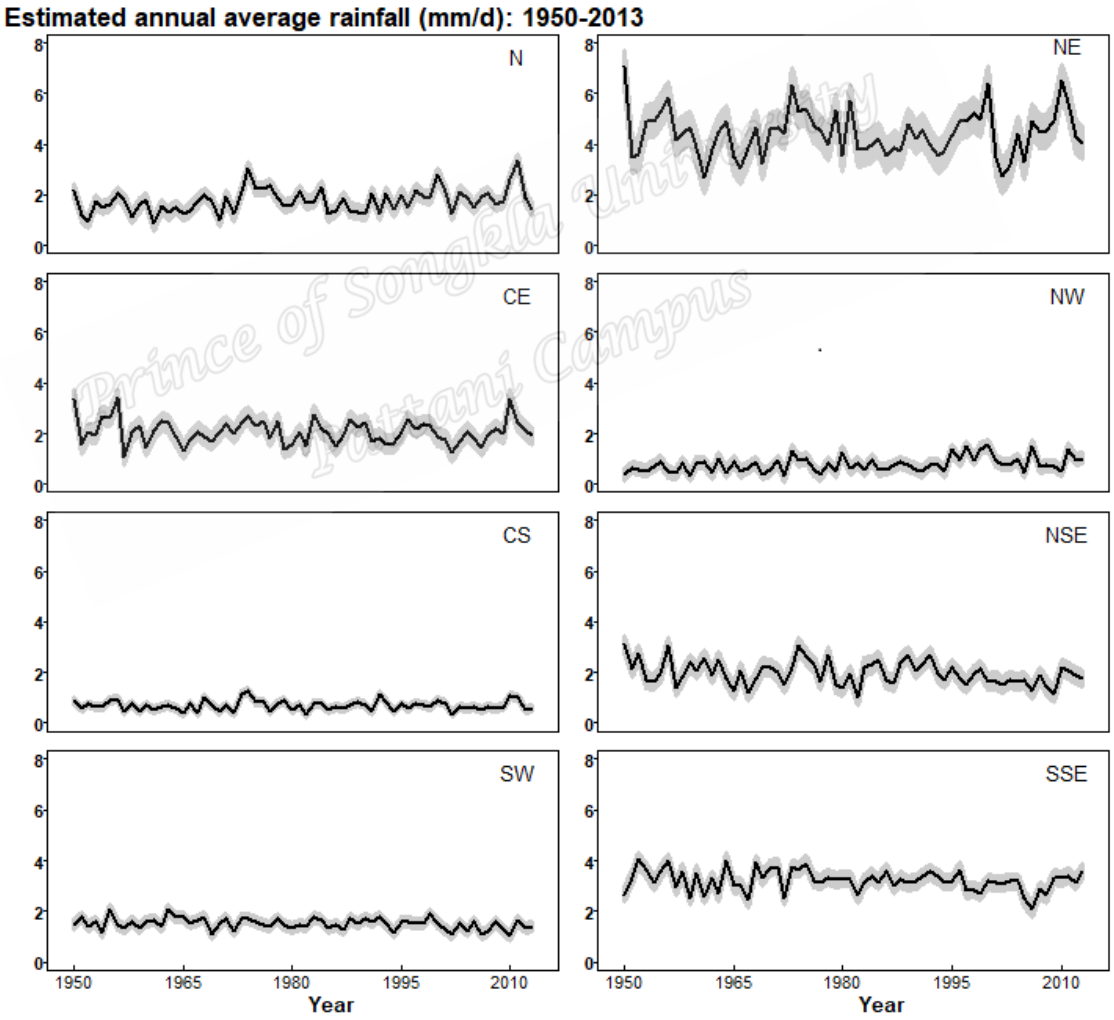


Figure 4.14 The estimated annual rainfall (black line) and the 95% confidence interval (shaded) for factor regions from 1950-2013

CHAPTER 5

The probability of rainfall occurrence in Australia

This chapter reports on the analysis of the probability of 5-day rainfall occurrence in Australia during 1950-2013. The analysis was done by using logistic regression.

5.1 Modelling results

The logistic regression was fitted to 5-day rainfall mean in all the 105 observational stations during 1950-2013 to predict the probability of occurrence. Each model was made up of two predictors (period and year factors).The results also appear in the author's third article (Appendix III) Term-wise addition of the predictors in the model showed a reduction in the deviance residual. The reduction in the deviance residual show that the addition of each variable was essential in the modelling process ($p<$ 0.001). Analysis of the variables of each level revealed that the period factors were more influential relative to the year factor in most of the models.These findings were observed especially during the wettest period of the year in the study areas. For instance, 47 parameters of the period factors were significant while only 6 of the year factors were significant in Darwin.

In Weipa, 50 covariates of the period factors were significant while only five covariates were significant concerning the annual factors. The model in Tambo recorded 80 significant parameters (highest) with 47 periods and 37-year factors followed by Renmark with 71 significantparameters of 31 periods and 40-year factors. On the other hand, Calliope had the lowest number of significant parameters (42), with 3periods and 39-year factors.

Analysis of the AIC values revealed values ranging from 3402.7 to 6142. The least value was observed in Darwin (3402.7) followed by Weipa with a value of (3818.7). However, the highest AIC value was observed in Tambo (6142.0) followed by Renmark (6034.3).

The contingency table for the classification of the occurrence and non-occurrence probability of rainfall by the confusion matrix using the logistic regression models are given in Table 5.1. In Darwin, analysis of the classification accuracy by the confusion matrix indicated that the logistic regression correctly predicted 86.3 and 85.1% of non-occurrence and occurrence of rainfall respectively using the training dataset with a global accuracy of 85.8%. The area under the ROC (AUC) was 0.93 (Figure 5.1). It indicates exceptional discrimination by the logistic model. These events resulted in omission and commission errors for of 13.7 and 14.9% correspondingly. The Hosmer and Lemeshow goodness-of-fit test revealed adequate fit of the logistic regression to the data (χ^2 = 2.42, *df* = 8, *p*-value = 0.97). The accuracy of the classification was also assessed with the validation dataset. Concerning the validation dataset, the confusion matrix revealed that the logistic regression correctly predicted 86.6 and 82.5% of non-occurrence and occurrence of rainfall respectively. The omission and commission errors for these events were 13.4 and 17.5% respectively with overall classification accuracy observed to be 84.8%.

The overall classification accuracy was observed to be 83.9% in Weipa, and the model correctly predicted 87.8 and 79.9% of non-occurrence and occurrence of rainfall respectively. The Hosmer and Lemeshow goodness-of-fit test revealed adequate fit of the logistic regression to the data (χ^2 = 3.26, *df* = 8, *p*-value = 0.92) signifying a good fit. The AUC was 0.91 (Figure 5.1) depicting exceptional

discrimination. The errors associated with these events were 12.2 and 16.1% respectively by means of the training dataset. The accuracy of the classification was also examined using the validation dataset. The confusion matrix shows that the regression models revealed the overall accuracy of 79.2, predicting correctly 86.6 and 72.0% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 13.4 and 28% respectively, which depict a good fit.

The Hosmer and Lemeshow goodness-of-fit test displayed sufficient fit of the regression to the data (χ^2 =11.64, *df* = 8, *p*-value = 0.17). The AUC was 0.91 (Figure 5.1) and the global accuracy of 83.6% (Table 5.1, using the training dataset), both of which indicated good fit and exceptional discrimination in Kowanyama. The model correctly predicted 82.0 and 83.6% of non-occurrence and occurrence of rainfall respectively. The errors of omission and commission for these events were 18.0 and 16.4% separately (training dataset). Moreover, the validation dataset revealed a global accuracy of 80.7% and predicted correctly 81.6 and 80.1% of non-occurrence and occurrence of rainfall respectively. The omission and commission errors for these events were 18.3 and 19.9% respectively.

In Coolawanyah, the Hosmer and Lemeshow goodness-of-fit test displayed sufficient fit of the regression to the data (χ^2 = 5.48, *df* = 8, *p*-value = 0.71) with AUC of 0.82 (Figure 3) indicating a good fit with excellent discrimination by the logistic regression model. The regression model predicted accurately 59.0 and 83.6% of non-occurrence and occurrence of rainfall respectively with a global accuracy of 80.7%. The errors associated with the classification of these events were 40.9 and 16.4% respectively (using the training dataset). The precision of the classificationwas also assessed with the validation dataset. The validation dataset gave an overall accuracy of 80.6% and

predicted correctly 57.9 and 83.4% of non-occurrence and occurrence of rainfall respectively. The errors associated with the classification of these events were 42.1 and 16.5% respectively.

The Hosmer and Lemeshow goodness-of-fit test for the logistic displayed (χ^2 = 5.10, *df* = 8, *p*-value = 0.75) in Tambo with AUC of 0.71 (Figure 5.1) signifying no indication of poor fit with sound discrimination. The results are good since here we know the model is undoubted, adequately specified. The classifications of the occurrence and non-occurrence indicated a global accuracy of 66.6%, predicting correctly 64.1 and 68.1% of the events respectively. The observed omission and commission errors for the events were 35.9 and 31.9% respectively (using the training dataset). The confusion matrix shows that the regression models revealed the overall accuracy of 58% predicting correctly 58.0 and 59.0% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events was 42.0 and 41.0% respectively (using the validation dataset).

In Calliope, 75.6% overall accuracy was observed in events classification, and the regression model correctly predicted 76.4 and 48.5% respectively of the events (Table 5.1). The errors associated with the classifications were 23.6 and 51.5% respectively with AUC of 0.68 (Figure 5.1) which indicates poor discrimination. The Hosmer and Lemeshow goodness-of-fit test for the logistic regression displayed (χ^2 = 9.70, *df* = 8, *p*-value = 0.29). Even though the Hosmer and Lemeshow goodness-of-fit test did not show any evidence of worst of fit by the logistic regression model, the AUC value revealed poor classification of the event (using the training dataset). Assessment of the validation dataset shows that the confusion matrix of the logistic regression model revealed the overall accuracy of 74.8% predicting correctly 76.2 and 28.8% of the

non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 28.8 and 71.2% respectively.

The overall precision of 75.9% was observed in Oakland in the classification by the confusion matrix of the logistic model, predicting correctly 75.9 and 75.9% of the non-occurrence and the occurrence events respectively (Table 5.1). The errors associated with these classifications are 24.1 and 24.1% respectively. The Hosmer and Lemeshow goodness-of-fit test for the logistic regression displayed (χ^2 = 1.94, *df* = 8, *p*-value = 0.98) signifying no indication of poor fit. The AUC was 0.83 (Figure 5.1) showing excellent discrimination of events classification. This is good since here we know the model is certainly properly specified using the training dataset. In assessing the validation dataset, the confusion matrix shows that the regression models revealed the overall accuracy of 71.4% predicting correctly 67.8 and 74.9% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 32.2 and 25.1% respectively.

The test of fitness of the model by Hosmer and Lemeshow displayed sufficient fit of the regression to the data (χ^2= 5.90, *df* = 8, *p*-value = 0.66) which indicates no evidence of poor fit. The AUC was 0.76 (Figure 5.1) which indicates sound discrimination of the events. The confusion matrix revealed that the regression model appropriately categorised 71.7% of all the observations in Ceduna. Of all the observations, 75.4% of the non-occurrence rainfall was correctly predicted while 64.2% of the rainfall occurrence was well predicted using the training dataset (Table 5.1). The omission and commission errors associated with the rainfall events were 24.6 and 35.8% respectively. On the other hand, by means of the validation dataset, the confusion matrix revealed that the regression model gave an overall accuracy of

65.9%, predicting correctly 70.1 and 57.0% of the non-occurrence and occurrence of rainfall respectively. The errors of 29.9 and 43.2% were associated with these events respectively(via the validation dataset).

In Renmark, the overall accuracy of the classification by the confusion matrix of the regression model was 67.6% while 68.8 and 66.1% of the non-occurrence and the occurrence events were correctly predicted respectively (Table 5.1). The AUC was 0.74 (Figure 5.1) indicating a sound classification of the events and the errors associated with the non-occurrence event was 31.2% while that of the occurrence event is 33.9%.

The Hosmer and Lemeshow goodness-of-fit test revealed adequate fit of the logistic regression to the data **(**$\chi^2 = 2.89$, $df = 8$, p-value = 0.94) using the training dataset.

In using the validation dataset to assess the model, the confusion matrix showed that the regression model predicted correctly 65.2 and 56.8% of the non-occurrence and the occurrence events respectively with an overall accuracy of 61.5%. The errors associated with these events were 34.8 and 43.2% respectively (using the validation dataset). The ROC curve for the logistic regression models for the stations is shown in Figure 5.1 In analysing the space in the ROC; the horizontal grey line depicts the random distribution of events to one of the two classes while the upper left corner signifies a perfect classification of the events.

Table 5.1 The contingency tables for the results of the confusion matrices of the logistic regression models for the nine stations

Predicted	Training				Validation			
	Rain	0	1	Percentage correct	Rain	0	1	Percentage correct
Darwin								
Observed								
		111						
	0	8	177	86.3	0	1122	173	86.6
	1	155	886	85.1	1	182	859	82.5
Overall percentage				85.8				84.8
Weipa								
Observed								
		104						
	0	8	145	87.8	0	957	148	86.6
	1	230	913	79.9	1	315	811	72.0
Overall percentage				83.9				79.2
Kowanyama								
Observed								
	0	790	173	820	0	785	177	81.6
			114					
	1	225	8	83.6	1	273	1100	80.1
Overall percentage				83.0				80.7
Coolawanyah								
Observed								
	0	167	116	59.0	0	154	112	57.9
			171					
	1	336	7	83.6	1	339	1715	83.4
Overall percentage				80.7				80.6
Tambo								
Observed								
	0	556	311	64.1	0	502	363	58.0
			100					
	1	469	0	68.1	1	602	867	59.0
Overall percentage				66.6				58.7
Calliope								
Observed								
		173						
	0	4	536	76.4	0	1729	541	76.2
	1	32	34	48.5	1	47	19	28.8
Overall percentage				75.6				74.8
Oakland								
Observed								
	0	884	280	759	0	789	375	67.8
	1	283	889	75.9	1	293	875	74.9
Overall percentage				75.9				71.4
Ceduna								
Observed								
		119						
	0	1	389	75.4	0	1107	472	70.1
	1	271	874	64.2	1	324	430	57.0
Overall percentage				71.7				65.9
Renmark								
Observed								
	0	904	410	68.8	0	855	456	65.2
	1	346	676	66.1	1	441	579	56.8
Overall percentage				67.6				61.5

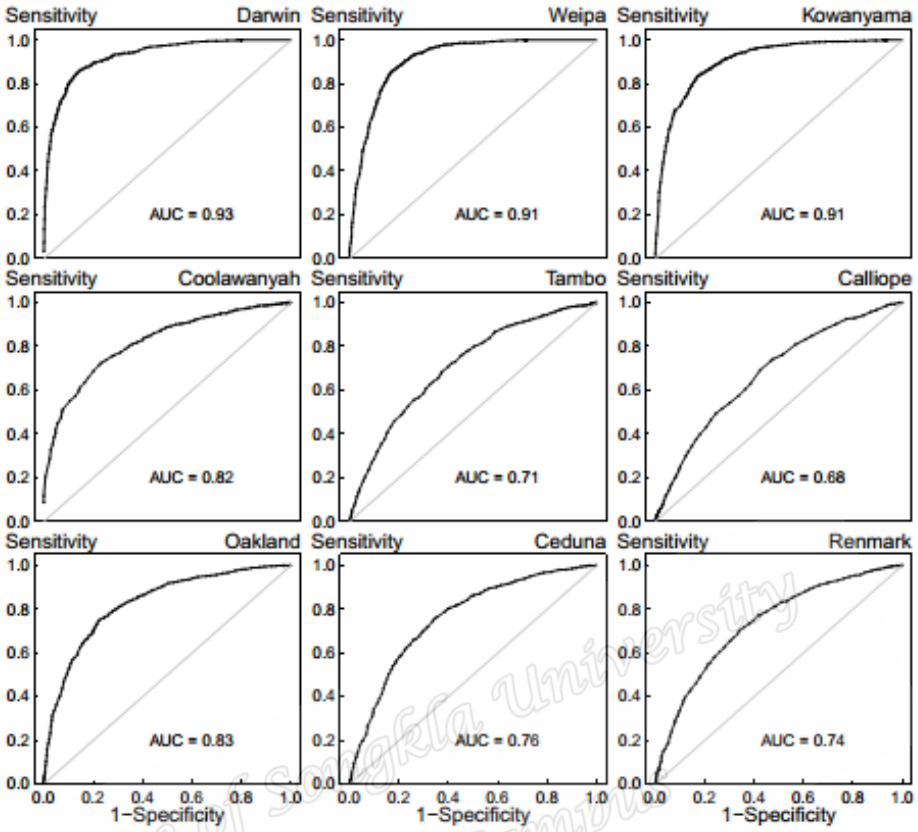


Figure 5.1. The ROC curves for the logistic regression models.

5.2 Adequacy of the models

The ROC plots are the curve lying between the two extremes (black curve). The nearer it lies to the upper left corner, the better the performance of the classifier to discriminate between two events. Moreover, the AUC is usually used to assess the performance of the classifier. The computed AUC values range from 0.68 to 0.92. The values in Darwin, Weipa and Kowanyama showed exceptional discrimination while the models in Coolawanya and Oakland revealed excellent discrimination of the

event (good fit of the models). Also, the models in Tambo, Ceduna and Renmark showed sound discrimination of the events while the model in Calliope revealed poor discrimination of the two events.

CHAPTER 6

Conclusions and Discussions

This chapter contains sum-ups of the main findings centred on the statistical modelling of the 5-day rainfall patterns and trends in Australia. It includes discussions of the results limitations and recommendations for further research of the study.

6.1 Conclusions

In this study, statistical models were applied to describe the period rainfall in Australia during 1950-2013. It is centred on 64 years daily aggregated data acquired from 92 stations distributed over Australia. The investigations by Factor analysis recognised eight factors, which relate to eight geological rainfall groups. The eight factors were satisfactory and represented around 52% of the observed correlations in the observation. The eight regions comprise north, northeast, central east, northwest, central south, north southeast, southwest and south-southeast. The rainfall from each region revealed different overall mean with numerous seasonal variations and was analysed using Gamma generalised linear and multiple regression models.

The period rainfall computed from each of the eight regions were then model by using these methods. Serial correlation was evident in the rainfall observations in successive periods and was reduced by using the AR(1) procedure in the modelling process. The two models fitted the data quite well in all the regions, predominantly in the tropical region (lower latitudes) where the internal atmospheric variability small relative to the forced change. Conversely, the results of the multiple regression models were observed to be superior to that of the Gamma generalised linear model in the north, northeast and the southwest regions.

The projected period and annual rainfall means by the multiple regression models showed three diverse categories and each of the categories show distinct monthly or annual pattern. In general, the eight recognised regions of the factor analysis show three main periods rainfall category with each having a unique dominating pattern. The highest observed annual rainfall amount was received in the northeast followed by the south-southeast. Conversely, very low rainfall amount was evident in the central south and the northwest. These low rainfall patterns were observed to be related throughout this period. The trend analysis by the linear regression model revealed substantial decreasing annual rainfall in the southwest and the north southeast regions. In disparity, substantial increasing annual rainfall trends were received in the north and the northwest regions.

Also, logistic regression has been applied to describe 5-day rainfall probability of occurrence during 1950-2013 in Australia. The viability of modelling rainfall occurrence using two easily acquire variables has been demonstrated. The fitted models have acceptable predictive precisions in predicting the probability of rainfall occurrence in Australia. The models perform quite well concerning the proportions of correct predictions for the occurrence or the non-occurrence rainfall that ranges between 48.5-86.6%.

6.2 Discussions

In general, this study classified rainfall variability of 92 meteorological stations into eight factors by factor analysis, which identified eight geographical regions. Analysis of the rainfall variability in the regions shows three main periods rainfall category with each having a unique dominating pattern. The understanding and knowledge of these patterns are vital in planning the rainfall resources in many regions of Australia.

However, 17 stations out of the 92 could not be put into any of group revealed by the factor analysis (they were mixed factors) which may be due to possible sub-classifications in some of the regions. These findings are consistent with similar findings in similar studies by Chambers (2003).

The detected decreasing annual rainfall trends revealed in this study in the southern regions might be as a result of a reducing austral autumn and winter rainfall in the southern areas of Australia mostlysouth Western Australia. The apparent significant decreasing trends in the SW regions is also consistent with the available literature. For instance, the IOCI (2002) also indicated the long-term drop in winter (May-October) rainfall of 15-20% since the 1970s in the southwestern corner of Western Australia. The average spring rainfall observed during 1997-2006 in the southeast of Australia was seen to be below average. The below average spring rainfall resulted in the problem of repeated drought in this part of Australia (Murphy & Timbal, 2008). Hope et al. (2010), discovered similar results in the analysis of the temporal relationship between rainfall variability in the SW and southeast of Australia.

A study by Timbal and Fawcett (2013) in the same way detected exceptional rainfall deficits in the southeastern Australia. The stated rainfall deficit was obvious in pre-winter and early winter rainfall. These falling trends could be partly because of changes in large-scale atmospheric circulation (Nicholls, 2006; Timbal & Fawcett, 2013) comprising a poleward movement of the westerly winds and increasing atmospheric surface pressure. Variations in anthropogenic greenhouse gases and ozone levels (Delworth & Zeng, 2014) and expansion of the Southern Hemisphere Hadley cell (Post et al., 2014) can affect the reducing rainfall trends in these regions.

Substantial increase in the trends in annual rainfall were obvious in the N and NW areas of Australia whereas an apparent non-significant rising trend at 95% confidence level was seen in the NE region. In exception of the N and NW regions, the NE region is frequently liable to strong weather conditions, for example, tropical cyclones that transport lots of convective clouds. These weather conditions may elucidate larger part of cloud cover in NE as correlated to that in the N and NW regions Cheung et al. (2015). It could also elucidate part of the upward rainfall trend in this region since cloud development is connected to rainfall.

The significant increasing annual rainfall trends in NW revealed in this study is also consistent with the earlier study by Taschetto and England (2009). Their study indicated that the increasing annual rainfall trend over northwest Australia is associated with more extreme deep convection triggered by variations in the monsoon trough.

The gamma GLM and the multiple linear regression models applied to describe the non-zero rainfall amount used the seasonal and the annual factors as predictors. These models captured the periodic seasonal patterns quite well so as the timings of periods. Relative to other regions, the highest amount of period rainfall in the seasonal span was observed in the NE region. This may be because of considerable number of temporal meteorological conditions such as fronts and low-pressure systems which possibly can disturb rainfall in this area.The area is also more liable to severe weather systems such as tropical cyclones together with a lot of convective clouds engrained, causing in a higher part of cloud cover relative to other regions (Cheung et al., 2015).Interestingly, the seasonal variations were observed to be higher in all the factor regions than their annual rainfall variations.

The fitted models reveal that the mean rainfall amount is higher in the NE followed by SSE and lowest in CS. Moreover, in N, NE, NW and CE, most of the annual rainfall was observed in the summer months, and these results were observed by Lavender and Abbs (2013) over Northern Australia and attributed it partially to the tropical cyclones and other low-pressure systems. Considering all the eight-factor regions, N, NE, NW and CE receives a substantial amount of rainfall from December-March (wettest months) while uniform monthly rainfall distributions were observed in CS and NSE. Similar results were seen by Hasan and Dunn (2010) on the analysis of monthly rainfall from 1912-1971 with Poisson and gamma models. On the other hand, SW and SSE have their wettest months between August and September.

Analysis of the Q-Q plots revealed that the fitted models described the data fairly well in all the factor regions but slight deviations at the upper and lower tails of some of the models. Similarly, the result was observed by Hasan and Dunn (2010) using simple Poisson-gamma models in the analysis of monthly rainfall in Australia and a study in Limpopo basin, Botswana by Kenabatho et al. (2012). Besides, the latter revealed that the observed departures at the lower and upper tails of some of the models might be data quality issues.

The logistic regression also predicted the rainfall probability of occurrence and non-occurrence quite well in all the regions. However, the predictive ability of the models could be enhanced with the inclusion of other climate variables such as mean temperature, mean sea level pressure, wind speed and direction in the modelling process.

6.3 Limitations and Recommendations

This study is based on the modelling of 5-daily rainfall using easily acquired predictors without considering other meteorological factors that affect rainfall such as, temperature, wind speed and sea level pressure. As population increases, factors such as land use change which results in urbanisation, which brings about changes in cloud condensation nuclei, low-level convergence due to turbulence, urban heat Islands which affect the development of rainfall. The presence of some of these factors in the modelling process may enhance it predictive adequacy.

Moreover, some part of the study area, especially the western desert had few observational stations, and besides, most stations did not have a long record of data continues for 64 years of study. Many stations had a lot of missing data and Satellite data could also be acquired to supplements the station data.

Even though the methods applied in this study delineated the periodic patterns and the timings of the 5-day rainfall quite well, further studies could compare other methods for modelling periodic functions such as cosine and sine and simple harmonic functions. Also, results from other methods such as the random forest and artificial neural networks in modelling the occurrence probability of the rainfall observations could also be compared.

Further studies could explore the mix-factors and examine the causes of these high seasonal rainfall patterns in the factor regions with climate variables such as (humidity, temperature) and climate indices/teleconnection. The methods used in this study could also be applied to other climatic variables such as temperature, solar radiations and wind speed.

References

Abdi, H. 2003. Factor rotations in factor analyses. Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks. CA, 792-795.

Allan, R.J. and Haylock, M.R. 1993. Circulation features associated with the winter rainfall decrease in southwestern Australia. Journal of Climate. 6, 1356-1367.

Ansell, T.J., Reason, C.J.C., Smith, I.N. and Keay, K. 2000. Evidence for decadal variability in southern Australian rainfall and relationships with regional pressure and sea surface temperature. International journal of climatology. 20 1113-1129.

Bärring, L. 1987. Spatial patterns of daily rainfall in central Kenya: Application of principal component analysis, common factor analysis and spatial correlation. International Journal of Climatology. 7, 267-289.

Bärring, L. 1988. Regionalization of daily rainfall in Kenya by means of common factor analysis. International Journal of Climatology. 8, 371-389.

Beecham, S., Rashid, M. and Chowdhury, R.K. 2014. Statistical downscaling of multi-site daily rainfall in a South Australian catchment using a Generalized Linear Model. International Journal of Climatology. 34, 3654-3670.

Boccolari, M. and Malmusi, S. 2013. Changes in temperature and precipitation extremes observed in Modena, Italy. Atmospheric Research. 122, 16-31.

Buishand, T.A. 1977. Stochastic Modelling of Daily Rainfall Sequences. Mededlingen Landbouwhoge school, Wageninge.

Bukantis, A. 2002. Application of factor analysis for quantification of climate-forming processes in the eastern part of the Baltic Sea region. Climate Research. 20, 135-140.

Chambers, L.E. 2003. South Australian rainfall variability and trends. BMRC Research Report Number 92, Commonwealth of Australia, Bureau of Meteorology.

Chandler, R.E. and Wheater, H.S. 2002. Analysis of rainfall variability using generalised linear models: a case study from the west of Ireland. Water Resources Research. 38, 1-11.

Chatfield, C. 1996. The Analysis of Time Series. Chapman & Hall: Melbourne, Australia.

Cheung, K.K., Chooprateep, S. and Ma, J. 2015. Spatial and temporal patterns of solar absorption by clouds in Australia as revealed by exploratory factor analysis. Solar Energy. 111, 53-67.

Cho, H.K., Bowman, K.P. and North, G.R. 2004. A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. Journal of Applied Meteorology and Climatology. 43, 1586-1597.

Coe, R. and Stern, R.D. 1982. Fitting models to daily rainfall data. Journal of Applied Meteorology. 21, 1024-1031.

Darand, M. and Daneshvar, M.R.M. 2014. Regionalization of precipitation regimes in Iran using principal component analysis and hierarchical clustering analysis. Environmental Processes. 1, 517-532.

Del Hoyo, L.V., Isabel, M.P.M. and Vega, F.J.M. 2011. Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data. European Journal of Forest Research. 130(6), pp.983-996.

Delworth, T.L., and Zeng, F. 2014. Regional rainfall decline in Australia attributed to anthropogenic greenhouse gases and ozone levels. Nature Geoscience. 7(8), 583-587.

Drosdowsky, W. 1993. An analysis of Australian seasonal rainfall anomalies: 1950-1987.II: temporal variability and teleconnection patterns. International Journal of Climatology. 13, 111-149.

England, M.H., Ummenhofer, C.C. and Santoso, A. 2006. Interannual rainfall extremes over southwest Western Australia linked to Indian Ocean climate variability. Journal of Climate. 19, 1948-1969.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern recognition letters. 27(8), 861-874.

Gabriel, K.R. and Neumann, J. 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. Quarterly Journal of the Royal Meteorological Society. 88, 90-95.

Haan, C.T., Allen, D.M. and Street, J.O. 1976. A Markov chain model of daily rainfall. Water Resources Research. 12, 443-449.

Hand, L.M. and Shepherd, J.M. 2009. An investigation of warm-season spatial rainfall variability in Oklahoma City: Possible linkages to urbanisation and prevailing wind. Journal of Applied Meteorology and Climatology. 251-269.

Hasan, M.M. and Dunn, P.K. 2010. A simple Poisson-gamma model for modelling rainfall occurrence and amount simultaneously. Agricultural and Forest Meteorology. 150, 1319-1330.

Hession, S.L., Moore, N. 2011. A spatial regression analysis of the influence of topography on monthly rainfall in East Africa. International Journal of Climatology. 31, 1440-1456.

Hope, P., Timbal, B., and Fawcett, R. 2010. Associations between rainfall variability in the southwest and southeast of Australia and their evolution through time. International Journal of Climatology. 30(9), 1360-1371.

Hosmer Jr., D. W., Lemeshow, S. 2000. Applied Logistic Regression, second ed. John Wiley and Sons. New York.

Husak, G.J., Michaelsen, J. and Funk, C. 2007. Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. International Journal of Climatology. 27, 935-944.

IOCI. 2002. Climate variability and change in south west Western Australia. Indian Ocean Climate Initiative Panel. Perth.

Ison, N.T., Feyerherm, A.M. and Bark, L.D. 1971. Wet period precipitation and the gamma distribution. Journal of Applied Meteorology. 10, 658-665.

Jauregui, E. and Romales, E. 1996. Urban effects on convective precipitation in Mexico City. Atmospheric Environment. 30, 3383-3389.

Kenabatho P.K., McIntyre, N.R., Chandler R.E. and Wheater, H.S. 2012. Stochastic simulation of rainfall in the semi-arid Limpopo basin, Botswana. International Journal of Climatology. 32, 1113-1127.

Kendall, M.G. 1975. Rank Correlation Methods, 4th ed. Charles Griffin, London.

Köppen, W. 1918. Klassifikation der klimate nach temperatur, niederschlag, und jahreslauf. Petermann's Mitteilungen. 64, 193-203.

Lavender, S.L. and Abbs, D.J. 2013. Trends in Australian rainfall: contribution of tropical cyclones and closed lows. Climate dynamics. 40, 317-326.

Li, T., Wang, B., Chang, C.P. and Zhang, Y. 2003. A theory for the Indian Ocean Dipole-Zonal mode. Journal of the Atmospheric Sciences. 60, 2119-2135.

Li, Y., Cai, W. and Campbell, E.P. 2005. Statistical modeling of extreme rainfall in southwest Western Australia. Journal of Climate. 18, 852-863.

Li, X., and Gao, S. 2011. Precipitation modeling and quantitative analysis. Springer Science & Business Media.

Lloyd-Hughes, B., and Saunders, M. A. 2002. Seasonal prediction of European spring precipitation from El Niño-Southern Oscillation and local sea-surface temperatures. International Journal of Climatology. 22(1), 1-14.

Mann, H.B. 1945. Non-parametric test against trend. Econometrica. 13, 245-259.

Mason, I. 1982. A model for assessment of weather forecasts. Australian Meteorological Magazine. 30(4), 291-303.

McCullagh, P., and Nelder, J. A. 1989. Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability.

Metz, C.E. 1978. Basic principles of ROC analysis. In Seminars in nuclear medicine. 8(4),283-298.

Murphy, B.F., and Timbal, B. 2008. A review of recent climate variability and climate change in southeastern Australia. International Journal of Climatology. 28(7), 859-879.

Nicholls, N. 2006. Detecting and attributing Australian climate change: a review. Australian Meteorological Magazine. 55(3), 199-211.

Oettli, P., and Chamberlin, P. 2005. Influence of topography on monthly rainfall distribution over East Africa. Climate Research. 28(3), 199-212.

Oke, T.R. 1988. The urban energy balance. Progress in Physical Geography. 12, 471-508.

Pitman, A.J., Narisma, G.T., Pielke, R.A. and Holbrook, N.J. 2004. Impact of land cover change on the climate of southwest Western Australia. Journal of Geophysical Research. 109.

Post, D.A., Timbal, B., Chiew, F.H., Hendon, H H., Nguyen, H. and Moran, R. 2014. Decrease in southeastern Australian water availability linked to ongoing Hadley cell expansion. Earth's Future. 2(4), 231-238.

Prasad, K., Dash, S.K. and Mohanty, U.C. 2010. A logistic regression approach for monthly rainfall forecasts in meteorological sub divisions of India based on DEMETER retrospective forecasts. International Journal of Climatology. 30, 1577-1588.

R Development Core Team. 2015. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available online: http://www.r-project.org [January 5, 2015].

Rana, A., Uvo, C.B., Bengtsson, L. and Sarthi, P.P. 2012. Trend analysis for rainfall in Delhi and Mumbai, India. Climate Dynamics. 38, 45-56.

Rencher, A.C. 2002. Factor analysis. Methods of Multivariate Analysis, 2nd ed., John Wiley and Sons, Inc., New York. U.S.A. pp.408-450.

Shem, W. and Shepherd, J.M. 2009. On the impact of urbanization on summertime thunderstorms in Atlanta: Two numerical model case studies. Atmospheric Research. 92, 172-189.

Shepherd, J.M. 2005. A review of current investigations of urban-induced rainfall and recommendations for the future. Earth Interactions. 9, 1-27.

Sivakumar, B., Woldemeskel F.M. and Puente, C.E. 2014. Nonlinear analysis of rainfall variability in Australia. Stochastic Environmental Research and Risk Assessment. 28, 17-22.

Smith, I.N., McIntosh, P., Ansell, T.J., Reason, C.J.C. and McInnes, K. 2000. Southwest Western Australian winter rainfall and its association with Indian Ocean climate variability. International Journal of Climatology. 20, 1913-1930.

Srikanthan, R. and Pegram, G.G. 2009. A nested multisite daily rainfall stochastic generation model. Journal of hydrology. 371, 142-153.

Stern, R.D. and Coe, R. 1984. A model fitting analysis of daily rainfall data. Journal of the Royal Statistical Society, Series A. 147, 1-34.

Taschetto, A.S., and England, M.H. 2009. An analysis of late twentieth century trends in Australian rainfall. International Journal of Climatology. 29(6), 791-807.

Timbal, B., and Fawcett, R. 2013. A historical perspective on southeastern Australian rainfall since 1865 using the instrumental record. Journal of Climate, 26(4), 1112-1129.

Todorovic, P. and Woolhiser, D.A. 1975. A Stochastic model of n-day precipitation. Journal of Applied Meteorology. 141, 17-24.

Tongkumchum, P. and McNeil, D. 2009. Confidence intervals using contrasts for regression model. Songklanakarin Journal of Science and Technology. 31, 151-156.

Trenberth, K.E. 2011. Changes in precipitation with climate change. Climate Research. 47, 123-138.

Trenberth, K.E., Dai, A., Rasmussen, R.M. and Parsons, D.B. 2003. The changing character of precipitation. Bulletin of the American Meteorological Society. 84, 1205-1217.

Um, M.J., Yun, H., Jeong, C.S. and Heo, J.H. 2011. Factor analysis and multiple regression between topography and precipitation on Jeju Island, Korea. Journal of Hydrology. 410, 189-203.

Unkel, S., Trendafilov, N.T., Hannachi, A. and Jolliffe, I.T. 2010. Independent exploratory factor analysis with application to atmospheric science data. Journal of Applied Statistics. 37, 1847-1862.

Venables, W.N. and Ripley, B.D. 2002. Modern applied statistics with S, 4th ed., Springer. New York. U.S.A. pp. 319-322.

Wanishsakpong, W., and McNeil, N. 2016. Modelling of daily maximum temperatures over Australia from 1970 to 2012. Meteorological Applications, 23(1), 115-122.

Wardle, R.and Smith, I. 2004. Modeled response of the Australian monsoon to changes in land surface temperatures. Geophysical Research Letters. 31(16).

Wickramagamage, P. 2010. Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. International Journal of Climatology. 30, 1235-1245.

Wickramagamage, P. 2015. Spatial and temporal variation of rainfall trends of Sri Lanka. Theoretical and Applied Climatology. 1-12.

Appendix I

Statistical modelling for 5-day average rainfall in Australia during 1950-2013

Bright Emmanuel Owusu[1,2] Nittaya McNeil[1,3*]Mayuening Eso[1,3]

[1]*Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Mueang Pattani, 94000, Thailand.*

[2]*Department of Information and Communication Technology/Mathematics, Faculty of Science and Technology, Presbyterian University College Ghana.*

[3]*Centre of Excellence in Mathematics, Commission on Higher Education (CHE), Ministry of Education, Ratchathewi, Bangkok, 10400 Thailand.*

*Corresponding authors address

Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Mueang Pattani, 94000, Thailand.

Email: nittaya.ch@psu.ac.th

Abstract

Many regions across Australia have high rainfall variability which has various effects on water and food availability. This study examines rainfall patterns over consecutive 5-day periods during 1950-2013 using data from 92 observational stations in Australia. In the first model, factor analysis was used to classify the stations into distinct geographical regions. Gamma generalised linear model (GLM) was then fitted to describe the rainfall amount in each category with season and year factors as the predictors.

Factor analysis revealed eight factors, which represent eight geographical regions of Australia. Analysis of the similarities in the seasonal evolutions between regions revealed three seasonal rainfall groupings. The GLM models fitted the data quite well in all the areas and showed that 5-daily rainfall is significantly affected by the period of the year and its annual average in most of the regions. The models could be used to simulate rainfall data for the areas with inadequate rainfall records.

Keywords: Rainfall variability; Classification; Generalised Linear Models; Factor analysis.

1. Introduction

Many areas of the globe particularly the tropics are extremely susceptible to changes in rainfall patterns. A study by Trenberth, Dai, Rasmussen, and Parsons (2003) revealed that, for the past few decades, the global hydrological cycle has been undergoing significant changes, which include rainfall amount, frequency and duration.These changes in rainfall have resulted in severe floods or droughts which often have diverse effects on human life and health, food and provision of potable water, ecosystems and infrastructure **(**Herring et al., 2016).

Modelling and analysis of rainfall with appropriate statistical methods has diverse applications not just for predictive purposes but also in the hydrological analysis, agricultural production and meteorological planning. It also facilitates sound understanding of spatial-temporal rainfall distribution patterns. Thus, classifying and a model fitting analysis of various rainfall events such as the probability of occurrence, the mean amount in a month or year and distribution patterns will be beneficial to different stakeholders in their policy planning.

Many authors have analysed rainfall amount with various mathematical models. Some of the most used methods include artificial neural networks (Deo & Shahin, 2015), the *K*-Nearest-Neighbours (Toth, Brath, & Montanari, 2000). Other studies also used multiple linear regression (Shukla, Tripathi, Pandey, & Das, 2011) and support vector machines for regression (Lin & Jhong, 2015) methods. Bagirov, Mahmood, andBarton (2017) evaluated these model with clusters linear regression (CLR) in modelling rainfall data in Australia observed the superiority of CLR over the other models.

Rainfall observations are known to be skewed; the gamma generalised model has been found by many researchers (Coe & Stern, 1982; Stern & Coe, 1984; Kenabatho, McIntyre, Chandler, & Wheater, 2012) to fit rainfall observations reasonably well. Cai, Cowan and Thatcher (2012) have revealed rainfall reductions over semi-arid parts of the Southern Hemisphere, such as southern coastal Chile, southern Africa, and southeastern Australia. There have been various studies on different aspects of Australian rainfall. For instance, Drosdowsky (1993) has studied the seasonal rainfall and has revealed that it is highly variable throughout the country.

Trend analyses of rainfall (Chambers, 2003; Taschetto & England, 2009; Lavender & Abbs, 2013) have shown significant observed changes, with increases in the northwest and decreases in the east during the period of 1970-2009. These rainfall changes could be partially associated with tropical cyclones and other low pressure systems as revealed by Lavender and Abbs (2013). In similar studies (Srikanthan & Pegram, 2009; Beecham, Rashid, & Chowdhury, 2014) have used multi-site two parts model to describe the daily variability at some sites. They described the occurrences and amounts with different models and observed acceptable model fits in all the sites.

Hasan and Dunn (2010) have used a Poisson-gamma model to describe monthly rainfall observations for agricultural planning. They used monthly observations from 220 stations and observed acceptable model fits on modelling the mean rainfall amount and the probability of occurrence. Knowledge of the trends in the spatial-temporal rainfall variability and their physical explanations are essential in climate change assessment (Kenabatho et al., 2012).

Many scholars (Bärring, 1987, 1988; Wickramagamage, 2010; Um, Yun, Jeong, & Heo, 2011) have studied rainfall regions classification. Most of these studies applied

Principal Component Analysis (PCA) and factor analysis to describe the spatial-temporal variability based on classification into distinct regions. These methods reasonably delineated rainfall regions well in all their study areas. However, the latter gave easily interpretable results relative to the PCA. Chambers (2003) used PCA and cluster analysis to group rainfall into regions in South of Australia and observed that both methods gave similar clusters. Factor analysis models have widely been used in several meteorological applications to classify multiple outcomes, of data collected from stations spread over a region of considerable size such as a continent or nation (Bukantis, 2002; Unkel, Trendafilov, Hannachi, & Jolliffe, 2010; Cheung, Chooprateep, & Ma, 2015). The use of the method may be its ability to accounts for spatial correlations in the observations. The present paper fits appropriate statistical models to classify and describe the variability patterns of the 5-day rainfall amount in Australia.

2. Materials and Methods

This study used 64 years (1950-2013) daily aggregated rainfall data for 92 observational weather stations spread over Australia. The data were obtained from Australian Bureau of Meteorology (http://www.bom.gov.au/climate/data). These 92 stations were selected to give a sample covering the whole area as evenly as possible, and they have continuous rainfall records that extend over a period of 64 years). Data recorded on leap years were omitted to maintain regular year with 365 days observations, and thus each station was made up of 23,360 observations for the 64 years records. For each site, the 5-day average rainfall defines as the averages over successive 5-days at each site was computed and used in this study.

This choice in the statistical analysis has the following advantages; the proportion of missing data is substantially reduced, and daily patterns are represented graphically well. Correlation between data in consecutive periods is also considerably reduced, and the number of parameters in the model is reduced by a factor of 5, facilitating data management, computations and providing smoother model fits. Where data are missing, the average is estimated based on the number present. If all the observations in a 5-day period are missing, the GLM model, described in the next section, is used for the imputation of these missing values.

Statistical methods were used to analyse the data. The first method used factor analysis to classify the 92 weather stations into factors which correspond to different geographical regions. The factor model with m factors $(f_1, f_2, f_3, \ldots, f_m)$ can be written as;

$$y_t = \mu + \sum_{j=1}^{m} \lambda_j f_j, \qquad (1)$$

where y_t is the rainfall for time t, μ is the average rainfall amount of each station, λ_j is the factor loadings on the j^{th} factor while f_j is the j^{th} common factor. The model also provides uniqueness values to the variables where variables with high uniqueness values cannot be assigned to any factor (Rencher,2002).

The factor analysis technique was then applied to the correlation matrix of the 5-day period rainfall observations and the parameters were estimated by the maximum-likelihood method. The factors were rotated by the “Promax” one of the “oblique” rotation technique to achieve simple structure patterns of loadings, and each factor is identified by the estimated loadings (Venables & Ripley,2002). Stations were

assigned to factors if it has at least 0.33 loadings and it is the largest among the loadings of all the other factors (Hair, Anderson, Tatham, & Black, 1998).

The second model involves fitting gamma GLM to the non-zero rainfall amount for these regions revealed by the factor analysis to generate its amount on wet days. The shape parameter of the gamma distribution is assumed constant through data set for each region. For instance, if μ_i is the mean rainfall amount for any period i connected with predictor vector x_i, then the average 5-day rainfall amount is estimated by equation (2) where α is the coefficient vector.

$$\ln(\mu_i) = x_i\alpha. + \mu_{i-1}, \qquad (2)$$

where μ_{i-1} denote the AR(1) term. A significant concern in fitting models to time series data for example rainfall is the dependence among the responds variable (serial correlation) which violates the independence assumption. The serial correlation was minimised by the introduction of AR(1) term in the modelling. The predictors used in this study were the 5-day rainfall periods (seasonal), the annual daily rainfall (year factors) and the AR(1) term. Therefore, 73 seasonal, 64 year factors and the AR(1) terms were employed in the models. The treatment contrast is usually used to fit non-Gaussian GLMs to obtain model coefficients that are easy to interpret. The models were thus fitted using the treatment contrast which makes the first coefficient for each factor a reference; such that each coefficient represents a comparison of that coefficient with the first factor (Venables & Ripley, 2002).

The fitted models were assessed by the residuals quantile-quantile (Q-Q) plots. The quantile-quantile plot is one of the measures used to check the model fit at each

observation for GLMs. The sum contrasts (Tongkumchum & McNeil, 2009) were used to obtain 95% confidence intervals (CI) to compare the fitted model means with the overall rainfall means. This contrast gives criteria to classify levels of the factor into three groups, according to whether each relating CI exceeds, lie on, or is below the overall mean. All data analysis and graphical displays were carried out in R version 3.2.3 (R Development Core Team, 2013).

3.Results and discussion

Factor analysis grouped Australia into eight distinct geographical rainfall regions and these regions are made up of groups of stations that load with just one factor. The analysis was conducted on several factors, but the factor loadings showed that eight factors were adequate and accounted for 52% of the observed correlations in the data with correlation values of at least 0.34. Also, 15 stations out of the 92 correlated with factor 1 (F1), 13 with factor 2 (F2), 10 with factors 3, 4 and 6 (F3, F4, and F6) and 7 with factor 5 (F5). Six stations correlated with factor 7 (F7) and 4 with factor 8 (F8). A similar classification was revealed by Wanishsakpong and McNeil (2016) in the modelling of daily maximum temperature in Australia. However, 17 stations correlated to more than one factor in this study (mix-factors). The seventeen stations with mixed factor loadings were not used in the GLM modelling. These mix factor stations may be due to sub-groups within the regions as a result of their climate representations. Figure 1 depicts the map of Australia, locations of the observational stations used in the study and the eight regions as determined by the factor model. These regions include F1: North (N), F2: Central South (CS), F3: South-Southeast (SSE) and F4: Central East (CE). The remaining regions are F5: Southwest (SW), F6: Northwest (NW), F7: Northeast (NE) and F8: North-Southeast (NSE).

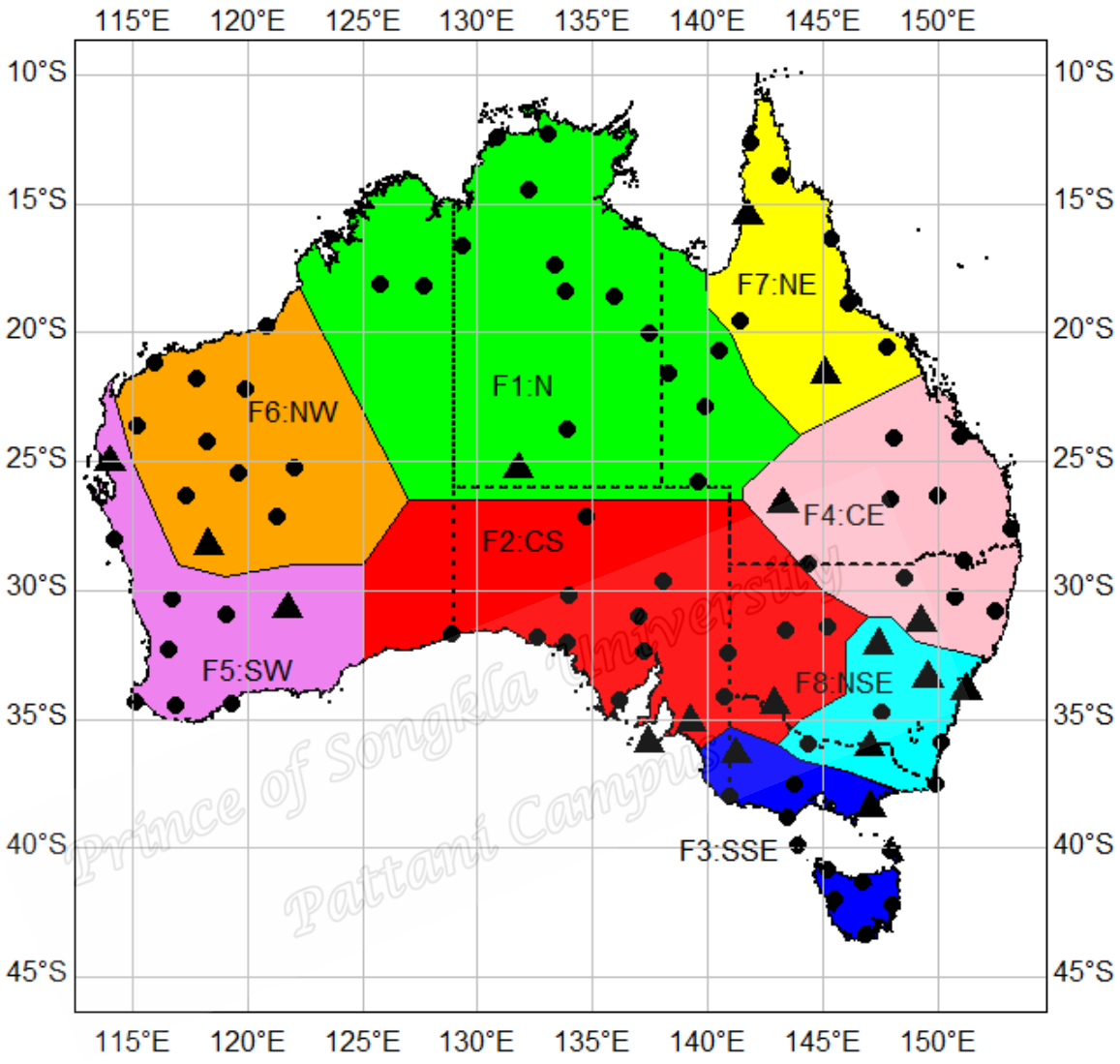


Figure 1. Map of rainfall factor regions of Australia during 1950-2013. The small size points show stations that correlated with only one factor and the triangle points show stations that correlated with more than one factor (mix factors).

Australian Bureau of Meteorology has categorised the climate in Australia into six regimes, namely tropical, grassland, desert, equatorial, temperate and subtropical group. These administrations were revealed using the modified Köppen scheme (Köppen, 1918), which identified climate boundaries with a mixture of natural

landscape topographies and parts of human experience. The tropical climate system is mainly in Northern Australia while the grassland climate system is to the south with the desert classification in central Australia. The equatorial area is the pointed part of Cape York, and Bathurst and Melville Islands in the north of Darwin, temperate regime mainly exist in southeastern Australia with the subtropical climate system in the more coastal areas of northeastern Australia.

However, the eight geographical regions revealed by the factor analysis do not follow the climate regime directly. Comparatively, the factor region N which is revealed by the factor analysis mostly comprised desert, grassland and tropical areas while the NE is made up of mostly the equatorial and the tropical areas. The CS stands out as part of the desert area while the SW, NSE and the SSE are mostly part of the temperate regime. The CE is observed to be a mixture of the grassland and subtropical categories. Besides, the NW comprised the grassland and the desert land categories.

3.1. Annual rainfall

The annual rainfall amount on 5-day period and its probability of occurrence defined as the likelihood of non-zero annual 5-day mean rainfall on any particular year in all the eight regions are shown in Figure 2. There are no distinct annual patterns revealed by the graph. Among all the regions, the NE received the highest average of 4.43 mm d^{-1} followed by SSE with a mean of 3.22 mm d^{-1} while the CE received an average of 2.08 mm d^{-1}.NSE, N and SW followed in that order with averages 1.98, 1.76 and 1.51 mm d^{-1} respectively. The NW received 0.79 mm d^{-1}

while the least of 0.67 mm d^{-1} was received in the CS. Also, the NE received considerably high annual rainfall relative to the other regions followed by the SSE.

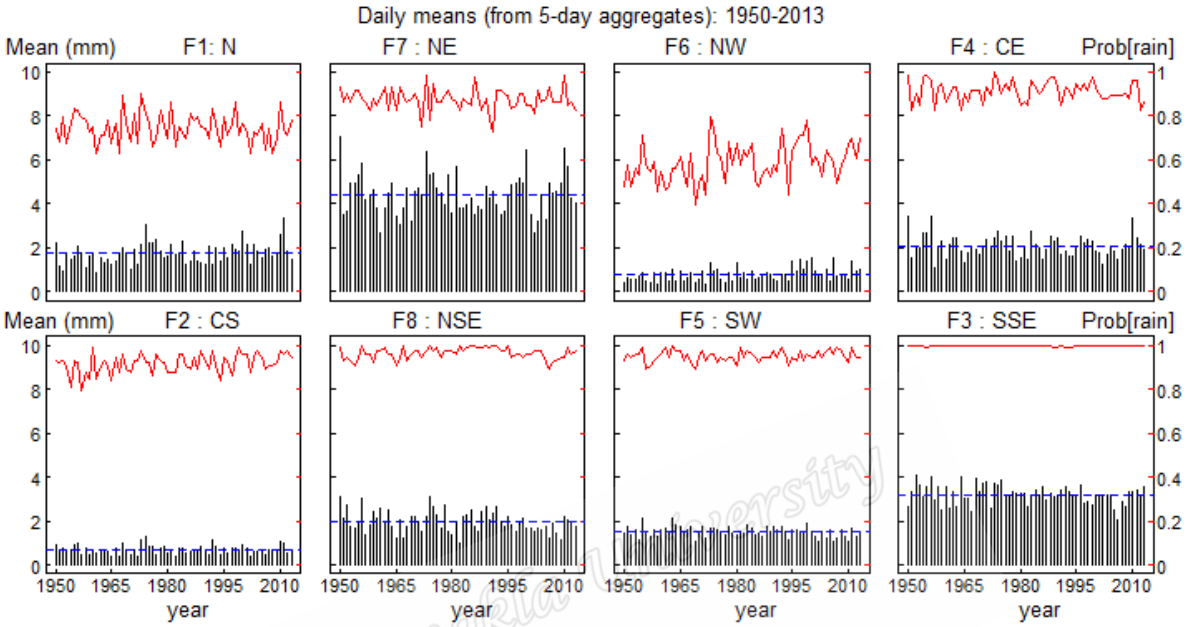


Figure 2. Annual rainfall amount and the probability of occurrences during 1950-2013. The black bars denote the rainfall amount; the horizontal blue dashed lines the overall mean, and the red curve denotes the probability of its occurring.

In contrast, the NW and CS received relatively low annual rainfall. Similar probability patterns in some of the regions were also evident (Figure 2). These probability patterns show the likelihood of obtaining the corresponding 5-day rainfall in that particular year.

The analysis of the probability curve revealed that SSE was anticipated to experienced rainfall throughout the year, whereas SW, NSE, CS and CE showed varying probabilities between 0.8-1.0. However, N and NE had a probability between 0.59 and 1.0. The least annual rain probability of 0.59 was observed with high variations throughout the 64 years in NW.

3.2. Seasonal rainfall

The observed seasonal rainfall amount and the probability of occurrence which is the likelihood of non-zero 5-day rainfall on any particular 5-day period of the year in Australia from 1950-2013 are shown in Figure 3. Each bar corresponds to 5-day average in each particular month. It revealed four distinct seasonal rainfall distribution patterns in Australia, but out of the four seasonal rainfall groups, some regions have similar patterns with different seasonal rainfall amount. Among the regions, the CS and NSE have similar seasonal patterns while the N and NE regions have a similar distribution.

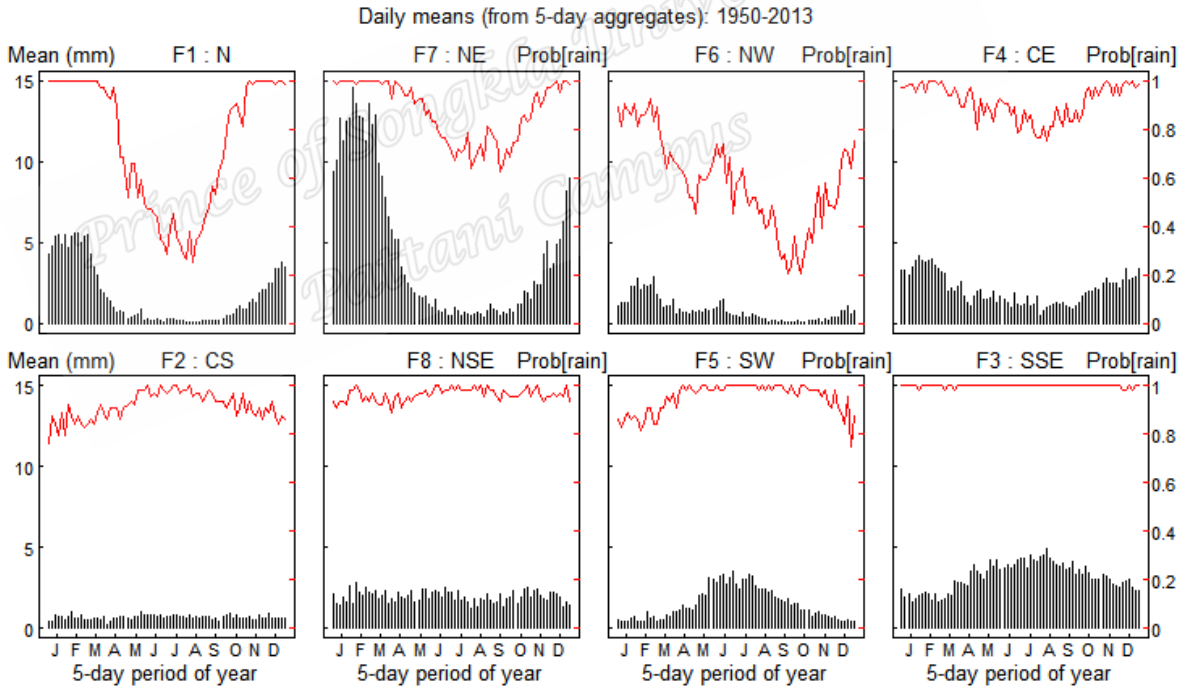


Figure 3. Seasonal rainfall amount and probability of occurrences in Australia during 1950-2013. The black bars denote the rainfall amount while the red curve indicates the probability of its occurrence.

Rainfall in these regions is mostly influenced by tropical cyclones that convey masses of convective clouds. Besides, NW and CE have similar distributions, but apparent two rainfall peaks were observed in NW. The apparent peaks in the seasonal rainfall in the NW may be due to summer monsoon and the cold winter fronts pushing northwards. Seasonal rainfall distribution was observed to be similar in the SW and SSE. These patterns had very low variations in the CS and NSE, but the remaining regions experienced high seasonal variations.

However, analysis of the similarities in the seasonal evolutions between regions by correlation analysis revealed three main seasonal rainfall groupings. This result is consistent with the earlier study by (Peel, Finlayson, & Mcmahon, 2007) which revealed three main climate patterns throughout Australia. The first group consisting of the N, NE, NW and the CE regions which were seen to be strongly positively correlated, have apparent summer monsoon rainfall with dry winters. The second group comprised the CS and the NSE which were also seen to be strongly correlated. These regions observe sporadic rainfall with the apparent seasonal pattern. The final group comprised the SW and SSE regions which have similar seasonal patterns with a correlation of 0.82. On the other hand, the SW and SSE have evident peak rainfall during the winter season. The observed peak in winter may be as a result of the Indian Ocean dipole weather system which is predominantly essential from June to October and spans much of the wet season in the southwest and southeast Australia where it has influence (winter frontal rain).

Seasonal rainfall probability of 1.0 was observed in the N and NE from December-March. These observed seasonal probability value of 1.0 shows that it is certain that it will rain during this period. However, this probability dropped to 0.3 in

the N in August and 0.6 in NE during September. In NW, rainfall probability varied from 0.2-0.9.

3.3. The analysis of the fitted gamma GLM models

The cube root transform of the response variable was used in fitting the models. In all regions. The predictors were seen to be significant in the patterns in the N, SSE, and CE but the models in SW, NW and NE were significantly affected by the seasonal effect only. In CS and NSE the models were seen to be significantly affected by the year factors.Figure 4shows the results of the fitted gamma GLM model of the rainfall of each region. The top panel shows the models in N, NE, NW, and CE while the bottom panel shows the models in CS, NSE, SW, and SSE. The grey points are the observed rainfall, and the black lines denote the fitted gamma model. One noticeable feature is the well-marked seasonal periodic pattern. The eight factor regions revealed different rainfall variability throughout the fitted models. The fitted models vary throughout the 64 years in all the regions with minimum variations observed in CS and NSE from 1950-2013 (the CS and NSE experiences uniform rainfall throughout the season). However, high temporal variations were observed in the N and NE. Besides, Figure 4 shows similar patterns of the fitted models in some of the regions. For instance, the mean rainfall patterns in N and NE were similar while that of NW was also similar to CE. Also, not much difference was observed between the patterns in CS and NSE, and the rainfall patterns in SSE and SW were similar.

The fitted gamma models were assessed using the residuals quantile-quantile (Q-Q) plots from the models (Figure 5). The residuals plots had low variations relative to the expected line of best fit that indicated the models fitted the data quite

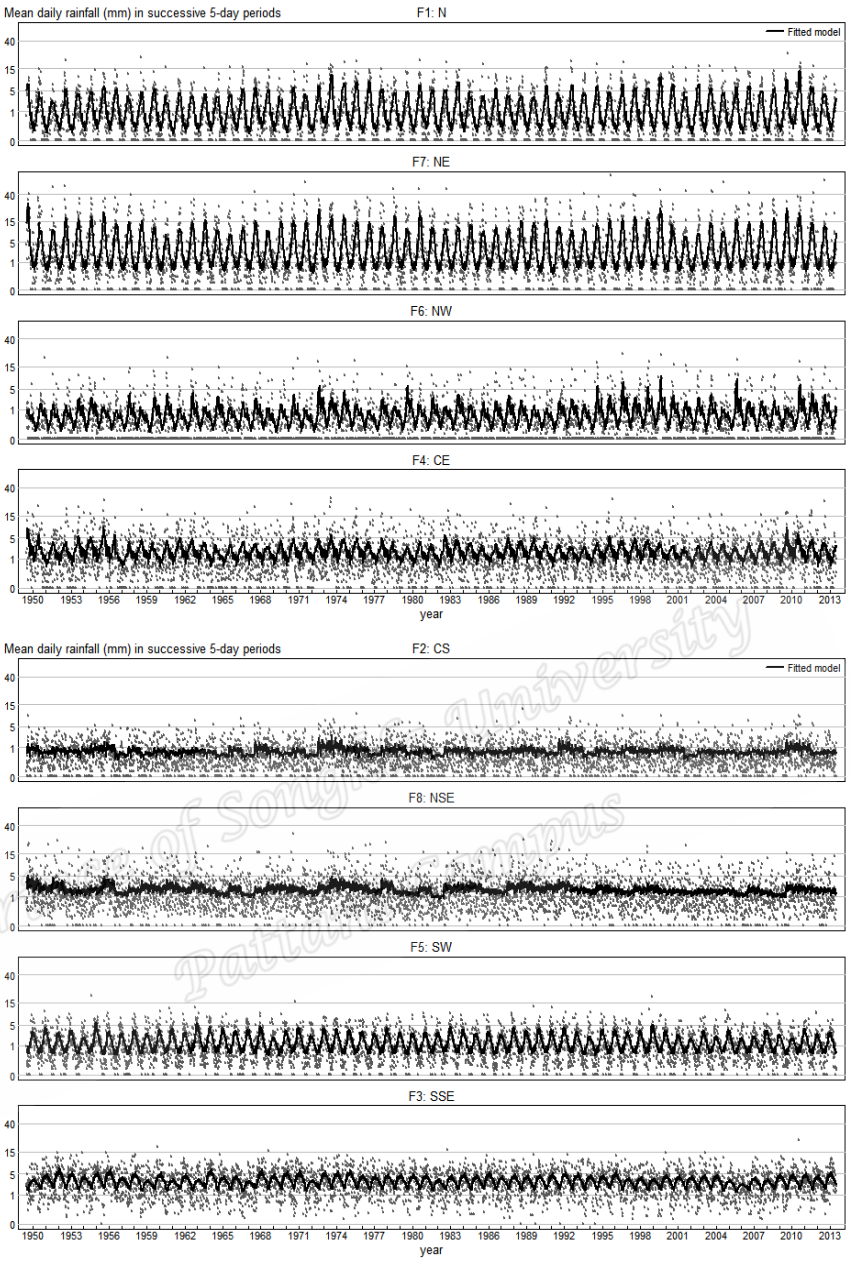


Figure 4. The results of the fitted gamma model for the 5-daily rainfall means in the eight factor regions in Australia during 1950-2013. The grey points denote the 5-day rainfall and the black line denote the fitted 5-day rainfall mean by the GLM.

well in most of the factor regions except little departures at the upper and lower tails of some of the models. These departures at the lower and upper tails of some of the

models may be as a result of rainfall extremes or other neglected influential predictors, which indicates that the gamma GLM models may only work well for non-extreme rainfall values. Similar results were seen by Hasan and Dunn (2010) using simple Poisson-gamma modelling of monthly rainfall in Australia, and a similar study by Kenabathoet al. (2012) in semi-arid Limpopo Basin in Botswana. However, Kenabatho et al. attributed the departures of the models from a normal distribution to the data quality. They established that GLM model is sensitive to data quality of rainfall observations and good quality data can improve the performance of the model. Also, the evaluation and **a**nalysis of the deviance residuals verse the fitted values of the models did not display any distinctive patterns and outliers signifying no indication of poor fit of the models. Interestingly, this study revealed that the GLM model performed better in modelling the observations on short-term period (considering the first thirty-two or last thirty-two years data separately). The model captured the periodic seasonal patterns quite well so as the timings of periods. Interestingly, the seasonal variations were observed to be higher in all the factor regions than their annual rainfall variations. The fitted models revealed significant mean rainfall amount in the NE followed by SSE and lowest in CS. Moreover, in N, NE, NW and CE, most of the annual rainfall was observed in the summer months, and these results were seen by Lavender and Abbs (2013) over north of Australia and attributed it partially to the tropical cyclones and other low pressure systems. Considering all the eight factor regions, N, NE, NW and CE receives a substantial amount of rainfall from December-March (wettest months) while uniform monthly rainfall distributions were observed in CS and NSE. Similar results were seen by Hasan and Dunn (2010) on the analysis of monthly rainfall from 1912-1971 with

Poisson and gamma models. On the other hand, SW and SSE have their wettest months between August and September.

3.4. The 95% confidence intervals for the adjusted means

The 95% confidence intervals for the adjusted means delineated the temporal variations of the fitted rainfall means (seasonal and annual) from the overall mean between 1950-2013 in Australia (Figure 6). All the models were within the 95% CI signifying that the models were represented quite well in all the regions but high variations in most parts. Seasonal rainfall increases sharply from January to February where it attains the maximum, decreases to a minimum in August and increases gradually until December in the N and NE.

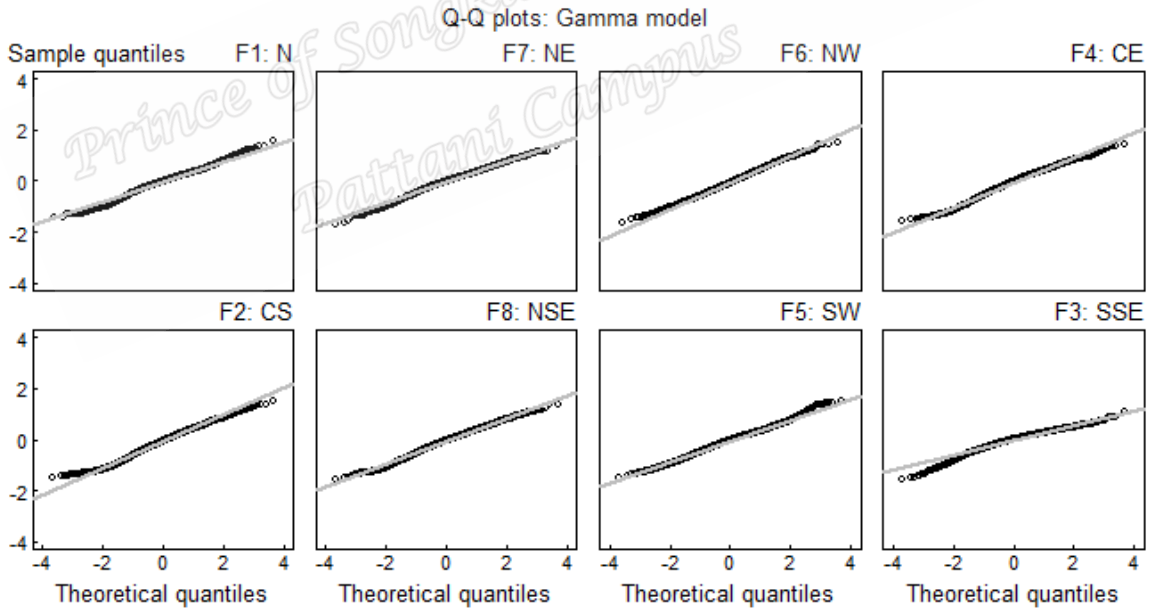


Figure 5. The residuals quantile-quantile (Q-Q) based on the gamma GLM for the fitted 5-day rainfall means for the eight regions during 1950-2013. The black lines denote the residual plots while the grey line indicates the line of best fit for the correct gamma model.

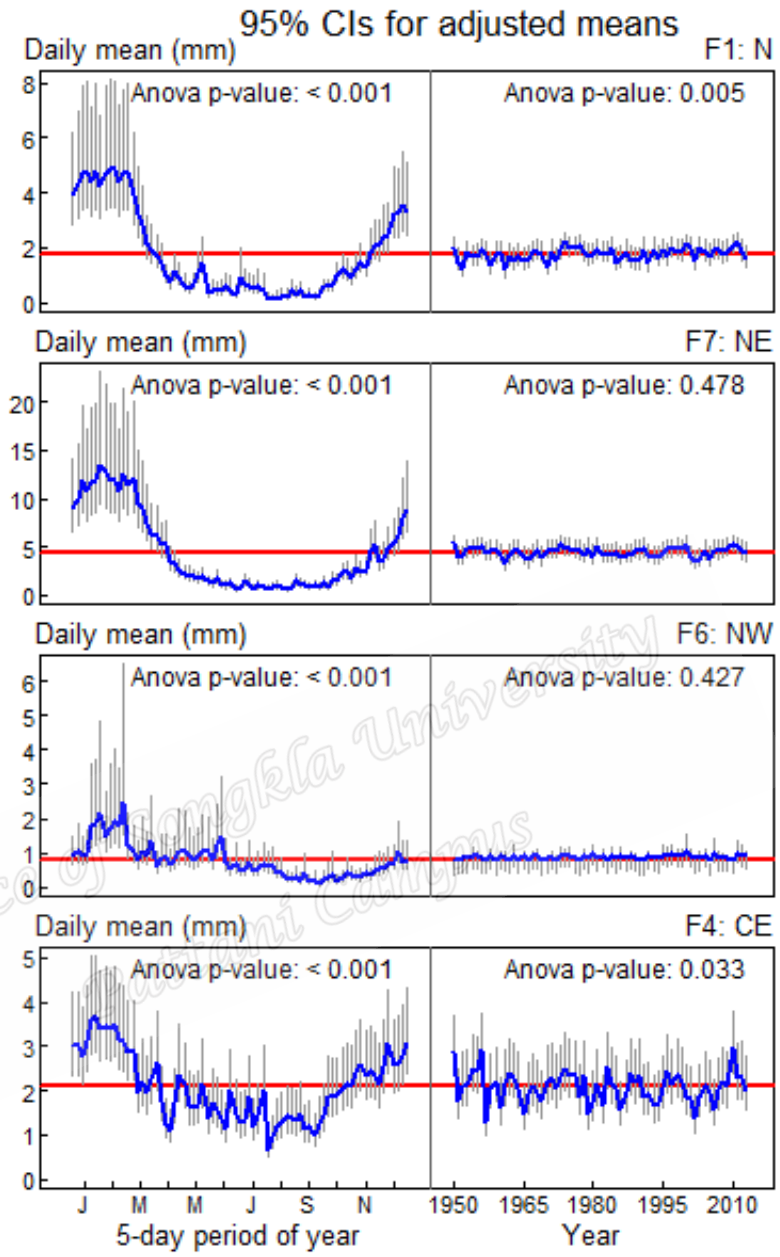


Figure 6a. The 95% confidence interval plots for the adjusted means of the 5-day rainfall from in N, NE, NW, and CE during 1950-2013. The left panels show the 5-day periods effects while the right panels show the annual effects. The horizontal red line indicates the overall mean and the blue curve is the fitted rainfall mean. The grey bars are the 95% CIs.

The NW receives above average rainfall mostly between January and June and experiences below average until December while in the CE below average rainfall mostly occurs between April and October. However, no clear seasonal patterns were observed in the CS and the NSE regions. In the SW and SSE regions, seasonal rainfall increases sharply from February to June and decreases from July to December. These seasonal patterns were significant in all the factor regions ($p< 0.001$) except CS and NSE regions (first and second panels of Figure 6b). Also, annual rainfall has a very low variation concerning the overall rainfall mean in the N, NE, NW and the SW regions ($p< 0.05$ in N, but $p> 0.05$ in the NE, NW and SW regions). Annual rainfall observed between the remaining regions varies considerably from the overall means during 1950-2013.

5. Conclusion

This study showed how a class of statistical models could be used to analyse daily rainfall to provide essential information on its variability patterns. In the first model, the estimated 5-day average rainfall based on the data in each station of the 64 years (1950-2013) collected from92 observational stations in Australia were grouped into distinct geographical regions by using factor analysis. The result showed that eight factors were adequate and accounted for about 52% of the observed correlations in the data, thus grouping Australia into eight geographical rainfall regions. The model revealed evident spatial rainfall variability during the 64 years. The analysis of similarities in the seasonal evolutions between the eight regions by correlation analysis revealed three seasonal rainfall distributions. These seasonal patterns were observed to be similar in the N, NE, NW and CE regions.

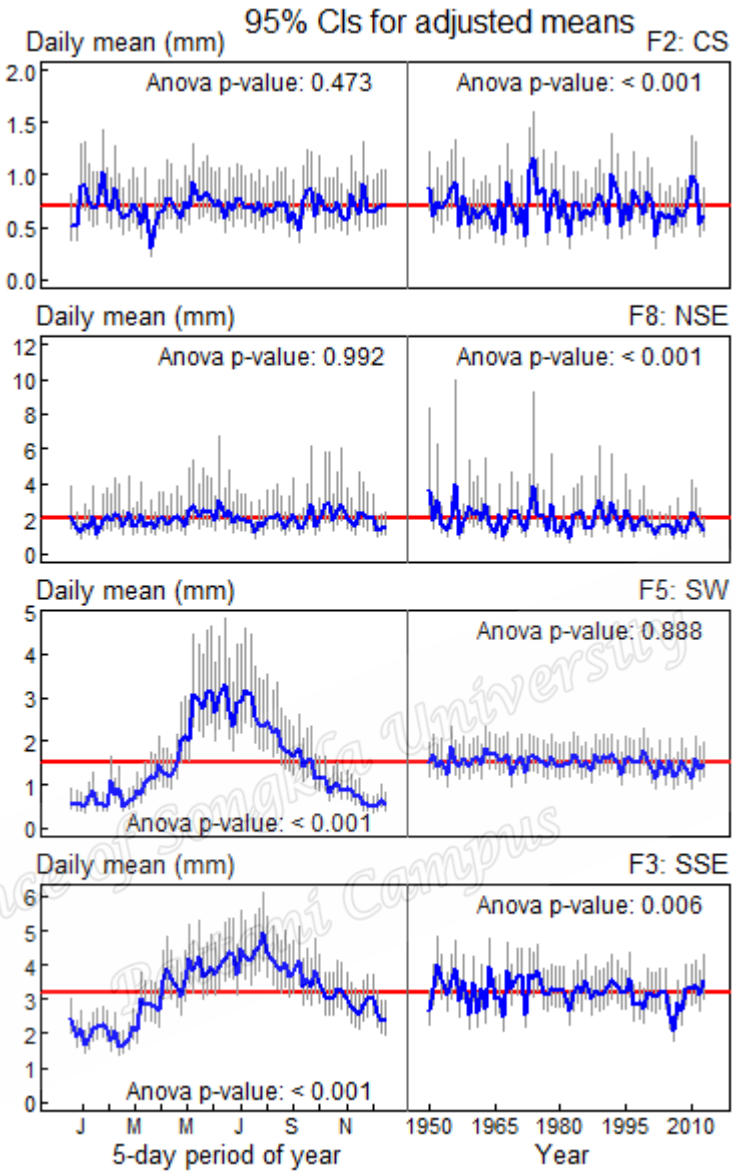


Figure 6b. The 95% confidence interval plots for the adjusted means of the 5-day rainfall in CS, NSE, SW and SSE during 1950-2013. The left panels show the 5-day periods effects while the right panels show the annual effects. The horizontal red line indicates the overall mean and the blue curve is the fitted rainfall mean. The grey bars are the 95% CIs.

The CS and NSE regions were seen to have similar seasonal distributions while that of SW and SSE were also observed to be related.

In general, these observed spatial patterns that comprise of eight regions, with three main seasonal groups with each having its unique dominating factor is essential in the planning of resources allocation in numerous areas of Australia.

The gamma GLM was then used to model the non-zero rainfall amount with the seasonal and the annual factors as predictors. The gamma GLM fitted the data quite well in most of the factor regions except little departures at the upper and lower tails of some of the models. The model could be used to simulate rainfall data for the areas with inadequate rainfall records.

Further studies could explore the mix factors, extreme rainfall values and examine the relationship between the seasonal rainfall in the factor regions with climate variables such as (humidity and temperature), climate indices/teleconnection and model the probability pattern of each region.

Acknowledgment

This work was supported by the Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission under grant no. THE-034. We are also grateful to BOM for the provision of the daily accumulated data. Our tremendous thanks go to Emeritus Prof. Dr. Don McNeil for his comments and support during this work.

References

Bagirov, A. M., Mahmood, A., & Barton, A. (2017). Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach, *Atmospheric Research*, 188, 20-29.

Bärring, L. (1987). Spatial patterns of daily rainfall in central Kenya: Application of Principal Component analysis, common factor analysis and spatial correlation. *International Journal of Climatology*, 7**,** 267-289. doi:10.1002/ jjoc.3370070306.

Bärring, L. (1988). Regionalization of daily rainfall in Kenya by means of common factor analysis. *International Journal of Climatology*, 8, 371-389. doi:10.1002/joc.3370080405.

Beecham, S., Rashid, M., & Chowdhury, R. K. (2014). Statistical downscaling of multi-site daily rainfall in a South Australian catchment using a Generalised Linear Model. *International Journal of Climatology*, 34, 3654-3670. doi:10.1002/joc.3933.

Bukantis, A. (2002). Application of factor analysis for quantification of climate forming processes in the eastern part of the Baltic Sea region. *Climate Research*, 20, 135-140. doi:10.3354/cr020135.

Cai, W., Cowan, T., & Thatcher, M. (2012). Rainfall reductions over Southern Hemisphere semi-arid regions: the role of subtropical dry zone expansion. *Scientific Reports*, 2, 1-5. doi:10.1038/srep00702.

Chambers, L. E. (2003). South Australian rainfall variability and trends. BMRC Research Report No. 92, Commonwealth of Australia.*Bureau of Meteorology Research Centre*, 51 pp.

Cheung, K. K., Chooprateep, S., & Ma, J. (2015). Spatial and temporal patterns of solar absorption by clouds in Australia asrevealed by exploratory factor analysis.*Solar. Energy*, 111, 53-67. doi:10.1016/j.solener.2014.10.014.

Coe, R., & Stern, R. D. (1982). Fitting models to daily rainfall data. *Journal of AppliedMeteorology*, 21, 1024 1031.

Deo, R., Shahin, M., (2015). Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in Eastern Australia. *Atmospheric Research*, 161, 65-81.

Drosdowsky, W. (1993). An analysis of Australian seasonal rainfall anomalies: 1950-1987.II: temporal variability and teleconnection patterns. *International Journal of Climatology*,13, 111-149. doi:10.1002/joc.3370130102.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate data analysis*. Prentice Hall International (UK) Limited: London.

Hasan, M. M., & Dunn, P. K. (2010). A simple Poisson-gamma model for modelling rainfalloccurrence and amount simultaneously. *Agricultural and Forest Meteorology*, 150,1319-1330. doi:10.1016/j.agrformet.2010.06.002.

Herring, S. C., Hoell, A., Hoerling, M. P., Kossin, J. P., Schreck, C. J III, & Stott, P. A., Eds. (2016). Explaining Extreme Events of 2015 from a Climate Perspective.Bulletin of the American Meteorological Society, 97 (12), S1-S145.

Kenabatho, P. K., McIntyre, N. R., Chandler, R. E., & Wheater, H. S. (2012). Stochastic simulation of rainfall in the semi-arid Limpopo Basin, Botswana. *International Journal of Climatology*,32, 1113-1127. doi:10.1002/joc.2323.

Köppen, W. P. (1918). Klassification der Klimate nach Temperatur, Niederschlag und Jahreslauf. *Petermmans*, 64, 243-248.

Lavender, S. L., & Abbs, D. J. (2013). Trends in Australian rainfall: contribution of tropicalcyclones and closed lows. *Climate Dynamics*, 40, 317-326. doi:10.1007/s00382-012-1566-y.

Lin, G.-F., & Jhong, B.-C. (2015). A real-time forecasting model for the spatial distribution of typhoon rainfall. *Journal of Hydrology*, 521, 302-313.

Peel, M. C., Finlayson, B. L., & Mcmahon T. A. (2007). Updated world map of theK¨oppen-Geiger climate classification. *Hydrology and Earth System Sciences*Discussions, European Geosciences Union, 11 (5), pp.1633-1644. Sciences Union, 11 (5), pp.1633-1644.

R Development Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL [Version 3.2.3]. Retrieved from *http://www.R-project.org*.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, (2nd ed). John Wiley and Sons, Inc., New York, U.S.A.

Shukla, R., Tripathi, K., Pandey, A., & Das, I. (2011). Prediction of Indian summer monsoon rainfall using Nino indices: a neural network approach. *Atmospheric Research*, 102(1-2), 99-109.

Srikanthan, R., & Pegram, G. G. S. (2009). A nested multi-site daily rainfall stochastic generation model. *Journal of Hydrology*, 371, 142-153. doi:10.1016/ j.jhydrol.2009.03.025.

Stern, R. D., & Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society, Series A*, 147, 1-34.

Taschetto, A. S., & England, M. H. (2009). An analysis of late twentieth century trends in Australian rainfall. *International Journal of Climatology*, 29, 791-807.doi:10.1002/joc.1736.

Tongkumchum, P., & McNeil, D. (2009). Confidence intervals using contrasts for regression model. *Songklanakarin Journal of Science and Technology*, 31, 151-156.

Toth, E., Brath, A., & Montanari, A. (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, 239 (1), 132-147.

Trenberth, K. E., Dai, A., Rasmussen, R. M., & Parsons, D. B. (2003). The changing character of precipitation. *Bulletin of the American Meteorological Society*, *84*(9),1205-1217. doi.org/10.1175/BAMS-84-9-1205

Um, M. -J., Yun, H., Jeong, C.-S., & Heo, J. -H. (2011). Factor analysis and multiple regressions between topography and precipitation on Jeju Island, Korea. *Journal of Hydrology*, 410, 189-203. doi:10.1016/j.jhydrol.2011.09.016.

Unkel, S., Trendafilov, N. T., Hannachi, A., & Jolliffe, I. T. (2010). Independent exploratory factor analysis with application to atmospheric science data. *Journal of Applied Statistics*, 37, 1847-1862. doi:10.1080/0266476090316 6280.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*, (4th ed). Springer, Queensland.

Wanishsakpong, W., & McNeil, N. (2016). Modelling of daily maximum temperatures over Australia from 1970 to 2012. *Meteorological Applications*, *23*(1), 115-122. doi: 10.1002/met.1536

Wickramagamage, P. (2010). Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. *International Journal of Climatology*, 30, 1235-1245. doi:10.1002/joc.1977.doi:10.1002/joc.1977.

Appendix II

Statistical modelling of daily rainfall variability patterns in Australia

Bright Emmanuel Owusu[1,2] * Nittaya McNeil[1]

[1]Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Mueang Pattani, 94000, Thailand

[2]Department of Information and Communication Technology/Mathematics, Faculty of Science and Technology, Presbyterian University College Ghana

ABSTRACT

The present study uses statistical methods specifically multiple regression (MR) and gamma generalised linear models (Gamma GLM) for the modelling and analysis of period rainfall variability in Australia during 1950-2013. The data from 92 observational stations was collected from Australian Bureau of Meteorology. Factor analysis was first used to group the 92 stations into eight rainfall regions. The consecutive 5-day mean rainfall in a year (period) in each region is then modelled using MR and the Gamma GLM. The results show that the models fitted the data quite well in all the regions, but the MR model did better than the Gamma GLM in some of the regions. The MR models revealed three rainfall groupings and each group has diverse rainfall pattern and trends. Significant decreasing annual rainfall trend was revealed in the southwest and the north southeast regions. In contrast, substantial increasing annual rainfall trends were found in the north and the northwest regions.

Keywords: Australia, factor analysis, gamma generalised linear models, multiple regression, rainfall.

INTRODUCTION

The increasing in severe weather conditions such as flood and drought in many areas of the globe have been attributed to climate change (O'Gorman & Schneider, 2009). Conversely, intensified extreme rainfall events are not the only concerns of climate change. Other features of rainfall such as the annual mean, occurrence and amount of events or the time of the year they occur, the sequence and duration of these events can also be affected (Kumar, 2013). Consistent water sources and supply are essential particularly to the increasing population of human society. They can be rigorously affected by long-term changes in rainfall. The fluctuations in the spatial distribution of rainfall will increase differences between dry and wet areas (Held & Soden, 2006; Allan et al., 2010). Therefore, modelling rainfall variability is essential for the determination of the various type of life in the wet tropical regions and mainly for production in dry areas in some countries.

Numerous authors have studied daily rainfall variability and patterns using various statistical techniques. One of the simplest models considered in the analysis and modelling of rainfall is MR. Mekanik et al. (2013) applied MR and Artificial Neural Networks (ANN) to estimate long-term seasonal spring rainfall in Victoria. The ANN models revealed greater correlation relative to MR models, showing that ANN models explained the pattern and trend of the rainfall well as compared to MR models.

In contrast, Jeong et al. (2012) evaluated MR, robust regression, ridge regression and ANN models to establish suitable transfer function in statistical downscaling models for the daily maximum and minimum temperatures and daily precipitation

occurrence and amounts. They revealed that the monthly MR, annual ANN and annual MR models performed better than the robust and ridge regression models based on the modified Akaike information criterion. In regression modelling, observations must be independent and normally distributed, and this assumption may not be valid for most time series data such as rainfall. Rainfall data are highly skewed even after aggregation, and in modelling such data, transformations are normally applied. Meng et al. (2007) applied a logarithmic transformation on monthly rainfall to attain normality. However, application of this transformation to the summer monsoon rainfall amounts in Asia by Mooley (1973) revealed poor results.

Alternatively, the transformation of the data to normality can be overlooked and it can directly be modelled using non-normal distributions. In the modelling of periods of non-zero rainfall amount, some authors have revealed that the Gamma GLM fits the data quite well (Coe & Stern 1982; Stern & Coe 1984; Kenabatho et al., 2012). Rainfall data between adjacent observational stations are spatially correlated, and numerous methods can be used to handle this concern. Factor analysis has currently gained attention in climate research to describe the variability of correlated observations with a possibly lesser number of unobserved variables called factors. This method has been used in studies such as analysis of the variability of temperature (McNeil & Chooprateep, 2014; McNeil & Chirtkiatsakul 2016; Wanishsakpong & McNeil, 2016 ;), solar radiation (Cheung et al., 2015) and rainfall (Wickramagamage, 2010). It was applied to put the period rainfall of the meteorological stations into groups.This study uses MR and Gamma GLM models in describing the variability and patterns of five-daily rainfall observations in Australia during 1950 to 2013.

MATERIALS AND METHODS

Daily accumulated rainfalls from 92 stations that are dispersed throughout Australia for the period between 1950 and 2013 were acquired from the Australian Bureau of Meteorology. The data are analysed by using statistical methods. The period rainfall for each station was computed to minimise the skewness and thus reduces the number of observations to 73 in each year after removing data on 29 February. Data on leap years were excluded, and this gave equal observations for all years. Moreover, the period rainfall was deemed appropriate to use because some of the stations had data on aggregation as far as five days. The computed mean was based on the number present in cases where data are missing. After computing the 5-day average rainfall, only one of the stations had more than 6% (7.3%) missing data. The remaining stations have mostly less than 1% missing data. However, after finding the 5-day average region wise, only one region had 0.2% missing value, and it was imputed by linear regression in fitting the MR model and GLM in fitting the Gamma GLM models. Factor analysis was initially used to put all the stations into groups, and this was done by conducting maximum likelihood factor analysis on the correlation matrix of the period rainfall of each station. The model with p factors can be written as:

$$y_q = \mu_q + \sum_{m=1}^{p} \lambda_q^{(m)} \phi^{(m)} \qquad [1]$$

where μ_q, are means over the five-days at each station, λ_q are the factor loadings and ϕ are the common factors. The factors were recognised via their loading values. Promax one of oblique's rotation methods was applied to control the loadings

(Venables & Ripley, 2002). A station is assigned to a factor if it loading is at least 0.33 (Hair et al., 1998).

The MR and Gamma GLM models were then fitted to the non-zero period rainfall observations in the groups revealed by the factor analysis. These methods used the mean period rainfall observations from each region as the dependent variable, the year and period factors as the independent variables in fitting the models. Brief accounts of these methods are given below:

Let μ_{kst} be the mean period rainfall observation for each region k (k,…, 8) at period s and year t, then the MR model in this paper is

$$h(\mu_{kst}) = \mu_{0k} + \sum_{i=1}^{64} \eta_{ik} x_{ti} \sum_{j=1}^{73} \beta_{jk} w_{sj} + h(\mu_{t(s-1)k}) + \varepsilon_{kst} \qquad [2]$$

where h(.) represents the dependent variable fourth root transformation which is defined as $h(\mu_{tsk}) = \sqrt[4]{\mu_{tsk}}$, ε_{tsk} are random error terms which are normally distributed having mean 0 and variance σ^2. x_{ti} are predictor for relating years *t* and year *i*, w_{sj} are predictor for relating periods s and period j. Thus, x_{ti} becomes 1 if $t = i$ and 0 elsewhere and w_{sj} becomes 1 if $s = j$ and 0 elsewhere. The year 1950 and period 1 are the starting year and period respectively. For region k, μ_{0k} denotes the mean period rainfall, which is the starting year and period, η_{ik} is the year *i* effect and β_{jk} the period *j* effect.Fitted values $h(\mu_{kst})$ from the model [2] have to be transformed backward to give fitted values of μ_{kst}. Thus the backwards transformed can be written as μ_{kst}. Let $\hat{\mu}$ be the fitted value for μ and $h(\mu)$ the above defined transformation. Then $\hat{\mu}$ can be estimated from the fitted value h($\hat{\mu}$) as $\hat{\mu} = (h(\hat{\mu}))^4$.

The probability density function of the Gamma distribution is normally written as

$$f(x \mid \alpha, \gamma) = \frac{1}{\Gamma(\alpha)} \gamma^{\lambda} x^{\lambda-1} e^{-\gamma x},\ x, \alpha, \gamma > 0 \qquad [3]$$

where α and γ are the shape and the scale parameters respectively and x is the non-zero period rainfall amount and α is assumed to be constant throughout the data set for each region. If μ_s is the mean rainfall amount for period s related to predictors ξ_s then the mean period rainfall amount is estimated by equation [4] where τ_0 is constant and τ's and ω are slopes.

$$\ln(\mu_s) = \tau_0 + \tau_1\xi_1 + \tau_2\xi_2 + \ldots + \tau_k\xi_k + \omega\mu_{s-1} \qquad [4]$$

R programming language is used for the data analysis and graphical displays (R Development Core Team, 2015).

RESULTS AND DISCUSSION

The factor model categorised the 92 stations into eight groups (factors) via the factor loadings. Table 1 gives the loadings of the factors organised in descending order within each factor and the uniqueness values excluding the mixed factors. The eight factors were made up of the following regions of Australia: factor 1 (F1) the north (N), factor 2 (F2) the central south (CS), factor 3 (F3) the south southeast (SSE), factor 4 (F4) the central east (CE). Factor 5 (F5) the southwest (SW), factor 6 (F6) the northwest (NW), factor 7 is the northeast (NE) and factor 8 (F8) is the north southeast (NSE). Figure 1 shows the stations as grouped by factor analysis. The classification reveals the spatial rainfall distribution of the 92 stations used in the study. A similar classification was uncovered by factor analysis (Cheung et al., 2015) in the analysis of

Table 1

The Factor Loading of the 92 Stations with the Identified Dominating Factors in Bold

Stations	Factor Loadings								Uniqueness
	F1	F2	F3	F4	F5	F6	F7	F8	
15085	**0.828**								0.381
15086	**0.810**								0.401
15015	**0.788**								0.464
37043	**0.744**			0.115	0.163	-0.132			0.435
14815	**0.731**	-0.145				0.101		0.109	0.405
15005	**0.728**				0.127	-0.106			0.510
2012	**0.703**				-0.119	0.205			0.400
14902	**0.690**	-0.145			-0.103		0.159	0.105	0.309
29008	**0.620**			0.119	0.116	-0.117	0.196		0.489
38003	**0.602**	0.111		0.192	0.156	-0.146			0.516
15590	**0.584**	0.316		0.106			-0.214	-0.107	0.512
3027	**0.562**	-0.114			-0.118	0.286			0.489
14015	**0.512**	-0.118			-0.219	0.107	0.205	0.104	0.349
38020	**0.347**	0.260				-0.114			0.746
14042	**0.499**	-0.112		-0.122	-0.178	0.107	0.325	0.122	0.280
18012	-0.204	**0.804**	0.148						0.353
16005		**0.804**							0.461
18002	-0.234	**0.775**				0.106			0.449
16031		**0.746**	-0.126			0.127			0.536
16000		**0.739**	-0.124						0.533
18091	-0.205	**0.712**	0.221						0.386
11003	-0.205	**0.655**				0.212			0.616
24003		**0.650**	0.176	-0.107				0.254	0.364
17031	0.198	**0.646**	-0.101						0.520
20017		**0.593**						0.187	0.566
16047	0.317	**0.518**					-0.161		0.604
46043		**0.434**		0.163				0.321	0.495
48115		**0.346**		0.249				0.259	0.578
98003			**0.886**						0.188
90015			**0.874**						0.210
91034	0.107		**0.860**				-0.102		0.293
91033	0.111		**0.772**						0.387
90059	-0.103	0.228	**0.755**						0.232
99005			**0.752**					0.207	0.382
94027		-0.184	**0.714**				-0.124	-0.162	0.557
97006		-0.174	**0.688**		-0.121		-0.126	-0.279	0.549
89002		0.197	**0.661**					0.290	0.263
92018			**0.444**					0.253	0.728
35029				**0.806**					0.414
41100				**0.803**				0.152	0.395
43020				**0.792**					0.377
54023				**0.742**			-0.173	0.297	0.373
48031				**0.721**				0.249	0.357
35065			0.115	**0.682**			0.298	-0.128	0.411
39020				**0.580**			0.344		0.534
40141	-0.145		-0.124	**0.559**			0.207		0.594
59000		-0.118	-0.254	**0.552**		0.109		0.199	0.520
44181	0.144	0.225		**0.380**		-0.112		0.168	0.569
10620		-0.169			**0.918**				0.282
8230		-0.126			**0.849**	0.137			0.333
8004					**0.809**	0.102			0.336

Table 1. (continue)

Stations	Factor Loadings								Uniqueness
	F1	F2	F3	F4	F5	F6	F7	F8	
9518					**0.775**	-0.156			0.299
12011					**0.736**	0.265			0.420
9594		0.165			**0.554**				0.563
7059						**0.724**			0.511
7103		0.113				**0.718**			0.514
12108		0.171				**0.674**			0.538
13002		0.183			-0.153	**0.637**			0.588
4006	0.175				-0.156	**0.628**			0.536
7049					0.267	**0.615**			0.526
5001	0.134				-0.116	**0.601**			0.572
5008						**0.594**			0.630
6029					0.129	**0.554**	0.107		0.651
4019	0.290				-0.101	**0.487**			0.615
33001		0.128		0.152			**0.741**		0.394
33013				0.284			**0.711**	-0.118	0.401
32001	0.163	0.125	-0.108				**0.698**		0.353
31062	0.181	0.129	-0.150	-0.201			**0.645**		0.471
27042	0.328			-0.132	-0.188		**0.582**		0.308
27005	0.261			-0.139			**0.477**		0.401
73025		0.146	0.218	0.124				**0.592**	0.352
69018			-0.193	0.268				**0.542**	0.565
84016			0.114	0.116				**0.530**	0.661
80044		0.333	0.329					**0.455**	0.348

spatial and temporal patterns of solar absorption by clouds and the modelling of daily maximum temperatures over Australia by Wanishsakpong and McNeil (2016).

Further concern for the modelling of time series observations is the dependency between the response observations, which contravene the independent errors assumption. Figure 2 is the autocorrelation function (ACF) which is used to evaluate the autocorrelation of mean period rainfall. Some of the sample lag values (the black vertical lines) are beyond the 95% confidence interval line (the horizontal dotted lines) which indicates significant dependencies among the response variables. The dependencies among the period rainfall in each region were minimised by adding the AR(1) term (Chatfield, 1996) to both the MR and Gamma GLM models (Figure 3). Greater part of the sample lag values are within the 95% confidence interval thus indicating a drastic reduction of serial autocorrelation in the models.

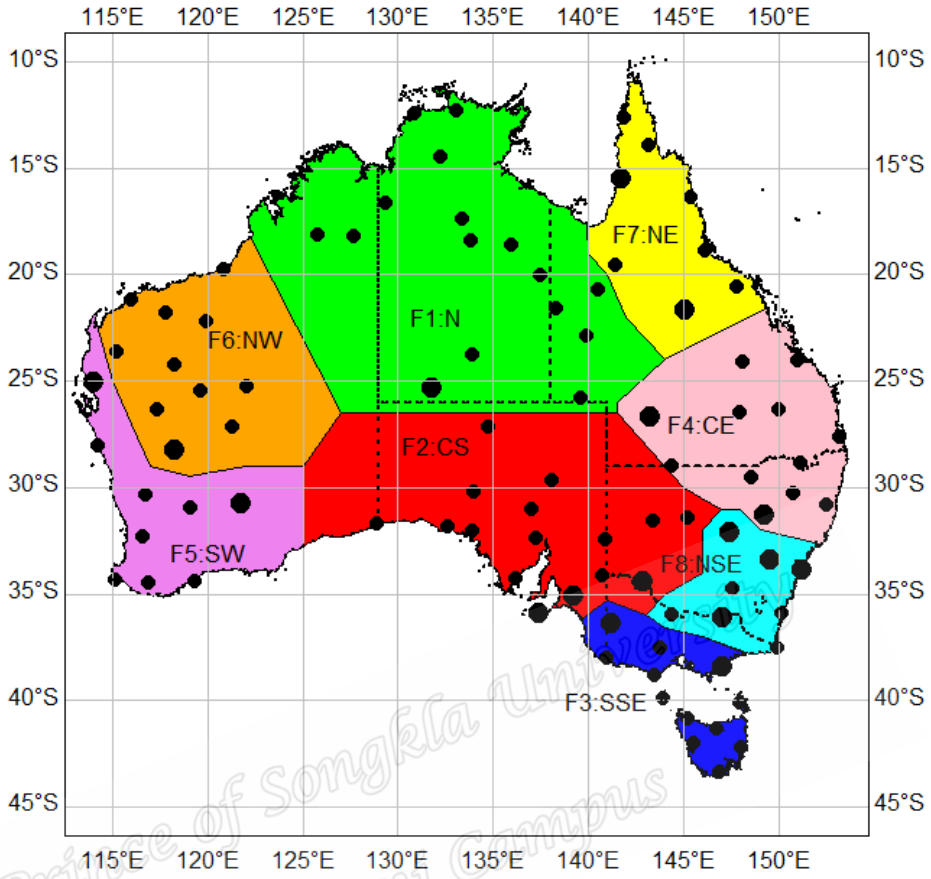


Figure 1. The locations of the stations used in the study and the groups (colours) revealed by factor analysis. The large points represent mixed factors. The state boundaries are also included (dotted lines).

Also, multicollinearity was assessed using the variance inflation factor (VIF) which is defined as $1/(1-R^2)$, where R^2 is the coefficient of determination. The response variables were said to have multicollinearity if VIF > 10 (Lin, 2008). The values of the VIF were below 5 in all the regions which revealed no indication of multicollinearity.

The pattern of period rainfall amount for all the eight regions as modelled by equation [2] and [4] are shown in Figure 4. Both models showed an obvious seasonal

periodic pattern and this is clearly seen especially in the N and NE regions where over 52% of the variations were explained by the models (coefficient of determination of 0.54 and 0.52 respectively). Moreover, the fluctuations and it timings in the data were well displayed by the models as well as the 64 years mean. However, the models underestimated the magnitude of the fluctuations.

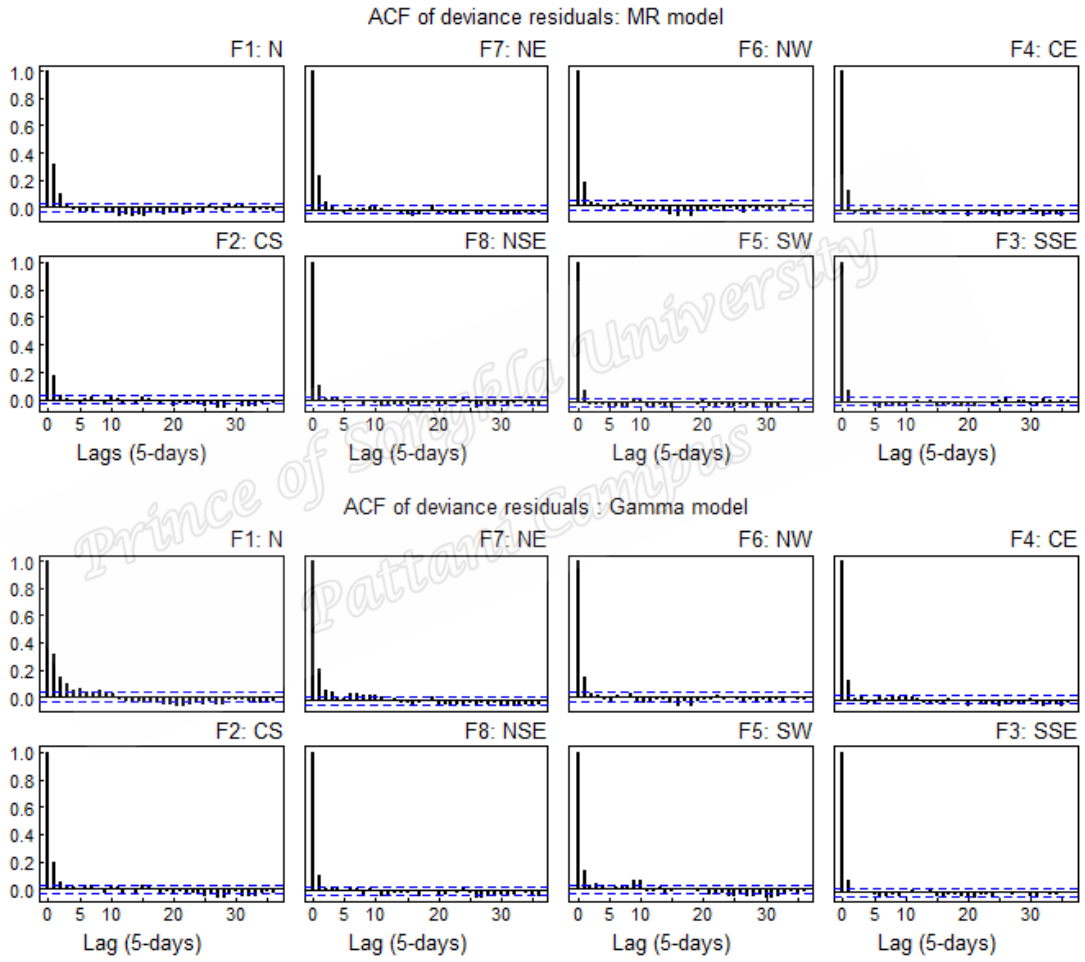


Figure 2. The autocorrelation function for the models. The top and bottom panels relate to before minimising serial correlation.

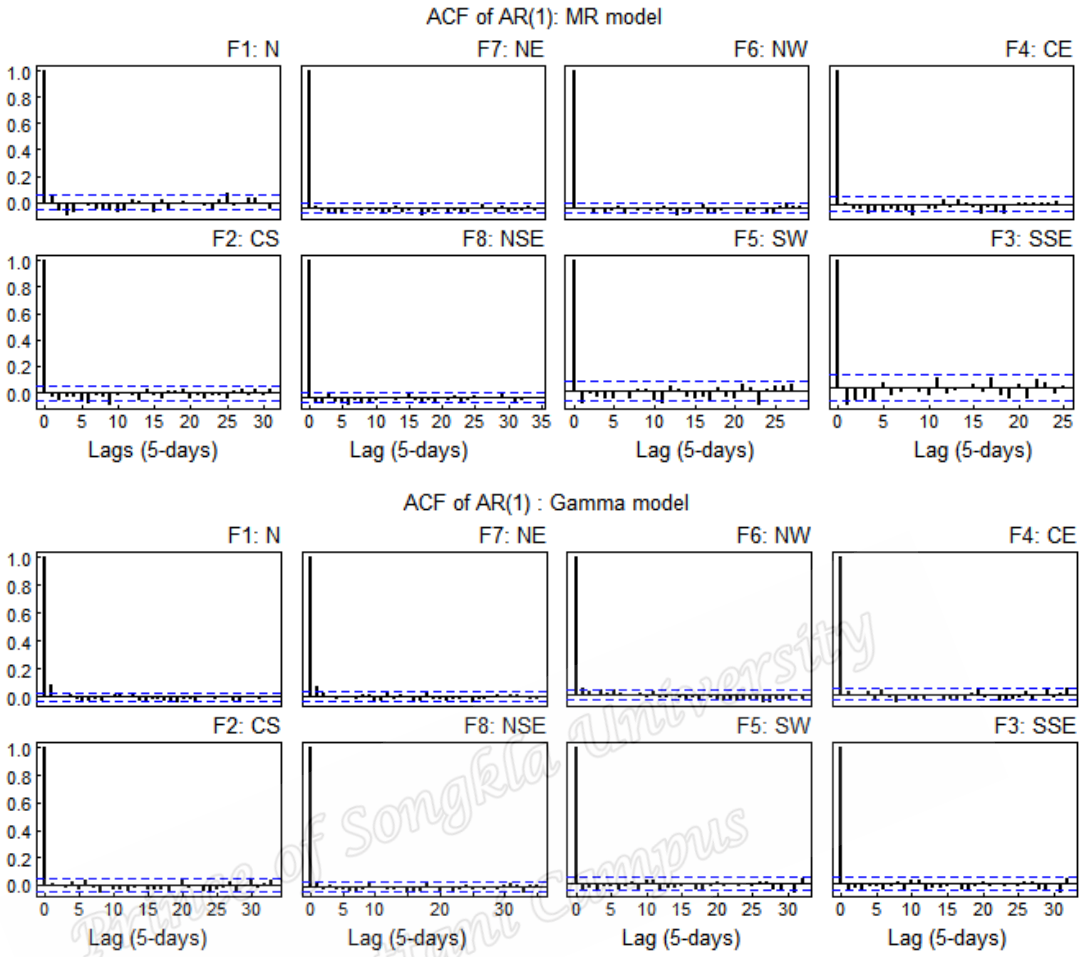


Figure 3. The autocorrelation function for the models. The top and bottom panels relate to after minimising serial correlation.

Relatively, there is not much difference between the models from both methods in the remaining regions as shown in Figure 4. The grey dots of Figure 4 represent the observed period's rainfall, while the black and red curves represent their estimated mean from models [2] and [4] respectively.The eight factor regions display diverse rainfall variability and trends for all the fitted models.The N and NE, which is in the tropical region of Australia, have similar period rainfall variation so as in the NW and the CE. The pattern in CS and NSE was also similar so as in the SW and SSE.

In general, the NE region relative to the other regions had the highest amount of period rainfall in the seasonal span, which can be due to a substantial number of temporal meteorological conditions such as fronts and low pressure systems, which possibly can disturb rainfall in this area.The area is additionally more susceptible to severe weather systems such as tropical cyclones together with a great deal of convective clouds entrenched, resulting in a greater proportion of cloud cover relative to other regions (Cheung et al., 2015).

The models were assessed by plotting residuals alongside the theoretical quantiles (Figure 5). The figure shows that the residuals plots had low variations with respect to the expected line

of best fit indicating that both models fitted the data quite well in the factor regions except little departures at the upper and lower tails of some of the models. The MR model seems to perform better than the Gamma GLM model in the N, NE and SW regions. Analysis of the deviance residuals and the fitted values of the models did not show any patterns and outliers.

Further, the MR was used to model the 64 years mean period (to show the monthly) and annual rainfall patterns for all the regions. This model was applied because it is one of the simplest methods used to describe monthly rainfall patterns (Hughes & Saunders, 2002; Oettli &Camberlin, 2005). It also fitted the period mean rainfall better than the Gamma GLM models in most of the regions.Figure 6 shows the results of the estimated 64 years periods rainfall variations for all the regions. Analysis of the graph indicated three main categories regarding the period rainfall estimates.

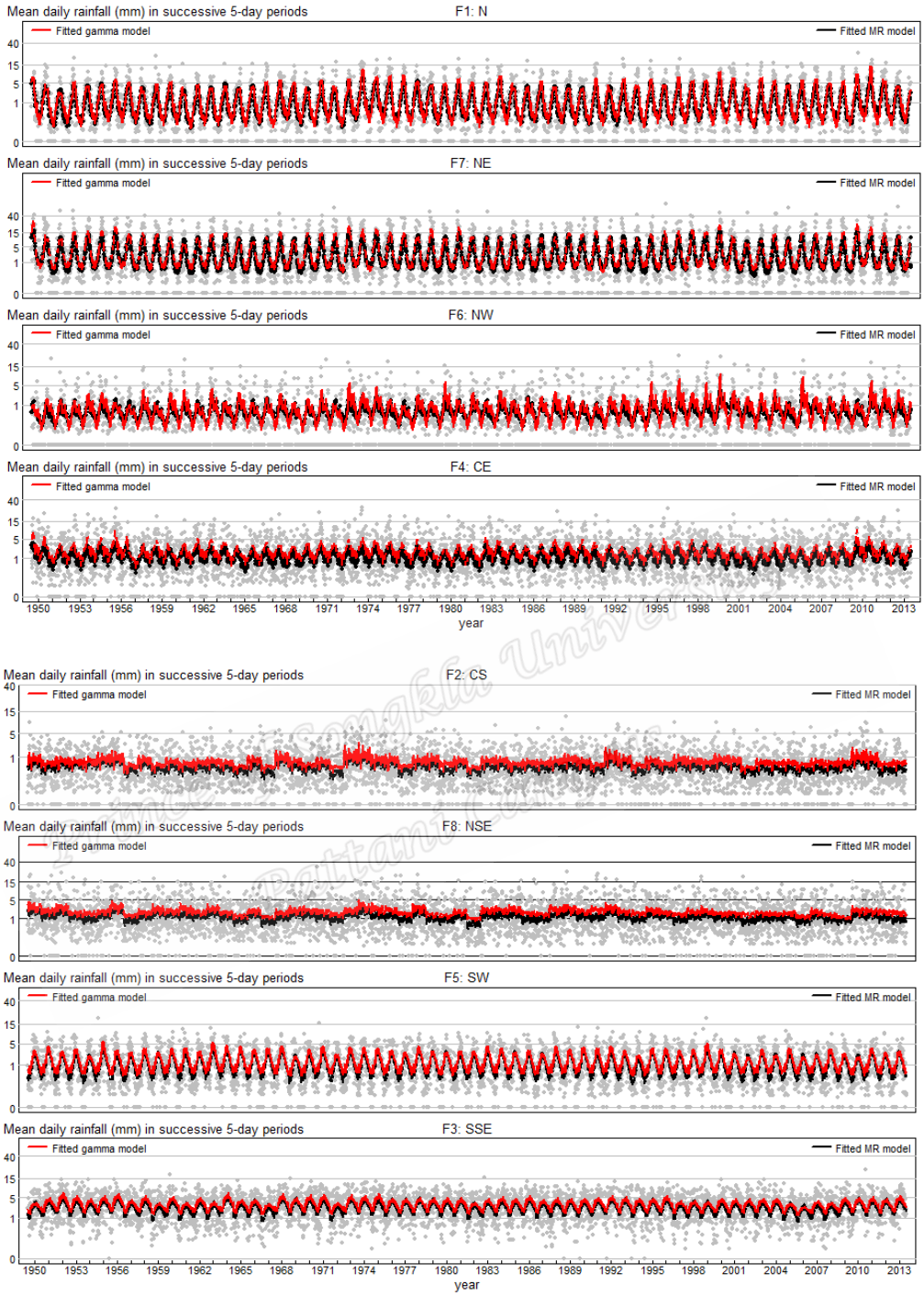


Figure 4. The plots of period rainfall and their estimated mean patterns in eight regions during 1950-2013.

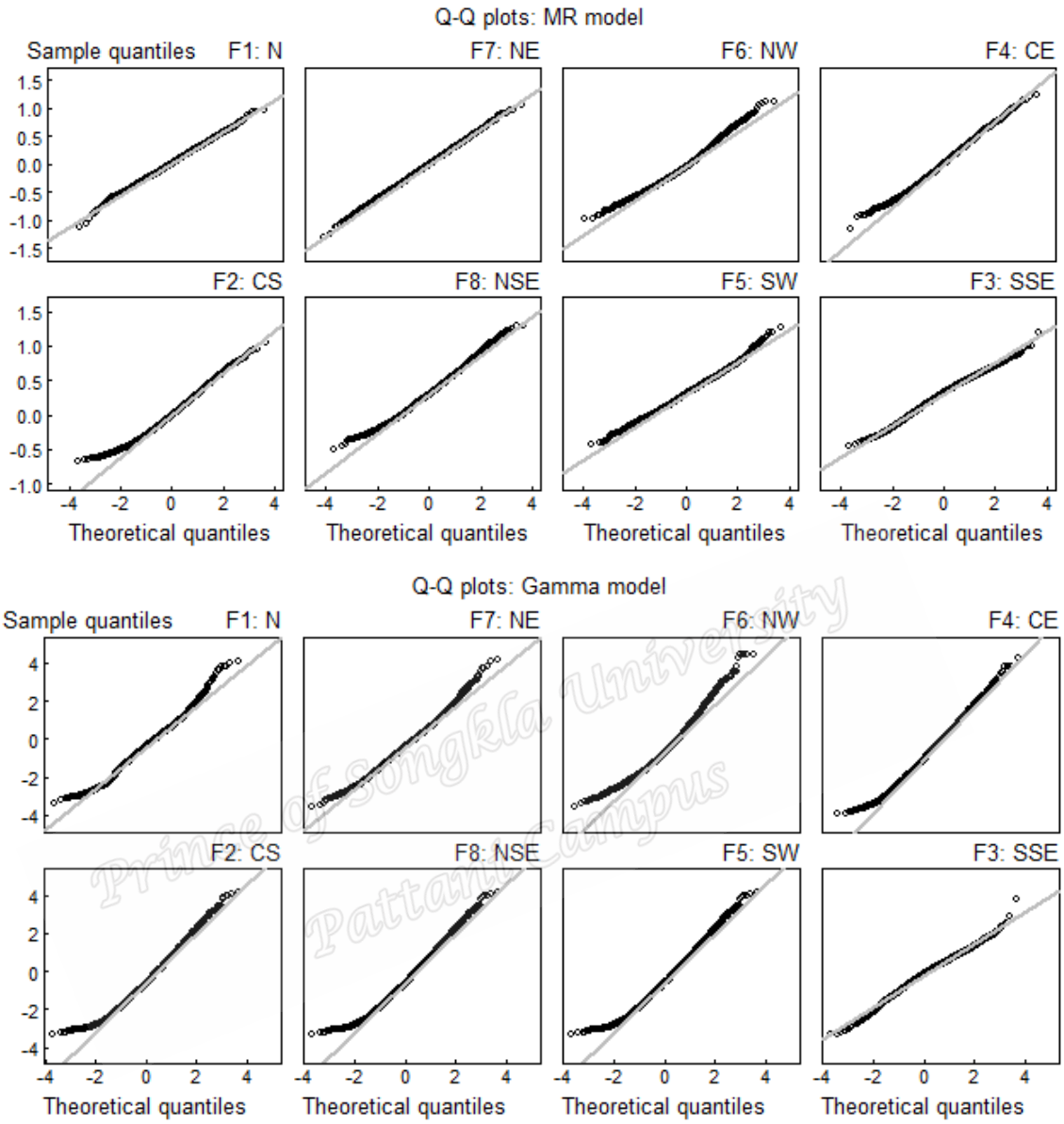


Figure 5. The residual quantile-quantile (Q-Q) plots for the models. The top graph shows that of the MR (response variable is the fourth root transformed) and the bottom shows the Gamma. Both models were filtered with AR(1) term.

The models in the N, NE, CE and NW regions can be put into a category (first group) possessing similar patterns while the models in the CS and NSE can also be put in a group (second group).

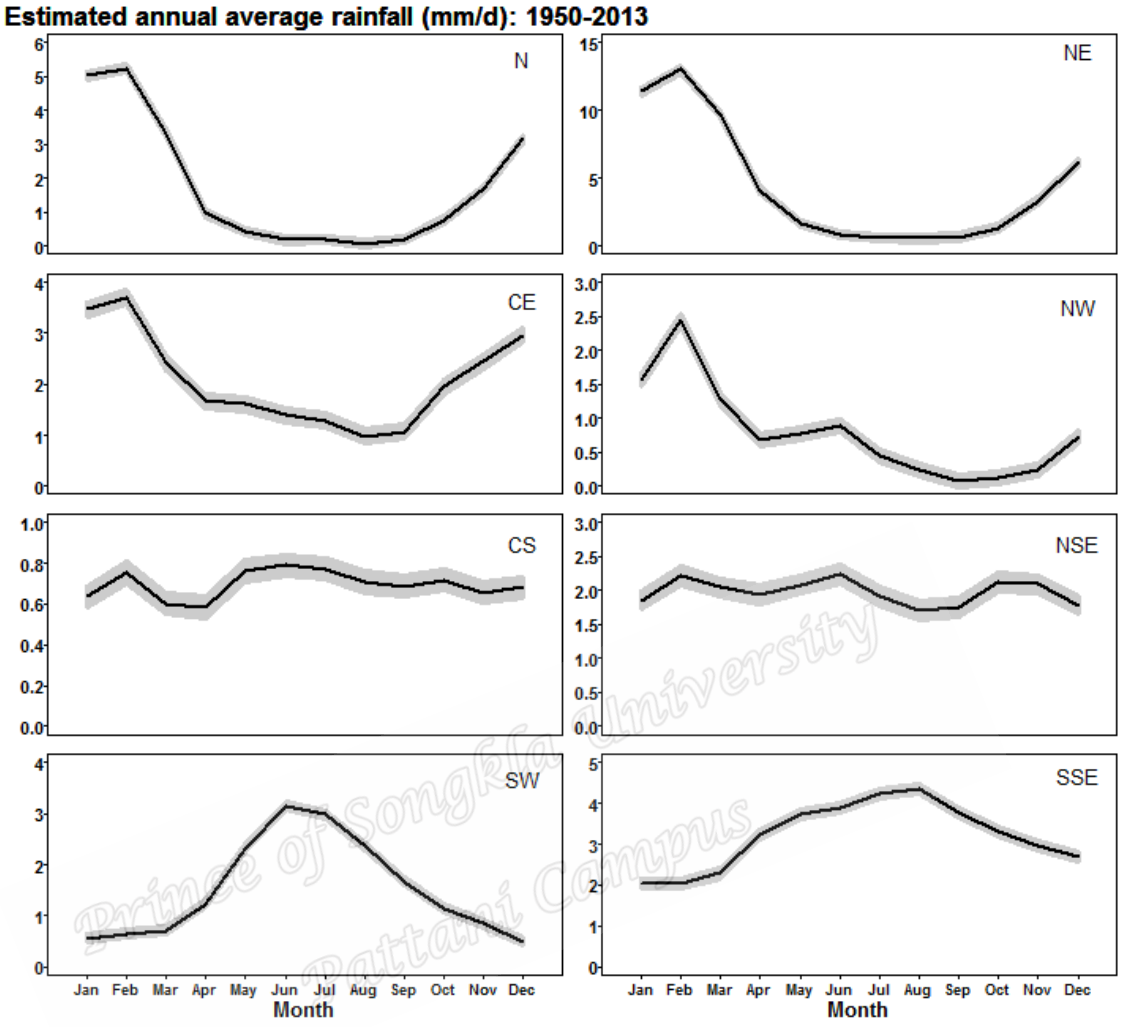


Figure 6. The estimated monthly rainfall (line) and the 95% confidence interval (shaded) in each month of a year in the factor regions.

The SW and SSE also form a group (final group) with similar patterns. In the first group, the period rainfall rises sharply from January to March, reaches its maximum between this period then decreases steadily until it minimum between August and September. The period rainfall observed in the second group is quite stable throughout the year (it varies from 0.6 to 2.2 mm d^{-1}). The largest estimated period rainfall in these regions occurred between May and July, but the minimum was observed during March and April in the CS and occurred between August and October in the NSE. On the other hand, the final group received their minimum period

rainfall between January and February. The period rainfall increased steadily from February until it attains the largest between June and August. Generally, the estimated period rainfall is different in each region: the NE, which comprises equatorial and tropical regions, has the highest in Austral summer during the active part of the monsoon while the mid-latitude region of SSE has the most during Austral winter.

Figure 7 shows the results of the estimated inter-annual rainfall pattern. In all eight regions, these patterns fluctuated between 0.3 and 7.1 mm d^{-1}, attained the lowest value between 1950 and 1975 in the NW, but the highest value occurred during 1950 and 1960 in the NE. In general, the highest annual rainfall values in most of the regions also occurred between 1950 and 1970 except in the N, CS and NW regions where it occurred between 1970 and 2012. The estimated annual rainfall amount was low especially in CS and the NW relative to the remaining regions.

Besides, examination of the 64 years annual rainfall time series by linear regression indicated negative and positive slopes for the regions (Table 2). Analysis of Table 2 revealed that the CE, CS, NSE, SW and SSE regions were characterised by negative slopes, which indicates decreasing trends. Significant decreasing trends were evident in the NSE and SW regions. The IOCI (2002) also reported long-term decline in winter (May-October) rainfall of 15-20% since the 1970s in the southwestern corner of Western Australia. Average spring rainfall during 1997-2006 in the southeast of Australia was revealed to be below average and resulted in the problem of recurring drought in this part of Australia (Murphy & Timbal, 2008). Hope et al. (2010), revealed similar results in the analysis of temporal relationship concerning rainfall variability in the SW and southeast of Australia.

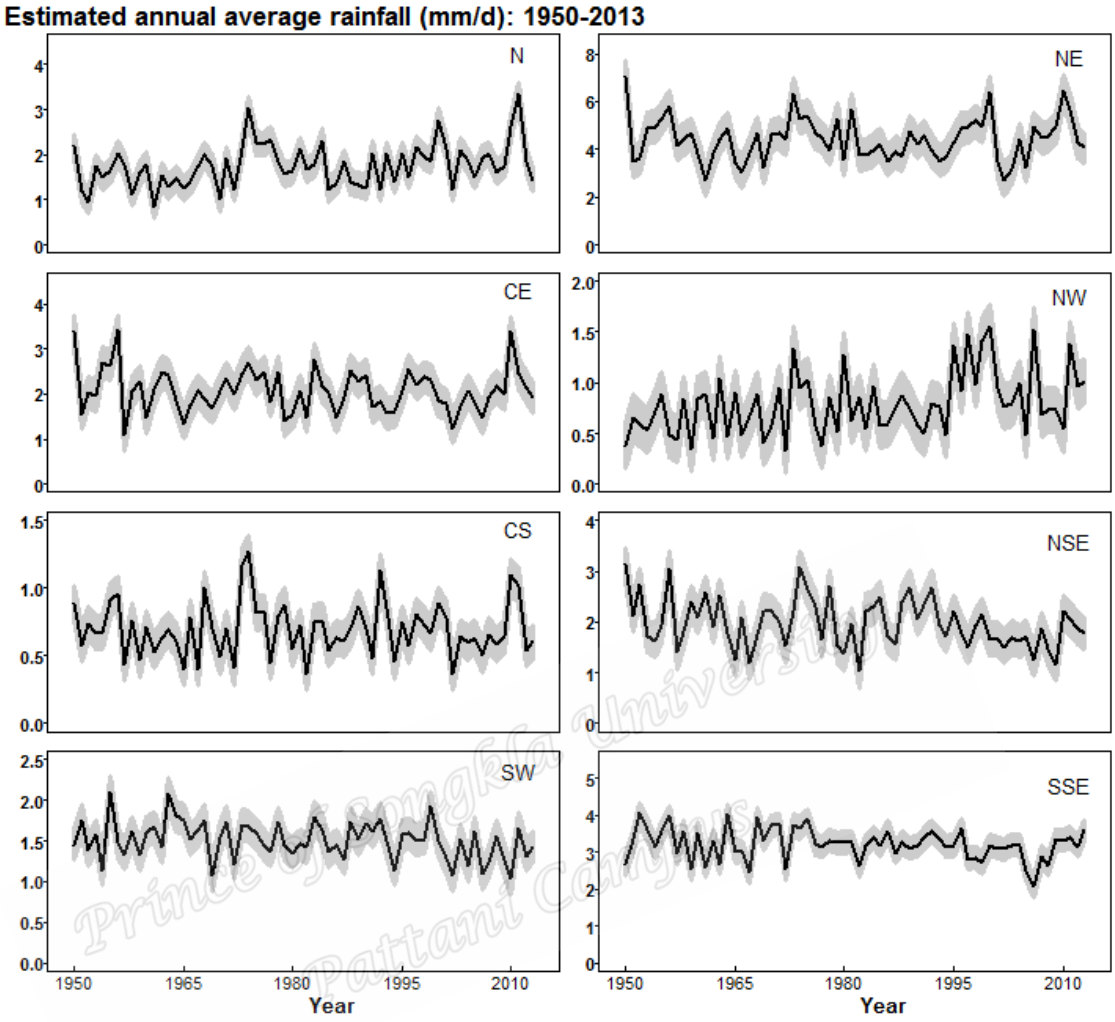


Figure 7. The estimated annual rainfall (line) and the 95% confidence interval (shaded) for factor regions from 1950-2013

Timbal and Fawcett (2013) also observed unique shortfalls in rainfall in southeastern Australia. The reported rainfall deficit was evident in pre-winter and early winter rainfall. These declining trends may be partly due to changes in large-scale atmospheric circulation (Nicholls, 2006; Timbal & Fawcett, 2013) including a poleward movement of the westerly winds and increasing atmospheric surface pressure. Changes in anthropogenic greenhouse gases and ozone levels (Delworth & Zeng, 2014) and expansion of the Southern Hemisphere Hadley cell (Post et al., 2014) can affect the decreasing rainfall trends in these regions.

Table 2

The Use of Linear Regression for Analysis of Annual Rainfall Trends

Region	N	NE	CE	NW	CS	NSE	SW	SSE
Slope	0.0076	0.0065	-0.0029	0.0001	-0.0001	-0.0075	-0.0035	-0.0050
p-value	0.0302	0.8757	0.4725	0.0026	0.7898	0.0235	0.0444	0.0964

In contrast, significant increasing annual rainfall trends were evident in the N and NW regions of Australia while a non-significant increasing trend at 95% confidence level was evident in the NE region (Table 2). Unlike the N and the NW regions, the NE region is mostly liable to intense weather conditions, for instance, tropical cyclones that bring about lots of convective clouds. This explains greater part of cloud cover in NE as related to that in the N and NW regions (Cheung et al. (2015). It could also explain part of the increasing rainfall trend in this region since cloud formation is related to rainfall.

CONCLUSIONS

This study uses statistical models in describing the period rainfall in Australia during 1950-2013. It is based on 64 years daily accumulated observations obtained from 92 stations distributed over Australia. Factor analyses identified eight factors, which correspond to eight geographical rainfall regions. The eight regions include north, northeast, central east, northwest, central south, north southeast, southwest and south southeast. The rainfall from each region exhibits diverse overall mean with various seasonal variations and was explored using Gamma generalised linear and multiple regression models.

These methods were applied to model the period rainfall obtained from each of the regions. The period rainfall observations in successive periods were serially correlated, and this was minimised by using the AR(1) technique in fitting the models.

Both models fitted the data quite well in all the regions, particularly in the tropical region (lower latitudes) where the atmospheric internal variability is small relative to the forced change. However, the multiple regression models did better than the Gamma generalised linear model in the north, northeast and the southwest regions.

The estimated period and annual rainfall means by the multiple regression models revealed three different groups and each of the groups show diverse monthly or annual pattern. In general, the eight identified regions of the factor analysis show three main periods rainfall category with each having a distinctive dominating pattern. The understanding and knowledge of these patterns are essential in planning the rainfall resources in numerous regions of Australia. The northeast received the highest annual rainfall amount followed by the south southeast. In contrast, central south and the northwest received low rainfall amount with similar pattern during this period. The decreasing annual rainfall observed in the southern regions may be due to a reduction in austral autumn and winter rainfall over the southern parts of Australia particularly south Western Australia.

ACKNOWLEDGEMENTS

The work was supported by Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries under grant no. THE-034. We are most grateful to Emeritus Prof. Don McNeil.

REFERENCES

Allan, R. P., Soden, B. J., John, V. O., Ingram, W. & Good, P. (2010). Current changes in tropical precipitation. Environmental Research Letters, 5(2), (p.025205).

Chatfield, C. (1996). The Analysis of Time Series. Chapman & Hall: Melbourne, Australia.

Cheung, K. K., Chooprateep, S., & Ma, J. (2015). Spatial and temporal patterns of solar absorption by clouds in Australia as revealed by exploratory factor analysis. Solar Energy, 111, 53-67.

Coe, R., & Stern, R. D. (1982). Fitting models to daily rainfall data. Journal of Applied Meteorology, 21(7), 1024-1031.

Delworth, T. L., & Zeng, F. (2014). Regional rainfall decline in Australia attributed to anthropogenic greenhouse gases and ozone levels. Nature Geoscience, 7(8), 583-587.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate data analysis, 5th. NY: Prentice Hall International.

Held, I. M. & Soden, B. J. (2006). Robust responses of the hydrological cycle to global warming. Journal of Climate, 19(21), pp.5686-5699.

IOCI. (2002). Climate variability and change in south west Western Australia. Indian Ocean Climate Initiative Panel, Perth.

Hope, P., Timbal, B., & Fawcett, R. (2010). Associations between rainfall variability in the southwest and southeast of Australia and their evolution through time. International Journal of Climatology, 30(9), 1360-1371.

Jeong, D. I., St-Hilaire, A., Ouarda, T. B. M. J., & Gachon, P. (2012). Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. Stochastic environmental research and risk assessment, 26(5), 633-653.

Kenabatho, P. K., McIntyre, N. R., Chandler, R. E., & Wheater, H. S. (2012). Stochastic simulation of rainfall in the semi-arid Limpopo basin, Botswana. International Journal of Climatology, 32(7), 1113-1127.

Kumar, P. (2013). Hydrology: seasonal rain changes. Nature Climate Change, 3(9), 783-784.

McNeil, N., & Chirtkiatsakul, B. (2016). Statistical models for the pattern of sea surface temperature in the North Atlantic during 1973-2008. International Journal of Climatology.

Lin, F. J. (2008). Solving multicollinearity in the process of fitting regression model using the nested estimate procedure. Quality & Quantity, 42(3), 417-426.

Lloyd-Hughes, B., & Saunders, M. A. (2002). Seasonal prediction of European spring precipitation from El Niño-Southern Oscillation and local sea-surface temperatures. International Journal of Climatology,22(1), 1-14.

McNeil, N., & Chooprateep, S. (2014). Modeling sea surface temperatures of the North Atlantic Ocean. Theoretical and applied climatology, 116(1-2), 11-17.

Mekanik, F., Imteaz, M. A., Gato-Trinidad, S., & Elmahdi, A. (2013). Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. Journal of Hydrology, 503, 11-21.

Meng, W. G., Zhang, Y. X., Dai, G. F., & Yan, J. H. (2007). The formation and development of a heavy rainfall mesoscale convective system along southern China coastal area J. Journal of Tropical Meteorology, 6, 000.

Mooley, D. A. (1973). Gamma distribution probability model for Asian summer monsoon monthly rainfall. Monthly Weather Review, 101(2), 160-176.

Murphy, B. F., & Timbal, B. (2008). A review of recent climate variability and climate change in southeastern Australia. International Journal of Climatology, 28(7), 859-879.

Nicholls, N. (2006). Detecting and attributing Australian climate change: a review. Australian Meteorological Magazine, 55(3), 199-211.

Oettli, P., & Camberlin, P. (2005). Influence of topography on monthly rainfall distribution over East Africa. Climate Research, 28(3), 199-212.

O'Gorman, P. A., & Schneider, T. (2009). The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. Proceedings of the National Academy of Sciences, 106(35), 14773-14777.

Post, D. A., Timbal, B., Chiew, F. H., Hendon, H. H., Nguyen, H., & Moran, R. (2014). Decrease in southeastern Australian water availability linked to ongoing Hadley cell expansion. Earth's Future, 2(4), 231-238.

R Development Core Team, 2015. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URLhttp://www.R-project.org**.**

Stern, R. D., & Coe, R. (1984). A model fitting analysis of daily rainfall data. Journal of the Royal Statistical Society. Series A (General), 1-34.

Timbal, B., Arblaster, J., Braganza, K., Fernandez, E., Hendon, H., Murphy, B., & Wheeler, M. (2010). Understanding the anthropogenic nature of the observed rainfall decline across South Eastern Australia. Centre for Australian Weather and Climate Research.

Timbal, B., & Fawcett, R. (2013). A historical perspective on southeastern Australian rainfall since 1865 using the instrumental record. Journal of Climate, 26(4), 1112-1129.

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S, 4th Edition, Springer, Queensland.

Wanishsakpong, W., & McNeil, N. (2016). Modelling of daily maximum temperatures over Australia from 1970 to 2012. Meteorological Applications, 23(1), 115-122.

Wickramagamage, P. (2010). Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. International Journal of Climatology, 30(8), 1235-1245.

Appendix III

Statistical modelling of 5-daily rainfall probability of occurrence in Australia during 190-2013

Bright Emmanuel Owusu[1,2] * Nittaya McNeil[1]

[1]Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Mueang Pattani, 94000, Thailand

[2]Department of Information and Communication Technology/Mathematics, Faculty of Science and Technology, Presbyterian University College Ghana

Abstract: Rainfall is an important meteorological parameter and part of the hydrological cycle. Spatial and temporal variations in it amount and occurrence can significantly affect the biological system and human culture. Modelling rainfall is essential for prediction and simulation purposes in numerous aspects of planning, agriculture, forestry, meteorology and hydrology. Two models are required to describe the two main features of rainfall: the occurrence and the amount.

This paper describes the occurrence probability of 5-daily rainfall observations collected from 105 meteorological stations in Australia. The observations spanning from 1950-2013 were collected from Australian Bureau of Meteorology. Logistic regression is applied to predict rainfall probability of occurrence in all the meteorological stations using nine observational rainfall stations as case studies. Weather conditions such as the 5-day period and annual rainfall factors were used as the predictors.

The fitted logistic regression models predicted the occurrence and non-occurrence 5-day rainfall events quite well with good accuracies. The predictors significantly affected the fitted models in all stations. Analysis of the levels of each of the predictors showed that the parameters of the 5-day period factors were more influential in all the models, particularly during the rainy season in most stations relative to that of the annual factors. The model could be used to simulate 5-day rainfall data for the stations with insufficient rainfall histories and could assist agronomist and various stakeholders in the planning of their various operations.

Keywords: Logistic regression, Rainfall occurrence, Receiver operating characteristics

Introduction

Daily rainfall observations are the primary meteorological inputs for the modelling of hydrological, agricultural, ecological and other environmental systems. As historical observational records give an insight of the underlying climate, stochastically produced sequences are used to assess the influence of climate variability on such systems.

Trenberth et al.(2003) revealed that, for the past few decades, the global hydrological cycle has been undergoing significant changes, which include rainfall amount and occurrence, frequency and duration. Rainfall distribution is highly variable over most of Australia throughout the year. For instance, the northern part experience a tremendous amount of annual rainfall during the summer half of the year (November-April) while the southeast and southwest has the most during Austral winter. Annual and seasonal rainfall variations are about 15-18% higher than any

other major agricultural nation (Cleugh et al., 2011; Mekanik and Imteaz, 2013; Walker and Mason, 2015).

Stochastic modelling of daily rainfall comprises two steps: first is the modelling of occurrence, and the second is the modelling of the non-zero amount. One of the arguably earliest most widely used, models for rainfall occurrence is the first-order Markov chain established by Gabriel and Neumann (1962) where the probability of rain or dry day is defined conditional only on the preceding day's rainfall state. Further application of this method is by (Bailey, 1964; Smith and Schreiber, 1973; Todorovic and Woolhiser, 1974; Chin, 1977; Haan et al., 1976; Richardson, 1981; Coe and Stern 1982; Stern and Coe, 1984; Wilks, 1989). The major setback of this technique is the ''short memory'' process models (where rainfall is only based on the past through the most recent day's occurrence) including under simulation of both the long dry spells as well as the variability at the interannual timescale.

However, these issues have been addressed by the application of higher-order Markov models, and Markov models considering exogenous climate variables as additional predictors (Wilks and Wilby, 1999). Other methods used in modelling rainfall events (occurrence and non-occurrence) has for some time, also been examined. These methods include alternating renewal process models (Green, 1964), Poisson models (Duckstein et al., 1972; Hasan and Dunn, 2010). Other methods include discrete autoregressive moving average models (Chang et al., 1984), and point procedure models (Kavvas and Delleur, 1981; Smith and Karr, 1983, 1985; Foufoula-Georgiou and Lettenmaier, 1987). Assessment of the differences among the numerous existing rainfall models has been given (Kavvas and Delleur, 1976; Waymire, 1981; Smith and Karr, 1985).

Logistic regression models have gained attention in modelling the probability of occurrence for some binary outcome variable using one or more continues or categorical variables as predictors (Ramos, 2017).It is arguably one of the most widely used methods for modelling rainfall occurrence, where the probability of a wet period at a specified site is a function of some predictors. Its application in daily rainfall probability modelling by (Stern and Coe, 1984; Chandler and Wheater, 2002; Kenabatho et al., 2012) predicted rainfall occurrences probabilities quite well in their respective areas. Prasad et al. (2010) applied logistic regression to model monthly rainfall occurrences in India, and the model performed well in capturing the extreme rainfall years and appear to work better than the direct model predictions of total precipitation in respect of such years.

Applequist et al. (2002) compared various linear and nonlinear statistical methodologies for probabilistic quantitative precipitation forecasting. These methods comprise linear regression, discriminant analysis, logistic regression, neural networks, and a classifier system. Their results indicated that logistic regression performs best among all the methodologies. Lo et al. (2007) conducted a probabilistic forecast of the onset of the north Australian wet season rainfall using logistic regression. They predicted probabilities that the inception of the rain would be later than the climatological mean date. Their prediction skills exceeded that of the climatological forecast in most areas of the north. While many studies have modelled rainfall occurrence specific periods of the year, the present study uses logistic regression to predict the probability of rainfall occurrence in Australia during the entire year.

Data and methods

In this study, long-term daily aggregated rainfall data for 105 observational weather stations spread over Australia from 1950 to 2013 were acquired from the Australian Bureau of Meteorology at (http://www.bom.gov.au/climate/data, Figure 1). These 105 stations were selected to give a sample covering the whole area as evenly as possible, and they have continued daily rainfall records that extend over a period of 64 years. However, there are no stations in the western desert, and most stations did not collect

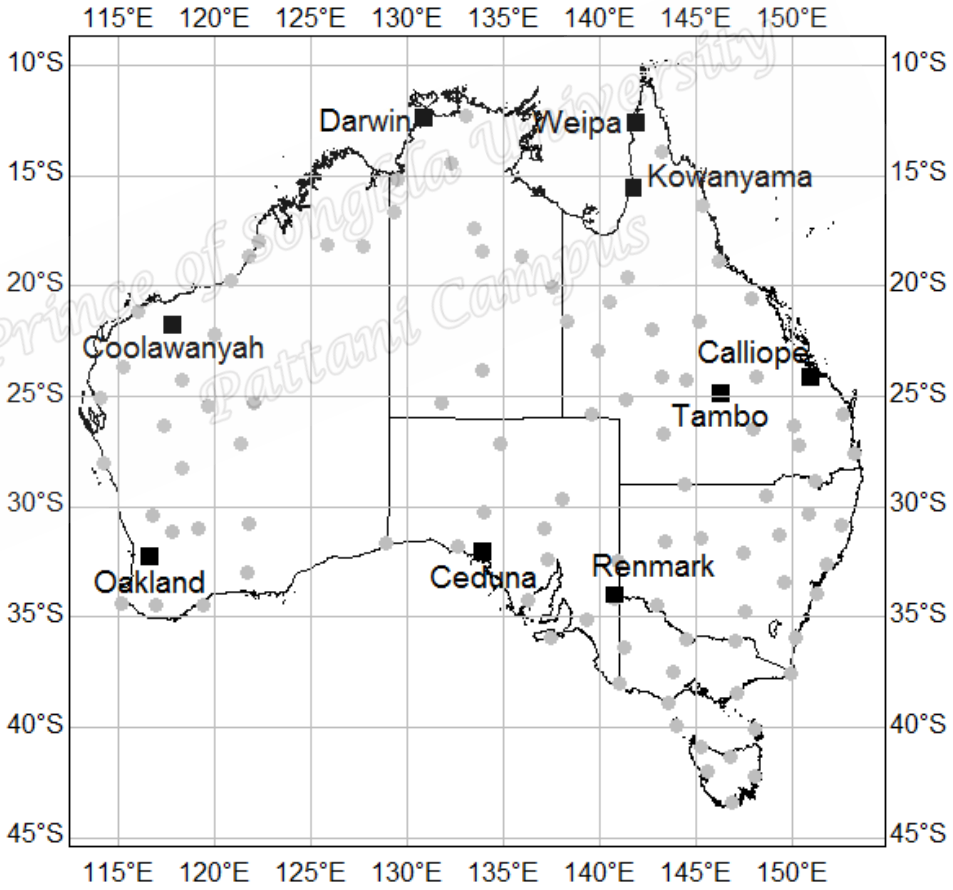


Figure 1. Location of the stations used for the study; with the nine case studies used in the paper named and indicated using black squares while the grey dots represent the remaining stations.

much data before 1950.Data recorded on leap years were excluded to maintain same observations for each year, and therefore each site is made up of 23,360 observations for the 64 years records.

Data from all the 105 stations well analysed with nine stations selected as case studies. The selected nine stations signify a variety of types of rainfall distributions and capture the major climate classification profiles concerning the Köppen climate classification (Köppen, 1918). These stations comprise different climatic regions of Australia with various rainfall distributions. Darwin and Weipa are characterised by tropical savanna climate with distinct wet and dry seasons with an annual rainfall of about 1,110.2 mm and over 2000 mm respectively.

Kowanyama also experiences tropical savanna climate with the wet season running from December to April and is characterised by frequent torrential downpours and high humidity.

During the dry season, almost no rainfall, warm to hot days with low humidity and the nights can get quite cold. Tambo experiences a hot semi-arid climate while Renmark and Ceduna have cold semi-arid climates and these areas receive little rainfall throughout the year. For instance, Ceduna has an annual rainfall of about 300 mm with hot, dry summers and cool, slightly wetter winters and July is the wettest month while Oakland experiences temperate oceanic climate.

Logistic regression model

The probability of 5-day rainfall occurrence and non-occurrence were predicted by using logistic regression. Logistic regression is a statistical modelling method that can be used to describe the association of numerous independent variables to a

dichotomous dependent variable. It is one of the generalised linear models with a binomial distribution for the response variable. The usual link function is a logit function, which in this study relates the logarithm of the odds ratio of the expected value of rainfall probability to linear predictors. It describes the occurrence probability of wet and dry days. To model the pattern of wet and dry days for each station using logistic regression, let p_i denote the probability of rain for the ith day in the data set, conditional on the variables x_i, then the model is given by:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \quad (1)$$

where α is a constant, β_i $(i = 1,2,3,...,k)$ are the slope coefficients of the model and x_i $(i = 1,2,3,...,k)$ are the predictors and k is the number of predictors. In fitting the models, the response variable was dichotomous with zero (0) representing a dry period (non-occurrence) and one (1) representing a wet period (occurrence) of rainfall. The first 50% of the rainfall observations (2336) were used as the training subset, and the remaining 50% of the observations (2336) was used as the validation subset. The predictors comprised 5-day periods and annual rainfall. The first predictor comprises of 73 categories while the latter consists of 64. All computations and analysis were done using R (2016).

Evaluation of the fitted Models

The global significance was evaluated by using the Hosmer and Lemeshow goodness-of-fit test (Hosmer et al. 1997; Hosmer and Lemeshow 2000). The confusion matrices and the Receiver Operating Characteristic (ROC) diagram were also used to assess the models. In examining the predictive ability of models with the confusion matrix, a

2×2 classification table of the observed with the predicted values for both the training dataset and the validation dataset was used to assess the classification accuracy. The ROC diagram is a discrimination-based graphical forecast verification display used to evaluate the goodness of fit of a model predicting a dichotomous outcome. It comprises a plot of true positive and false positive classification rates (1-specificity) for several probable thresholds of the model. Numerous indices of precision have been suggested to summarise ROC curves. Specifically, the area under the ROC curve (AUC) index is a standout amongst the most regularly utilised. A detailed account of the underlying principles of the ROC curve is given by Metz (1978), Hosmer and Lemeshow (2000) and Fawcett (2006).

The first application of ROC diagram in meteorology was by Mason (1982), and it is derived from signal detection theory in electrical engineering. A perfect forecast would have AUC value of 1, depicting forecast outcome as 1 (5-day rainfall) if $P_{ij} \geq c$ and 0 otherwise. The choice of c is made to draw a compromise between the number of the predicted and observed to ensure that equal weights are allocated to false positive and false negative prediction errors. A study by Vilar del Hoyo et al. 2011 revealed that an AUC value of 0.5 depicts no discrimination; a value between 0.5 and 0.69 shows poor discrimination; a value of 0.7-0.79 shows sound discrimination; a value of 0.8-0.9 indicates excellent discrimination, and a value of 0.9 or higher indicates exceptional discrimination.

Results and discussion

Summary of the descriptive analysis of the observed 5-day rainfall mean of the selected nine stations is provided in Table 1. Weipa has the highest rainfall mean of

5.23 mm followed by Darwin and Kowanyama with rainfall mean of 4.81 and 3.57 mm respectively. The least rainfall mean of 0.73 mm was recorded in Renmark. The highest rainfall variation occurred in Calliope followed by Coolawanyah with the coefficient of variation (CV) of 586 and 369 respectively. The comparison the 5-day average rainfall station wise revealed substantial spatial variability in the study area (Figure. 2). Among the wettest stations, Darwin Weipa and Kowanyama show similar patterns, with wet months from December to March.

In Oakland and Ceduna, the wettest months were observed from May to August. However, Renmark experiences uniform rainfall throughout the year. Analysis of the occurrence probability patterns of the mean rainfall amount of the stations revealed three main groupings. Darwin, Weipa and Kowanyama can be put in a group. They show similar rainfall occurrence probability patterns with average lying between the probability of 0.44-0.56 and higher probability values are observed between December and March. However, the rainfall occurrence probability is mostly below 0.2% throughout the remaining periods of the year.

The probability of rainfall occurrence in Coolawanyah, Tambo, and Calliope can also be put into a group. The average ranges from 0.46-0.76 and the occurrence probability are quite stable relative to the other stations.Oakland, Ceduna and Renmark can also be put in a group. They have similar probability patterns with the mean ranging from 0.48 to 0.62. The average occurrence probability is mostly above 0.2% throughout the year with higher probabilities values occurring between May and August.However, the 5-day rainfall patterns in Renmark seem to be consistent throughout the year to Oakland and Ceduna.

Table 1Summary statistics of the 5-day rainfall for the dataset (1950-2013), for the nine selected stations

Station Name	Station number	Lat (°C)	Lon (°C)	Elev (m)	Mean (mm)	Median (mm/d)	% missing	CV (%)
Darwin	14015	-12.42	130.89	30	4.81	0.20	0.00	198.96
Weipa	27042	-12.63	141.88	20	5.23	0.16	2.20	200.02
Kowanyama	29038	-15.48	141.75	10	3.57	0.00	0.00	230.84
Coolawanyah	5001	-21.80	117.81	360	1.07	0.00	0.30	368.69
Tambo	35069	-24.88	146.26	395	1.49	0.00	0.00	241.68
Calliope	39020	-24.02	150.97	58	0.97	0.00	0.40	586.00
Oakland	10620	-32.30	116.64	240	1.365	0.00	0.1	187.56
Ceduna	18012	-32.07	133.94	15	0.80	0.16	0.1	190.74
Renmark	24003	-34.17	140.75	20	0.73	0.76	0.1	217.96

Note: Lat (latitude), lon (longitude), Elev (elevation) % (percentage)

Relative to the 5-day means, station-wise comparison of the annual mean rainfall did not show any considerable variability in the study area throughout the 64 years. There were no apparent patterns of the average annual rainfall and little variations in the occurrence probabilities (Figure 3).The annual means were higher in Weipa, Darwin and Kowanyama relative to the other stations. Low annual rainfall means can be observed in Oakland, Ceduna and Renmark.

Modelling results

The logistic regression was fitted to 5-day rainfall mean in all the observational stations during 1950-2013 to predict the probability of occurrence. Each model was made up of two predictors (period and year factors).Term-wise addition of the predictors in the model showed a reduction in the deviance residual. The decrease in deviance residual indicates that the addition of each variable was essential in the modelling process (p-value < 0.001).

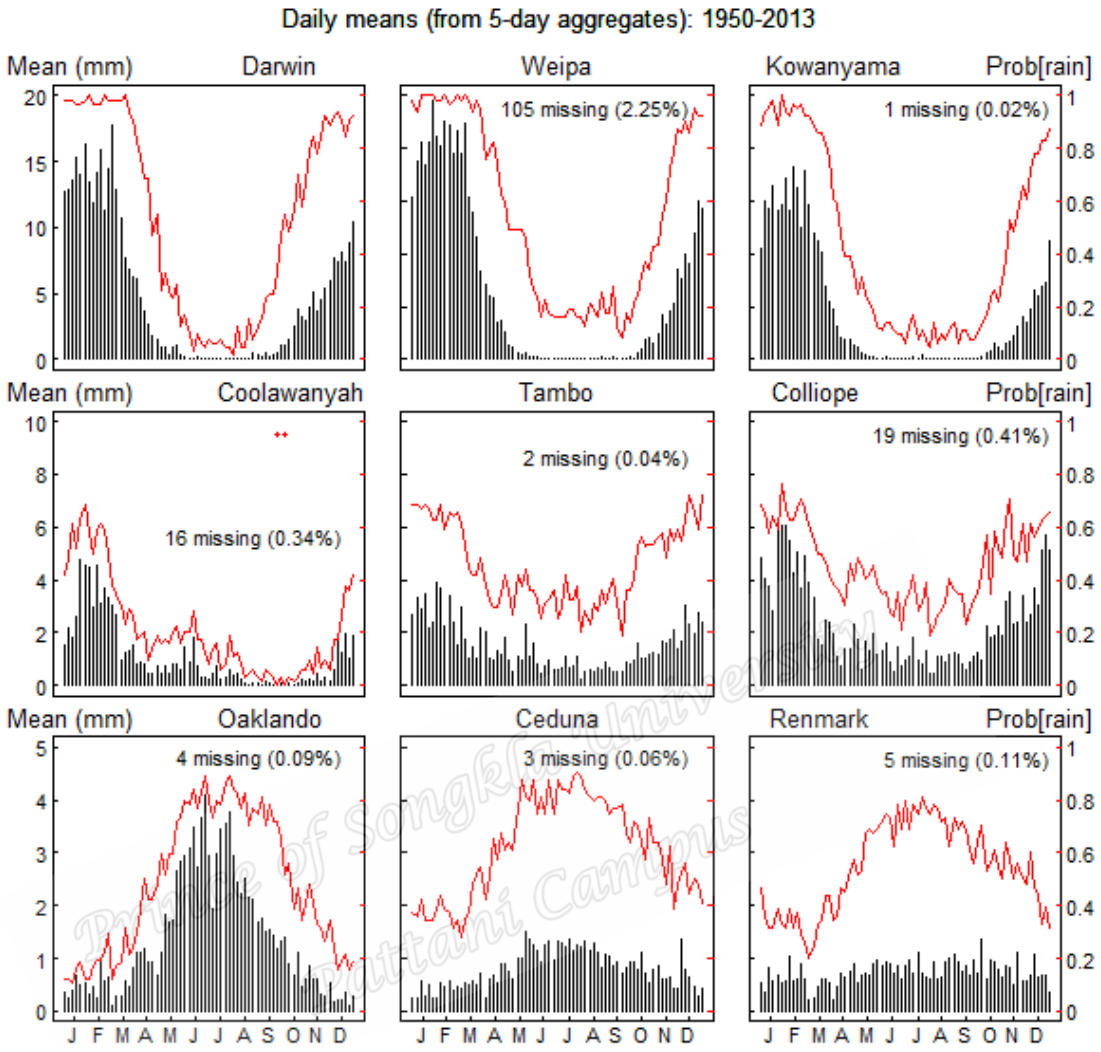


Figure 2. 5-daily rainfall means and the probability of occurrence.

Analysis of the variables of each level revealed that the period factors were more influential relative to the year factors in most of the model especially during the wettest period of the year in the study areas. For instance, 47 parameters of the period factors were significant while only 6 of the year factors were significant in Darwin. In Weipa, 50 covariates of the period factors were significant while only five covariates were significant concerning the annual factors.

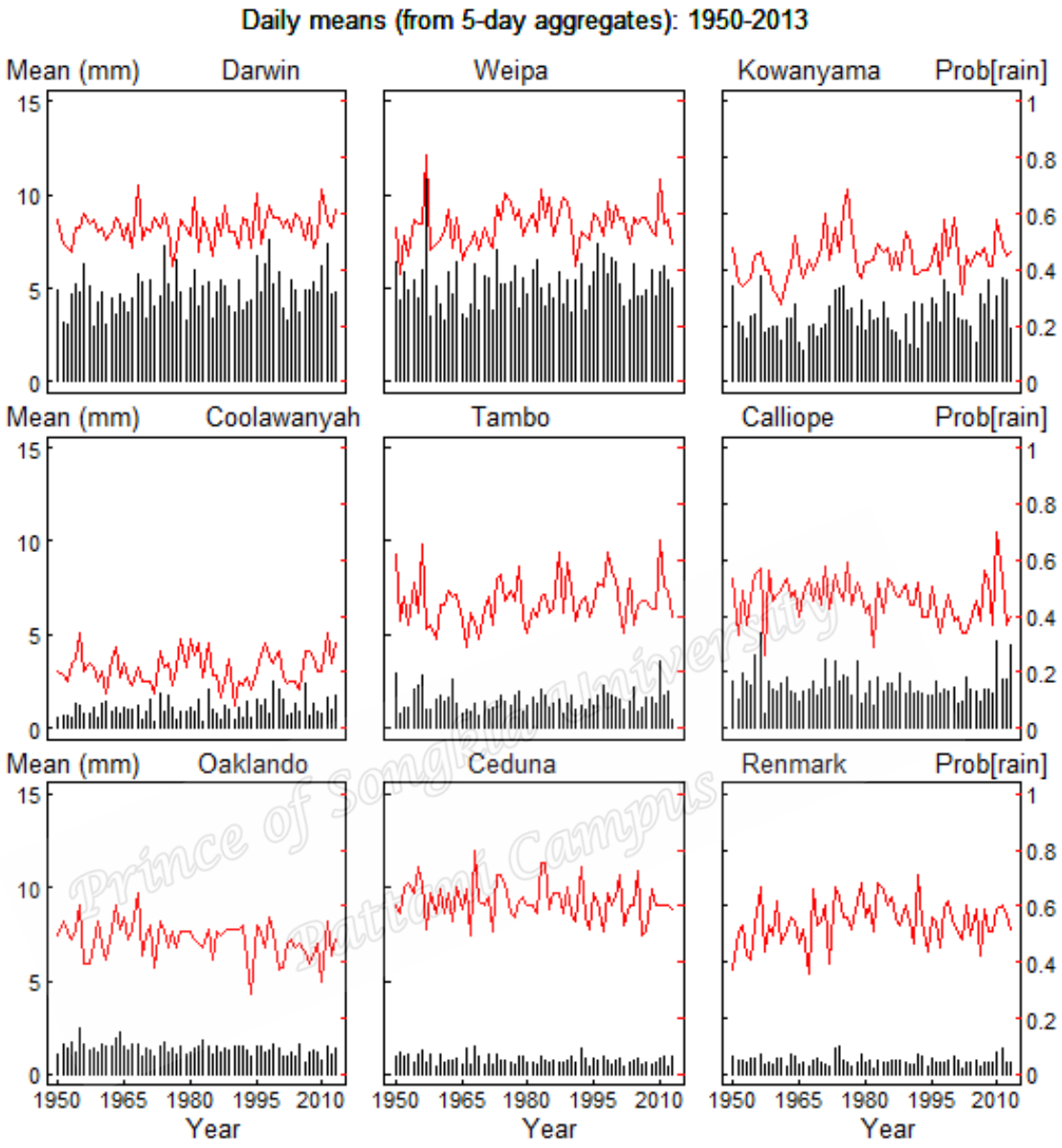


Figure 3. Annual rainfall and occurrence probability patterns.

The model in Tambo recorded 80 significant parameters (highest) with 47 periods and 37-year factors followed by Renmark with 71 significant parameters of 31 periods and 40-year factors. On the other hand, Calliope had the lowest number of these parameters (42) with three periods and 39-year factors (the result is not shown here).

The contingency table for the classification of the occurrence and non-occurrence probability of rainfall by the confusion matrix using the logistic regression models are given in Table 2. In Darwin, analysis of the classification accuracy by the confusion matrix indicated that the logistic regression correctly predicted 86.3 and 85.1% of non-occurrence and occurrence of rainfall respectively using the training dataset with a global accuracy of 85.8%. The area under the ROC (AUC) was 0.93 (Figure 3). The values of both the global accuracy and the AUC indicates exceptional discrimination by the logistic model. The omission and commission errors for these events were 13.7 and 14.9% respectively. The Hosmer and Lemeshow goodness-of-fit test revealed adequate fit of the logistic regression to the data ($\chi^2 = 2.42$, $df = 8$, p-value = 0.97). The classification accuracy was also assessed using the validation dataset. Concerning the validation dataset, the confusion matrix revealed that the logistic regression correctly predicted 86.6 and 82.5% of non-occurrence and occurrence of rainfall respectively. The omission and commission errors for these events were 13.4 and 17.5% respectively with overall classification accuracy observed to be 84.8%.

The overall classification accuracy was observed to be 83.9% in Weipa, and the model correctly predicted 87.8 and 79.9% of non-occurrence and occurrence of rainfall respectively. The Hosmer and Lemeshow goodness-of-fit test revealed adequate fit of the logistic regression to the data ($\chi^2 = 3.26$, $df = 8$, p-value = 0.92) signifying a good fit. The AUC was 0.91 (Figure 3) depicting exceptional discrimination. The errors associated with these events were 12.2 and 16.1% respectively with using the training dataset. The classification accuracy was also examined using the validation dataset. The confusion matrix shows that the regression

models revealed the overall accuracy of 79.2, predicting correctly 86.6 and 72.0% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 13.4 and 28% respectively, which depict a good fit.

The Hosmer and Lemeshow goodness-of-fit test displayed sufficient fit of the regression to the data ($\chi^2 = 11.64$, *df* = 8, *p*-value = 0.17). The AUC was 0.91 (Figure 3) and the global accuracy of 83.6% (Table 3, using the training dataset), both of which indicate good fit and exceptional discrimination in Kowanyama. The model correctly predicted 82.0 and 83.6% of non-occurrence and occurrence of rainfall respectively. Omission and commission errors for these events were 18.0 and 16.4% respectively (training dataset). On the other hand, the validation data set revealed a global accuracy of 80.7% and predicted correctly 81.6 and 80.1% of non-occurrence and occurrence of rainfall respectively. The omission and commission errors for these events were 18.3 and 19.9% respectively.

In Coolawanyah, the Hosmer and Lemeshow goodness-of-fit test displayed sufficient fit of the regression to the data ($\chi^2 = 5.48$, *df* = 8, *p*-value = 0.71) with AUC of 0.82 (Figure 3) indicating a good fit with excellent discrimination by the logistic regression model. The regression model predicted accurately 59.0 and 83.6% of non-occurrence and occurrence of rainfall respectively with a global accuracy of 80.7%. The errors associated with the classification of these events were 40.9 and 16.4% respectively (using the training dataset). The classification accuracy was also assessed using the validation dataset. The validation data set gave an overall accuracy of 80.6% and predicted correctly 57.9 and 83.4% of non-occurrence and occurrence

of rainfall respectively. The errors associated with the classification of these events were 42.1 and 16.5% respectively.

The Hosmer and Lemeshow goodness-of-fit test for the logistic displayed ($\chi^2 = 5.10$, *df* = 8, *p*-value = 0.75) in Tambo with AUC of 0.71 (Figure 3) signifying no indication of poor fit with sound discrimination. The results of the Hosmer and Lemeshow test and the value of the AUC revealed a good model since here we know the model is sufficiently stipulated. The occurrence and non-occurrence classifications indicated a global accuracy of 66.6%, predicting correctly 64.1 and 68.1% of the events respectively. The observed omission and commission errors for the events were 35.9 and 31.9% respectively (using the training dataset). The confusion matrix shows that the regression models revealed the overall accuracy of 58% predicting correctly 58.0 and 59.0% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events was 42.0 and 41.0% respectively (using the validation dataset).

In Calliope, 75.6% overall accuracy was observed in events classification, and the regression model correctly predicted 76.4 and 48.5% respectively of the events (Table 2). The errors associated with the classifications were 23.6 and 51.5% respectively with AUC of 0.68 (Figure 3) which indicates poor discrimination. The Hosmer and Lemeshow goodness-of-fit test for the logistic regression displayed ($\chi^2 = 9.70$, *df* = 8, *p*-value = 0.29). Even though the Hosmer and Lemeshow goodness-of-fit test did not show any evidence of worst of fit by the logistic regression model, the AUC value revealed poor classification of the event (using the training dataset). Assessment of the validation dataset shows that the confusion matrix

of the logistic regression model revealed the overall accuracy of 74.8% predicting correctly 76.2 and 28.8% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 28.8 and 71.2% respectively.

The overall accuracy of 75.9% was observed in Oakland in the classification by the confusion matrix of the logistic model, predicting correctly 75.9 and 75.9% of the non-occurrence and the occurrence events respectively (Table 2). The errors associated with these classifications are 24.1 and 24.1% respectively. The Hosmer and Lemeshow goodness-of-fit test for the logistic regression displayed ($\chi^2 = 1.94$, *df* = 8, *p*-value = 0.98) signifying no indication of poor fit. The AUC was 0.83 (Figure 3) showing excellent discrimination of events classification. The model fitted the data well since here we know the model is certainly properly specified using the training dataset. In assessing the validation dataset, the confusion matrix shows that the regression models revealed the overall accuracy of 71.4% predicting correctly 67.8 and 74.9% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 32.2 and 25.1% respectively.

The Hosmer and Lemeshow goodness-of-fit test displayed sufficient fit of the regression to the data ($\chi^2 = 5.90$, *df* = 8, *p*-value = 0.66) which indicates no evidence of poor fit. The AUC was 0.76 (Figure 3) which indicates sound discrimination of the events. The confusion matrix revealed that the regression model correctly classified 71.7% of all the observations in Ceduna. Of all the observations, 75.4% of the non-occurrence rainfall was correctly predicted while 64.2% of the rainfall occurrence was well predicted using the training dataset (Table 2). The omission and commission

errors associated with the rainfall events were 24.6 and 35.8% respectively. On the other hand, using the validation dataset, the confusion matrix revealed that the regression model presented an overall accuracy of 65.9%, predicting correctly 70.1 and 57.0% of the non-occurrence and occurrence of rainfall respectively. The errors associated with these events were 29.9 and 43.2% respectively (using the validation dataset).

In Renmark, the overall accuracy of the classification by the confusion matrix of the regression model was 67.6% while 68.8 and 66.1% of the non-occurrence and the occurrence events were correctly predicted respectively (Table 2). The AUC was 0.74 (Figure 3) indicating a sound classification of the events and the errors associated with the non-occurrence event was 31.2% while that of the occurrence event is 33.9%.The Hosmer and Lemeshow goodness-of-fit test revealed adequate fit of the logistic regression to the data ($\chi^2 = 2.89$, $df = 8$, p-value = 0.94)using the training dataset. In using the validation dataset to assess the model, the confusion matrix showed that the regression model predicted correctly 65.2 and 56.8% of the non-occurrence and the occurrence events respectively with an overall accuracy of 61.5%.The errors associated with these events were 34.8 and 43.2% respectively (using the validation dataset).The ROC curve for the logistic regression models for the stations is shown in Figure 4. In analysing the space in the ROC, the horizontal grey line depicts the random distribution of events to one of the two classes while the upper left corner signifies perfect classification of the events. The ROC plots are the curve lying between the two extremes (black curve). The nearer it lies to the upper left corner, the better the performance of the classifier to discriminate between two events.

Table 2. The contingency tables for the results of the confusion matrix of the logistic regression models for the nine stations

	Predicted							
	Training				Validation			
	Rain	0	1	% correct	Rain	0	1	% correct
Darwin								
Observed								
	0	1118	177	86.3	0	1122	173	86.6
	1	155	886	85.1	1	182	859	82.5
Overall percentage				85.8				84.8
Weipa								
Observed								
	0	1048	145	87.8	0	957	148	86.6
	1	230	913	79.9	1	315	811	72.0
Overall percentage				83.9				79.2
Kowanyama								
Observed								
	0	790	173	820	0	785	177	81.6
			114				110	
	1	225	8	83.6	1	273	0	80.1
Overall percentage				83.0				80.7
Coolawanyah								
Observed								
	0	167	116	59.0	0	154	112	57.9
			171				171	
	1	336	7	83.6	1	339	5	83.4
Overall percentage				80.7				80.6
Tambo								
Observed								
	0	556	311	64.1	0	502	363	58.0
			100					
	1	469	0	68.1	1	602	867	59.0
Overall percentage				66.6				58.7
Calliope								
Observed								
	0	1734	536	76.4	0	1729	541	76.2
	1	32	34	48.5	1	47	19	28.8
Overall percentage				75.6				74.8
Oakland								
Observed								
	0	884	280	759	0	789	375	67.8
	1	283	889	75.9	1	293	875	74.9
Overall percentage				75.9				71.4
Ceduna								
Observed								
	0	1191	389	75.4	0	1107	472	70.1
	1	271	874	64.2	1	324	430	57.0
Overall percentage				71.7				65.9
Renmark								
Observed								
	0	904	410	68.8	0	855	456	65.2
	1	346	676	66.1	1	441	579	56.8
Overall percentage				67.6				61.5

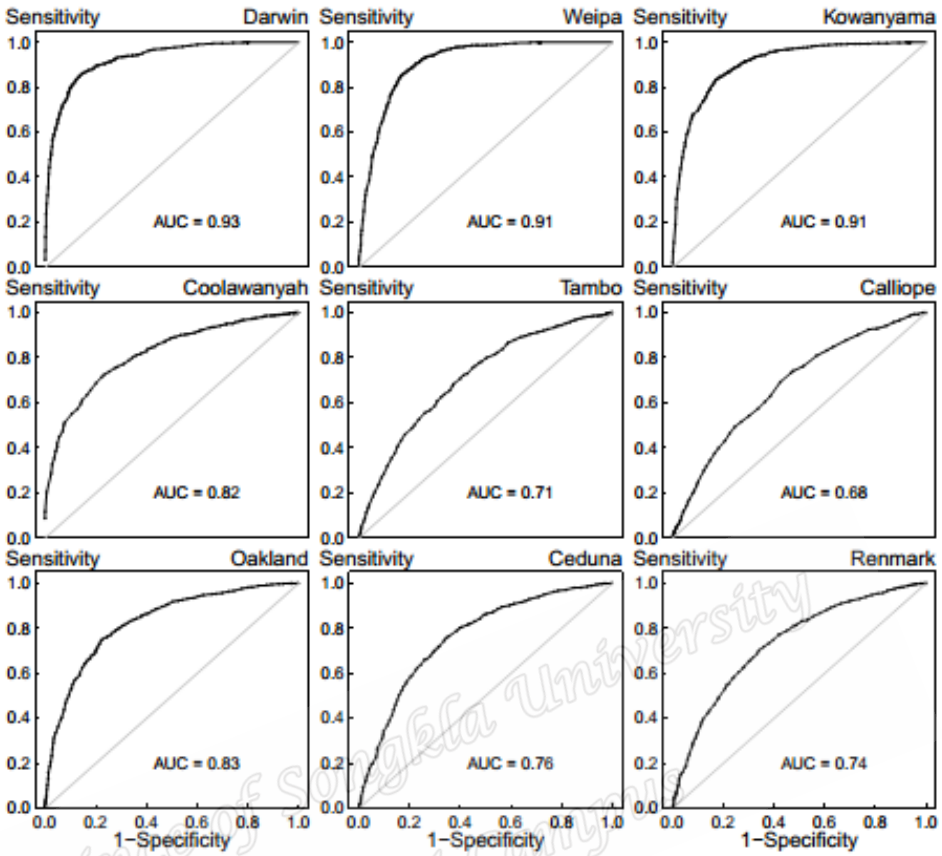


Figure 3. The ROC curves for the logistic regression models.

Moreover, the AUC is usually used to assess the performance of the classifier. The computed AUC values range from 0.68 to 0.92.

The values in Darwin, Weipa and Kowanyama showed exceptional discrimination while the models in Coolawanya and Oakland revealed excellent discrimination of the event (good fit of the models). Also, the models in Tambo, Ceduna and Renmark showed sound discrimination of the events while the model in Calliope revealed poor discrimination of the two events.

Conclusion

In this study, logistic regression has been applied to describe 5-day rainfall probability of occurrence during 1950-2013 in Australia. The viability of modelling rainfall occurrence using two easily acquire variables has been demonstrated. The fitted models have acceptable predictive precision in predicting the probability of rainfall occurrence in Australia. The models perform quite well concerning the proportions of correct predictions for the occurrence or the non-occurrence rainfall that ranges between 48.1-86.5%. However, the predictive ability of the models will possibly be enhanced with the inclusion of other climate variables such as mean temperature, mean sea level pressure, wind speed and direction.

Acknowledgements

This work was supported by Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries under grant no. THE-034. We are also grateful to Emeritus Prof. Don McNeil.

References

Applequist, S., Gahrs, G. E., Pfeffer, R. L., & Niu, X. F. (2002). Comparison of methodologies for probabilistic quantitative precipitation forecasting. Weather and Forecasting, 17(4), 783-799.

Bailey, N.T.J., 1964. The Elements of- Stochastic Processes. John Wiley, New York Chin, E. H.(1977). Modeling daily precipitation occurrence process with Markov chain. Water Resources Research, 13(6), 949-956.

Chandler, R. E., & Wheater, H. S. (2002). Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland. Water Resources Research, 38(10).

Chang, T. J., Kavvas, M. L., & Delleur, J. W. (1984). Daily precipitation modeling by discrete autoregressive moving average processes. Water Resources Research, 20(5), 565-580.

Chin, E. H. (1977). Modeling daily precipitation occurrence process with Markov Chain. Water Resources Research, 13(6), 949-956.

Cleugh, H., Cleugh, H., Smith, M. S., Battaglia, M., & Graham, P. (2011). Climate Change: science and solutions for Australia. CSIRO.

Coe, R., & Stern, R. D. (1982). Fitting models to daily rainfall data. Journal of Applied Meteorology, 21(7), 1024-1031.

Del Hoyo, L. V., Isabel, M. P. M. and Vega, F.J. M. (2011). Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data. European Journal of Forest Research, 130(6), 983-996.

Duckstein, L., Fogel, M. M., & Kisiel, C. C. (1972). A stochastic model of runoff-producing rainfall for summer type storms. Water Resources Research, 8(2), 410-421.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Foufoula- Georgiou, E., & Lettenmaier, D. P. (1987). A Markov renewal model for rainfall occurrences. Water Resources Research, 23(5), 875-884.

Gabriel, K. R., & Neumann, J. (1962). A Markov chain model for daily rainfall occurrence at Tel Aviv. Quarterly Journal of the Royal Meteorological Society, 88(375), 90-95.

Green, J. R. (1964). A model for rainfall occurrence. Journal of the Royal Statistical Society. Series B (Methodological), 345-353.

Haan, C. T., Allen, D. M., & Street, J. O. (1976). A Markov chain model of daily rainfall. WaterResources Research, 12(3), 443-449.

Hosmer Jr., D. W., Lemeshow, S. (2000). Applied Logistic Regression, second ed. John Wiley & Sons, New York.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine, 16(9), 965- 980.

Kavvas, M.L., & Delleur, J.W., (1976). A point stochastic model for continental droughts. Hydrological Sciences. 21 (3), 407-417.

Kavvas, M. L., & Delleur, J. W. (1981). A stochastic cluster model of daily rainfall sequences. Water Resources Research, 17(4), 1151-1160.

Kenabatho, P. K., McIntyre, N. R., Chandler, R. E., & Wheater, H. S. (2012). Stochastic simulation of rainfall in the semi- arid Limpopo basin, Botswana. International Journal of Climatology, 32(7), 1113-1127.

Koppen, W. (1918). Klassifikation der klimate nach temperatur, niederschlag, und jahreslauf. Petermann's Mitteilungen, 64, 193-203.

Lo, F., Wheeler, M. C., Meinke, H., & Donald, A. (2007). Probabilistic forecasts of the onset of the north Australian wet season. Monthly Weather Review, 135(10), 3506-3520.

Mason, I. (1982). A model for assessment of weather forecasts.Australian Meteorological Magazine, 30(4), 291-303.

Mekanik, F., Imteaz, M. A., Gato-Trinidad, S., & Elmahdi, A. (2013). Multiple regression andArtificial Neural Network for long-term rainfall forecasting using large scale climate models. Journal of Hydrology, 503, 11-21.

Metz, C. E. (1978). Basic principles of ROC analysis. In Seminars in nuclear medicine, 8(4),283-298.

Prasad, K., Dash, S. K., & Mohanty, U. C. (2010). A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts. International Journal of Climatology, 30(10), 1577-1588.

Ramos, H. M., Ollero, J., & Suárez-Llorens, A. (2017). A new explanatory index for evaluating the binary logistic regression based on the sensitivity of the estimated model. Statistics & Probability Letters, 120, 135-140.

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. Water resources research, 17(1), 182-190.

Smith, J. A., & Karr, A. F. (1983). A point process model of summer season rainfall occurrences. Water Resources Research, 19(1), 95-103.

Smith, J. A., & Karr, A. F. (1985). Parameter Estimation for a Model of Space- Time Rainfall. Water Resources Research, 21(8), 1251-1257.

Smith, R. E., & Schreiber, H. A. (1973). Point processes of seasonal thunderstorm rainfall: 1. Distribution of rainfall events. Water Resources Research, 9(4), 871-884.

Stern, R. D., & Coe, R. (1984). A model fitting analysis of daily rainfall data. Journal of theRoyal Statistical Society. Series A (General), 1-34.

Trenberth, K. E., Dai, A., Rasmussen, R. M., & Parsons, D. B. (2003). The changing character of precipitation. Bulletin of the American Meteorological Society, 84(9), 1205-1217.

Todorovic, P., & Woolhiser, D. A. (1974). Stochastic model of daily rainfall. In Proceedings of the Symposium on Statistics in Hydrology, USDA Miscellaneous Publication (Vol. 1275, pp.32-246).

Walker, R., & Mason, W. (Eds.). (2015). Climate change adaptation for health and social services. CSIRO Publishing.

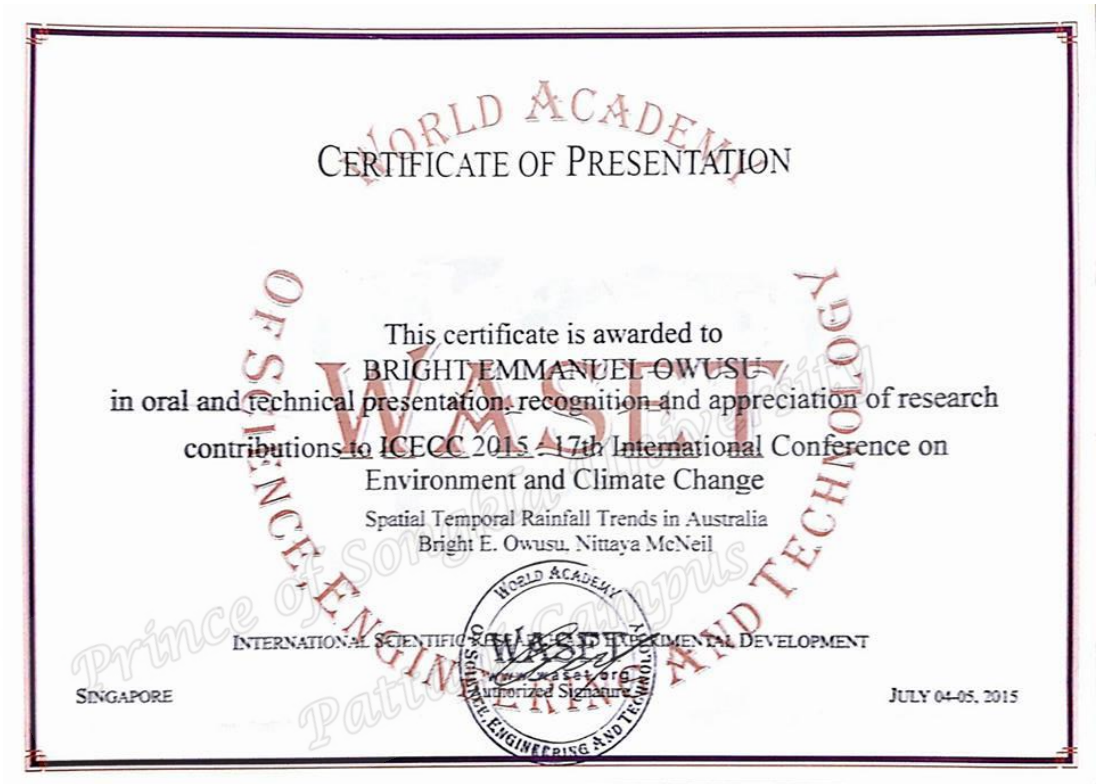
Waymire, E. D., & Gupta, V. K. (1981). The mathematical structure of rainfall representations:1. A review of the stochastic rainfall models. Water resources research, 17(5), 1261-1272.

Wilks, D. S. (1989). Conditioning stochastic daily precipitation models on total monthly precipitation. Water Resources Research, 25(6), 1429-1439.

Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. Progress in physical geography, 23(3), 329-357.

Appendix IV

Presenting Conference Paper



WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY

CERTIFICATE OF PRESENTATION

This certificate is awarded to
BRIGHT EMMANUEL OWUSU
in oral and technical presentation, recognition and appreciation of research contributions to ICECC 2015 - 17th International Conference on Environment and Climate Change

Spatial Temporal Rainfall Trends in Australia
Bright E. Owusu, Nittaya McNeil

INTERNATIONAL SCIENTIFIC RESEARCH AND ENVIRONMENTAL DEVELOPMENT

SINGAPORE

Authorized Signature

JULY 04-05, 2015

World Academy of Science, Engineering and Technology
International Journal of Environmental and Ecological Engineering Vol:2, No:7, 2015

Spatial Temporal Rainfall Trends in Australia

Authors: Bright E. Owusu, Nittaya McNeil

Abstract: Rainfall is one of the essential quantities in meteorology and hydrology. It has important impacts on people'sdaily life, and excess or inadequate of it could bring tremendous losses in the economy and cause fatalities. Population increase around the globe tends to have a corresponding increase in settlement and industrialization. Some countries are affected by flood and drought occasionally due to climate change, which disrupts most of the daily activities. Knowledge of trends in spatial and temporal rainfall variability and their physical explanations would be beneficial in climate change assessment and to determine erosivity. This study describes the spatial-temporal variability of daily rainfall in Australia and their corresponding long-term trend during 1950-2013. The spatial patterns were investigated by using exploratory factor analysis and the long-term trend in rainfall time series were determined by linear regression, Mann-Kendall rank statistics and the Sen's slope test. The exploratory factor analysis explained most of the variations in the data and grouped Australia into eight distinct rainfall regions with different rainfall patterns. Significant increasing trends in annual rainfall were observed in the northern regions of Australia. However, the northeastern part was the wettest of all the eight rainfall regions.

Keywords: climate change, explanatory factor analysis, Mann-Kendall and Sen's slope test, rainfall.
Conference Title: ICECC 2015: 17th International Conference on Environment and Climate Change
Conference Location: Singapore, SG
Conference Dates: July 04-05, 2015

Vitae

Name: Mr. Bright Emmanuel Owusu

Student ID 5720330004

Educational Attainment:

Degree	Name of institution	Year of Graduation
BSc. (Mathematics)	Kwame Nkrumah University of Science and Technology (KNUST)	2004
MSc. (Mathematics)	Kwame Nkrumah University of Science and Technology (KNUST)	2009

Scholarship Awards during Enrolment

1. The Scholarship Awards for Master and Ph.D. Studies by Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries (TEH-AC) under grant number THE/034.
2. Graduate School Research Scholarship, Prince of Songkla University, Songkhla, Thailand.

Work-Position and address:

Lecturer

Department of Information and Communication Technology / Mathematics
Presbyterian University College, Ghana.

List of Publication:

Owusu, B E., and McNeil, N. 2017. Statistical modelling of daily rainfall variability patterns in Australia. Pertanika Journal of Science and Technology. (Accepted for publication on Sept. 27, 2017).

Owusu, B E., McNeil, Nand Eso, M. 2018.Statistical modelling for 5-day average rainfall in Australia during 1950-2013.SongklanakarinJournal of Science and Technology (Under Review).

Owusu, B E., and McNeil, N. 2018. Statistical modelling of 5-daily rainfall probability of occurrence in Australia during 190-2013. Hydrological Science Journal (Under Review).

Proceedings:

Owusu, B E., and McNeil, N. 2015.Spatial-Temporal Rainfall Trends in Australia. 17th International Conference on Environment and Climate Change, July 4-7, 2015, Singapore (Oral Presentation).